

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN
COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on
Methodological Issues of Environment
Statistics

(Jerusalem, Israel, 11-14 October 1999)

Working paper No.2

METHODOLOGY FOR EPIDEMIOLOGIC INVESTIGATION OF RARE DISEASES

Rina Chen
Impulse Dynamics, Tirat Hacarmel, Israel

Abstract

Several inherent difficulties are involved in the statistical methods designed for detection and investigation of clustering of rare diseases (e.g. cancer). Because of these, both types of statistical error are large. An exploratory approach to be used in the initial epidemiologic investigation is therefore needed. Such an approach is presented here. The approach incorporates several techniques, each is based on the inter-event time interval. One of the incorporated methods is to be applied to events observed subsequent to the alarm, other methods are to be applied to the alarming data. Although the methods are based on statistical derivations, their application is intended for use in ad hoc situations. Therefore, the analyses should be considered as a way to weigh the evidence in support of one or another of the possible interpretations associated with the alarm rather than as testing for significance.

Examples in which this methodology has been applied are given. These examples demonstrate: 1. The efficiency of this approach in revealing "hidden" cluster embedded in the data set. 2. The interpretation of a detected cluster provided by the suggested analyses.

I. INTRODUCTION

1. Health professionals are often required to conduct an investigation regarding what seems to be a cluster of cancer cases in a community. The requirement may stem from alarm elicited among the community members, or from statistical analyses of data observed in that or in a related community. Several inherent difficulties are involved in the statistical methods designed for detection or investigation, of clustering of rare events. Because of these difficulties both types of statistical errors are involved in these analyses. The type I error is inflated when the analyses are carried out in ad hoc situations. The low efficiency of the methods is related to the facts that the data are sparse, the incubation period is long and the impact of each carcinogen is minor.

2. In view of these difficulties, the need for an epidemiologic investigation methodology for rare diseases, is evident. Like the methodology associated with investigation of contagious diseases, a lot can be learned from the temporal pattern of the diagnoses. The temporal pattern of cases in the classic investigation is evaluated by the changes in the incidence rates measured in consecutive days or weeks. In contrast, not much can be learned from the temporal pattern of the incidence rate of a chronic disease. This is because of the small number of diagnoses that are spread over a long period. I therefore suggest the inter-event time interval as the relevant statistic in the rare event epidemiologic investigation. In this presentation I will show how this time interval can be expressed in terms of either one of two standardized time units. The advantage of measuring the interval in a standardized unit (rather than in a calendar time unit) is two fold, first it enables control changes in the size and/or profile of the population at risk, in the analyses. Second, the distribution of each of these standardized time intervals is known and enables tests of significance, etc.

3. The suggested methodology is based on three analyses. One is applied to data observed after the alarm (either one of the two confirmatory techniques) and two are applied to the alarming data (the CUSCORE test of significance and the cumulative q interval curve). I will describe each of the methods and present examples in which analyzes of the alarming data sets were made. In general, the methodology can indicate that the cluster is an incidental rare event or that it is related either to an exposure or to another cause, such as the introduction of a new medical device that enables earlier diagnosis.

II. METHODS

The RI

4. The RI (Relative Interval)¹ measures the observed time interval in one of the two standardized units. This unit is defined as the expected waiting time until diagnosis. The number of months in each unit changes with respect to changes in the size and/or profile of the population at risk..

5. In the simple case, when the population at risk is constant in size and profile, the number of events is assumed to follow the Poisson distribution, and the inter-event time interval follows the exponential distribution. When the population's size and/or profile changes over time, the distribution of the number of events is time dependent Poisson and the inter-event time

interval is exponential. In both cases the RI is of exponential distribution with parameter $\lambda=1$.

6. In general, the RI is the expected number of diagnoses within the specific interval. For example, if we expect 1 diagnosis every 11 months, then an interval of 33 months is translated to $RI=3$ (since 3 diagnoses are expected within that 11 month interval). When the number of events is a Poisson variate with a constant parameter, the RI is simply calculated as the ratio between the observed (w) and the expected interval ($E(w)$) time interval. (Hence, its notation abbreviated from the Relative Interval).

The q interval^{2,3}

7. The q interval is defined as the null probability of no event within the actually observed inter-event time interval. As mentioned above, the RI is Poisson ($\lambda=1$), thus if the interval observed between the $i-1$ th and the i th event is expressed as RI, then for the i -th event, $q=\exp(-RI)$.

8. It should be noted that although the q interval is calculated as a probability, it is actually (like the RI) a random number. This is because the associated time interval (in terms of months or in terms of the RI units) is a random variable.

9. Because q is evaluated as a cumulative distribution (of the exponential distribution), its null distribution is uniform over the 0-1 interval. As such its expected value is 0.5 and its variance is $1/12$. Clearly then, under stable conditions we expect that consecutive q intervals that are larger than 0.5, are randomly allocated among those that are smaller than 0.5. But, if some of the diagnoses are associated with a relatively new exposure, then a cluster of large q intervals is expected.

10. It should be noted that increased incidence rate leads to smaller RI and to larger q.

Confirmatory analyses

11. Two techniques were suggested for confirmatory analyses¹. Both are based on the RIs of the first five diagnoses made subsequent to the alarm. One of the techniques is based on the median RI, the other on the mean RI. In both techniques the test statistics is compared with each one of the two reference values, t_1 and t_2 . The alarm is confirmed if the test statistic is below t_1 , it is rejected if it is above t_2 , and judgement is reserved if it is between t_1 and t_2 .

12. The use of two critical values (t_1 and t_2) corresponds to two separate tests. This enables early decision when the results of the five events strongly support either confirmation or rejection of the alarm, but judgement is delayed in other cases.

13. The advantage of the median over the mean technique is that sometimes we may be able to confirm or reject the alarm even before the fourth diagnoses. Since the largest value out of the three observed intervals, is the maximum possible value while the smallest value is the minimum possible value for the median, the alarm is confirmed if the maximum RI is shorter than t_1 and rejected if the minimum RI is longer than t_2 . Thus, in terms of early decision, the median is preferable over the mean as the relevant

statistic. However, the probability of confirming a true alarm is somewhat higher for the mean based technique.

Evaluation of the reference values for the median based technique

14. Each of the t_1 and t_2 parameters are calculated according to the binomial distribution. For the t_1 evaluation, each interval is defined as either short or long. Accordingly, the probability that a RI is shorter than t_1 is:

$$\text{EQ (1)} \quad p=1-\exp(-t_1).$$

For a given significance size P , the probability that the median is below t_1 is:

$$\text{EQ (2)} \quad P=\Pr(x>2)=\sum_{i=3}^5 \binom{5}{3} p^i q^{5-i}$$

where x is the number of short RIs, p and q are the respective probabilities for a short and for a long interval. In order to evaluate t_1 we first find p that satisfies our predetermined P and then evaluate t_1 . In the analyses of our studies, we aimed at confirming 25% of the false alarms (i.e., $P=0.25$), accordingly we used $p=0.3595$; and $t_1=\ln(1-p)=0.4455$.

15. The evaluation of t_2 can be carried out in a similar way to that of t_1 . If we aim at 0.975 probability of confirming an alarm when actually the rate is twice the baseline rate, we define short and long RI according to t_2 and using equation (2). Thus for setting $\Pr(x>2)=0.025$, we find $p=0.8534$ and since the RI is now exponential distribution with twice the baseline rate,

$$\text{EQ. (3)} \quad 0.8534=1-\exp(-2t_2)$$

hence $t_2=0.9600$.

Evaluation of the reference values for the mean based technique

16. The evaluations of t_1 and of t_2 are based on the assumption that the statistic $2n\bar{w}/E(w)$ has a chi-square distribution with $2n$ degrees of freedom, where \bar{w} is the average observed time interval and $E(w)$ is its expectation (i.e., $1/\lambda$). Hence 10 times the relative interval has a chi-square distribution with 10 degrees of freedom⁴. (The chi-square distribution of 10 times the RI is derived if we consider $E(w)/2$ rather than $E(w)$ as the time unit). Thus in order to confirm only 25% of the false alarms,

$$\Pr(10(\bar{w}/E(w))>C_{10,0.75})=0.75,$$

where $C_{10,75}$ is the chi-square critical value for 10 degrees of freedom and 0.75 is the probability of not confirming a false alarm. Hence,

$$\text{EQ.(3)} \quad t_1=C_{10,0.75}/10=0.6737,$$

For 0.025 probability of rejecting a true alarm when the rate is twice the baseline, we have

$$\Pr(10(\bar{w}/(E(w)*2)<t_2))=0.025$$

$$EQ.(4) \quad t_2 = C_{10,0.025} / (10 \cdot 2) = 1.0242.$$

The 2 in the denominator arises because the null $E(w)$ rather than the actual expected value.

The CUSCORE test of significance⁵

17. The test is aimed at detection of a cluster that is embedded in a larger data set. It is based on defining each interval as either short or long. An interval is considered short if $RI < k$. Hence, short intervals become more frequent when the rate is increased. A score is attached to each interval. The score of the first interval is either 1 or 0. It is 1 if the first interval is short and 0 if it is long. For each of the next events, the score increases by 1 if the RI is short and reduced by 1 if it is long, but it never assumes a negative value. If the score of the event $i-1$ is 0 and the RI of the event i is long, the score of event i remains 0 (rather than reduced to -1). The test is considered significant if the score equals 5 for any of the S events in the data set. The k parameter that defines an interval as short depends on the number (S) of events in the analyzed data set and on the level of significance. The k values for given S and significance level are presented in Table 1 in the Appendix.

The temporal pattern of diagnoses

18. In general, a cluster may result from either one of the following:
 1. Exposure to a local carcinogen. 2. Realization of a rare occurrence. 3. Diagnosis at an earlier stage (because of a new medical device, for example).

19. As mentioned above, in disease cluster analyses both types of statistical errors are large. On one hand, a detected cluster may be spurious. On the other hand a real cluster may be disguised because of one or more of several reasons (e.g., the long latent period of the disease; limited number of individuals that are sensitive to the exposure, etc.). The clustered cases may therefore constitute only part of the data. As a result, the overall count of cases may be close to that expected even though some of the cases are clustered.

20. The temporal pattern of the q interval may reveal a "hidden" cluster on one side, and may indicate which of the three possibilities is a likely explanation when a cluster is encountered. When the alarm is caused by exposure to a local carcinogen, the data are expected to show a single cluster. When the clustering is an incidental event, the data may have more than just one cluster of cases. When clustering is due to enhanced diagnoses, a single cluster (occupying perhaps the midst of the data set) is expected. This clustering may be followed by an incidence rate that is even lower than that expected. Thus, the temporal pattern of the incidence rate may indicate which of the possible explanations to the alarm has a sound basis.

III. EXAMPLES

21. The first two examples demonstrate situations where the suggested approach reveals a cluster although the overall count of events is close to or even lower than that expected. The third example demonstrates a situation where the temporal pattern of the events rejects the assumption that the apparent clustering is due to an exposure.

22. Example 1 presents diagnoses of chronic myeloid leukemia in a city in Israel. The observed and expected numbers of diagnoses are presented by 10 or 11 years period in Table 1. The comparison between the observed and the expected number of events, barely indicates increase in the risk starting in the second decade. Testing the significance we get $p < 0.16$ for the total number over the 31 years (13 vs. 9.5 expected), and $p < 0.07$ for the last two decades (11 observed vs. 6.6 expected). The results of the CUSCORE test is significant as seen from the data presented in Table 2. These data include: the dates of diagnosis, the RI, the q and the running score. A significant score was attained for the case diagnosed on 11/86. (It should be pointed that for some cases, the month of the diagnosis date is missing. For these cases the range within which the time interval is confined was calculated. The RI was considered as the middle point in that interval). These cases are listed as the first ones in the year but their RI was based on all possibilities. The observed and expected cumulative q intervals curves are depicted in Figure 1. It is clear from this figure that the slope of the curve increased from the 4-th event (i.e., in about the middle of 1977).

23. Example 2 presents data that were given to me as a correct data set, but was later found that to be based on erroneous baseline rates. Nevertheless, I show these data in order to demonstrate a situation where there is a cluster within the data set although the overall count is smaller than expected. The data were related to colon cancer deaths among a group of workers that are being followed up.

24. Table 3 presents the observed and the expected number of deaths during a 15 year period and during each of 7.5 year period. The expected number of deaths is smaller than that expected in each period (i.e., during the 15 year period (1978-92) and in each of the two sub-periods). The expected number during 1978-June 1989 is 5.26 as compared to 3 observed. The expected number during the second period is 9.23 as compared with 7 observed. In real situation a cluster of cases may be embedded within the data because of several reasons. The long latent period and the healthy worker effect may be responsible to low rate of the disease at early period; the limited proportion of individuals that are sensitive to the exposure, may result in low rates at a later period.

25. Table 4 presents for each individual, the date of death and the associated RI, q and the score. Figure 2 presents the cumulative q curves. The CUSCORE results are significant. The q or the RI clearly indicate a cluster of the last five individuals. The smaller than expected number of events is related to the small q interval of the first five cases. Because of the healthy worker effect, we may encounter such a situation in a workplace.

26. Thus, in this case the Cuscore revealed a cluster that could not be revealed by comparing the observed and the expected number of cases.

27. Example 3 is related to a residential community in the US. The residents in this area, complained of high CNS cancer rates. An investigation carried out reported that the number of cases in that area is indeed significantly larger than expected. The investigators suggested that exposure to pollution emitted from a close by plant, is responsible to the cluster of deaths. However, another conclusion is definitely more acceptable if we consider the temporal pattern of the events. The data of this example are presented in Table 5. Inspection of the q intervals leads one to the

conclusion that whatever caused this surplus of events, it can not possibly be related to an exposure to a local carcinogen. The q intervals are indeed larger than 0.5 (as expected for stable conditions) for the 6-th to the 12-th cases. But since all these 7 cases were diagnosed in a single year, 1985, whereas the number of diagnoses made at each of the other 8 years was at most 3, the excessive number can not possibly be related to an exposure. A more reasonable interpretation is that for some reason (e.g. introduction of a new medical diagnostic device) diagnoses were made earlier as of 1984.

References

- 1 Chen R., Connelly R.R and Mantel N. Analysing post-alarm data in monitoring system in order to accept or reject the alarm. *Statistics in Medicine*, 12,1807-1812, 1993.
- 2 Chen, R. The cumulative q interval as a starting point in disease cluster investigation, *Statistics in Medicine*, In print.
- 3 Chen R., Iscovich J. and Goldbourt U. Clustering of leukemia cases in a city in Israel. *Statistics in Medicine*, 16,1873-1887,1997.
- 4 Cox D.R. and Oakes D. *Analysis of Survival Data*, Chapman and Hall, 1984,(p.38).
- 5 Chen R. and Goldbourt U. Analysis of data associated with seemingly temporal clustering of a rare disease. *Methods of Information in Medicine*, 37,26-31, 1998.

Table 1 - Expected and observed number of chronic myeloid leukemia (CML) cases in a city, by period, 1960-90

Period	Expected	Observed	Expected/yr
1960-70	2.9	2	0.26
1971-80	3.0	6	0.30
1981-90	3.6	5	0.36
Total	9.5	13	

Table 2 - Diagnosis dates and q intervals of CML cases in a city during 1960-90.

Case no.	Diagnosis date	RI	q	Score
1	10/66	1.65	0.193	0
2	-/67	0.21	0.812	1
3	6/72	1.44	0.237	0
4	-/77	1.38	0.251	0
5	-/77	0.12	0.884	1
6	3/77	0.12	0.884	2
7	-/78	0.29	0.749	3
8	5/80	0.61	0.542	4
9	7/84	1.40	0.246	3
10	-/85	0.35	0.708	4
11	11/86	0.50	0.607	5*
12	4/87	0.15	0.859	
13	10/88	0.55	0.576	

* Significant at 5% (using $k=0.417$)

Figure 1 - Observed and expected cumulative q interval of CML cases by chronological order of diagnosis.

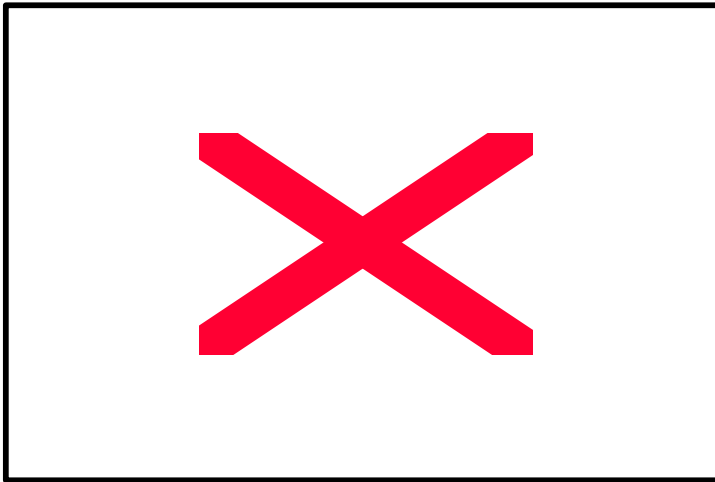


Table 3 - Number of deaths of colon cancer in a workplace (fictitious), 1978-92

Period	Expected	Observed
1/1979-6/1985	5.26	3
7/1985-12/1992	9.23	7
Total	14.49	10

Table 4 - Dates and time intervals of deaths of colon cancer in a workplace (fictitious), 1978-92

Case no.	Date of Diagnosis	RI	Observed q	Score
1	7/1979	0.813	0.443	0
2	9/1983	2.893	0.055	0
3	6/1984	0.627	0.534	1
4	3/1986	1.618	0.198	0
5	8/1991	6.535	0.001	0
6	9/1991	0.118	0.888	1
7	10/1991	0.118	0.888	2
8	10/1991	0	1.000	3
9	1/1992	0.357	0.699	4
10	3/1992	0.247	0.781	5*

- Significant at 5% (Using $k=0.474$)

Figure 2- Observed and expected cumulative q interval of deaths from colon cancer a fictitious example), 1978-92

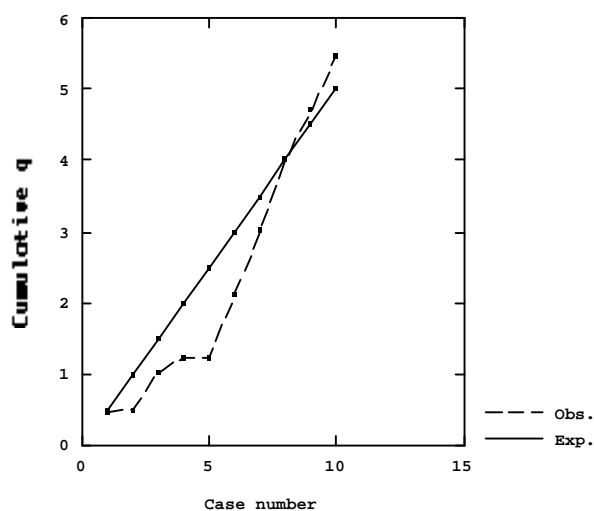


Table 5 - Dates of diagnosis and associated intervals of CNS cases in a residential community

Case #	Diagnosis date	RI	q	Score
1	1/82	0.048	0.953	1
2	1/82	0.048	0.953	2
3	7/82	0.580	0.560	3
4	6/83	1.063	0.345	2
5	12/84	1.740	0.176	1
6	4/85	0.387	0.679	2
7	8/85	0.387	0.679	3
8	8/85	0.048	0.953	4
9	8/85	0.048	0.953	5*
10	10/85	0.193	0.824	
11	10/85	0.048	0.953	
12	11/85	0.097	0.908	
13	12/86	1.257	0.285	
14	3/87	0.350	0.704	
15	7/88	1.933	0.145	
16	7/89	1.450	0.235	
17	1/90	0.725	0.484	
18	3/90	0.242	0.785	

* Significant at 5% (Using $k=0.368$)

APPENDIX

Table 1- Value of k by the number of events and significance level

No. of events S	Significance level	
	0.05	0.10
5	0.797	0.997
6	0.707	0.889
7	0.590	0.727
8	0.546	0.673
9	0.501	0.612
10	0.474	0.579
11	0.450	0.546
12	0.433	0.525
13	0.417	0.504
14	0.405	0.488
15	0.394	0.473
16	0.384	0.461
17	0.376	0.450
18	0.368	0.441
19	0.362	0.432
20	0.355	0.424
21	0.350	0.417
22	0.345	0.411
23	0.340	0.405
24	0.336	0.399
25	0.331	0.394
26	0.328	0.389
27	0.324	0.385
28	0.321	0.380
29	0.317	0.376
30	0.314	0.373