MEASUREMENT OF TEMPORO-SPATIAL CLUSTERING OF DISEASE

A Thesis

Presented to

The Faculty of Graduate Studies

University of Manitoba

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

by

Paul Tin-gen Ma

February 1969

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I wish to express sincere appreciation to Dr. K. Subrahmanian, my major advisor, and to Dr. G. I. Paul for their guidance throughout the course of the work which enabled me to complete this thesis.

To Dr. N. W. Choi, Associate Professor, Department of Social and Preventive Medicine, University of Manitoba, a special acknowledgement of gratitude for providing the material upon which this study was based, as well as for his enthusiastic encouragement throughout the entire study.

SUMMARY OF THE THESIS

Ederer-Myers-Mantel approach is one of the statistical models for detecting the occurrence of disease which might be clustered in time and space. The modifications of this model are suggested and the comparisons among them are discussed.

1,111 congenital malformation cases were reported in Manitoba between June 1, 1964 to December 31, 1967. Mothers of 109 congenital malformations of neural-tube defects (live birth and still birth) were the residents of Metropolitan Winnipeg. All the necessary information about those cases were obtained from Manitoba Congenital Anomalies Registry reports and from the hospital medical records.

The malformed babies, with the same conception date, appear to occur grouped together in space and time because of the influence of suspect agents. This hypothesis is tested by the Ederer-Myers-Mantel model and its modifications. Using the same data, Ederer-Myers-Mantel approach and modified models result in different conclusions about the case aggregations within the time-space clusters. Some possible explanations for such disparities in conclusions are presented.

CHAPTER I

MOTIVATION FOR THE PROBLEM

In medical research, the etiology of some diseases such as cancer and congenital defects is not yet well established. The study about the occurrence of a disease in clusters in terms of time and space provides us with many clues to the etiology. It is usually found that for the diseases at rare occurrence, the usual independent study of time and space clustering may not necessarily provide us with very satisfactory ideas about the etiology of the disease.

Recently, the observance of congenital malformation incidences in clusters within small areas during a short period of time has been suggested by some workers as evidence for a viral-like spread of the diseases. However, the following areas also are indicative of having some possible etiologic involvement with congenital malformations.

Nutrition and Chemicals

It has been shown experimentally in rats that excessive Vitamin A causes congenital deformities.[13] Hydrocephalus has also been produced by Vitamin A deficiency[20] or Vitamin B_{12} deficiency.[34] It is suspected that, with great mechanization in food processing and packaging, with more and more chemical preservations added to

keep food fresher for a longer period of time, the possible number of unsuspected teratogenic agents might be consistently increasing.

Radiation

Radiation of the pelvic region during pregnancy has long been known to be teratogenic to the fetus.[11] Gentry[10] found increased likelihood of deformity in areas of high background radiation, especially in areas where drinking water was from a well or spring. New York State was divided into areas of probable and improbable inherent geographic radiation and noted that rates of congenital malformations were higher for the probable. So this information suggested radiation as a good area for further study. Leguene and co-workers demonstrated the presence of an extra chromosome in cultured connective tissue cells of mongols, strongly suggesting a chromosomal abnormality as the basic defect in this condition. This finding strongly suggested that if there are any environmental factors producing the sequence of events leading to the chromosomal defects in mongolism, they must be acting on parental gametes or on the fertilized ovum not later than several days after fertilization.

Infection and Disease

There have been many attempts to relate a specific disease or infection to the incidence of congenital malformations both experimentally and epidemiologically with many and varied results. Diseases that have been investigated include influenza[5], rubella, measles, mumps, chicken pox[29], poliomyelitis, T.B., and kidney infection[24]. Diabetes also has been implicated in the etiology of

malformations.

A survey of mongoloid births in Australia has suggested that there was an injurious agent (or agents) which may have been infectious[6]. Stoller and Collmann[31] also presented evidence for an association between aggregations of Down's syndrome births and epidemics of infectious hepatitis nine months earlier. Pleydell[26] also suggested that the clustering of Down's syndrome birth in England was associated with streptococcal infection during pregnancy.

In 1963, several clusters of childhood leukemia cases came to the attention of the United States Public Health Service. The most striking of these was an aggregation of eight cases in children under fifteen in Niles, Illinois, where less than two cases were expected from the national average. A virus has been suspected as the cause of these outbreaks[8].

Drugs

The area of drugs appeared most promising. Many authors submitted evidence that certain drugs are teratogenic. Among these drugs is aminopterin[27] - a cytoloxic drug used as an abortifacient. Thiouracil[33] could cause the birth of malformed babies. There is also evidence of penicillin-streptomycine, actomycin and tetracyline teratogenicity in man. These effects might depend upon the time and dosage of administration[3]. Lucey et al[16] lists aspirin, vitamins, antihistamines, antimetics, antibiotics, and contraceptive foam as being used by his study group in women during their first trimester who produce babies with spina bifida. Beside these, insulin, thyroid

hormone, quinine, anticancer drugs and antituberculin drugs were also suspected as the cause of congenital malformations.

Altemus and Ferguson[1] have suggested that drugs and disease such as antibiotics, thalidomide, and German measles can cause birth defects. Also, there is a great debate on the possible genetic effects of increased radiation that man is being subjected to during the age of atomic power and testing. These combinations have encouraged many public health and research groups into seeking a better understanding of the problems in this area.

Thus it is commonly accepted that besides genetic factors, radiation, infection and disease, nutrition, drugs, and other environmental factors might play a very important role as the etiology of the congenital malformations.

If all these factors, or one factor, were introduced into a confined area for a controlled time period, and if they were the main cause of congenital malformations, then the incidence would indicate time and space clustering. The hypothesis is that significant temporal-spacial clusters are formed by the introduction of agents such as micro-organisms, drugs, radiation, and so on, and the distribution of the incidence of these congenital malformations support the hypothesis. Several approaches have been suggested recently to study the case aggregations both in time and space, generally associated with the study of childhood leukemia. As a disease of rare occurrence, such as leukemia and congenital malformations appear to arise at random rather than as clear-cut epidemics, the validity of the results of past studies needs to be improved. It is necessary

to devise sensitive and effective procedures for the study of the distribution of diseases which appear to be random in time and space and also for non-random cases.

CHAPTER II

LITERATURE REVIEW

Kellet (1937) was perhaps the first one to point out that leukemia could come in clusters because the analysis of 63 leukemia cases admitted into the Royal Victoria Infirmary over a period of five years (1932-1936) tended to support this assumption. From that time on, several approaches have been presented for testing the time-space clusters of leukemia and other diseases. Most of the techniques used for evaluating the time-space clustering are based on looking for the numbers of "paired-cases" that are time-space related in a defined space-time criterion out of C_2^n possible pairs, "n" being the total observed number of cases in a certain region. Under the hypothesis of no clustering, a statistical test of significance of the observed number of pairs out of the C_2^n possible pairs will enable us to conclude whether the time-space clustering is due to a non-random process.

Pinkel and Nefzger (1960) studied the time-space aggregation of 95 leukemia cases in the Buffalo, N.Y. area over a 14-year period. The area they studied is 42.67 square miles. The assumption of a uniform density of population was made. The criterion of the time-space proximity of cases was 0.35 miles and 2 years. Among the 95 cases that Pinkel and Nefzger found, 20 were arranged into 10 pairs and 3 cases into a triple which occurred within the criterion defined above. The obtained arrangement of cases would occur at random

with a probability of 0.15. The frequency of time-space proximity (12 pairs) as defined here looks impressive on a map but whether the statistical evaluation suggests a notion of residential aggregation is uncertain. At best, it can only suggest such a factor since the probability of 0.15 does not reach the usual level for the rejection of statistical hypothesis. Besides questioning this probability model, the assumption of a uniform density of population in that area could lead us to a wrong conclusion. Thus this procedure cannot be generally recommended for evaluation of clusters in time and space.

Knox (1963) suggested a method of examining the interactions between time and space concentrations. The method proposed is the examination of all possible pairs from "n" cases or a selection of them to see whether short geographical distances are positively correlated with short time intervals. Knox assumed that X (the number of "paired-cases" that are time-space related in a defined time and space unit) is a Poisson variable so a test between X and E(X) can be established to decide if the interactions of time-space do exist.

Applying this model to the study of the cases of childhood leukemia in Northumberland and Durham in England, 96 cases were found from 1951-1960 in that area by Knox. A 2 x 2 table was formed according to time and space criterion. The hypothesis that the time-space clusters are independent is rejected from this application.

Pike(1966) tried to verify Knox's assumption of the Poisson distribution of variable X. He allocated, at random, the 96 time -

co-ordinates of 96 cases to the 96 space co-ordinates and counted the number of pairs which were adjacent both in defined time and space criterion. This was repeated 2,000 times and the comparison between the observed frequencies and the expected frequencies were made. ($x_4^2 = 0.75$) Thus, the Poisson approximation is seen to be quite reliable in this case.

Barton and David (1966) derived the first and second moments for the X variable and applied them to the example which was studied by Knox. They found: $E(X) = 0.833$

$$\sigma^2 = 0.802$$

Therefore, the $E(X) \approx \sigma^2$ is another further confirmation of the adequency of X in Poisson and it can be applied to similar instances.

Barton and David (1966) suggested applying a test they had previously devised for studying the random points on a plane. The device is one developed in analogy with the analysis of variance F test. Suppose there are "n" points in a plane, between any two of which we can measure a distance, and get the sum of squared distances between the C_2^n possible pairs. If the "n" points can be divided into subgroups, similar sums of squares can be obtained within each subgroup. By subtraction, a residual for between subgroups is determined. A Q test can be established based on the between and within subgroups ratio. Normal approximation can be used under certain conditions so that $\left(\frac{Q-1}{\sigma_Q}\right)$ may be referred to normal tables (if it exceeds 1.96 in absolute value, it is significant at the 5 per cent level). This device is applied to the disease incidence data by dividing the study period into sub-intervals, each sub-interval identifying a sub-group for this new test, and analysing the distribution of the points in

space. Specific objective rules for identifying the sub-intervals are given. Applied to the Knox data, the procedure divides Knox's 96 cases of leukemia into 35 groups and $\frac{(Q-1)}{\sigma_Q} = 0.452$. This is far from significant and shows that leukemia cases in Northumberland and Durham do not appear to cluster in time and space. A contrasting application is provided by cases of an undeniable epidemic disease, measles. Barton and David have used the data of measles at the two Southall schools. There are 104 cases at each school over a period of 1034 days. The Q statistic value is 0.21758 and $\sigma_Q = 0.01849$, so that $\frac{Q-1}{\sigma_Q} = 42.32$ which is highly significant. The measles cases in the two schools are clustering both in time and space. The test likewise does not require knowledge of the population at risk, its age structure and so on. On the other hand, where such ancillary information is available, the test as it stands will not be able to make use of this information and so will not be fully efficient. This emphasizes the rather drastic simplicity of the test which suggests that it will require a large number of cases to give a reasonable chance of detecting real interaction. Particularly since such interaction, in the nature of things, is not very marked.

Lundin et al (1966) presented an approach to study the temporal relationships among cases which were geographically proximate deaths in a metropolitan area over a 10 year period.

A total 2,190 deaths of Leukemia and Lymphoma in Pittsburgh and Allegheny county, Pa., were observed in the previous ten years (1953-1962). 350 study pairs were formed according to the following criteria;

1) Same street: Residences in the same street which are 100 or less numbers from the residence of an index death.*

2) Neighborhood: Residence in the same street and the two nearest parallel streets which are 100 or fewer house numbers from those of the index death and in the two nearest intersecting streets within 100 house numbers of the intersections.

3) *Index death: The earliest death among all those whose addresses meet the defined spatial proximity.

4) Pair: Two spatially proximate decedents consisting of an index and a subsequent death.

Under the above definitions of spatial proximity, the first death (index death) occurring in an aggregate of two or more decedents was noted and paired with subsequent death in that aggregate. To avoid overlapping of areas of spatial proximity, a restriction required that index deaths in the same street be 200 or more house numbers apart. In addition, subsequent deaths in aggregates were not permitted to pair with later deaths. Thus each subsequent decedent was allowed to pair with only one index death.

In each set of deaths among the aggregates, the subsequent deaths occurred during an interval of months m between the date of the index death and the end of the survey period. Assuming a constant probability of a subsequent death over this interval, one would expect $1/m$ subsequent death each month. Expected value can be calcu-

lates for each value of m and added over all pairs. The statistical significance of the aggregation of the subsequent deaths both in time and space can be detected from the deviation of the observed pairs and the corresponding expected pairs. For the total of 350 pairs from 2,190 cases in Pa. from 1953 to 1962. The deviations of all observed deaths from expectations was statistically significant at the 0.005 level ($\chi^2_9 = 18.2$).

All the approaches stated above are mainly based on analyzing the significance between the observed and expected "paired cases" which fall in the defined criterion. Recently, Ederer et al (1964) have preferred another approach. This does not use the "paired case" techniques, but is based on the analysis of the observed defined clusters and the expected one which can be obtained by combinatorial probability under the hypothesis of a random allocation of the cases. (i.e., of no clustering).

The definition of the time-space clustering should be established before the survey and the criterion of time-space clustering should be clearly defined. The whole region that is to be surveyed can be divided into numbers of space units under the criterion of space clustering. Then the distribution of the time clustering of the cases in each time unit can be analyzed. If the assumption is made that the cases occurring r times in each unit are distributed at random, the expected number of time-space clusters can be derived for each unit where r cases occurred. So the sum of observed clusters, statistically significantly greater than the corresponding sum of expected clusters, will be taken as evidence of the tendency of

the case aggregation both in time and space.

Ederer, Myers and Mantel (1964) used the model to detect the time-space clustering for 333 leukemia cases (1945-1959) in 169 towns of Connecticut. No important departures from randomness in leukemia incidence were found. But the same method had been applied to the two known virus diseases, Poliomyelitis and infectious hepatitis. Clustering in time and space of these incidences was easily detected.

Stark and Mantel used the data from 2,431 Down's syndrome births during 1950-1964 in Michigan. The E-M-M approach was applied in order to find:

1) whether there was a year of peak occurrence.
2) whether there was a season of peak occurrence when the 15 years of data were combined.
3) whether there were months of peak Down's syndrome incidence within seasons.

Stark and Mantel were unable to demonstrate clustering within the separate counties of lower Michigan by years, seasons or months within seasons.

During a defined initial base-line year interval, incidence of leukemia was distributed on a tracted map of a city. Tracts were defined as being "with leukemia" or "without leukemia" according to whether their residents had contributed one or more cases during the initial year period. If the occurrence of leukemia is governed predominately by the laws of chance, no significant difference should be expected in the subsequent incidence rates of the two