SMALL AREA CLUSTERING OF CASES OF PNEUMOCOCCAL BACTEREMIA

.

JP Metlay, MD, PhD
T Smith, PhD
N Koizumi, PhD
C Branas, PhD
E Lautenbach, MD
NO Fishman, MD
PH Edelstein, MD

Center for Health Equity Research and Promotion, VA Medical Center, Departments of Medicine, Biostatistics and Epidemiology, Engineering, and Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA

ABSTRACT

Background. While recent studies have identified the existence of neighborhood-level factors that influence the carriage of *S. pneumoniae*, no studies have extended this observation to determine whether substantial small area variation exists in the risk of pneumococcal disease.
Methods. Data from population based surveillance for bacteremic pneumococcal pneumonia in the five-county region surrounding Philadelphia was analyzed for the existence of geographic clustering of cases of pneumococcal bacteremia. Detection of case clusters was accomplished using a spatial analytic method and confirmed with a second, distinct method..
Results. Over the period 2002-2004, there were 608 cases of pneumococcal bacteremia , of which 538 cases were geocoded. Case cluster analysis using both analytic methods demonstrated the existence of geographically distinct clusters within the region. The identification of clusters of pneumococcal disease was not explained by the racial or age distribution of the underlying population.
Conclusions. Cases of pneumococcal bacteremia demonstrate geographic clustering not accounted for by the underlying density and race-age distribution of the population. The identification of neighborhood-level factors underlying this clustering may have important implications for efforts to control pneumococcal disease and other respiratory pathogens.

Key Words: pneumococcal infections, small-area analysis, geographic information systems

INTRODUCTION

Streptococcus pneumoniae is a major cause of morbidity and mortality, responsible for a spectrum of diseases ranging from otitis media to meningitis. One of the more severe forms of pneumococcal infection is pneumococcal bacteremia, which continues to have a high mortality rate in adults despite the development of multiple effective antimicrobials. Thus, targeting of prevention strategies, particularly for high risk populations, remains a major public health approach to reducing pneumococcal morbidity and mortality. Prior research has emphasized that individual level factors strongly influence the risk of pneumococcal disease, including extremes of age, socioeconomic status, immunosuppression, and chronic heart and lung diseases.(1-6) However, an important parallel line of research has begun to uncover community-level factors that may influence the risk of transmission of pneumococcal bacteria, potentially shedding light on important prevention strategies.(7)

It is well known that invasive pneumococcal disease risk varies widely on a global scale.(8) To some extent, large-area geographic variation likely reflects climatic issues that are also observed in analyzing season trends in pneumococcal disease, leading to increased infection during periods of increased transmission of respiratory viruses that facilitate pneumococcal infection.(9) However, whether pneumococcal disease risk varies over small geographic areas is less well known. The identification of such small area variation in disease risk could ultimately shed light on other environmental factors, such as housing conditions and social interactions that influence pneumococcal disease risk and could reveal prevention strategies for the future.

Most studies of small area variation in pneumococcal disease risk have utilized existing geographic boundaries to calculate disease risk per unit area or analyze community-level risk factors.(1, 7, 10) However, these boundaries may have very little relationship to the social and environmental structures that influence pneumococal disease transmission. In contrast, the present methods of "hot spot" analysis allow small area variation in disease risk to be studied without utilizing any assumptions about existing geographic boundaries. Such approaches may thus provide substantially greater insight into the existence of small area variations in disease risk that can shed light on community-level factors promoting or impeding pneumococcal disease transmission. Moreover, such approaches permit adjustment for types of population heterogeneity that may underlie some of the observed small area variation in disease risk.

The aims of this study were: 1) to examine the geographic distribution of invasive pneumococcal disease in a five county region of Pennsylvania to determine whether geographic clusters of disease exist and 2) to analyze the extent to which such case clustering can be explained by characteristics of the underlying population.

MATERIALS AND METHODS

Study population

This study was conducted within the five-county region surrounding Philadelphia, PA: Bucks, Chester, Delaware, Montgomery, and Philadelphia counties. The adult population (age ≥ 18 years) of this region is 2,881,132 (US Census 2000). At the start of the surveillance period, there were 46 acute care hospitals serving this region. Of these, 43 hospitals participated in this study. Two of the remaining three hospitals were small hospitals outside Philadelphia county that

were closed to external studies and one was a larger academic hospital that was unable to participate (but was projected to account for < 2 percent of all cases in the region).

Subjects

For population surveillance, the case definition was restricted to adults ≥ 18 years with at least one blood culture drawn within 48 hours of hospital admission with growth of *S. pneumoniae*; residence in one of the five counties; and confirmation in our laboratory that the bacterial isolate was *S. pneumoniae* (see below). Cases were further restricted based on physician reports to those cases with radiographic evidence of an acute respiratory infection. These restriction criteria reflected the aims of the parent study to explore risk factors and outcomes for adults with bacteremic pneumococcal pneumonia. Exclusion criteria for the case-control study included evidence of bacterial meningitis (CSF growth of *S. pneumoniae* or CSF findings compatible with bacterial meningitis) or hospitalization within 10 days preceding the index hospitalization.

Cases were identified by microbiology laboratory personnel at each participating hospital. Validation of this surveillance system was established by comparing the number of cases identified by the research team to the total number of pneumococcal bacteremia cases reported to the Philadelphia Health Department under a mandatory reporting system in 2002. The study surveillance system identified 97% of the cases reported to the Philadelphia Health Department (all of the non-identified cases came from one of the non-participating hospitals). Whenever laboratory personnel identified a blood culture with growth of *S. pneumoniae*, research staff contacted the physician of record to determine subject eligibility. Eligible subjects (or proxies in cases of mental incompetence or death) were then approached for study enrollment at a time determined by the treating physician (typically after hospital discharge). Subjects were mailed informational study materials and then contacted by phone to provide consent for study participation and complete a telephone interview. This study was approved by the institutional review board at the University of Pennsylvania and each of the participating hospitals.

Microbiological data collection

Pneumococcal blood isolates were transported to a central laboratory at the Hospital of the University of Pennsylvania for analysis. Isolates were re-identified to confirm that they were pneumococci on the basis of colony morphology and hemolytic activity, Gram stain appearance, catalase reaction, bile solubility, and optochin susceptibility. (11)

Spatial cluster analyses

The residential addresses of all subjects in the study were geocoded (that is, assigned geographical coordinates) using ArcView 9.0 (ESRI) software and the StreetMaps USA reference database.

Two distinct methods of spatial cluster analysis were employed. First the existence of residential address data allowed individual case locations to be analyzed as a point pattern. Here a variant of the *point cluster analysis* of Rushton and Lolonis (1996) was employed,(12) in which a grid of reference points was constructed over the region at about one half mile on center. To test for clustering at various scales, a range of radii, d, were selected, and approximate circular regions, R_i, about each grid point, i, were constructed by including all city blocks within centroid distance d of point i. This allowed exact block-level census data to be used for

the analysis. The simplest null hypothesis considered for testing purposes was that each resident in the total region, R, is equally likely to be a case (refinements of this hypothesis are considered in the next section). If the total populations in R_i and R are denoted by n_i and n, respectively, then under this hypothesis, the probability that a randomly sampled case occurs in R_i is simply, $p_i = n_i / n$. Hence if the total case counts in R_i and R are denoted by c_i and c, respectively, and if individual cases are assumed to be statistically independent events, then the probability of observing c_i cases in R_i given c cases in R is given by the Binomial probability

$$(1) \qquad P(c_i \mid c) = \frac{c!}{c_i!(c-c_i)!} p_i^{c_i} (1-p_i)^{c-c_i}$$

The appropriate *p-value*, P_i, for a (one sided) test of clustering at i is thus simply the probability of observing as many cases as c_i under this null hypothesis, i.e.,

$$(2) \qquad P_i = \sum_{k=c_i}^{c} P(k \mid c)$$

This procedure was programmed in Matlab, and grids of p-values were computed for a range of selected radii. Results presented here are limited to the use of a radius of one half mile, which produced the most visually coherent group of clusters. *A priori*, we defined a meaningful cluster of cases as one which included at least 5 cases within the half mile radius since smaller values generated substantially larger numbers of clusters due to the overall rarity of invasive pneumococcal infection.

One shortcoming of the above procedure is that the identification of cluster size is somewhat *ad hoc* in that it is based largely on a visual inspection of the plotted results for selected radii. However there is a second well known method of *regional cluster analysis* due to Besag and Newell (1991) that seeks to determine a single "most significant" clustering of regions.(13) As a sensitivity analysis to examine the robustness of our findings, we chose to construct a version of Besag-Newell at the census tract level. The basic idea is to start with a given census tract i and to test for significant clustering in this tract in the same manner as above, where the relevant region is now tract i, denoted by R_{i0}. But rather than stopping here, one can then "grow" a larger region by adjoining to tract i the adjacent tract j_1 closest to i in centroid distance. For this larger region, say R_{i1} with total population, $n_{i1} = n_i + n_{j_1}$, and case count, $c_{i1} = c_i + c_{j_1}$, one may again test for clustering by setting $p_{i1} = n_{i1} / n$, and replacing p_i and c_i in (1) above with p_{i1} and c_{i1}, respectively. If the resulting p-value in (2) is denoted by P_{i1}, then by successively adjoining closest tracts to the existing cluster, one can produce a set of p-values, $\{P_{ij} : j = 0,1,..,J\}$, for a sequence of successively larger clusters, $j = 0,1,..,J$, where P_{i0} is the original p-value for tract i alone. The *most significant i-cluster* is then taken to be the cluster $R_i \equiv R_{ij}$ with the lowest p-value, $P_i \equiv P_{ij} = \min\{P_{ik} : k = 1,..,J\}$. Hence the *most significant cluster* in R is then taken to be the smallest of these, namely the minimum p-value cluster for that tract i^* with

(3) $P_{i^*} = \min_i P_i$

Adjustment for population heterogeneity

Because pneumococcal disease risk is known to vary with specific individual level characteristics, it was necessary to account for heterogeneity in the distribution of these characteristics in order to eliminate these factors as explanations for any observed small area variation in disease risk. To account for population heterogeneity with respect to risk of bacteremic pneumococcal pneumonia, it is necessary to refine the simple null hypothesis of "homogeneous risk" used above. Here the expected case rates for various population subgroups at the city-block level (as subgroups of either the one half mile radii applied for the point cluster analysis or census tracts used for the Besag-Newell approach) were estimated using absolute annual incidence rates of invasive pneumococcal disease for population subgroups in 2002 provided by the Centers for Disease Control through their Active Bacterial Core Surveillance system.(14) While these absolute rates apply to all invasive pneumococcal disease (including non-pneumonia associated cases), the underlying assumption was that the relative rates would be the same for population subgroups when restricted to cases of bacteremic pneumococcal pneumonia. The underlying size of the subgroup populations within each city-block were estimated using 2000 Census data. In particular, if the population, n_b, of block b is partitioned into relevant subpopulations, $(n_{bk} : k = 1,..,K)$ with $n_b = \Sigma_k n_{bk}$, and if the reported annual incidence rate for invasive pneumococcal disease in subpopulation *k* is denoted by r_k, then the total number of cases expected in block b is given by

(4) $e_b = \sum_{k=1}^{K} r_k n_{bk}$

Hence the corresponding *risk-adjusted null hypothesis* is simply that the probability, p_i, of observing a randomly sampled case in region R_i is now proportional to the number of cases, $e_i = \Sigma_{b \subset R_i} e_b$, expected in that region, i.e.,

(5) $$p_i = \frac{e_i}{e_R} = \frac{\sum_{b \subset R_i} e_b}{\sum_j \left(\sum_{b \subset R_j} e_b \right)}$$

This refinement into subpopulations was carried out in two stages. First the population was partitioned by race, using only white and black (since only five cases did not self-identify into one of these two groups). The point cluster analysis in (1) and (2) was then carried out using (4) and (5) for these 2 subpopulations. Next, this classification was further stratified by age, using five age groups as reported in the CDC data on population rates of disease (18-34, 35-49, 50-64, 65-79, ≥ 80 years).(14) The above analysis was then repeated for this partition into 10 subpopulations.

Of note, this approach is limited to exploring those individual level characteristics that are available both in terms of known expected population rates of disease and in terms of distribution in the population. Moreover, simultaneous analysis of multiple individual level factors requires knowledge of *joint* disease risks. In effect, these constraints resulted in our focusing only on adjusting for individual effects of age, race and age/race in the cluster analyses.

For all point analyses, the figures display the calculated p-value grids and do not reflect the actual point locations of any observed cases. In particular, each dot on the map is a grid point that is within half a mile of at least five cases, and for which this cluster of cases is significant in the p-value range shown. For the Besag-Newall calculation, the figure reflects the census aggregations of consecutively larger p values.

RESULTS

Case identification

Over the period April 1, 2002 through March 31, 2004, we identified a total of 608 adult cases of bacteremic pneumococcal pneumonia in the 5 county region surrounding Philadelphia (Bucks, Chester, Delaware, Montgomery and Philadelphia counties). The overall adult population annual risk of disease was 10.6 per 100,000. Of 608 total subjects, 545 had some geographical information about their residence (geocoding candidates). Out of the 545 candidates, 4 candidates provided only city or county information (one subject in each of Bucks, Chester, Montgomery, and Philadelphia counties). These 4 cases were subsequently removed from the analysis. Among the remaining 541 with complete address information, 3 addresses (two in Philadelphia county and one in Bucks county) could not be geocoded due to non-existent information and were also removed from the analysis. The final dataset, therefore, consisted of 538 total subjects of which, by county, 63 were from Bucks, 42 from Chester, 78 from Delaware, 105 from Montgomery and 250 from Philadelphia Counties.

Cluster analysis under homogeneous population risk

Figure 1 displays the initial cluster analysis applying a half-mile radius as the bandwidth. Of note, multiple approaches varying the bandwidth failed to detect any significant clusters outside of Philadelphia County. Therefore, all subsequent analyses focused on Philadelphia County alone.

The overall hot spot analysis identified several regions with significant geographic clusters over the two year observation period. In order to confirm the existence of multiple clusters, we analyzed the same data using a second hot spot detection technique, specifically the Besag Newell method, sequentially removing clusters from the analysis as they were identified. In this method, the four most significant clusters are identified (Figure 4) in locations that coincide with four clusters identified in the first method (Figure 1), confirming the existence of multiple discrete hot spots.

Cluster analysis under heterogeneous population risk

The analyses above all examined the underlying adult population distribution under the assumption of homogeneous risk. Of note, the patients located in the observed geographic clusters were almost all self-identified as black. Thus, it was natural to ask whether this cluster could be accounted for by the underlying race distribution in the region. To answer this question, the above point cluster analysis was repeated generating expected case distributions based on block-level counts of blacks and whites separately. The results in Figure 2 show that while all clusters are somewhat less significant than before, the most significant clusters in Figure 1 remain. This is partly due to the fact that while these clusters are mostly black, there are other communities in the region with dense black populations but with relative few cases. This

analysis was repeated using the Besag-Newell procedure at the tract level, and produced essentially the same results (not shown) as the point-cluster analysis.

Next, the degree of underlying population heterogeneity was further refined by including age/race categories (with five age categories and two race categories yielding a total of 10 subgroups) (Figure 3). Here the significance levels again seem to be reduced to some degree; but the most significant clusters in Figures 1 and 2 remain. Again, a Besag-Newell analysis for this case produced essentially the same results (not shown) and added further confirmation to these findings.

DISCUSSION

Our analyses demonstrate that the incidence of invasive pneumococcal disease exhibits significant geographic heterogeneity over small areas. This geographic heterogeneity is not fully explained by selected individual level demographic factors that are known to modify the risk of pneumococcal infection. In addition, the specific location of disease clusters was confirmed by a second analytic approach.

Prior studies on pneumococcal disease have emphasized individual level heterogeneity in disease risk. On a more global scale, it is well established that certain populations experience above average disease risk.(8) However, it is not clear whether such elevated risk reflects endogenous characteristics of inhabitants of these regions or exogeneous environmental factors that mediate disease risk. It is well established that pneumococcal disease displays seasonal and climatic variation,(9) which may be an important consideration in interpreting global and national patterns of pneumococcal disease.

However, small area variations in disease risk are unlikely to reflect variation in climatic factors. More likely, if these results are confirmed, the existence of pneumococcal disease clusters suggests that small area variation in community characteristics influence the spread of pneumococcal bacteria among hosts in an area. Such relevant factors could include the structure of housing and shared living spaces and, especially, the relationship of day care centers and schools within a community. Prior research confirms that as the proportion of children in a community attending day care increases, the probability of pneumococcal carriage increases among both attendees and non-attendees of the day care centers.(7) Our results expand on this observation by indicating that risk of invasive disease can vary across communities in a small geographic region. It is possible that as the density of pneumococcal carriage increases within a small area, the incidence of invasive pneumococcal disease increases proportionately.

One limitation of this study is that we were able to control for only a small number of potential endogenous characteristics of patients that could influence the observed geographic distribution of disease. For the approaches adopted in this study, we were limited to exploring only those factors for which expected population rates were known and for which distributions of the characteristics were known in the source population. Thus, we could not explore the impact of the distribution of clinical factors (e.g., chronic heart and lung disease) or social factors (e.g., day care center utilization) on the observed small area heterogeneity. .

A second limitation is that we can not eliminate the possibility that the observed geographic heterogeneity reflects heterogeneous distribution of pneumococcal clones with variable invasive potential. Thus, the introduction of a particularly virulent clone in one community may lead to a higher observed case rate in that area due to the chance introduction of

that clone. Similar outbreaks have been observed in more closed environments. In this regard, it is notable that only two percent of all patients in this study were nursing home residents at the time of infection and only two patients were residents in the same nursing home in Philadelphia, making it unlikely that outbreaks in closed environments explain our results(data not shown). Future work will specifically examine molecular heterogeneity of isolates to distinguish mechanisms of clonal spread from other community-level factors that influence disease transmission.

In summary, we have observed significant geographic clusters for bacteremic pneumococcal pneumonia within a single urban region, a level of small area variation in pneumococcal disease risk not previously observed. Such clusters may provide important opportunities for identifying community-level factors that modify disease transmission risk—such factors could ultimately be valuable for designing public health interventions in the face of emerging pathogens. Moreover, better understanding of the multi-level nature of pneumococcal disease risk may play an important role in targeting limited resources to the highest risk populations.

REFERENCES

1. Chen FM, Breiman RF, Farley M, Plikaytis B, Deaver K, Cetron MS. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumoniae infections. Am J Epidemiol 1998;148(12):1212-8.

2. Nuorti JP, Butler JC, Farley MM, et al. Cigarette smoking and invasive pneumococcal disease. Active Bacterial Core Surveillance Team. N Engl J Med 2000;342(10):681-9.

3. Filice GA, Darby CP, Fraser DW. Pneumococcal bacteremia in Charleston County, South Carolina. Am J Epidemiol 1980;112(6):828-35.

4. Breiman RF, Spika JS, Navarro VJ, Darden PM, Darby CP. Pneumococcal bacteremia in Charleston County, South Carolina. A decade later. Arch Intern Med 1990;150(7):1401-5.

5. Lipsky BA, Boyko EJ, Inui TS, Koepsell TD. Risk factors for acquiring pneumococcal infections. Arch Intern Med 1986;146(11):2179-85.

6. Talbot TR, Hartert TV, Mitchel E et al. Asthma as a risk factor for invasive pneumococcal disease. N Engl J Med 2005;352(20):2082-90.

7. Huang SS, Finkelstein JA, Lipsitch M. Modeling community- and individual-level effects of child-care center attendance on pneumococcal carriage. Clin Infect Dis 2005;40(9):1215-22.

8. Fedson DS, Scott JA. The burden of pneumococcal disease among adults in developed and developing countries: what is and is not known. Vaccine 1999;17 Suppl 1:S11-8.

9. Talbot TR, Poehling KA, Hartert TV, et al. Seasonality of invasive pneumococcal disease: temporal relation to documented influenza and respiratory syncytial viral circulation. Am J Med 2005;118(3):285-91.

10. Huang SS, Finkelstein JA, Rifas-Shiman SL, Kleinman K, Platt R. Community-level predictors of pneumococcal carriage and resistance in young children. Am J Epidemiol 2004;159(7):645-54.

11. Koneman E, Allen S, Janda W, Schreckenberger P, Winn Jr. W. Color atlas and textbook of diagnostic microbiology. 5th ed. Philadelphia: Lippincott-Raven; 1997.

12. Rushton G, Lolonis P. Exploratory spatial analysis of birth defect rates in an urban population. Statistics in Medicine 1996;15:717-726.

13. Besag J, Newell J. The detection of clusters in rare diseases. Journal of the Royal Statistical Society, Series A. 1991;154:143-155.

14. Flannery B, Schrag S, Bennett NM, et al. Impact of childhood vaccination on racial disparities in invasive *Streptococcus pneumoniae* infections. JAMA 2004;291(18):2197-203.

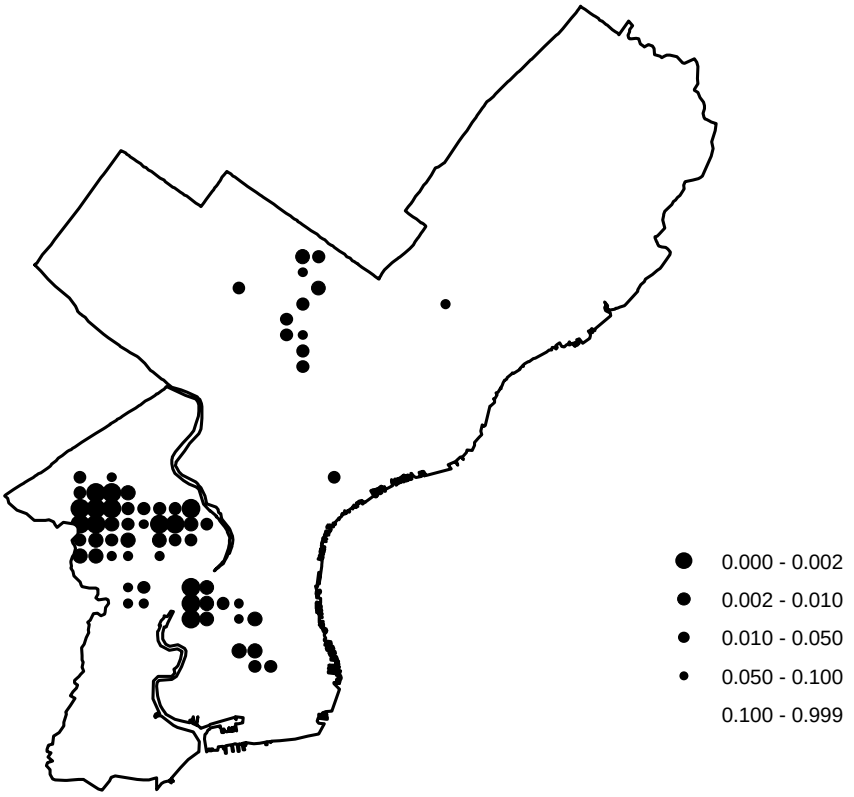

Figure 1. Point location cluster analysis under an assumption of homogeneous adult population risk. The figure depicts the Philadelphia county region. A grid of 0.5 mile dimensions is overlaid on the figure and we calculate the p values associated with each cluster analysis at each grid intersection. The analysis is conducted for 0.5 mile radii, excluding clusters with < 5 cases within the circle. Of note, the location of the plotted p values does not represent actual case locations. P-values > 0.1 are not plotted on the map.

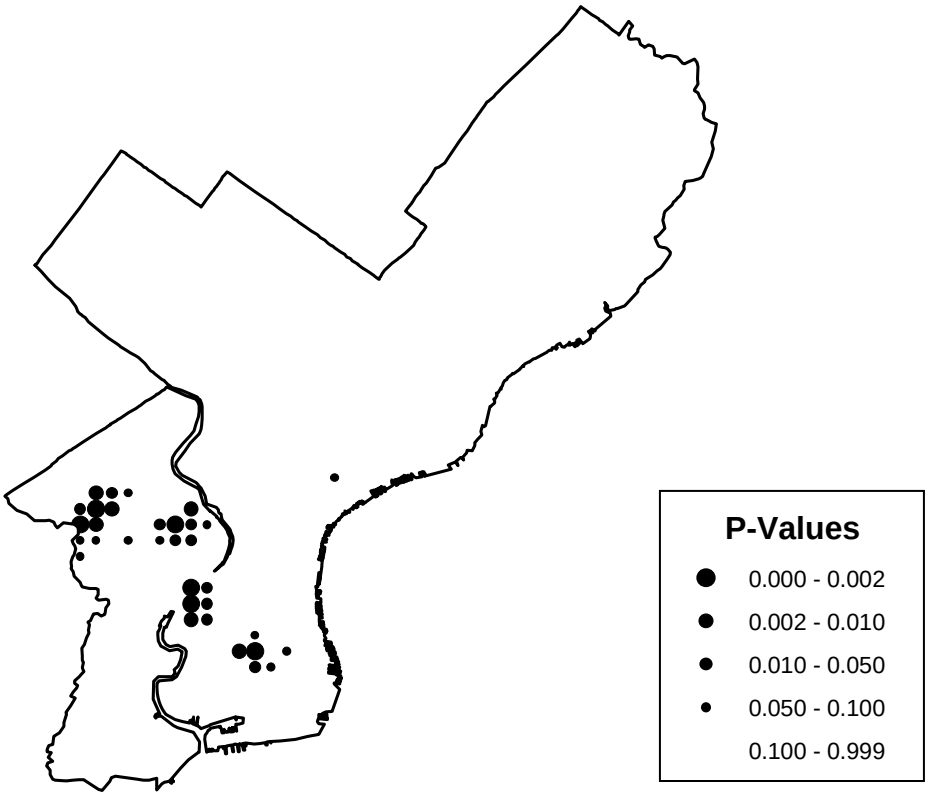

Figure 2. Point location cluster analysis adjusting for the black and white adult population distribution in Philadelphia. The figure depicts the Philadelphia county region. A grid of 0.5 mile dimensions is overlaid on the figure and we calculate the p values associated with each cluster analysis at each grid intersection. The cluster analysis adjusts for the underlying black and white adult population distribution from the Census data and applies absolute race specific rates of disease based on CDC national surveillance data from 2002. The analysis is conducted for 0.5 mile radii, excluding clusters with < 5 cases within the circle. Of note, the location of the plotted p values does not represent actual case locations. P values > 0.1 are not plotted on the map.

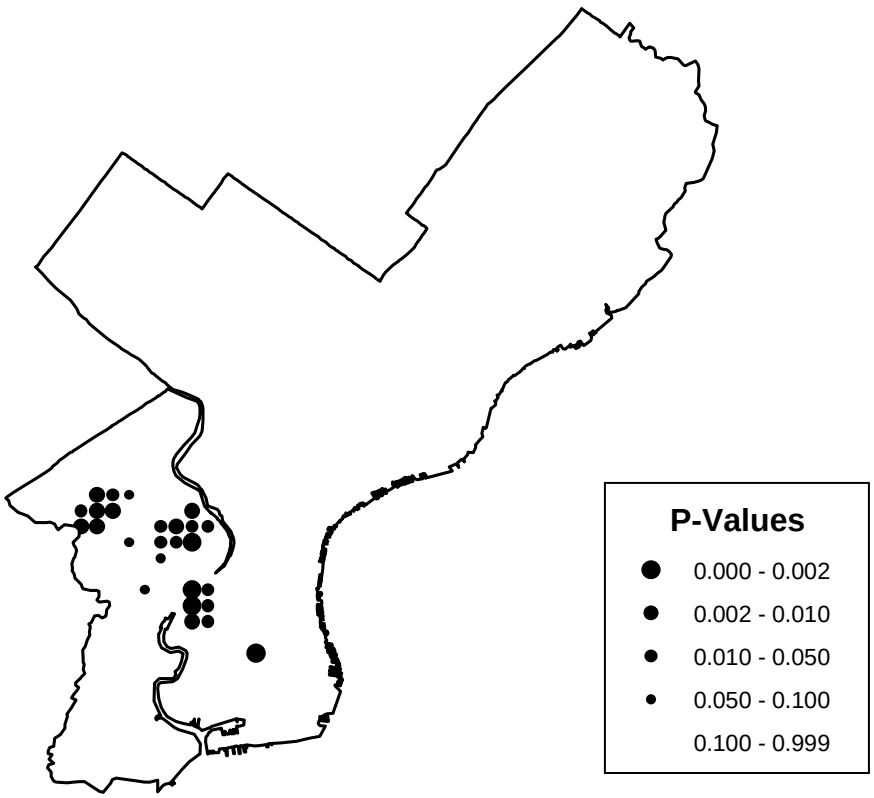

Figure 3. Point location cluster analysis adjusting for the underlying age and racial distribution of the population. The figure depicts the Philadelphia county region. A grid of 0.5 mile dimensions is overlaid on the figure and we calculate the p values associated with each cluster analysis at each grid intersection. The cluster analysis adjusts for the underlying black and white adult population distribution in age strata available from the Census data and applies absolute age by race specific rates of disease based on CDC national surveillance data from 2002. The analysis is conducted for 0.5 mile radii, excluding clusters with < 5 cases within the circle. Of note, the location of the plotted p values does not represent actual case locations. P values > 0.1 are not plotted on the map.

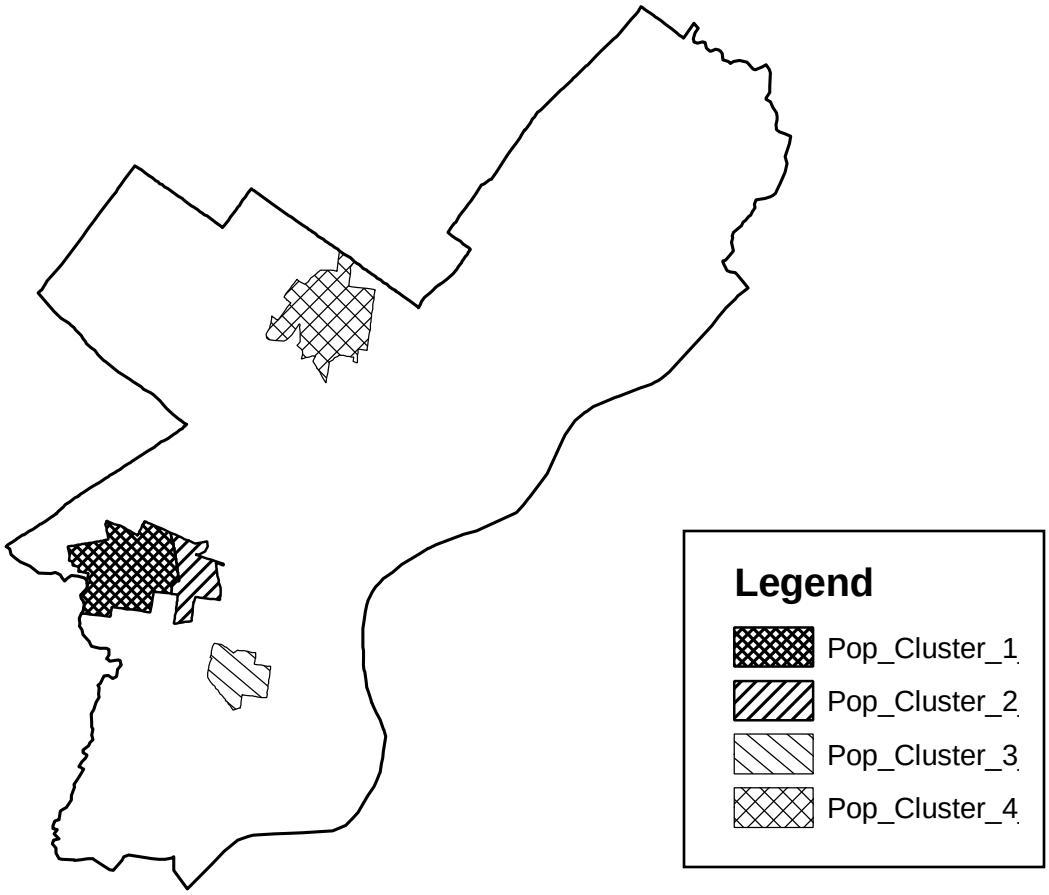

Figure 4. Besag Newell cluster analysis under an assumption of homogenous adult population risk. The figure depicts the Philadelphia county region. Cluster analysis is conducted at the census tract level, examining all possible combinations of contiguous census tracts to identify the clusters with the lowest probability of observation under an assumption of homogenous risk in the population. As each cluster is identified, it is removed from subsequent analyses. In this figure, the four clusters with the lowest probability of observation are depicted. The p values associated with each cluster are as follows: Population_Cluster_1 (p=.0000000165), Population_Cluster_2 (p=.0000475), Population_Cluster_3 (p=.000035), Population_Cluster_4 (p=.00088). Note, the sequential removal procedure does not guarantee that each successive cluster has a larger p value than the preceding cluster—hence, the p value for the third identified cluster is slightly smaller than the p value for the second identified cluster. The analysis excludes clusters with < 5 cases per combination of census tracts.