An Index for Cancer Clustering
by Toshiro Tango*

This paper generalizes the index for temporal clustering proposed by Tango in two ways: it allows for nonuniform population distributions across the study period and it is applicable to the detection of disease clustering in space where there are variations in population distribution among categories of the confounding factor such as age and sex. Applications are illustrated with 1833 cases of mortality from uterine cancer in the Tokyo metropolitan area during 1978–1982.

Introduction

The investigation of disease clustering in space, in time, or in both is an important aspect of epidemiological studies in order to find clues to the causative mechanism of the disease in question. For example, the evidence of space-time clustering suggests that individual cases of disease are closely related in both space and time, as is often found in the case with infectious diseases. It has been stated on several occasions that childhood leukemia occurs in clusters in both space and time in many of the studies, which indicates the possibility of viral etiology. Therefore, tests for the detection of space-time clustering have been the subject of considerable research in recent years (1–6).

In the study of chronic disease such as cancer, on the other hand, those tests for space-time clustering may not be adequate because cases of chronic disease may be close in space, but they are unlikely to be close in time because of long and variable periods between exposure and diagnosis. Thus, tests for space clustering may be more adequate in this case. However, previous tests for space clustering (6–8) have been derived under the unrealistic assumption that the population at risk is fairly uniform across the region. Therefore direct use of those tests would produce spurious evidence of space clustering.

This paper presents a test statistic for the detection of disease clustering in space or in time as an extension of the index C for temporal clustering proposed by Tango (9) which can adjust differences in population distribution among categories of the confounding factor such as age and sex. Recently, Whittemore et al. (10) proposed a test having the capability of adjusting variations in population distribution among demographic subgroups at different disease risk. However, their procedure, based on the statistic that is essentially identical in form to the index C, is shown to be less adequate than the method proposed in this paper.

An Index for Time Clustering

Tango (9) proposed an index C for disease clustering in time

$$C = r^t A r$$

(1)

where \(nr^t = (n_1, \ldots, n_m)\), \(n = n_1 + \ldots + n_m\), denote a vector of observed frequencies in \(m\) successive time intervals, which is assumed to be a random sample from the uniform multinomial distribution. Hence, asymptotically,

$$\sqrt{n}(r - m^{-1}l) \overset{\text{d}}{\sim} N(0, m^{-2}V(m^\frac{1}{2}))$$

(2)

where

$$V(x) = \Delta(x) - 11^t$$

(3)

and \(\Delta(x)\) is the \(m \times m\) diagonal matrix based on the vector \(\bar{x}\) and \(1\) is the \(m\)-dimensional vector of one. The entries \(\alpha_{ij}\) of \(\bar{m} \times m\) symmetric matrix \(A\) are arbitrary known measures of closeness between \(i\)th and \(j\)th interval with property \(\alpha_{ii} = 1\) and \(\alpha_{ij}\) is a monotonically nonincreasing function of \(d_{ij}\), the time distance between \(i\)th and \(j\)th interval. This index attains its maximum value of 1 if and only if \(n_i = n\) for some \(i\) and \(n_j = 0\) for \(j \neq i\). A natural selection for the form of the distance \(d_{ij}\) may be

$$d_{ij} = |i - j|.$$  

(4)

Although the choice of the form of \(d_{ij}\) may be variable depending on the situation, an exponential form

$$a_{ij} = \exp(-d_{ij})$$

(5)

has been considered.

The asymptotic distribution function of the index \(C\) under the hypothesis of no clustering in time has been, at first, derived using expansion in a series of central chi-square distribution (9):

*Division of Theoretical Epidemiology, Department of Epidemiology, The Institute of Public Health, 6-1 Shirokanedai 4-chome, Minato-ku, Tokyo 108, Japan.
Pr{C < c} = \sum_{j=0}^{\infty} \alpha_j Pr\{x_{m-1+j}^2 < (c - h)/\beta\} \tag{6}

where \(x_i^2\) denote the chi-square variable with \(g\) degrees of freedom. We shall omit the details on the parameters \(\alpha_j, h,\) and \(\beta\) here. However, this formula was not so easy to use in a simple way for more general cases.

Recently, Tango (11) suggested that a better approximation for the distribution of \(C\) may be obtained by standardizing \(C\) with

\[ T = (C - E(C)) / \sqrt{\text{Var}(C)} \tag{7} \]

and approximating it with one central chi-square distribution, i.e., the \(p\)-value for the observed value \(c\) of the index \(C\) can be approximated by

\[ Pr\{C > c\} = 1 - I\left(\frac{v + T\sqrt{2v}}{2}, \frac{v}{2}\right) \tag{8} \]

where

\[ I(x, \phi) = \int_0^x \frac{1}{\Gamma(\phi)} e^{-t^\phi-1}dt \tag{9} \]

is the incomplete gamma function and

\[ E(C) = m^{-2}[1^tA 1 + n^{-1}\text{tr}(AV(m1))] \tag{10} \]

\[ \text{Var}(C) = m^{-4}n^{-1}[d 1^tAV(m1)A 1 + 2n^{-1}\text{tr}(AV(m1))2] \tag{11} \]

and \(v\) is the degrees of freedom of approximated chi-square distribution and is given by

\[ v = 8[\sqrt{\beta_i}(C)]^{-2} \tag{12} \]

where \(\sqrt{\beta_i}(C)\) is the skewness of the index \(C\) and given by

\[ \sqrt{\beta_i}(C) = \frac{8[3 1^t(AV(m1))^2A 1 + n^{-1}\text{tr}(AV(m1))3]}{\sqrt{2n}[d 1^tAV(m1)A 1 + 2n^{-1}\text{tr}(AV(m1))2]^{1.5}} \tag{13} \]

For convenience in practical applications, the approximated upper 100\(\alpha\) percentiles \(T_\alpha\) of standardized clustering index \(T\) are given in Table 1 as a function of the skewness value \(\sqrt{\beta_i}(C)\).

### Extension of the Index

In this section we shall extend the index \(C\) so that it is applicable to disease clustering in time or in space where the overall population at risk is not uniform across the region or where there are differences in population distributions among categories of confounding factors such as age.

Let \(m\) indicates the number of points in time or in space called regions. Let \(n_i\) and \(E_i\) \((i = 1, \ldots, m)\) denote the observed number of cases and the expected number of cases in the \(i\)th region, respectively. Then, as a proper index which can measure the relative intensity of disease incidence or mortality for the \(i\)th region, the so-called O-E ratio can be used:

\[ \frac{n_i}{E_i} = \frac{\text{observed number}}{\text{expected number}} \]

One example of this quantity is the well known SMR (standardized mortality ratio), which is frequently used in epidemiological studies. Using the above quantity, an extended index can be introduced:

\[ G = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{n_i n_j}{E_i E_j} a_{ij} = q^t Aq \tag{14} \]

where \(a_{ij}\) is the same form defined by Eq. (5) and \(d_{ij}\) may be the Euclidean distance properly scaled between the \(i\)th region and the \(j\)th region for the case of space clustering problem, \(E_i\) can be computed by combining all the regions (i.e., take the standard population to be the entire population being studied), and

\[ q^t = \left(\frac{n_1}{E_1}, \ldots, \frac{n_m}{E_m}\right) \tag{15} \]

In fact, when \(E_i = E_j\) for all \(i, j\), then \(E_i = n/m\) and

\[ G = m^2 C. \tag{16} \]

Therefore, it can be said that the index \(C\) is reasonably extended to \(G\) which can accommodate the variations in the confounding factor distributions over the region. Furthermore, under the hypothesis of no clustering, we have

\[ \lim_{n/m \to \infty} E(G) = 1^t A 1. \tag{17} \]

First, let us consider the problem of disease clustering in time or in space where only the differences are the population size across the region. Let \(E_i\) denote the population size in the \(i\)th region. Then, the vector...
(n₁, ..., nₘ) can be assumed to be a random sample of size n from a nonuniform multinomial distribution with parameter p' = (p₁, ..., pₘ), where pᵢ = ξᵢ/Σₖξₖ > 0, for i = 1, ..., m. In this case, we have

\[ E_i = n p_i , \]  
(18)

and, asymptotically,

\[ \sqrt{n} (q - 1) \sim N(0, V(\mathbf{p}_{\text{inv}})), \]  
(19)

where

\[ \mathbf{p}_{\text{inv}} = (p_{1}^{-1}, \ldots, p_{m}^{-1})'. \]  
(20)

Second, consider the problem of disease clustering in space where the population size is, of course, different over the region with variations in the distributions of the confounding factor such as age.

Let K denote the number of categories in the confounding factor and let ξᵢ and nₖ denote the population size and the observed number of cases, respectively, for the ith region and the kth category of the confounding factor. Under the hypothesis that there occurs no clustering and the disease incidence rate changes across the categories of the confounding factor, the vector of the observed frequencies (n₁₁, ..., nₖₖ) for the kth category of the confounding factor can be a random sample of size n⁺ₖ = n₁₁ + ... + nₖₖ from a nonuniform multinomial distribution with parameter p'ₖ = (p₁₁, ..., pₖₖ) where pᵢₖ = ξᵢₖ/Σⱼξⱼₖ > 0, for i = 1, ..., m and k = 1, ..., K. For this case,

\[ E_i = \sum_{k=1}^{K} n_{+k}p_{i+k} \]  
(21)

and, asymptotically,

\[ \sqrt{n}(q - 1) \sim N(0, W), \]  
(22)

where W = (wᵢⱼ) is the m × m matrix with element

\[ wᵢⱼ = \begin{cases} \frac{1}{E_i} - \frac{1}{E_j} \sum_{k=1}^{K} n_{+k}p_{i+k}, & \text{for } i = j \\ \frac{-n}{E_j} \sum_{k=1}^{K} n_{+k}p_{i+k}p_{j+k}, & \text{for } i \neq j \end{cases} \]  
(23)

and n = n₁ + ... + nₖ. Needless to say, when K = 1, W = V(\mathbf{p}_{\text{inv}}). Therefore, the mean, variance and skewness values for the index G are shown to be similar in form to Eqs. (10), (11), and (13), respectively, i.e.,

\[ E(G) = 1'A 1 + n^{-1}tr(AW), \]  
(24)

\[ \text{Var}(G) = n^{-1}[41'AWA1 + 2n^{-1}tr((AW)^2)] \]  
(25)

and

\[ \sqrt{\bar{\beta}_1(G)} = \frac{8[31'(AW)^2A1 + n^{-1}tr(AW)^3]}{\sqrt{n}[41'AWA1 + 2n^{-1}tr((AW)^2)]^{1.5}} , \]  
(26)

Consequently, the procedure of approximating the asymptotic distribution of the index G under the hypothesis of no clustering can be done exactly in the same way as that for the index C, i.e., we can use the approximation of Eq. (8) where

\[ T = \frac{(G - E(G))}{\sqrt{\text{Var}(G)}} \]  
and \( v = 8(\sqrt{\bar{\beta}_1(G)})^{-2}. \)  
(27)

Needless to say, we can use Table 1 to read the approximated upper 100α percentiles of T for the extended index G.

On the other hand, Whittemore et al. (10) proposed a test statistic identical in form to the unadjusted index C even for the above-stated situation and approximated it with normal distribution. Clearly, the statistic C itself cannot be a standardized measure. Furthermore, their test has poorer power compared with the test based on the index G since they have used the matrix A as a measure of distance (11), and the normal approximation to the asymptotic distribution of the index G should be cautious because it almost always has a substantial amount of positive skewness, which was examined by Tango (11) for the detection of time clustering; it will be investigated for the detection of space clustering in detail by simulation study in the next section.

Simulation

To investigate the goodness of approximation by chi-square distribution, we performed the following Monte Carlo simulation. Situations considered here are that there are differences in the overall population size across the region, i.e., K = 1.

Step 0: As an entire population Ω, we shall consider the set of 400 points in two dimensional space defined as

\[ \Omega = \{X = (u,v) : u = 1, \ldots, 20, \]  
\[ v = 1, \ldots, 20\} \]

where each point X = (u,v) constitute the centroid of the region.

Then, repeat the following procedure, step 1 to step 3, 100 times.

Step 1: Take random sample with size m = 100 (regions) from the set Ω and assume that sampled points (X₁, ..., X₁₀₀) constitute the whole region under study. The distance \( d_{ij} \) between \( X_i \) and \( X_j \) is defined as

\[ d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}. \]

Step 2: Take m random numbers from \( N(10,2^2) \), say \( (r_1, \ldots, r_m) \). Then the value \( r_i \) is assigned to \( \xi_i (i = 1, \ldots, m) \), the population size for the ith region \( X_i \), and compute

\[ p_i = \xi_i / \sum_{j=1}^{100} \xi_j. \]

Step 3: For \( N = 20(20)100 \) and \( K = 1 \), compute the skewness value \( \sqrt{\bar{\beta}_1(G)} \) and the difference
between two kurtosis values, kurtosis value $\beta_2(G)$ and its approximated value $\beta_2(x_G^v) = 3 + 12v$, where

$$\beta_2(G) = \frac{48}{n^3} \left( \frac{1}{n} \sum (AW)^3 - 1 \right)$$

(28)

The results are given in Table 2 showing that the asymptotic distribution of the index have a substantial amount of positive skewness and that the chi-square approximation is fairly good.

Table 2. Results of Monte Carlo simulation (100 trials) for examining goodness of approximation of the distribution of the index $G$ to a chi-square distribution in the case of space clustering described in the fourth section ($m = 100$ and $K = 1$).

| N  | Mean  | Median | SD    | Minimum | Maximum |
|----|-------|--------|-------|---------|---------|
| 20 | 0.3989 | 0.3659 | 0.0144 | 0.1570  | 0.4500  |
| 40 | 0.4266 | 0.4121 | 0.0101 | 0.1572  | 0.4857  |
| 60 | 0.4830 | 0.4792 | 0.0082 | 0.1815  | 0.6107  |
| 80 | 0.5699 | 0.5671 | 0.0065 | 0.2385  | 0.7101  |
| 100| 0.7481 | 0.7754 | 0.0061 | 0.2650  | 1.1806  |

*Skewness indicates skewness value $\beta_2(G)$. $^b$Difference indicates the difference between two kurtoses, i.e., $[\beta_2(G) - (3 + 12v)]$, where $v$ is the adjusted degrees of freedom given in Eq. (27).

Applying our test to examine the level of clustering among $n = 1833$ cases of mortality from uterine cancer occurred in Tokyo metropolitan area during 1978 to 1982. Table 3 shows that the female distribution of population by age, number of deaths from uterine cancer and its SMR, $n_i/E_i$, and the latitude and longitude of the geographical centroid in each of $m = 23$ wards. The population numbers for each of $K = 7$ age groups in each of 23 wards were obtained from the 1980 Japanese census. The distance between any two different wards, $d_{ij}$, was calculated in kilometers using the following approximate formula applicable to Tokyo metropolitan area:

$$d_{ij} = \sqrt{[110.92 \times (u_i - u_j)^2 + 90.152 \times (v_i - v_j)^2]} \text{ (km)}$$

where $u_i$ and $v_i$ indicate the latitude and longitude of

| Ward   | Age         | Deaths from uterine cancer | Geographical centroid |
|--------|-------------|-----------------------------|-----------------------|
|        | Number | SMR | Latitude | Longitude |
| Chiyoda| 3937 | 4004 | 139.871 | 139.755 |
| Chuo   | 5749 | 6575 | 139.850 | 139.797 |
| Minato | 17159 | 19062 | 139.786 | 139.736 |
| Shinjuku| 33767 | 29944 | 139.736 | 139.718 |
| Bunkyo | 18123 | 16253 | 139.656 | 139.707 |
| Taito  | 12853 | 14595 | 139.604 | 139.724 |
| Sumida | 17370 | 19098 | 139.572 | 139.707 |
| Koto   | 27727 | 38313 | 139.475 | 139.709 |
| Shinagawa | 31340 | 29415 | 139.448 | 139.709 |
| Meguro | 27411 | 24525 | 139.438 | 139.709 |
| Ohta  | 54216 | 56012 | 139.466 | 139.709 |
| Setagaya| 76547 | 67062 | 139.428 | 139.709 |
| Shibuya| 26883 | 21850 | 139.402 | 139.709 |
| Nakano | 35716 | 28712 | 139.382 | 139.709 |
| Sugimino | 56188 | 44649 | 139.365 | 139.709 |
| Toshima| 29638 | 24117 | 139.348 | 139.709 |
| Kita   | 32271 | 33604 | 139.331 | 139.709 |
| Arakawa| 15054 | 15310 | 139.314 | 139.709 |
| Itabashi| 43394 | 43823 | 139.297 | 139.709 |
| Nerima | 45218 | 47755 | 139.280 | 139.709 |
| Adachi | 43501 | 55543 | 139.263 | 139.709 |
| Katsushika| 32210 | 34701 | 139.246 | 139.709 |
| Edogawa| 36605 | 46306 | 139.229 | 139.709 |

Total deaths 11 69 228 366 501 436 212 1833

Table 3. Female distribution of population by age, number of deaths from uterine cancer, and its standard mortality ratio in each of 23 Tokyo metropolitan wards during 1978 to 1982. Also shown is the latitude and longitude of the geographical centroid of each of wards.

Table 4. Results of the application of the clustering index $G$ to 1833 cases of mortality from uterine cancer in 23 Tokyo metropolitan wards during 1978 to 1982.

| Scale parameter $\lambda$ | 1   | 2   | 3   | 4   |
|---------------------------|-----|-----|-----|-----|
| $G$                       | 28.356 | 46.907 | 74.07 | 103.629 |
| $E(G)$                    | 25.964 | 44.217 | 69.668 | 97.623 |
| $\text{Var}(G)$           | 0.655 | 2.636 | 7.056 | 13.451 |
| Skewness                  | 0.318 | 0.211 | 0.156 | 0.126 |
| $T$                       | 1.007 | 1.657 | 1.657 | 1.638 |
| $p$-value*                | ~0.05 | ~0.05 | ~0.05 | ~0.05 |

*p-value is read from Table 1.
the geographical centroid of the ith ward, respectively. Maximum and minimum distance was 21.443 km and 1.555 km, respectively. As to the closeness measure, we considered
\[ a_{ij} = \exp\left( -\frac{d_{ij}}{\lambda} \right) \]
where \( \lambda \) is a scale parameter. Large \( \lambda \) will give a test sensitive to large clusters and small \( \lambda \) will give a test sensitive to small clusters. When \( \lambda = 2 \), for example, we have
\[ G = 46.907, \quad E(G) = 44.217, \quad \text{Var}(G) = 2.636, \quad \sqrt{\beta_1(G)} = 0.211 \]
and
\[ T = \frac{G - E(G)}{\sqrt{\text{Var}(G)}} = \frac{46.907 - 44.217}{\sqrt{2.636}} = 1.657 \]
If we have a good computer program for the incomplete gamma function, we can obtain the approximated \( p \)-value for the observed value of \( T \) from Eq. (8). But here, we shall use Table 1 for simplicity. By referring to the row of skewness = 0.2 (an approximation of 0.211), we can read
\[ T_{0.05} = 1.70, \quad T_{0.01} = 2.47, \quad T_{0.001} = 3.38. \]
Therefore, the \( p \)-value of \( T = 1.657 \) is slightly greater than 0.05, indicating a weak but approximately significant evidence of clustering (\( p = 0.05 \)). Results for several values of \( \lambda \), summarized in Table 4, are very similar one another. Therefore we can make an inference that some kind of space clustering may have occur for the mortality from uterine cancer during 1978 to 1982 in metropolitan Tokyo. Visual inspection of the map of SMR illustrated in Figure 1 suggests that a clustering occurs in the east of Tokyo such as Arakawa (SMR = 124), Taitoh (SMR = 119), Sumida (SMR = 122), Koto (SMR = 122), and Edogawa (SMR = 118). The result might provide a motivation for further investigation of etiologic clues that may explain the clustering of uterine cancer in this area.

Computing time for these statistics required about 4 min of NEC PC 9801 (VX 21) CPU time using a BASIC computer program that is available from the author upon request.

The author thanks S. Hashimoto for his collaboration in providing data on the population and the latitude and longitude in each of Tokyo

FIGURE 1. Standardized mortality ratio from uterine cancer in Tokyo's 23 metropolitan wards during 1978–1982.
This work was supported in part by a grant in aid for scientific research (grant no. 62530019) from the Ministry of Education, Science and Culture of Japan.

REFERENCES
1. Pinkel, D., and Nefzger D. Some epidemiological features of childhood leukemia in the Buffalo, N.Y. Area. Cancer 12: 351–358 (1959).
2. Heath, C. W., Jr., and Hasterlik, R. J. Leukemia among children in a suburban community. Am. J. Med. 34: 796–812 (1963).
3. Knox, G. Epidemiology of childhood leukemia in Northumberland and Durham. Br. J. Prev. Soc. Med. 18: 17–24 (1964).
4. Ederer, F., Myers, M. H., and Mantel, N. A statistical problem in space and time: do leukemia cases come in clusters? Biometrics 20: 626–638 (1964).
5. David, F. N., and Barton, D. E. Two space-time interaction tests for epidemicity. Br. J. Prev. Soc. Med. 20: 44–48 (1966).
6. Mantel, N. The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209–220 (1970).
7. Clark, P. J., and Evans, F. C. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35: 445–453 (1954).
8. Lewis, M. S. Spatial clustering in childhood leukemia. J. Chronic Dis. 33: 703–712 (1980).
9. Tango, T. The detection of disease clustering in time. Biometrics 40: 15–26 (1984).
10. Whittemore, A. S. Friend, N., Brown, B. W., and Holly, E. A. A test to detect clusters of disease. Biometrika 74: 631–635 (1987).
11. Tango, T. Asymptotic distribution of an index for disease clustering. Biometrics, in press.