Spatial and Spatio-temporal Epidemiology 3 (2012) 297–310

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Performance of cancer cluster Q-statistics for case-control residential histories

Chantel D. Sloan [a], Geoffrey M. Jacquez [c,d], Carolyn M. Gallagher [a], Mary H. Ward [e], Ole Raaschou-Nielsen [f], Rikke Baastrup Nordsborg [f], Jaymie R. Meliker [a,b,*]

[a] Department of Preventive Medicine, Stony Brook University, Stony Brook, NY, USA
[b] Graduate Program in Public Health, Stony Brook University, Stony Brook, NY, USA
[c] BioMedware, Inc., Ann Arbor, MI, USA
[d] State University of New York at Buffalo, Buffalo, NY, USA
[e] Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD, USA
[f] Danish Cancer Society, Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 25 August 2010
Revised 4 September 2012
Accepted 10 September 2012
Available online 24 September 2012

Keywords:
Geographic information systems
Residential mobility
Space–time clustering

ABSTRACT

Few investigations of health event clustering have evaluated residential mobility, though causative exposures for chronic diseases such as cancer often occur long before diagnosis. Recently developed Q-statistics incorporate human mobility into disease cluster investigations by quantifying space- and time-dependent nearest neighbor relationships. Using residential histories from two cancer case-control studies, we created simulated clusters to examine Q-statistic performance. Results suggest the intersection of cases with significant clustering over their life course, Q_i, with cases who are constituents of significant local clusters at given times, Q_{it}, yielded the best performance, which improved with increasing cluster size. Upon comparison, a larger proportion of true positives were detected with Kulldorf's spatial scan method if the time of clustering was provided. We recommend using Q-statistics to identify when and where clustering may have occurred, followed by the scan method to localize the candidate clusters. Future work should investigate the generalizability of these findings.

1. Background

Disease cluster investigations often result in null findings (Schulte et al., 1987), prompting some to argue there is little value in studying clusters of health events (Rothman, 1990). This perception can be attributed to several factors: (1) historically, cluster investigations were limited to pre-identified subjectively defined disease clusters, as opposed to systematic examination of representative incidence data; (2) residential histories and thus disease latency were ignored; and (3) cases are typically aggregated into arbitrary geographic units, making results ecologic in nature, and subject to the modifiable area unit problem (Meliker et al., 2009; Rothman, 1990). In recent years, epidemiologists have collected detailed address information as part of a residential history for the purpose of geocoding and mapping residences, thereby permitting systematic examination of disease patterns over the life-course. Accurate and complete historical residence locations can be used to overcome the three main limitations described above.

Statistical approaches for investigating space–time patterns are being developed to aid in the analysis of geocoded residential history data in epidemiologic studies. While dozens of approaches are available for quantifying patterns on disease maps (e.g. Besag and Newell, 1991; Cuzick and Edwards, 1990; Kulldorff and Nagarwalla, 1995; Kulldorff et al., 2006; Tango and Takahashi, 2005; Turnbull et al., 1990; Waller and Turnbull, 1993; Waller et al., 1995), most of these tests were developed for spatially static datasets and do not account for mobile populations. Recently, several

* Corresponding author at: Graduate Program in Public Health, Stony Brook University, USA. Tel.: +1 631 444 1145.
E-mail address: jaymie.meliker@stonybrook.edu (J.R. Meliker).

1877-5845/$ - see front matter
http://dx.doi.org/10.1016/j.sste.2012.09.002

methods have been developed for investigating space–time patterns in mobility data (Jacquez et al., 2005; Sabel et al., 2009; Webster et al., 2006). Our group has been involved in development of Q-statistics for evaluating space–time clustering in residential histories of case-control data (Jacquez et al., 2005, 2006). The Q-statistics utilize nearest neighbor calculations to evaluate local and global clustering at any moment in the life-course of the residential histories of cases relative to the residential histories of controls.

Given the exploratory nature of space–time clustering investigations, these Q-statistics can be re-calculated whenever a participant changes residence and thus can result in hundreds or thousands of local test statistics depending on the mobility of the population. As a result of these multiple tests, interpreting statistical significance can be a challenge. Bonferonni-type corrections are known to be overly conservative (Hochberg, 1988), with alternatives making use of simulation studies of receiver operating curves and false discovery rate (FDR) adjustments (Kleinman and Abrams, 2006; Narum, 2006; Read et al., 2007; Caldas de Castro and Singer, 2005; Catelan and Biggeri, 2010). Approaches accounting for multiple testing that combine information from two of the Q-statistics, Q_i and Q_{it} are presented. These new approaches combine Q_{it} and Q_i to identify whether individuals with significant clustering over their life course co-occur in space and time.

The number of nearest neighbors (k) used to calculate a Q-statistic is a user-defined parameter. The selection of k is important, as a k too large can result in over-smoothing and a failure to detect smaller clusters, while a k too small may result in an inability to differentiate true and false positives (Cuzick and Edwards, 1990). Certain cluster and population characteristics influence the most appropriate k and performance of the Q-statistics; these include cluster density (number of cases relative to number of controls in the cluster region), size of the overall population, size of the cluster, and mobility of the population. Though we could not exhaustively evaluate each characteristic here because this would require hundreds of thousands of simulations, we did explore a range of different populations and geographies using different cluster sizes within multiple regions. In this report we create a series of clusters across a range of these cluster and population characteristics to examine performance of Q-statistics and sensitivity of results to choice of k nearest neighbors. Simulations are run using residential history data from two large case-control studies in the United States (US) and Denmark. Our objective is to use these simulations to (1) guide development of protocols for using and interpreting Q-statistics such that researchers can differentiate true local space–time clusters from false positives, and (2) to provide guidance on specification of the appropriate number of k nearest neighbors to use in an analysis.

2. Methods

2.1. Background on Q-statistics

Jacquez et al. (2005) develop global and local tests for case-control clustering of residential histories. Readers unfamiliar with Q-statistics may wish to refer to the original work; these are described briefly here. Q-statistics rely on a matrix representation that describes how spatial nearest neighbor relationships change through time. A person's residential history is represented as a space–time thread using a step function (Fig. 1).

To identify the location and timing of significant clustering, the following spatially and temporally local case-control cluster statistic is used:

$$Q_{i,t}^{(k)} = c_i \sum_{j=1}^{k} \eta_{i,j,t}^{(k)} c_j. \tag{1}$$

This quantity is the count, at time t, of the number of k nearest neighbors of case i that are cases, and not controls. Individuals i and j have case-control identifiers, c_i and c_j defined to be 1 if and only if a participant i is a case, and 0 otherwise. N is the total number of participants (cases and controls) in a study. The term $\eta_{i,j,t}^{(k)}$ is a binary spatial proximity metric that is 1 when participant j is a k nearest neighbor at time t of participant i; otherwise it is 0. Since a given individual i may have k unique nearest neighbors, the $Q_{i,t}^{(k)}$ statistic is in the range 0–k. When i is a control, $Q_{i,t}^{(k)} = 0$. When i is a case, low values indicate cluster avoidance (e.g. a case surrounded by controls), and large values indicate a cluster of cases. When $Q_{i,t}^{(k)} = k$, at time t all of the k nearest neighbors of case i are cases. The user must specify the value for k before a statistic is calculated; guidelines on the specification of k is a topic of this research.

We also wish to calculate a subject-specific statistic that integrates through time (Eq. (2)).

When integration is accomplished over a subject's residential history we think of this as a "life-course" statistic that assesses a tendency to have other cases, rather than controls nearby over the life-course

$$Q_i^{(k)} = \int_{t=t_0}^{T} Q_{i,t}^{(k)}\, dt. \tag{2}$$

A time-specific statistic that provides an overall measure of case clustering when all of the participants are considered together is given in Eq. (3). It is the sum, over all cases, of the subject-specific and time-specific measure of case clustering in Eq. (1)

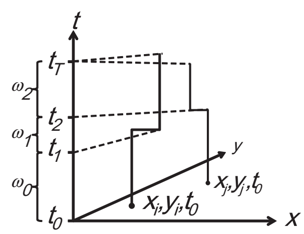

Fig. 1. Residential histories as space–time step functions. The axes x and y define a geographic domain (e.g. longitude and latitude decimal degrees), the t axis represents time (e.g. date). The study extends from time t_0 to time t_T. The residential histories for persons i and j are shown as step functions through space–time. For example, person i begins the study residing at location x_i, y_i, t_0. They remain at that geographic coordinate until the instant before time t_1, when they move to x_i, y_i, t_1. The duration of time they reside at this first place of residence is ω_0.