# Performance of Cancer Cluster Q-statistics for Case-Control Residential Histories


Jaymie R Meliker

Graduate Program in Public Health, Department of Preventive Medicine
Consortium for Inter-Disciplinary Environmental Research (CIDER)
Stony Brook University (SUNY)

Co-investigators: Chantel D. Sloan[1], Geoffrey M. Jacquez[2], Carolyn M Gallagher[1], Mary H Ward[3], Rikke Baastrup[4], Ole Raaschou-Nielsen[4]

[1]Stony Brook University (SUNY), [2]BioMedware, Inc., [3]US National Cancer Institute, [4]Danish Cancer Society



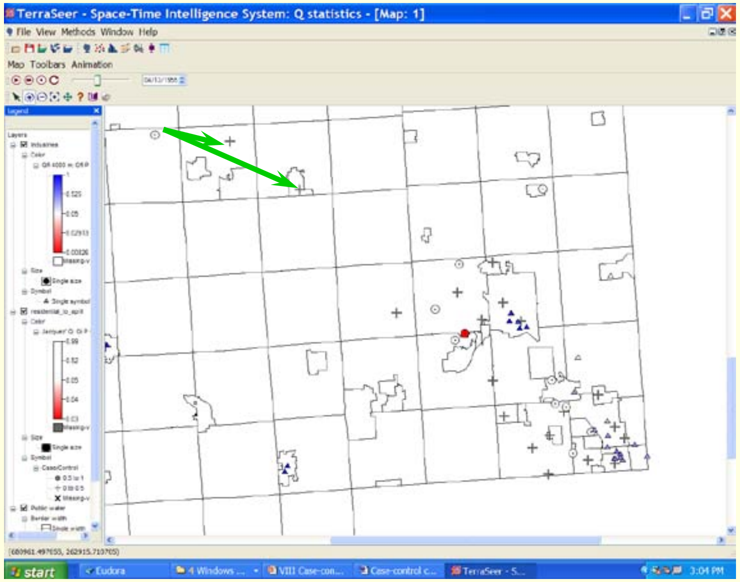
Funding: NIEHS R21 ES015501, and NCI R03 CA125827, R44 CA117171-02

# Statement of Problem

- Goal: Use cancer clusters to generate valuable hypotheses for diseases with largely unknown etiology
  - Most cancer cluster investigations ignore disease latency, using locations at time of diagnosis or death
  - Recent statistical advances have begun to investigate clustering in mobile populations
    - Spatial generalized additive models
    - Jacquez’s Q-statistics
  - Few performance evaluations have been conducted on these new statistics
    - Multiple testing through time is a large concern

# Q-statistics for Case-Control Populations

- Rely on a matrix representation that describes how spatial nearest neighbor relationships change through time
- Space-time extension of Cuzick-Edwards' Test
- User must specify number of nearest neighbors
  - Neighbors that are cases are then counted around each case
    - Repeated every time there is a change in location



# Q-statistics cont'd

- Different versions:
  - $Q_{ikt}$: When and where is there local clustering around a case?
    - Assesses clustering around each case every time there is a change in residence
  - $Q_{kt}$: When is there global clustering of cases?
    - Assesses global clustering at each time slice
  - $Q_{ik}$: Is there clustering surrounding a case, on average, throughout his/her mobility history?
    - Assesses clustering around a person through time; Sum of $Q_{ikt}$
  - $Q_k$: Is there global clustering, overall across all cases, in the residential histories?
    - Assesses whether, in general, clustering is present

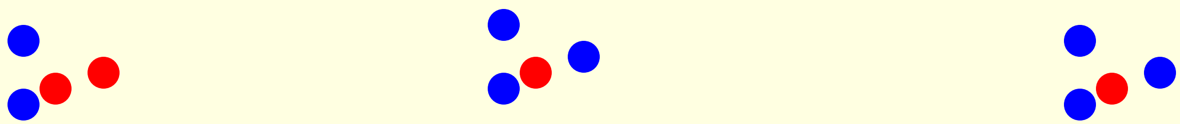
Focused versions are also available

# Procedure for Evaluating Significance

- Step 1. Calculate Q-statistic (Q*) for the observed data.
- Step 2. Reallocate the case-control identifier $c_i$ over the participants using approximate randomization, and calculate Q-statistic:
  - consistent with the desired null hypothesis
  - holding the observed number of cases fixed
  - holding the locations and attributes fixed

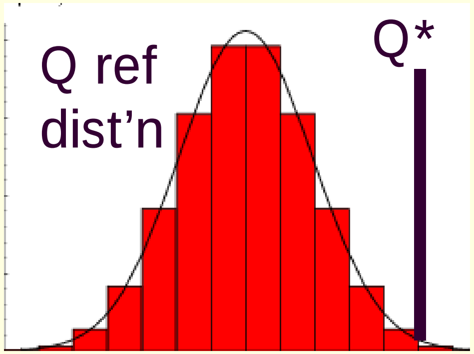
Repeat many times (e.g., 999) to create a reference distribution

● Case
● Control



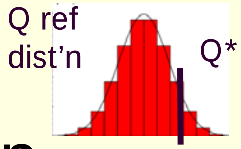
Observed | Randomization #1 | Randomization #2

- Step 3. Compare Q* to this reference distribution to evaluate the statistical probability of observing Q*.



# Comments on Q-statistics and Evaluating Significance

- Must run many randomizations to resolve small p-values




  - Time-consuming → lessens likelihood of p-value correction such as false discovery rate[1]

- Can we identify a p-value to use as a cut-off for significance (in light of multiple testing)?

- Can we determine which Q-statistic(s) to use to identify a cluster?

[1]Caldas de Castro M, Singer BH. Controlling the false discovery rate: A new application to account for multiple and dependent tests in local statistics of spatial association. Geogr Anal 2005; 38: 180-208.

# Analytic Plan

- Blank slate: many approaches could be used
- Simulated clusters were created to examine Q-statistics’ performance
  - Used actual mobility histories from studies of NHL in US, and testicular cancer in Denmark
- Examine whether Q-statistics identify simulated clusters, and differentiate them from false positives

# Simulated Clusters

- Iowa
- California
- Central Denmark

- Reflected a variety of space-time cluster characteristics

Table 1. Characteristics of the Cluster Regions

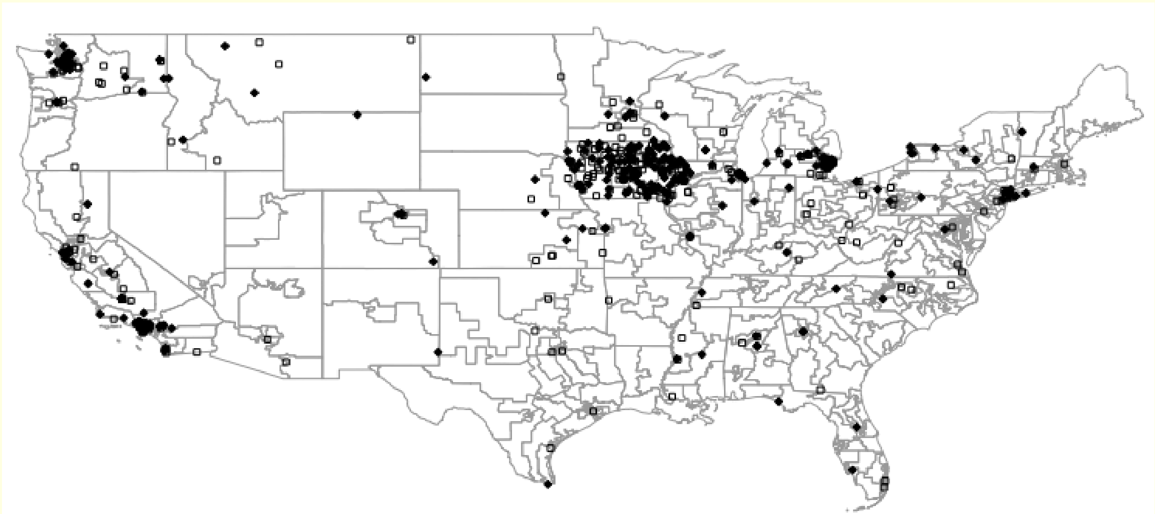
| | | Number of Cases | Cluster Size[a] | Cluster Density[b] | Case Mobility[c] |
|---|---|---|---|---|---|
| US Case-Control Dataset, Clusters Created in 1960 | 1000 residential histories | | | | |
| | | 5 | 1.0% | 100% | 99% |
| | | 12 | 2.4% | 100% | 90% |
| | Iowa | 18 | 3.6% | 95% | 83% |
| | | 27 | 5.4% | 90% | 84% |
| | California | 43 | 8.6% | 63% | 47% |
| | 2378 residential histories | | | | |
| | | 6 | 0.3% | 75% | 87% |
| | Iowa | 14 | 0.6% | 70% | 80% |
| | | 23 | 1.0% | 66% | 84% |
| | | 33 | 1.4% | 69% | 78% |
| Danish Case-Control Dataset, Clusters Created in 1971 | 6594 residential histories | | | | |
| | | 11 | 0.3% | 89% | 50% |
| | | 41 | 1.1% | 84% | 74% |
| | | 90 | 2.6% | 82% | 70% |
| | | 127 | 3.7% | 81% | 80% |

[a]Cluster Size: Percent of cases in cluster out of total number of cases in study

[b]Cluster Density: Percent of cases in cluster region out of total number of cases and controls in cluster region from 1960-1975 in US dataset, 1971-1980 in Danish dataset.

[c]Case Mobility: Percent of person-years of cases in cluster region out of maximum possible person-years from 1960-1975 in US dataset, 1971-1980 in Danish dataset

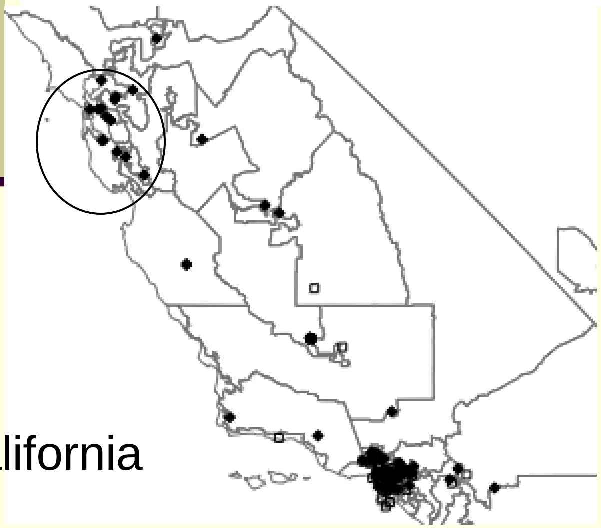
# US Cluster Regions



Case

Control

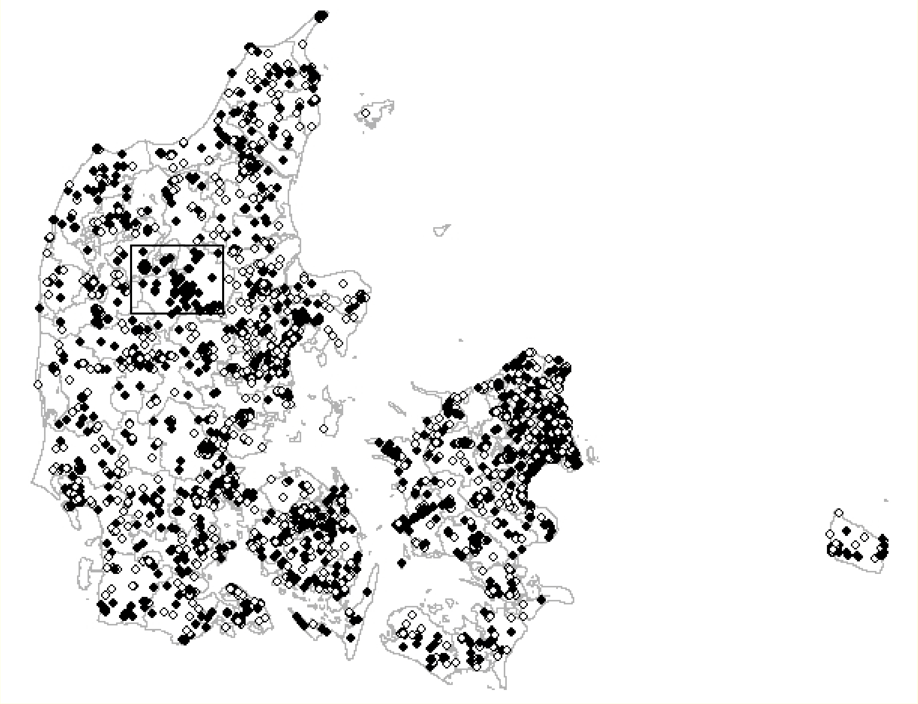
Locations in 1960



California

Iowa

# Danish Cluster Region



Locations in 1971

# Results - Summary

- Using k=5, 10, 15 (and 20)
- Global $Q_k$: significant (p=0.05) for only 1 of the 31 analyses of simulated clusters
- Local $Q_{ikt}$: significant (p=0.01 or smaller) even for very small clusters, but unable to differentiate true clusters from false positives
- Global $Q_{kt}$: time-slice global; also conservative like $Q_k$
- Local $Q_{ik}$: best able to identify true clusters and differentiate them from false positives.
- Combining $Q_{ik}$ and $Q_{ikt}$ showed best performance

# Local Clusters Significant for both $Q_{ik}$ (p=0.001) and $Q_{ikt}$ (p=0.05)

| | Cluster Region | Number of Cases in Cluster | No. of Nearest Neighbors | True Positives[a] | False Positives[b] | Max. Size of False Positive Cluster[c] |
|---|---|---|---|---|---|---|
| US Case-Control Dataset | Iowa, 500 cases, 500 controls | N=0 (purely random) | k=5 | N/A | 0 | 0 |
| | | | k=10 | N/A | 1 | 1 |
| | | | k=15 | N/A | 0 | 0 |
| | | N=5 | k=5 | 0 | 2 | 1 |
| | | | k=10 | 0 | 1 | 1 |
| | | | k=15 | 0 | 1 | 1 |
| | | N=12 | k=5 | 0 | 0 | 0 |
| | | | k=10 | 0 | 1 | 1 |
| | | | k=15 | 0 | 0 | 0 |
| | | N=18 | k=5 | 1 | 0 | 0 |
| | | | k=10 | 5 | 0 | 0 |
| | | | k=15 | 2 | 0 | 0 |
| | | N=27 | k=5 | 11 | 0 | 0 |
| | | | k=10 | 16 | 0 | 0 |
| | | | k=15 | 11 | 1 | 1 |
| | | | K=20 | 0 | 0 | 0 |
| | California, 500 cases, 500 controls | N=43 | k=5 | 0 | 0 | 0 |
| | | | k=10 | 2 | 1 | 1 |
| | | | k=15 | 0 | 0 | 0 |
| | California + Iowa, 500 cases, 500 controls | N=43 in Cal. N=27 in Iowa | k=10 | 6, in both Iowa and Cal. cluster regions | 0 | 0 |

Calif cluster:
Greater size,
lower density,
lower mobility

Each row presents results of one suite of Q-statistic analyses.

# Local Clusters Significant for both $Q_{ik}$ (p=0.001) and $Q_{ikt}$ (p=0.05) Cont'd

| Cluster Region | Number of Cases in Cluster | No. of Nearest Neighbors | True Positives[a] | False Positives[b] | Max. Size of False Positive Cluster[c] |
|---|---|---|---|---|---|
| Iowa, 1189 cases, 1189 controls | | k=5 | N/A | 2 | 1 |
| | N=0 (purely random) | k=10 | N/A | 1 | 1 |
| | | k=15 | N/A | 0 | 0 |
| | | k=5 | 0 | 3 | 2 |
| | N=5 | k=10 | 0 | 4 | 2 |
| | | k=15 | 0 | 4 | 1 |
| | | k=5 | 0 | 2 | 1 |
| | N=12 | k=10 | 0 | 2 | 1 |
| | | k=15 | 0 | 3 | 2 |
| | | k=5 | 3 | 2 | 2 |
| | N=18 | k=10 | 2 | 2 | 2 |
| | | k=15 | 1 | 3 | 2 |
| | | k=5 | 3 | 2 | 2 |
| | N=27 | k=10 | 3 | 3 | 2 |
| | | k=15 | 2 | 3 | 2 |

Did not perform as well differentiating clusters of smaller density from false positives.

Fairly consistent across choice of k-nearest neighbors.

Maximum size of false cluster never exceeds 2 individuals.

# Local Clusters Significant for both $Q_{ik}$ (p=0.001) and $Q_{ikt}$ (p=0.05) Cont'd

| | Cluster Region | Number of Cases in Cluster | No. of Nearest Neighbors | True Positives[a] | False Positives[b] | Max. Size of False Positive Cluster[c] |
|---|---|---|---|---|---|---|
| Danish Case-Control dataset | Viborg, Denmark, 3297 cases, 3297 controls | N=0 (purely random) | k=5 | N/A | 0 | 0 |
| | | | k=10 | N/A | 0 | 0 |
| | | | k=15 | N/A | 1 | 1 |
| | | | k=20 | N/A | 2 | 1 |
| | | N=11 | k=5 | 0 | 0 | 0 |
| | | | k=10 | 0 | 0 | 0 |
| | | | k=15 | 0 | 0 | 0 |
| | | | k=20 | 0 | 0 | 0 |
| | | N=41 | k=5 | 0 | 0 | 0 |
| | | | k=10 | 2 | 0 | 0 |
| | | | k=15 | 3 | 0 | 0 |
| | | | k=20 | 5 | 2 | 1 |
| | | N=90 | k=5 | 1 | 0 | 0 |
| | | | k=10 | 2 | 1 | 1 |
| | | | k=15 | 10 | 0 | 0 |
| | | | k=20 | 11 | 3 | 1 |
| | | N=127 | k=5 | 5 | 0 | 0 |
| | | | k=10 | 6 | 0 | 0 |
| | | | k=15 | 22 | 1 | 1 |
| | | | k=20 | 32 | 4 | 1 |

Performs better for larger clusters (but still not that large: size ~2-4%!).
Some differences across choice of k-nearest neighbors.

# Supplementary Analyses

- We ran FDR p-value adjustment on two of the simulated clusters (only 2 because time-consuming), using 9999 randomizations to create reference distribution

| Cluster Region | Number of Cases in Cluster | No. of Nearest Neighbors | True Positives[a] | False Positives[b] | Max. Size of False Positive Cluster[c] |
|---|---|---|---|---|---|
| Iowa, 500 cases, 500 controls | N=18 | k=10 | 5 | 0 | 0 |
| | FDR results | | 0 | 0 | 0 |
| | N=27 | k=10 | 16 | 0 | 0 |
| | FDR results | | 5 | 0 | 0 |

Suggests FDR is more conservative than combined $Q_{ik}$, $Q_{ikt}$ approach.

# Conclusions

- These are the first simulation analyses of Q-statistics and provide several insights into their performance:
  - Ability to detect cluster is sensitive to # of cases, cluster size, density, and population mobility
  - Global $Q_k$ is conservative, unable to detect localized clusters
  - Local $Q_{ik}$ and $Q_{ikt}$ were able to identify strong true clusters, occasionally without false positives, using a critical value for $Q_{ik}$ of $p=0.001$ and examining $Q_{ikt}$ ($p \leq 0.05$) only among those individual cases significant for $Q_{ik}$.
  - Choice of k not critical for these ranges of cluster characteristics

# Conclusions cont'd

- Recommendation from these limited simulations:
  - Begin analyses using k=10 or k=15 neighbors
  - A cluster of three significant ($Q_{ik}$, $Q_{ikt}$) individuals or larger can be called a true positive and is a good starting point for follow-up studies
    - Only useful for distinguishing dense, large, low mobility clusters
    - Misses smaller, lower density, less persistent clusters
    - At this stage in development of Q-statistics, we feel this is an acceptable compromise since it limits inquiry into false positives, thereby conserving limited resources for more thorough investigations of true clusters
    - Are implementing this rule set with these (non-simulated) datasets

# Future Work

- Generalizability uncertain: differences such as edge effects, population density, mobility patterns, case-control ratio, and cluster shape, size, and density
- At this juncture, we recommend user conducts similar sets of simulation analyses on each dataset to determine the best criteria (p-values, number of *k* nearest neighbors) for identifying true positive clusters
  - In time we hope a consistent rule set will emerge
  - Alternatively, could explore wide library of potential clusters, datasets, and geographies to derive more empirical rule-set(s) and sensitivity to cluster characteristics; this would take a very long time.
- Comparing results of Q-statistics with other recently developed methods for mobile populations (Sabel et al., 2009; Webster et al., 2006) is also important

# Thank you!

Contact details:
jrmeliker@gmail.com

The perspectives are those of the authors and do not necessarily represent the official position of the funding agency.

We thank Patricia Hartge, ScD, principal investigator of the NCI-SEER NHL case-control study for use of the data.

We also acknowledge the contribution of the staff and scientists at the SEER centers of Iowa, Los Angeles, Detroit, and Seattle for the conduct of the study’s field effort. The NCI-SEER study was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Public Health Service (contracts N01-PC-65064, N01-PC-67008, N01-PC-67009, N01-PC-67010, N02-PC-71105).