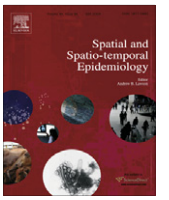
Spatial and Spatio-temporal Epidemiology 3 (2012) 235–242

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Statistical power of disease cluster and clustering tests for rare diseases: A simulation study of point sources

Sven Schmiedel [a,b,*], Maria Blettner [b], Joachim Schüz [c]

[a] Institute for Cancer Epidemiology, Danish Cancer Society, Strandboulevarden 49, 2100 Copenhagen, Denmark
[b] Institute for Medical Biostatistics, Epidemiology and Informatics, University Medical Center of Johannes Gutenberg University Mainz, Obere Zahlbacher Straße 69, 55131 Mainz, Germany
[c] International Agency for Research on Cancer, Section of Environment and Radiation, 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France

ARTICLE INFO

Article history:
Received 24 February 2011
Revised 21 December 2011
Accepted 29 February 2012
Available online 7 March 2012

Keywords:
Clustering
Cluster
Power
Denmark
Aggregation

ABSTRACT

Two recent epidemiological studies on clustering of childhood leukemia showed different results on the statistical power of disease cluster and clustering tests, possibly an effect of spatial data aggregation. Eight different leukemia cluster scenarios were simulated using individual addresses of all 1,009,332 children living in Denmark in 2006. For each scenario, a number of point sources were defined with an increased risk ratio at centroid, decreasing linearly to 1.0 at the edge; aggregation levels were administrative units of Danish municipalities and squares of 5, 12.5 and 25 km^2. Six statistical methods were compared. Generally, statistical power decreased with increasing size of aggregated units. In our scenarios, statistical tests based on individual data usually had lower statistical power than the best test based on aggregated data. In conclusion, spatial aggregation does not necessarily blur a clustering effect; this depends on the nature of clustering and the aggregated units.

1. Background

In two recent epidemiological studies, we investigated spatial clustering of childhood leukemia in Denmark (Schmiedel et al., 2011) and Germany (Schmiedel et al., 2010). We found clustering for the subgroup of 2–6-year-old children with acute lymphoblastic leukemia in Denmark but no clustering of any subgroup in Germany. The reason for the difference in these results might be that we relied on aggregated data in Germany (based on administrative units), whereas the geographical coordinates of all addresses of each child were available in Denmark. We therefore decided to study the effect of spatial data aggregation on the statistical power of disease cluster and clustering tests using the individual data available in Denmark, which gave us a unique opportunity to simulate clustering with real point data. The simulation was based on a fixed number of randomly spread circular point sources, each described by a pre-specified radius with an increased relative risk (RR) above 1.0 at the centroid. The RR decreased linearly to 1.0 at the edge of the circle. In addition to the individual data, we also used the administrative units of Denmark ("kommunes") for aggregation. Furthermore, we build artificial aggregated units by overlaying the map of Denmark with a grid based on quadratic squares of 5, 12.5 and 25 km^2.

Studies on the statistical power of disease clustering tests published by Kulldorff et al. (2003) and Song and Kulldorff (2003) addressed the power of different disease clustering tests based on aggregated data by using two cluster processes: 'hot-spot clustering' and 'global chain clustering'. They focused on differences in the performance of the tests by population density (clusters in rural, mixed or urban areas) and did not analyze whether aggregation blurs the clustering of a disease in comparison with real

* Corresponding author. Permanent address: Statistisches Bundesamt, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany. Tel.: +49 611 752275; fax: +49 3018 106442275.
E-mail addresses: Sven.Schmiedel@destatis.de (S. Schmiedel), blettner-sekretariat@imbei.uni-mainz.de (M. Blettner), schuzj@iarc.fr (J. Schüz).

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.sste.2012.02.011

Table 1
Characteristics of eight cluster scenarios and simulation.

Scenario	No. of potential cluster	Risk ratio at centroid (RR_0)[a]	Radius of a potential cluster (d_{max}) (km)	Average No. of potential clusters without cases	No. of addresses out of 1,009,332 in potential clusters (quantiles)			Incidence rate per 100,000 inside and outside potential clusters
					10%	50%	90%	
1	1	10	20	0.0	1338	3454	10,650	96.8, 4.7
2	5	5	10	0.2	12,929	22,839	54,002	23.8, 4.5
3	5	10	5	0.1	2913	5648	14,630	46.0, 4.7
4	20	5	5	2.2	17,663	28,191	53,611	22.6, 4.5
5	30	5	3	11.8	9955	16,226	28,399	20.3, 4.7
6	50	2	5	15.3	52,044	72,325	112,814	9.2, 4.7
7	100	2	5	32.6	110,939	145,466	204,476	8.9, 4.3
8	200	2	5	74.9	225,590	280,809	343,133	8.3, 3.7

[a] Decreasing linearly to background risk at edge of potential cluster.

point data, changing the performance of the test. The effect of different aggregation levels on the results of disease clustering tests was analyzed by Gregorio et al. (2005) for the spatial scan statistic and by Kulldorff et al. (2006) for nine different statistical tests using cancer data. However, in both publications the focus was on the results of the statistical tests using different levels of aggregation without investigating the individual level. Liu et al. (2009) analyzed six different putative hazard tests (testing on adverse health effects around fixed locations). Their simulation on a unit square area focused on different distributions of background risk. They investigated the individual level and did not compare the behavior of statistical power when using different aggregation levels.

Our aim was to analyze the performance, measured as statistical power, of the disease clustering tests that are widely used in analyzing the incidence of rare diseases at different levels of spatial aggregation including the individual level.

2. Methods

The simulation study was based on real addresses in Denmark. All children aged less than 15 years in 2006 and registered in the Central Population Registry (n = 1,009,332) were included. The probability of a child having leukemia was assumed to be 5 per 100,000 on the basis of the approximate incidence rate of childhood leukemia in Denmark (Svendsen et al., 2007). This corresponds to 50 expected cases in 2006. In order to increase the number of cases, we expanded the study period to 20 years. As we did not account for changes of place of residence, the simulation can be interpreted as a study in a stable population, in which children reaching the age of 15 years were replaced by newborns. On average, ($1{,}009{,}332 \times (5/100{,}000) \times 20$=) 1009 children were simulated as having the disease.

2.1. Data simulated under the null hypothesis (H_0)

A uniformly distributed random number between 0 and 1.0 was assigned to each address. If this number was smaller than the cumulative incidence rate over 20 years ($5/100{,}000 \times 20 = 0.001$), the child residing at the respective address was considered to be a 'case'. This simulation of cases was used as a reference to create one realization of the test statistics. We repeated this process 1000 times in order to evaluate the distributions of all test statistics under H_0 and used these distributions to calculate critical values.

2.2. Data simulated under alternative hypotheses (H_1)

The simulation was performed for eight different H_1 scenarios, gradually increasing from a single to 200 potential clusters in Denmark, with the centers of the potential clusters randomly (complete spatial randomness) spread over Denmark. Within a circle of radius d_{max}, an elevated risk ($RR(d)$) was determined. The RR was fixed ($RR_0 > 1$) at the center of the circle, decreasing linearly to 1.0 at the edge:

$$RR(d) = \begin{cases} RR_0 - (RR_0 - 1)\frac{d}{d_{max}} & , \quad \text{where } d \leqslant d_{max} \\ 1 & , \quad \text{where } d > d_{max} \end{cases}$$

The different scenarios and their parameters are listed in Table 1. The word 'potential' was used as our process defined centers in which the chance of additional cases in the vicinity was increased. This, however, does not necessarily lead to additional cases.

In order to keep the number of cases constant between H_0 and H_1, every time a case inside a potential cluster was produced, a randomly chosen case outside the potential cluster(s) was deleted. The individual tests and how their statistical power were calculated is described in more detail in Section 2.4.

2.3. Aggregation levels

For the simulation, we used the following aggregation levels:

1. No aggregation. Individual data (addresses) were used.
2. The n = 225 administrative units in Denmark defined before 2007, called 'kommunes' (municipalities), were used as the aggregation level (Fig. 1a).
3. Denmark was split into quadratic squares of arbitrarily defined different sizes for three further scenarios of aggregation levels: 5 km^2 (n = 2063, around 40% of the average size of a municipality in Denmark), 12.5 km^2