NORTH DAKOTA STATEWIDE CANCER REGISTRY

Cancer Cluster Inquiry/Investigation Protocol[1]

Introduction

A "cluster" is an unusual aggregation, real or perceived, of health events that are grouped together in time and space and that are reported to a health agency. With respect to cancer clusters, the health event is an unusual aggregation of cancer incidence and/or mortality within a specific community. The following points highlight key concepts learned from past cluster investigations performed throughout the world:

> --In many reports of cluster investigations, a geographic or temporal excess in the number of cases cannot be demonstrated.
> --When an excess is confirmed, the likelihood of establishing a definitive cause-and-effect relationship between the health event and an exposure is slight. A cluster may be useful for generating hypotheses but is not likely to be useful for testing hypotheses. Frequently, the issues raised by a cluster cannot be definitively answered by the investigation per se; they require an alternative epidemiological approach.
> --From a public health perspective, the perception of a cluster in a community may be as important as, or more important than, an actual cluster. In dealing with cluster reports, the general public is not likely to be satisfied with complex epidemiological or statistical arguments that deny the existence or importance of a cluster. Achieving rapport with a concerned community is critical to a satisfactory outcome, and this rapport often depends on a mutual understanding of the limitations and strengths of available methods.
> [MMWR, July 27, 1990 / 39 (RR-11); 1-16]

Despite these learned lessons, an integrated and systematic plan is needed for responding to reports of cancer clusters. This plan must recognize the social dimensions of the reported cancer cluster as well as the epidemiologic analysis and tests for statistical significance. The following cancer cluster inquiry/investigation protocol is a four-stage process:

1. initial response
2. assessment
3. major feasibility study, and
4. etiologic investigation.

This four-stage process is taken from a Centers for Disease Control and Prevention (CDC), MMWR article titled "Guidelines for Investigating Clusters of Health Events", July 27, 1990. Each step of the protocol provides opportunities for collecting data and making decisions. Although this protocol may not always be followed sequentially, it provides a systematic plan with points at which the decision may be made to terminate or continue the investigation.

Organizational Requirements - Participants and Committees

A citizen reporting a potential cancer cluster must be assured that an appropriate person will be notified and timely action will be taken. The following participants and committees comprise the organizational structure of this protocol:

[1] The following is reprinted from "Guidelines for Investigating Clusters of Health Events", MMWR, July 27, 1990 / 39 (RR-11); pgs. 1-23. This reprint has been edited to meet the needs of the University of North Dakota Medical School and Health Sciences Pathology Department and the North Dakota Statewide Cancer Registry (NDSCR).

1. Principal Investigator: This individual with stature in the Pathology Department, University of North Dakota School of Medicine and Health Sciences serves as the identifiable point of responsibility for all cancer cluster investigations.
2. State Cancer Registry Co-Program Directors: Coordinates the investigation processes and communications relevant to the cancer cluster investigation. This individual serves as the initial contact person for a citizen reporting a potential cancer cluster.
3. State Cancer Registry Research Analyst: Performs quantitative and statistical analysis of data relevant to the cancer cluster investigation.
4. Ad Hoc Cluster Investigation Advisory Committee: This ad hoc committee provides oversight, guidance, and advice to the Principal Investigator and State Cancer Registry Co-Program Directors. The main focus of this committee is overseeing the cancer cluster investigation process rather than on making technical decisions. This ad hoc committee may include representatives from:
    - the North Dakota Department of Health Environmental Health State Epidemiologist,
    - the North Dakota Department of Health Public Information Officer,
    - the North Dakota Department of Health Data Processing Coordinator (Vital Records),
    - the North Dakota Department of Health Environmental representative,
    - the North Dakota Department of Health Local Health Coordinator, and
    - the North Dakota Department of Health GIS representative.
    - The committee may also include representatives from other government agencies, concerned citizens' groups, the media, and selected individuals as determined as necessary.

These above individuals participate in various aspects of a cancer cluster investigation. The following investigation protocol describes each individual's role.

Systematic Cluster Investigation Protocol - 4 Stage Process

Stage 1 - Initial Contact

Purpose: To collect information from the person(s) or group(s) first reporting perceived cluster.

The initial contact is crucial. The caller should be referred quickly to the State Cancer Registry Co-Program Directors who in turn will contact the registries research analyst. Most reports of potential cancer clusters can be successfully closed at the time of initial contact. This first encounter is the best opportunity for communication with the caller about the nature of the cancer cluster investigation protocol. The State Cancer Registry Co-Program Directors with assistance of the registries research analyst is responsible for the following activities during this time of initial contact:

A. Gather identifying information on the caller, unless anonymity is requested: name, address, telephone number, and organization affiliation, if any. If anonymity is requested, advise the caller that the inability to follow up may hinder further investigation.

B. Gather initial data on the potential cluster: suspected health event(s), suspected exposure(s), number of cases, geographic area of concern, time period of concern, and how the caller learned about the cluster.

C. Obtain identifying information on persons affected: name, sex, age (or birth date, age at diagnosis, age at death), occupation, race, diagnosis, date of diagnosis, date of death, address (or approximate geographic location), telephone number, length of time in residence at site of interest, contact person (family, friend) and method for contact, and physician contact. In some instances, the health official may choose not to collect identifying information during the first contact but instead to gather it during several contacts.

D. Discuss initial impressions with a caller. The following frequently arise:

--A variety of diagnoses speaks against a common origin.

--Cancer is a common illness (with a one in three lifetime probability). The risk increases with age, and cases among older persons are less likely to be true clusters.

--Major birth defects are less common than cancer but still occur in 1%-2% of live births.

--Length of time in residence must be substantial to implicate a plausible environmental carcinogen because of the long period of latency required for most known carcinogens.

--Cases that occurred among persons now deceased may not be helpful in linking exposure to disease because of the lack of information on exposure and because of possible confounding factors.

--Rare diseases may occasionally "cluster" in a way that is statistically significant, but such an occurrence may be a statistical phenomenon not related to exposure.

E. Request further information on cases, obtain more complete enumeration, and plan a follow-up telephone contact, as needed.

F. Assure the caller that he or she will receive a written response. (Often, the written response simply confirms what has already been communicated by telephone.)

G. Maintain a log of initial contacts, whether they are made in writing, by telephone, or in person. The log should include the date, time, caller identification, health event, exposure, and geographic area (see Appendix B for sample paper log/file). Follow-up contacts should be logged in as well, with a brief note as to purpose and result.

H. Notify the Principal Investigator about the contact.

Outcome: If initial contact suggests further evaluation is needed (e.g. single and rare disease entity, plausible exposure, or plausible clustering), then proceed to Stage 2. If initial contact permits satisfactory closure, then prepare a written correspondence summarizing the call/case for the caller and for the advisory committee.

Stage 2 - Assessment

Once the decision has been made to proceed with an assessment, an important step is to separate two concurrent issues: whether an excess has actually occurred and whether the excess can be linked etiologically to some exposure. The first issue usually has precedence, and it may or may not lead to the second. This stage initiates a mechanism for evaluating whether an excess has occurred. Three separate elements are identified:

(Stage 2a) a preliminary evaluation to assess quickly from the available data whether an excess may have occurred;

(Stage 2b) case evaluation to assure that a biological basis exists for further work;

(Stage 2c) an occurrence investigation for the purpose of obtaining a more detailed description of the cluster through case finding, interaction with the community, and descriptive epidemiology.

Additionally, a review of the scientific literature and consultation with other investigators are useful. These activities are often interrelated and may occur in parallel.

Stage 2a. Preliminary evaluation

Purpose: to provide a quick, rough estimate of the likelihood that an important excess has occurred.

The Research Analyst is responsible for the following activities during the preliminary evaluation:

A. Determine the appropriate geographic area and the period in which to study the cluster.

B. Determine which cases will be included in the analysis. A helpful step may be to tabulate frequencies of health events and to look at related descriptive statistics.

C. Determine an appropriate reference population. Occurrence rates (or other statistics) calculated for the cluster should be compared with those for a reference population in order to identify an excess number of cases.

D. If the number of cases is sufficient, and if a denominator is available (e.g., population of a community, number of children in school, or number of employees in a workplace), calculate occurrence rates, standardized morbidity/ mortality ratios, or proportional mortality and/or incidence ratios (see Section 4). Compare the calculated statistic with that for the reference population to assess significance. Chi-square tests and Poisson regression are also commonly used techniques for comparing proportions.

E. If the number of cases is not large enough to obtain meaningful rates, or if denominator data are unavailable, use one of the statistical tests developed to assess space, time, or space-time clustering (See Section on Statistical and Epidemiologic Techniques).

Outcome: If the preliminary evaluation suggests an excess occurrence, proceed to case evaluation. If the preliminary evaluation suggests no excess, respond to the caller, indicating findings and advising that no further investigation is needed. If the preliminary evaluation shows no excess but the data suggest an occurrence of biologic and public health importance, decide if further assessment is warranted. A decision to proceed further at this point should not be based solely on an arbitrary criterion for statistical significance.

Stage 2b. Case evaluation

Purpose: to verify the diagnosis.

The State Cancer Registry Co-Program Directors are responsible for the following activities during the case evaluation:

A. Verify the diagnosis by contacting the responsible physicians or by referring to the appropriate health-event registry. Verification is often a multi-step process, involving initial contact with the patient, family, or friends and subsequent referral to the responsible physicians to obtain permission to examine the records.

B. If possible, obtain copies of relevant pathology reports or medical examiner's report.

C. Obtain histologic reevaluation if needed.

D. Notify the Principal Investigator, the Chief Medical Officer, the Public Health information Officer and the Local Health Department Coordinator about the findings of the case evaluation stage (stage 2b).

Outcome: If cases are verified and an excess is confirmed, proceed to Stage 2c, the occurrence evaluation (which already may be under way). If some (or all) of the cases are not verified and an excess is not substantiated, respond to the caller, outlining findings and advising that further evaluation is not warranted. If some of the cases are not verified but biologic plausibility persists and the data are suggestive, consider initiating or continuing the occurrence evaluation.

Stage 2c. Occurrence evaluation

Purpose: to design and perform a thorough investigation to determine if an excess has occurred and to describe the epidemiologic characteristics.

The occurrence evaluation is meant to define the characteristics of the cluster, often requiring a field investigation. This evaluation begins with a written protocol that outlines the costs and provides information on data collection, the methods to be used, and the plan of analysis. The Chief Medical Officer is responsible for coordinating the following activities. The Ad Hoc Cluster Investigation Advisory Committee can be helpful at this point in the cluster investigation process.

A. Identify and convene the Ad Hoc Cluster Investigation Advisory Committee.

B. Determine the most appropriate geographic (community) and temporal boundaries.

C. Ascertain all potential cases within the defined space-time boundaries.

D. Identify the appropriate databases for both numerator and denominator and their availability.

E. Identify statistical and epidemiologic procedures to be used in describing and analyzing the data.

F. Perform an in-depth review of the literature, and consider the epidemiologic and biologic plausibility of the purported association.

G. Assess the likelihood that an event-exposure relationship may be established.

H. Review available environmental data; perform rapid environmental assessment.

I. Assess community perceptions, reactions, and needs.

J. Complete the proposed descriptive investigation.

K. Notify the Principal Investigator, the Chief Medical Officer, the Public Health Information Officer, the Local Health Department Coordinator, and members of the Ad Hoc Cluster Investigation Advisory Committee about the findings of the occurrence evaluation stage (stage 2c).

Outcome: If an excess is confirmed and the epidemiologic and biologic plausibility is compelling, proceed to Stage 3, the major feasibility study. If an excess is confirmed but no relationship to an exposure is apparent, terminate the investigation and inform the persons concerned of the possible risks/no risks involved. If an excess is not confirmed, terminate the investigation and report findings to the caller.

Stage 3 - Major Feasibility Study

Purpose: to determine the feasibility of performing an epidemiologic study linking the health event and a putative exposure.

The major feasibility study examines the potential for relating the cluster to some exposure. All of the options for geographic and temporal analysis should be considered, including the use of cases that were not part of the original cluster and are of a different geographic locale or time period. The major feasibility study is determined by the Chief Medical Officer with consultation

from the Ad Hoc Cluster Investigation Advisory Committee. Things to consider when undertaking a generic major feasibility study include:

A. Consult with the Centers for Disease Control and Prevention (CDC) Cancer Control Program officials.

B. Review the detailed literature search with particular attention to known and putative causes of the outcome(s) of concern.

C. Consider the appropriate study design, with attendant costs and expected outcomes of alternatives (e.g., a consideration of sample size, the appropriateness of using previously identified cases, the geographic area and time period concerned, and the selection of controls).

D. Determine what data should be collected on cases and controls, including physical and laboratory measurements.

E. Determine the nature, extent, and frequency of and the methods used for environmental measurements.

F. Delineate the logistics of data collection and processing.

G. Determine the appropriate plan of analysis, including hypotheses to be tested and power to detect differences; assess the epidemiologic and policy implications of alternative results.

H. Assess the current social and political ambiance, giving consideration to the impact of decisions and outcomes.

I. Assess the resource implications and requirements of both the study and alternative findings.

Outcome: If the feasibility study suggests that an etiologic investigation is warranted, proceed to Stage 4. The investigation may require extensive resources, however, and the decision to proceed will be related to the allocation of resources. If the feasibility study suggests that little will be gained from an etiologic investigation, summarize the results of this process (by now rather extensive) in a report to the caller and all other concerned parties. In some circumstances the public or media may continue to demand further investigation regardless of cost or biologic merit. The effort devoted to community relationships, media contacts, and advisory committee interaction will be critical for an appropriate public health outcome.

Stage 4 - Etiologic Investigation

Purpose: to perform an etiologic investigation of a potential disease-exposure relationship.

The primary purpose of the study is to pursue the epidemiologic and public health issues that the cluster generated--not necessarily to investigate a specific cluster. In that context, this step is a standard epidemiologic study, for which all the preceding effort has been preparatory. The etiologic investigation is determined by the Chief Medical Officer with consultation from the Ad Hoc Cluster Investigation Advisory Committee. The circumstances of most epidemiologic studies tend to be unique; therefore the following serves as a generic one-step guide:

A. Using the major feasibility study as a guide, develop a protocol, and implement the study.

Outcome: The results of an etiologic investigation are expected to contribute to epidemiologic and public health knowledge. This contribution may take a number of forms, including the demonstration that an association does or does not exist between exposure and disease, or the confirmation of previous findings.

Statistical and Epidemiologic Techniques

The approach taken to investigate a suspected cluster of health events depends on the nature of the cluster, the data available, and the questions being asked, including the following:

--Do the health events cluster in space or time alone, in space and time simultaneously, or in neither?
--What are the spatial and temporal boundaries of the cluster?
--What are the characteristics of the health events (e.g., acute or chronic disease, long or short latency period, and known or unknown etiology)?
--What data are available for the health event (e.g., case counts, disease rates, or data on each event, such as place of residence and time of onset of disease or death)?
--What data are available to describe the population at risk?

A number of problems are encountered in the study of clusters. The health events being investigated (often morbidity or mortality) are usually rare, and increases of these events tend to be small and may occur over a long period. Another issue that complicates the investigation is that some clusters occur by chance. Information on the population at risk or on the expected rates often is not available. A further complicating factor for methods using aggregated data is that health events occur in space and time continua, thus yielding optional and sub-optimal units for displaying a pattern. The choice of a geographic area that is too small or too large, or of a time period that is too short or too long, may result in insufficient statistical power to indicate a cluster. Many of the articles referenced in the Appendix contain informative discussions about issues that can compromise application of statistical methods in investigations of clusters.

Standard statistical and epidemiologic techniques for assessing excess risk can often be used to evaluate reported clusters. Tabulating frequencies of the health event and examining related descriptive statistics is a useful first step in the evaluation. Mapping the data is also helpful. If the number of cases is sufficient and population data are available, examination of rates (possibly age-, race-, and sex-adjusted), standardized mortality/morbidity ratios, or proportional mortality ratios may determine whether there is an excess number of events. If the number of health events is too small to show meaningful rates, pooling across geographic areas or time may be possible. Combinatorial methods are often used for small amounts of data. Other commonly used statistical approaches include chi-square tests of observed versus expected frequencies (based on the Poisson distribution for low-frequency data) and Poisson regression (used for comparison of rates). Confidence intervals may be calculated for point estimates.

Whether the rate for a geographic area or time period is excessive may be determined by comparing it with rates of other areas or times. If a spatial cluster is being assessed, the occurrence in the geographic area can be compared with that in adjacent areas (e.g., a census tract with surrounding census tracts) or with other areas of similar size (e.g., a county with other counties in a state). Alternatively, the rate for an area can be compared with that of a larger area (e.g., the rate for a city with that of the surrounding county). If a temporal cluster is being assessed, the occurrence in that time period can be evaluated in the context of previous or subsequent periods. When such comparisons are made, the referent population must be chosen carefully to ensure its appropriateness. Mortality and morbidity data for referent populations are available from state and national vital statistics systems or registries such as cancer and birth defect registries. Population data are available from the Bureau of the Census. A county-level file with both mortality and population data for 1968-1985 (the Compressed Mortality File) is available for public use from the National Technical Information Service.

If the above standard approaches cannot be used in an investigation of clusters because the number of health events is too small, data on the population at risk are unavailable, or space-time

clustering is suspected, numerous statistical tests are available for use in detecting spatial, temporal, and space-time clusters. Although some of these tests may not be familiar to investigators and may require the preparation of more data than required by standard techniques, many of the tests are simple to understand and use. Numerous methods for studying clusters have been reviewed (22,23). Brief descriptions and critiques of some of these techniques are presented in the Appendix.

Most of the tests reviewed in the Appendix use data on individual cases of health events, although a few employ aggregated data such as frequency counts or rates. Information generally required for each case is location of the case (often the geographic coordinates of place of residence) and date of onset of the disease (or injury) or of death. Most of the tests based on aggregated data assume that the number of health events that occur in an area and/or time period follows a Poisson distribution. The tests do not usually require knowledge of the distribution of the population at risk. Instead, they may assume that the population at risk remains constant over time, and they offer special considerations for differing population sizes. The reporting rate for the health event is also assumed to be constant.

The assumption of minimal population shifts over time is frequently violated. More subtly, subgroups of the population with different levels of risk may not remain constant over the time period of interest. Violations of these assumptions can lead to spurious results. An additional problem is encountered when investigators study the occurrence of health events over a long period, i.e., the problem posed by migration. Migration tends to decrease the chance of detecting clustering; however, certain tests account for non-uniformity of or changes in the population (24-26). As an alternative, adjustments for the size of the population at risk (to account for population changes during the study period) can be made before testing.

In addition to the techniques described in the Appendix, other approaches in use or under investigation for the analysis of clusters include the quality control measure known as the cumulative sum, or cusum, technique (27), the sets technique (28), nearest-neighbor procedures (29,30), and nonlinear and Bayesian time series methods. Normal-theory confidence intervals and bootstrap-prediction intervals for detecting frequencies of disease occurrence above those expected have been explored (31).

Because of the diverse and complicated nature of clusters, there is no omnibus test for assessing them. Investigators are advised to perform several related tests and to report the results that are most consistent with validated assumptions. This process will be aided by the use of CLUSTER, an IBM PC-compatible software program that will soon be commercially available and will offer investigators a choice of statistical procedures to use when investigating clusters.

References

22. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res 1967;27:209-20.

23. Klauber MR. Space-time clustering analysis: a prospectus. Proceedings of a SIMS Conference on Epidemiology, Alta, Utah, July 8-12, 1974.

24. Weinstock MA. A generalized scan statistic test for the detection of clusters. Int J Epidemiol 1981;10:289-93.

25. Whittemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. Biometrika 1987;74:631-5.

26. Cuzick JC, Edwards R. Spatial clustering for inhomogeneous populations. Proceedings of the Joint Statistical Meetings, Washington, D.C., August 7, 1989.

27. Woodward RH, Goldsmith PL. Cumulative sum techniques. ICI Monograph No. 3, Oliver and Boyd.

28. Chen R. A surveillance system for congenital malformations. Journal of the American Statistical Association 1978;73:323-7.

29. Cliff AD, Ord JK. Spatial processes, models and application. London: Pion, 1981.

30. Diggle PJ. Statistical analysis of spatial patterns. London: Academic Press, 1983.

31. Stroup DF, Williamson GD, Herndon JL, Karon JM. Detection of aberrations in the occurrence of notifiable diseases surveillance data. Statistics in Medicine 1989;8:323-9.

APPENDIX A: Summary of Methods for Statistically Assessing Clusters of Health Events

The following summaries are provided as a resource to investigators who may become involved with the statistical aspects of reported clusters of health events and who are not likely to have a direct effect on the day-to-day management of the clusters.

TEMPORAL CLUSTERING

Ederer, Myers, and Mantel approach

Ederer, Myers, and Mantel (1) developed a test for temporal clustering using a cell-occupancy approach. They divided the time period into k disjoint subintervals. Under the null hypothesis of no clustering, the n cases are randomly distributed among the subintervals (i.e., are multinomially distributed). The test statistic m is the maximum number of cases occurring in a subinterval. If the health event is rare and of unknown etiology, m is summed over several locations and time periods. The sum is tested by using a single degree of freedom chi-square test. Ederer, Myers, and Mantel (1) and Mantel, Kryscio, and Myers (2) provide tables of the exact null distribution of m for selected values of k and n.

Scan Test

Naus proposed a test of temporal clustering that is known as the scan test (3). The test statistic, the maximum number of cases observed in an interval of length t, is found by "scanning" all intervals of length t in the time period (resulting in overlapping intervals). In certain cases, this approach is intuitively more appealing than the disjoint interval approach of Ederer, Myers, and Mantel (1), but more complicated mathematically. However, situations exist for which the disjoint interval approach is the more satisfactory choice. Statistical significance of the scan test is assessed by using tables of p-values calculated by Naus (4) and Wallenstein (5) for selected interval lengths, time lengths, and sample sizes. Unfortunately, the computations necessary to obtain other exact p-values for the scan statistic are complex and often not feasible. However, Knox and Lancashire (6) have derived a set of relatively simple formulas for an approximation to the exact p-value.

Naus compared the power of the scan test with that of the Ederer, Myers, and Mantel test and concluded that if the scanning interval is small and the data are continuous over the interval, the scan test is the more powerful of the two (7). Weinstock proposed a generalization of the scan test that adjusts for changes in the population at risk (8).

Bailar, Eisenberg, and Mantel Test of Temporal Clustering

Bailar, Eisenberg, and Mantel suggested a test of temporal clustering based on the number of pairs of cases in a given area that occur within a specified length of time d of each other (9). The numbers of close pairs occurring in q areas are summed. The test statistic is assumed to be approximately normally distributed.

Larsen Test

Larsen, Holmes, and Heath developed a rank order procedure for detecting temporal clustering (10). The time period is divided into disjoint subintervals that are numbered sequentially (i.e., ranked). The test statistic K is the sum of absolute differences between the rank of the subinterval in which a case occurred and the median subinterval rank. Small values of K indicate unimodal clustering. Generally, the K statistics for multiple geographic areas are summed. The resulting statistic is asymptotically normal with simple mean and variance. This test is sensitive only to unimodal clustering; it cannot distinguish multiple clustering from randomness.

Tango Clustering Index

Tango developed a test of temporal clustering based on the distribution of counts in disjoint equal time intervals (11). The test is useful when the data are grouped. The test statistic (cluster index) is a quadratic form involving the relative frequencies in each interval and a measure of distance between intervals. The clustering index obtains a maximum value of 1 when all cases occur in the same interval. Although the statistic is easy to calculate, the asymptotic distribution using Tango's formula is not. However, Tango will provide upon request an algorithm written in BASIC to obtain the asymptotic distribution.

Whittemore and Keller showed that the distribution of Tango's index is asymptotically normal with simple mean and variance (12).

SPATIAL CLUSTERING

Geary Contiguity Ratio

Geary developed a test of spatial clustering that assesses whether rates for adjacent areas are more similar than would be expected if they were randomly distributed among the geographic areas (13). The test statistic, the contiguity ratio, is the ratio of the sum of mean squared differences between rates for pairs of adjacent areas to the weighted sum of mean squared differences between rates for all pairs of areas. If the rates are geographically distributed at random, the contiguity ratio is close to one; otherwise, it is less than one. Geary derived an expression for the approximate variance of the ratio. If the number of areas is not too small, the ratio is asymptotically normally distributed. Hechtor and Borhan provide another computational formula for the statistic (14).

Ohno, Aoki, and Aoki Test

Ohno, Aoki, and Aoki (15) and Ohno and Aoki (16) developed a simple test for spatial clustering that uses rates for geographic areas (e.g., census tracts, counties, or states) rather than data for individual cases. The test assesses whether the rates in adjacent areas are more similar than would be expected under the null hypothesis of no clustering.

For this test, the rate for each area is classified into one of n categories, and each pair of adjacent areas is identified. The test statistic is the number of adjacent concordant pairs--i.e., the number of pairs of areas that are adjacent and have rates in the same category. An overall clustering measure (summed across all categories) can be obtained as well as category-specific clustering measures. The observed number of adjacent concordant pairs is compared with the expected number by using a chi-square test. Ohno, Aoki, and Aoki provide a simple formula for calculating the expected number of pairs (15).

Grimson Test

Grimson, Wang, and Johnson proposed a test of spatial clustering for use in detecting clusters of geographic areas designated as high risk (17). The null hypothesis is that high-risk areas are randomly distributed within a larger area and do not cluster.

Given n high-risk areas, the test statistic is the number of pairs of high-risk areas that are adjacent to each other. This statistic is equivalent to the category-specific statistic from Ohno, Aoki, and Aoki (15). Grimson et al. recommended using a simple Monte Carlo simulation to obtain p-values for the test statistic (17).

Whittemore Test

Whittemore, Friend, Brown, and Holly developed a test for spatial clustering across geographic areas that adjusts for different distributions of population subgroups across the region (18). Thus, the test requires population data. The test statistic is the mean distance between all pairs of cases, and can be expressed as a generalization of Tango's clustering index--i.e., a quadratic form involving relative frequencies from subgroups and a matrix of distances between pairs of areas. The statistic is asymptotically normal (mean and variance derived), and the test has good power when disease rates for all subgroups are elevated in the same areas. Power is poor when areas with elevated rates vary for subgroups. The test also has poor power when clusters occur in more than one area. The test can be adapted to detect temporal clustering when the distance matrix represents distances between pairs of time intervals.

Cuzick and Edwards Test

Cuzick and Edwards proposed a test for spatial clustering that applies to populations with non-uniform population density (19). The test involves drawing a set of controls from the population and combining them with the cases. Cuzick and Edwards propose two nearest-neighbor tests. The statistic for the first test is the number of persons in the case group whose nearest neighbor also is in the case group. The second test statistic is the sum of the number of cases among the K nearest neighbors for each person who is in the case group. This second test will be more powerful when a few large clusters exist, whereas the first test is more powerful when many small clusters are involved. Cuzick and Edwards provide formulas for the mean and variance and establish asymptotic normality for the test statistics.

SPATIAL AND TEMPORAL CLUSTERING

Pinkel and Nefzger Cell Occupancy Approach

In 1959, Pinkel and Nefzger proposed a cell occupancy approach to test for spatial-temporal clustering (20). Assuming that r cases are randomly allocated to m space-time cells, these investigators developed an exact test for determining the probability of observing k "close" cases (i.e., cases occurring within a specified distance and length of time of each other).

For this test, the study area and time period are divided into space-time cells based on the space and time distances used to define closeness. The test is sensitive not only to space-time clustering but also to spatial clustering or temporal clustering alone, a property that is not desirable (21).

Knox 2 x 2 Contingency Table Test

Knox developed a space-time clustering test that involves dichotomizing the spatial and temporal dimensions (22,23). A 2 x 2 contingency table is formed by classifying the n(n-1)/2 pairs of cases as close in space and time, close in space only, close in time only, or close in neither space nor time.

The test statistic X, the observed number of pairs close in both space and time, is assumed to be approximately Poisson (since although pairs are dependent, X is small compared with the total number of pairs).

Barton and David concluded that, although use of the Poisson approximation is appropriate in some situations, in general it could yield misleading results (24). Mantel outlined methodology for obtaining the exact permutational distribution of X (21).

Barton and David Points-on-a-Line Approach

Barton, David, and Herrington (25) and David and Barton (26) adapted an earlier test (27) for use in detecting space-time interaction. The test, analogous to analysis of variance, involves the ratio of within-group variance to overall variance. Pairs of cases separated in time by less than a specified length of time are formed into time clusters (i.e., treatment groups).

The test statistic Q is the ratio of the average squared geographic distance between pairs of cases within clusters to the average squared distance between all pairs of cases. Under the null hypothesis of no space-time interaction, one would expect this ratio to be 1. When clustering is present, Q is smaller than 1. To assess significance, David and Barton suggested using a randomization test to determine the exact distribution of Q (26). Since calculation of the exact distribution often is not feasible, Barton and David suggested using a beta approximation when the number of cases is small and a normal approximation when the number of cases is large (28). When the number of clusters is large, Q is approximately normally distributed; otherwise, an F approximation is more appropriate.

An advantage of Barton and David's test is that actual distances are used, and the only arbitrariness is in the selection of the critical time point. A disadvantage of the test is that the small distances, which are of most interest, have less influence on the statistic than do the large distances. In fact, the large distances may so dominate the statistic that they mask any clustering.

Mantel Generalized Regression Approach

Mantel developed a "generalized regression" approach to the detection of clustering in space and time (21). The test statistic Z is the sum over all pairs of cases of a function of the distance between two cases multiplied by a function of the time between two cases. Knox's test can be derived as a special case of Mantel's test. Mantel recommended using reciprocal transformations of the distances to increase the influence of close distances and decrease the influence of long distances. Mantel (21) and Siemiatycki (29) concluded that the test has low power if no transformation is made.

A constant must be added to the distances before making the reciprocal transformation because of the possibility of very small or zero time and/or space distances. Unfortunately, the constants chosen influence the value of the test statistic and the outcome of the test of significance if the normal approximation is used. Mantel suggested that, for best results, the constants be close to the expected distances between close pairs. Glass, Mantel, Guns, and Spears (30) and Siemiatycki (29) found that as the size of the constants increases, the test statistic tends to decrease.

A test of statistical significance is obtained by obtaining the exact randomization distribution of Z, by using Monte Carlo simulation to obtain an approximation to the distribution of Z, or by assuming that Z is asymptotically normally distributed (Mantel derived expressions for the measured variance) (21). Klauber (31) and Siemiatycki (29) found the distribution of Z to be highly skewed and showed that although the use of the normal approximation is appropriate when Z is highly significant or nonsignificant, its use is inappropriate when Z has borderline significance.

One asset of Mantel's test is that actual space and time distance are used, thus avoiding arbitrary cutpoints and loss of information. Another advantage to this approach is its applicability to two or more samples (31,32).

Pike and Smith Extension to Knox Test

Pike and Smith extended Knox's test to diseases with long latent periods by defining a geographic area and period of time of infectivity and susceptibility (33). Pairs of cases are

considered close in space if their geographic areas of infectivity and susceptibility overlap, and close in time if their periods of infectivity and susceptibility overlap. The test statistic is the number of pairs close in both space and time.

Lloyd and Roberts Test

Lloyd and Roberts outlined a test for either spatial or temporal clustering that Smith and Pike noted in 1974 can be viewed as a special case of Knox's test (34). Lloyd and Roberts suggested using the number of pairs among all possible pairs of cases that are close in time (or in space) as the test statistic. A test of significance is obtained by calculating the mean number of close pairs for sets of randomly selected controls and by assuming a Poisson distribution with this mean. Smith and Pike indicated that the randomization distribution of the test statistic could be obtained, and they suggested that matched controls be used in the procedure (35).

References

1.Ederer F, Myers MH, Mantel N. A statistical problem in space and time: do leukemia cases come in clusters? Biometrika 1964;20:626-38.

2. Mantel N, Kryscio RJ, Myers MH. Tables and formulas for extended use of the Ederer-Myers-Mantel disease clustering procedure. Am J Epidemiol 1976;104:576-84.

3. Naus JI. The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association 1965;60:532-8.

4. Naus JI. Some probabilities, expectations, and variance for the size of the smallest intervals and largest clusters. Journal of the American Statistical Association 1966;61:1191-9.

5. Wallenstein S. A test for detection of clustering over time. Am J Epidemiol 1980;111:367-72.

6. Knox EG, Lancashire R. Detection of minimal epidemics. Stat Med 1982;1:183-9.

7. Naus JI. A power comparison of two tests of non-random clustering. Technometrics 1966;8:493-517.

8. Weinstock MA. A generalized scan statistic test for the detection of clusters. Int J Epidemiol 1981;10:289-93.

9. Bailar III JC, Eisenberg H, Mantel N. Time between pairs of leukemia cases. Cancer 1970;25:1301-3.

10. Larsen RJ, Holmes CL, Heath CW. A statistical test for measuring unimodal clustering: a description of the test and of its application to cases of acute leukemia in metropolitan Atlanta, Georgia. Biometrics 1973;29:301-9.

11. Tango T. The detection of disease clustering in time. Biometrics 1984;40:15-26.

12. Whittemore A, Keller JB. A letter to the editor. On Tango's index of disease clustering in time. Biometrics 1986;42:218.

13. Geary RC. The contiguity ratio and statistical mapping. Incorp Statist 1954;5:115-45.

14. Hechtor HH, Borhan NO. Mortality and geographic distribution of atherosclerotic heart disease. Public Health Rep 1965;80:11-24.

15. Ohno Y, Aoki K, Aoki N. A test of significance for geographic clusters of disease. Int J Epidemiol 1979;8:273-81.

16. Ohno Y, Aoki K. Cancer deaths by city and county in Japan: a test of significance for geographic clustering of disease. Soc Sci Med 1981;15D:251-8.

17. Grimson RC, Wang KC, Johnson PWC. Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome. Soc Sci Med 1981;15D:287-93.

18. Whittemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. Biometrika 1987;74:631-5.

19. Cuzick JC, Edwards R. Spatial clustering for inhomogeneous populations. Proceedings of the Joint Statistical Meetings, Washington, DC, August 7, 1989.

20. Pinkel D, Nefzger D. Some epidemiological features of childhood leukemia in the Buffalo, NY area. Cancer 1959;12:351-7.

21. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Research 1967;27:209-20.

22. Knox G. Detection of space-time interactions. Applied Statistics 1964;13:25-30.

23. Knox G. Epidemiology of childhood leukemia in Northumberland and Durham. Br J Prev Soc Med 1964;18:17-24.

24. Barton DE, David FN. The random intersection of two graphs. In: David FN, ed. Research papers in statistics. New York: John Wiley & Sons Inc, 1966:45-59.

25. Barton DE, David FN, Herrington M. A criterion for testing contagion in time and space. Ann Hum Genet 1965;29:97-103.

26. David FN, Barton DE. Two space-time interaction tests for epidemicity. Br J Soc Med 1966;20:44-8.

27. Barton DE, David FN. The analyses of chromosome patterns in the normal cell. Ann Hum Genet 1962;25:323-32.

28. Barton DE, David FN. Randomization basis for multivariate tests. I. The bivariate case: randomness of n points in a plane. Bull Int Stat Inst 1962;i39 II:455-67.

29. Siemiatycki J. Mantel's space-time clustering statistic: computing higher moments and a comparison of various data transforms. Journal of Statistical Computation and Simulation 1978;7:13-31.

30. Glass AG, Mantel N, Guns FW, Spears GFS. Time-space clustering of childhood leukemia in New Zealand. J Natl Cancer Inst 1971;47:329-36.

31. Klauber MR. Two sample randomization test for space-time clustering. Biometrics 1971;27:129-42.

32. Klauber MR. Space-time clustering tests for more than two samples. Biometrics 1975;31:719-26.

33. Pike MC, Smith PG. Disease clustering: a generalization of Knox's approach to the detection of space-time interactions. Biometrics 1968;24:541-6.

34. Lloyd S, Roberts CJ. A test for space clustering and its applications to congenital limb defects in Cardiff. Br J Prev Soc Med 1973;27:186-91.

35. Smith PG, Pike MC. A note on a "close pairs" test for space clustering. Br J Prev Soc Med 1974;28:63-4. *Representatives of the Association of State and Territorial Health Officials, the Agency for Toxic Substances and Disease Registry, and CDC, consisting of Carl W. Armstrong, M.D., David H. Culver, Ph.D., Richard L. Ehrenberg, M.D., Patricia A. Honchar, Ph.D., Dedun Ingram, Ph.D., Jeffrey A. Lybarger, M.D., Stanley I. Music, M.D., Richard B. Rothenberg, M.D., Karen K. Steinberg, Ph.D., Stephen B. Thacker, M.D., and G. David Williamson, Ph.D. **Timothy E. Aldrich, Ph.D., Alan P. Bender, D.V.M., Valerie Beral, M.D., Glyn G. Caldwell, M.D., Beth Fiore, Roger C. Grimson, Ph.D., Clark W. Heath, Jr., M.D., Dennis M. Perotta, Ph.D., and Lowell E. Sever, Ph.D.

APPENDIX B: Sample Paper Log/File for Recording Initial Contact (Stage 1) Information

NDSCR
NORTH DAKOTA STATEWIDE CANCER REGISTRY

CANCER CONCERN IDENTIFICATION FORM

NORTH DAKOTA STATEWIDE CANCER REGISTRY
University of North Dakota School of Medicine and Health Sciences
Department of Pathology
SFN 51874 (Revised 11-12)

Inquiry received from:	
Facility/Clinic	
Mailing Address	
City, State, Zip	
Phone	

NDSCR Use Only

Date Received: _____

Date Inquiry Closed: _____

Completed by: _____

Issues of Concern	
ND Geographic area cancer located	
Length of time involved with cancer (number of months/years)	
Info on persons affected -	
Name	
Sex	
Race	
Date of birth	
Occupation *[parents occupation if reporting on a child]*	
Contact Person *[family/frield]*	
Address - street, city, county	
Length of time living at address at time of diagnosis	
Telephone number	
Type of cancer	
Diagnosis date	
Suspected exposure	
Physician Name	
Medical facility/Physician address	
Other information	