ACUTE LEUKAEMIA IN NEW ENGLAND
AN INVESTIGATION INTO THE CLUSTERING OF CASES IN TIME AND PLACE

BY
M. MERRINGTON, Ph.D.
Lecturer in Statistics, University College, London
AND
C. C. SPICER, M.R.C.S.
Director, M.R.C. Computer Unit, London

The problem of discovering evidence for clustering of cases in place and time has only recently been investigated by epidemiologists and statisticians. Acute leukaemia has sometimes been described as a disease which does occur in clusters (Heath and Hasterlik, 1963; Kellett, 1937; Knox, 1964; Mainwaring, 1966; Meighan and Knox, 1965). Conversely, it has also been reported that the disease does not occur in clusters (Ager, Schuman, Wallace, Rosenfield, and Gullen, 1965; Barton, David, and Merrington, 1965; Clemmesen, Busk and Nielsen, 1952; Ederer, Myers and Mantel, 1964; Ederer, Myers, Eisenberg, and Campbell, 1965; Lock and Merrington, 1967; Lundin, Fraumeni, Lloyd, and Smith, 1966; Stark and Mantel 1967). We report here the statistical analysis on some 540 cases of acute leukaemia occurring in Maine, Massachusetts, New Hampshire, and Vermont. The records of acute leukaemia in Connecticut were not available to us but they have been described elsewhere (Ederer et al., 1965).

MATERIAL
The death certificates or tumour registry records of all cases of acute leukaemia were obtained from the State Houses in Augusta, Maine; Boston, Massachusetts; Concord, New Hampshire; and Burlington, Vermont. The information needed for the statistical analysis was the date of onset of the disease and the address of the place where the patient resided at that date. The date of onset was not always recorded and Table I gives the number of cases of acute leukaemia available for analysis.

Thus out of a total of 717 there were 174 records where it was not possible to obtain the date of onset of the disease. It should be noted that if the death occurred in one state but the address of the patient when the disease first began was in another, then this patient was allocated to the latter state.

STATISTICAL ANALYSIS
Several statistical techniques are now available for determining whether a group of patients who develop a particular disease at about the same time also live near each other.

Two methods have been applied to the cases of acute leukaemia in New England. Each state was considered separately, although when New Hampshire and Vermont were considered together for one of the tests this only led to a blurring of the results.

The first method, devised by David and Barton (1966), is worked out on the basis of testing whether the spatial pattern is random within a group of cases in time. Initially all the cases are written down chronologically. They are then divided into groups in time according to a rule, which must be decided before the statistical test is calculated. Unless there is some particular reason for choosing a special interval, it is usual to choose the groups with several different time intervals between them and to repeat the same statistical analysis for each interval. The next step is to find the distances between all cases. This is done by first allotting \((x, y)\) coordinates to each patient's address and then calculating the distances by simple algebra. The David and Barton
statistical test criterion is the ratio of the variation of the distances within time groups to the variation of all the possible distances between the cases. If the cases are randomly distributed this ratio will be unity. The closer they are to each other within a time group the smaller the ratio will be. As a yardstick for deciding whether clustering exists, we may subtract its expected value, unity, from the ratio calculated from the data and divide by the standard deviation. If this criterion is less than –2 it can be concluded that there is some clustering effect. The method of calculating is as follows:

Suppose there are \( N \) persons suffering from the disease and they are divided according to some time interval into \( n \) groups. Let \( (x_{it}, y_{it}) \) be the co-ordinates of the \( i \) th person in the \( t \) th group and \( n_t \) the number of persons in the \( t \) th group. Write

\[
k_{st} = \frac{1}{n_t} \sum_{i=1}^{n_t} (x_{it} - \bar{x})^2, \quad k'_{st} = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{it} - \bar{y})^2,
\]

and

\[
m_{ab} = \frac{1}{N} \sum_{t=1}^{n} \sum_{i=1}^{n_t} (x_{it} - \bar{x})^a (y_{it} - \bar{y})^b,
\]

where \( \bar{x}, \bar{y} \) are the overall means of \( x \) and \( y \) respectively. Then the test criterion is

\[
Q = \left( \frac{N-1}{N-n} \right) \left[ 1 - \frac{1}{N(m_{20} + m_{20})} \sum_{t=1}^{n} n_t (k_{st}^2 + k'_{st}^2) \right]
\]

If we consider the distribution of \( Q \) obtained by taking all possible ways in which the \( N \) observations in space can be assigned to the \( n \) groups, it is found that

\[
E(Q) = 1
\]

and

\[
\text{var } Q = \left( \frac{N-1}{N-n} \right)^2 \frac{1}{N^{(4)}} \left\{ \frac{m_{20} + 2m_{20} + m_{04}}{(m_{20} + m_{02})^2} \right. \\
\left. - 2 \left\{ \frac{(m_{20})^2 + 2m_{11}^2 + m_{04}}{(m_{20} + m_{02})^2} \right\} \frac{[N(\Delta - (N-2)(N-n)]}{N(\Delta - (N-2)(N-n)} \\
- \frac{N(\Delta - (N-2)(N-n))}{N(\Delta + 2(N-n)(n-1)(N-n))} \right) \\
\]

with \( \Delta = \sum_{t=1}^{n} \left( \frac{1}{n_t} - \frac{n}{N} \right) \).

The randomization distribution of \( Q \) may be approximated by a \( B \)-function as described by Barton et al. (1965). But the data available in the present investigation are not such that it is appropriate to use any refined statistical techniques. For our purposes, it may be assumed that \( Q \) is normally distributed with standard deviation \( \sqrt{\text{var } Q} \). Under the null hypothesis of no clustering the expected value of \( Q \) is unity, and under the alternative hypothesis of clustering the observed value of \( Q \) will be less than one. Thus we may calculate

\[
(Q - 1)/ \sqrt{\text{var } Q}
\]
as a yardstick for deciding whether clustering exists. If this ratio is less than –2 it can be concluded that there is some clustering effect. Table II gives the detailed results of the application of this test to the victims of acute leukaemia in New England.

Since none of the results is less than –2, it can be seen that there is no evidence of clustering among the 543 cases of acute leukaemia in Maine, New Hampshire, Vermont, and Massachusetts.

The second method used for detecting clustering is that devised by Knox (1964). All cases are considered in pairs, their distance apart is calculated, and their time apart is noted. If the number of cases with a short distance and a short time apart is larger than would be expected if they were randomly distributed, then clustering must occur. Whatever the specific distances and times used for this test, no excessive numbers of pairs of cases were found close to each other in space and time. Thus, no evidence for clustering among the cases of acute leukaemia in New England was provided by this second statistical test.

**DISCUSSION**

It has already been stated that the date of onset was not recorded for about one quarter of the victims of acute leukaemia, so that these individuals cannot be included in the calculations of the statistical analysis. Ideally, each of these cases should be traced individually in order to ascertain the date of onset from the hospital, doctor or family. We are hoping that it will be possible to carry out these investigations at least for the cases in Vermont and New Hampshire. Knox (1964) was not able to establish the date of onset of 10 cases out of 185 of all types of leukaemia in Northumberland and Durham, England. For two of them he took one month before the date of admission to hospital as the date of onset and for the other eight patients, who died without traceable records, he took six months before the date of death as the date of onset.
In the present investigation an attempt was made to allocate a date of onset randomly. This allocation was based on the distribution of the duration of acute leukaemia observed in the New England states. By associating a numerical scale with the frequencies of this distribution, a random number can be chosen to allocate a date of onset. The $Q$ test was then calculated, with the results shown in Table III.

The results in this Table are mainly similar to those given in Table II. The test criterion is not less than $-2$ and there is no evidence for the clustering of cases. However there are two exceptions. The 309 cases in Massachusetts, of which 77 had the date of onset allocated randomly, appeared to be clustered in place when they occurred with intervals of four to six days between them. This is not evidence for clustering, but it does indicate that if it were possible to establish the dates of onset of all the cases, it might then be possible to show whether clustering existed or not.

For some of these data the test for clustering was calculated for children of 15 years and under. Results were obtained similar to those already found and given in Table II. Since the date of onset was not available for one quarter of the cases the division into young and old according to the onset date could not be made with certainty.

Till, Hardisty, Pike, and Doll (1967), in their search for evidence of clustering in childhood leukaemia in London, divided their cases into the principal sub-groups of the disease (lymphoblastic and myeloblastic and unspecified) and applied Knox's statistical test for clustering to the sub-groups. No statistically significant clustering was found. Weak evidence of clustering was found for cases of lymphoblastic leukaemia; and the strongest evidence was found when the date of birth was considered with the place of residence of the parents.

For the data from New England, the grouping of cases of acute leukaemia into clinical types was not considered reasonable in the test for clustering because the diagnoses were carried out in widely different areas and at different times. There was unlikely to be any consistency in the differentiation of lymphatic and myelocytic acute leukaemia and
indeed over one third of the cases in this investigation were listed in the records as unspecified acute leukaemia. It has also been suggested that the diagnosis of acute lymphatic leukaemia may not have been made on morphological grounds but may have been based on the age of the patient.

Although the present study gives no evidence of clustering it cannot be said to exclude the possibility. However, any clustering that exists must be of an altogether lower order than that seen in infectious diseases such as jaundice and poliomyelitis which the David-Barton test can readily demonstrate (Barton et al., 1965). Furthermore, the mathematical studies of Barton, David, Fix, Merrington, and Mustacchi (1967) have shown that the test used is quite powerful in relation to a realistic model of clustering.

The basic data cannot be considered to be perfect particularly in view of the incomplete recording of dates of onset, but there is no obvious reason why their imperfections should obscure rather than emphasize any clustering that may have occurred.

**Summary**

The records of 543 victims of acute leukaemia in four New England states between 1957 and 1964 were obtained mainly from death certificates. Two statistical criteria for testing the association of cases in time and place were calculated. With the records available, no clustering of cases was detected. The data were limited because the dates of onset were not obtainable for one quarter of the cases.

We wish to acknowledge with thanks the help given by many doctors and medical records officers, in particular Mr. R. D. Lavallee, State Registrar of Massachusetts, Mrs. M. M. Colby, State Registrar of New Hampshire, and Dr. D. E. Bergstrom, Chief of the Vermont Public Health Statistics. Dr. Miriam D. Manning kindly gave us access to her records of acute leukaemia in Maine and New Hampshire. The calculations were carried out on the ICT Atlas computer at the University of London Institute of Computer Science; we wish to acknowledge gratefully the expert assistance given us by Mrs. Pat Hills, Miss Norma Pateman, and Miss Pat Williamott in processing these data. Mr. Peter Smith of the Medical Research Council, kindly carried out the calculations for the Knox test. One of us (M.M.) owes grateful thanks to the Milton Fund, Harvard University, who gave her a grant towards the cost of this investigation.

**REFERENCES**


