Research paper

Spatial aspects of MRSA epidemiology: a case study using stochastic simulation, kernel estimation and SaTScan.

L. BASTIN*†, J. ROLLASON‡, A. HILTON‡, D. PILLAY§, C. CORCORAN§, J.ELGY†, P.LAMBERT‡, P.DE§, T.WORTHINGTON‡ AND K.BURROWS‡

† School of Engineering and Applied Science, University of Aston, Birmingham, UK

‡ School of Life and Health Sciences, University of Aston, Birmingham, UK

§ Good Hope Hospital NHS Trust, Sutton Coldfield, Birmingham, UK

Correspondence *L. Bastin. Email: l.bastin@aston.ac.uk

The identification of disease clusters in space or space-time is of vital importance for public health policy and action. In the case of meticillin-resistant *Staphylococcus aureus* (MRSA) it is particularly important to distinguish between community and healthcare-associated infections, and to identify reservoirs of infection. 832 cases of MRSA in the West Midlands (UK) were tested for clustering and evidence of community transmission, after being geo-located to the centroids of UK unit postcodes (postal areas roughly equivalent to Zip+4 zip code areas). An age-stratified analysis was also carried out at the coarser spatial resolution of UK

Census Output Areas. Stochastic simulation and kernel density estimation were combined to identify significant local clusters of MRSA ($p<0.025$), which were supported by SaTScan spatial and spatio-temporal scan. In order to investigate local sampling effort, a spatial 'random labelling' approach was used, with MRSA as cases and MSSA (meticillin-sensitive *S. aureus*) as controls. Heavy sampling in general was a response to MRSA outbreaks, which in turn appeared to be associated with medical care environments. The significance of clusters identified by kernel estimation was independently supported by information on the locations and client groups of nursing homes, and by preliminary molecular typing of isolates.

In the absence of occupational/lifestyle data on patients, the assumption was made that an individual's location and consequent risk is adequately represented by their residential postcode. The problems of this assumption are discussed, with recommendations for future data collection.

Keywords: MRSA; Stochastic simulation; kernel density; SaTScan; random labelling.

AMS Subject Classification: [62H12 (Estimation): 62H30 (Classification and discrimination; cluster analysis): 62M30 (Spatial processes): 91B72 (Spatial models)]

1 Introduction

1.1 SPATIAL EPIDEMIOLOGY

Clustering of disease in space and/or time due to environmental factors or infection events is a vital topic to inform public health protection and policy. The application of spatial analysis in epidemiology is well established, with many well-tested techniques for visualisation and exploration of disease pattern, and a variety of methods for assessing cluster significance. The locations at which disease events are recorded may be points which accurately represent the location at which the event occurred, or they may be *aggregated* locations such as administrative polygons, or population-weighted centroids. The aggregation process is generally forced by three constraints:

a) the availability and cost of precise location data to which disease events can be mapped, such as individual street addresses.

b) the pre-aggregated nature of the demographic data (such as census variables) to which the disease events must be related in order to estimate rates, rather than counts.

c) legal/ethical data protection and anonymity requirements, which demand that demographic data are aggregated to a level where the individual can no longer be easily traced.

A major issue for epidemiological cluster analysis is the heterogeneity of underlying determinants such as population density, meaning that patterns are rarely tested against a null hypothesis of truly complete spatial randomness (CSR) but rather against simulated patterns based on what we know of those underlying factors. To generate these simulations, a model of disease incidence (often Poisson-based) is applied to the available population data. Models of infectious process and flow are outside the scope of this paper.

The work reported here addressed for *Staphylococcus aureus* isolates (MSSA, meticillin-sensitive *S. aureus* or MRSA, meticillin-resistant *S. aureus*) processed by a district general hospital in North Birmingham, UK, in

the one-year period between 01/09/2004 and 31/08/2005 (see Figure 1). Important questions in the community epidemiology of MRSA are:

a) Do reservoirs of infection exist outside hospitals and other health care settings, and can they be identified?

b) Can community transmission be distinguished from health care associated ('nosocomial') spread?

c) Are specific strains of MRSA more clustered in space than we would expect?

This paper aims to address question a) directly, and to identify some useful pointers towards tackling question b). Currently, molecular typing of the data used in this study is ongoing, and will be the basis for explicit models to address question (c), with particular emphasis on 'hyper-transmissible' strains of MRSA.

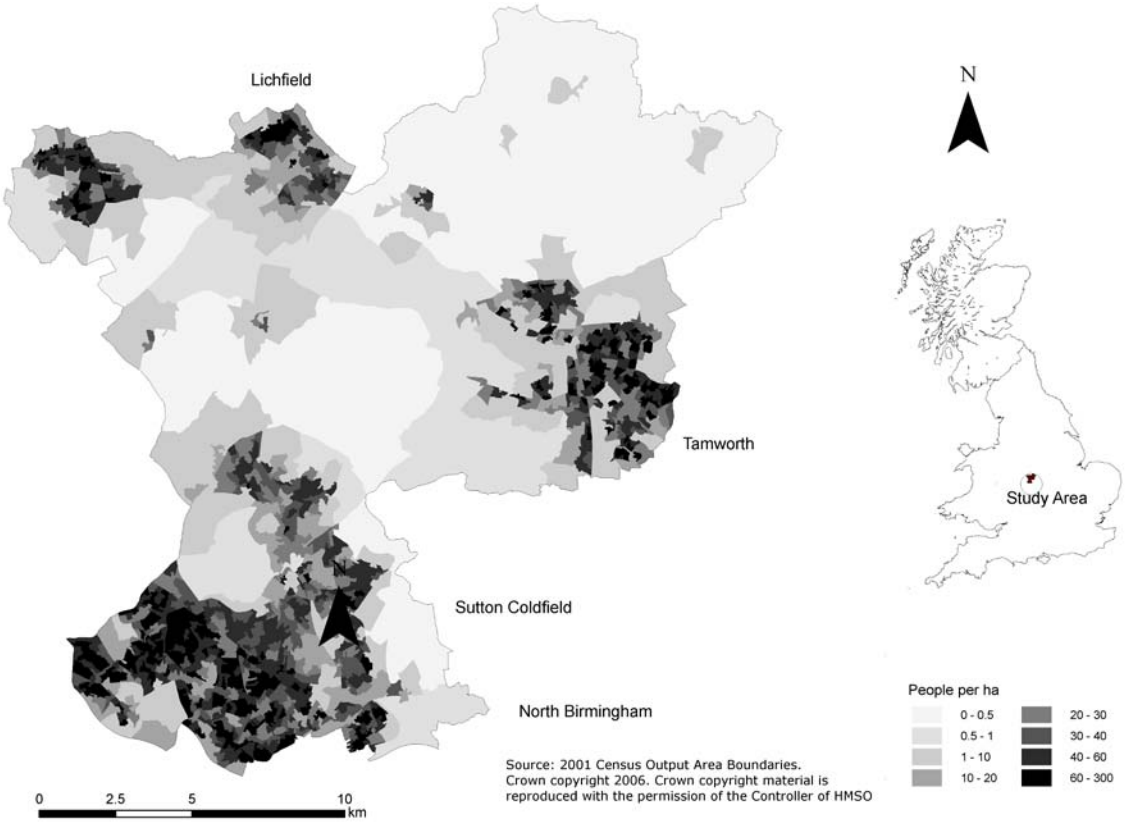

Figure 1. Geographic study bounds, showing population density and context within the UK. Population density is shown at the level of Census OA, for 2001.

The work was undertaken in order to assess the utility of several cluster analysis techniques for a longer-term spatial analysis of MRSA cases and subtypes, and particularly those cases which can be attributed to community transmission. The aim was to identify suitable techniques for integration into an exploratory visualisation framework, and to identify supplementary data which would be required for such an approach.

On the basis of public health importance, permission for this study was given by the Medical Director of Good Hope Hospital NHS Trust and the Director of Public Health of North Birmingham Primary Care Trust. Formal ethical approval was not required as patient identifiers were deleted from the datasets. To protect patient confidentiality while using spatially-referenced data, database extracts were made in a two-stage process. This ensured that postcoded records did not record date of birth, but only the broader age category (under 65, 65-85, 85+) of the patient.

1.2 MRSA – A CASE STUDY FOR CLUSTERING ANALYSIS

Staphylococcus aureus is a clinically important bacterium causing a wide range of clinical infections from superficial boils to systemic infections such as pneumonia, bacteraemia, phlebitis, meningitis, UTI, osteomyelitis, sepsis and endocarditis. Treatment is usually effective in a majority of cases where the bacterial strain is susceptible to a variety of antibiotics including meticillin (also known by the US Adopted Name of 'methicillin'); these strains are referred to as meticillin-sensitive *Staphylococcus aureus*, (MSSA). Meticillin-resistant strains of *Staphylococcus aureus* (MRSA) were first observed shortly after meticillin was introduced in the 1960s, and by the 1970s these had already spread worldwide. Limited treatment options exist for MRSA infection, due to its resistance to multiple antibiotics. Attributable mortality data is difficult to obtain, as those who die from MRSA are usually already ill and often their existing illness, rather than MRSA is designated as the underlying cause of death. MRSA-associated mortality in 2002 in the UK was recorded

as 248 cases and MRSA was estimated to be involved in 0.07% of all deaths in 2004; by comparison, the far more prevalent MSSA was recorded as the cause of death in 394 cases in 2002 (Griffiths *et al*., 2004).

MRSA is generally perceived by the public to be associated with poor application of hygiene standards in hospitals. However, since the 1980s, cases have been observed outside the hospital environment and within the community (Saravolatz *et al.*, 1982). Although many Primary Care Trusts publish summary MRSA statistics for selected hospitals, there is little definitive information on MRSA transmission and carriage rates outside large medical institutions. However, evidence is beginning to accumulate which implicates community-associated MRSA infection (Okuma *et al*., 2002; Holmes *et al*., 2005). In the absence of exhaustive public screening, *Staphylococcus aureus* is estimated to be carried by 30-50% of the adult population, with 10-20% being persistently colonised and 60% being intermittent carriers (Kluytmans *et al*., 1997). MRSA carriage is pragmatically assumed to be around 1% (Mainous *et al.*, 2006).

There is a notable distinction between the molecular sub-types of MRSA that prevail in medical environments and in the community. SCC*mec* I, II and III clones are characteristic of hospitals and display multiple resistance to antibiotics (Ito *et al*., 2001). SCC*mec* IV and V clones are commonly observed outside hospital environments; they show restricted antibiotic resistance, often being limited to beta-lactam antibiotics (Ma *et al*., 2002, Ito *et al*., 2004) but increased virulence (Baba *et al*., 2002). Community acquired MRSA are noted to be predominant in patients with no hospital connection (Chambers, 2001), no predisposing hospital risk (Herold *et al*., 1998) or who have been isolated between the first 48-72 hours of hospital admission.

1.3 AREA-BASED DISEASE DATA AND ANALYSIS TECHNIQUES

Disease data are sometimes attached to systematically defined sub-regions of a study area, such as electoral wards or clinic catchment areas. The primary benefit of an area-based approach is the ability to construct occurrence rates or frequencies based on known populations or susceptibilities within the areal unit. The

disadvantages stem from the 'prism-like' nature of the areal units, (implying uniform rates across each area), and from the sensitivity of the sampled spatial variation to the 'Modifiable Areal Unit Problem' (MAUP). For example, depending on the relative grain and placement of the areal unit boundaries, a significant cluster surrounded by sparse events may be divided and diluted to appear insignificant. Techniques such as regression also suffer from data aggregation and its potential masking of genuine relationships (the 'ecological inference' problem, as reviewed by Gelman *et al*., 2001).

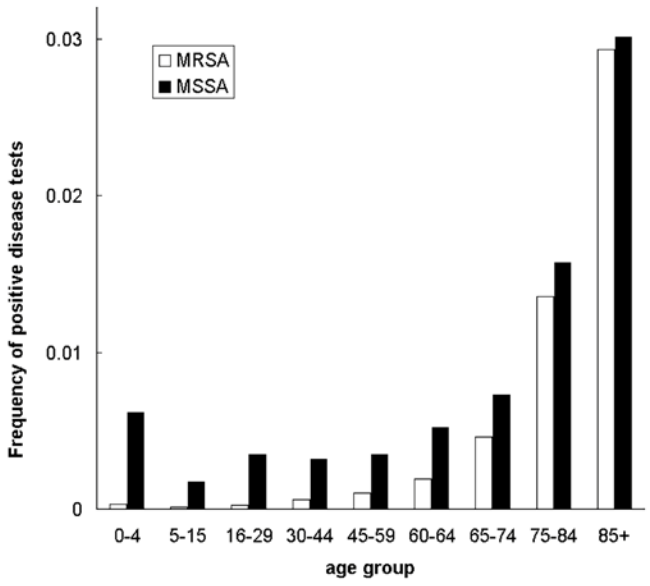

Figure 2a. Incidence of MRSA and MSSA by age group.

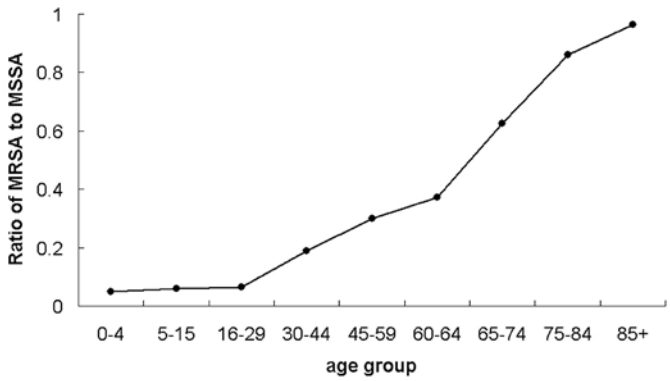

Figure 2b. Ratio of MRSA to MSSA cases, by age group.

Individual patient records for this study held information on the unit postcode of the patient's home address. The UK unit postcode is part of a coding system created and used by the UK Mail Service for the sorting and delivery of mail. The unit postcode (made up of between 5 and 7 characters) represents groupings of delivery points roughly equivalent to the US Zip +4 zip code classification. In the context of this study, some of the independent variables required for analysis (e.g., population age structure) were only available at Census Output Area level (henceforth, 'Output Area' will be abbreviated to 'OA'), forcing aggregation of unit postcode-level data to a coarser spatial resolution. Within the study area, each unit postcode covered a mean

of 19 domestic delivery points, (s.d. ± 16.1), and each OA contained a mean of 301 people (s.d. ± 62.6). Aggregation to OA population centroids also allowed a direct consideration of age-structured population, rather than an estimation of population from the number of domestic delivery points. This was very important, since the relative prevalence of MRSA varied with age (see Figures 2a and 2b). The case data themselves were thus handled spatially as points, but it should be noted that the effects of MAUP still apply to the underlying data.

1.3.1 LISA (Local indicators of spatial association)

A LISA aims to measure 'the extent of spatial clustering of similar values around an observation' (Anselin, 1995) and thus to identify pockets of local non-stationarity. LISAs such as the local Moran's *I* allow rates within areas to be assessed for clusters (i.e., associations of high or low rates) or outliers (i.e., low rates surrounded by high rates, or vice versa). In order to assess the context of each area, a matrix is used which records the 'adjacency' of each pair of polygons. This matrix may contain simple Boolean values ('touching/not touching'), distance between polygons, or a subtler contiguity value, based on the relative length of the shared boundary segment (as described, for example, in Haining, 1990). For this study, Moran's *I* was calculated for MRSA rates (using ArcMap 9.1 and a Boolean adjacency matrix), since a consideration of polygon adjacency, rather than centroid distance, may mitigate modifiable areal unit problems.

1.4 Point-based disease data and analysis techniques

The local observed intensity of a point set may be a function of a global trend across a region, local correlations between points or with environmental determinants, noise, or a combination of all three. The aim of spatial epidemiology is to tease out and model these elements so that reliable meaning can be extracted from the observed patterns.

The last two decades have seen substantial work in the GIS arena on the robust statistical analysis of point patterns (e.g., Openshaw *et al*., 1987, Diggle and Chetwynd, 1991, Besag and Newell, 1991). Useful summaries of issues and techniques for point pattern analysis can be found in Cressie (1993), while Gatrell *et al.* (1996) specifically review point pattern analysis techniques for epidemiological applications. The most relevant issues for this paper are:

a) *Spatial aggregation of data to the point centroids of administrative areas*. The original data were aggregated to unit postcode level, but, in order to make use of available demographic data, the spatial resolution was further coarsened by aggregation to census OA population centroids. One particular danger here is that the local intensity of a point pattern may be artificially boosted by the 'snapping' of surrounding points to a single centroid, implying sharper peaks of case frequency than actually occur at any point in space. The stochastic simulations in this study mitigate this effect by simulating cases at the same aggregated centroids, potentially generating similar spikes in the simulations. This 'cools' and moderates the peaks caused by aggregation.

b) *Correction for edge effects*. In choosing a finite study area for point pattern analysis, points just outside this boundary will not be included in any calculations. These 'edge effects' bias most estimators of pattern, and unless otherwise stated, all analyses in this study are edge-corrected.

1.4.1 Kernel estimation. Most statistical descriptions of point patterns sample the data by stepping across the map, selecting subset 'windows' of points and assessing the distribution of points within that area. In kernel estimation, points closest to the window's centre contribute most to the calculated index of intensity. The selection of kernel bandwidth is extremely important so that a balance can be achieved between conforming too tightly to noise and 'spikes' in the data, and obliterating important local variation altogether. An alternative to using a fixed kernel is adaptive kernel estimation, where bandwidth is varied according to

local point intensity so that a required minimum number of points are sampled (e.g., Brunsdon 1995), broadening and flattening the kernel where data are sparse. A fixed quartic kernel can be defined by the equation

$$\lambda_\tau (s) = \Sigma\ 3/\pi\tau^2\ \ (1 - h^2/\tau^2) \text{ for each } h_i < \tau \text{ ; otherwise } 0 \qquad (2)$$

where $\lambda\tau$ is the cumulative 'intensity' of neighbouring points, calculated for a kernel centred on point *s*, and τ is the bandwidth radius. The potential contribution of data point s_i to λ_τ varies smoothly, depending on the distance h_i between s_i and *s*, peaking at distance 0 and decreasing to 0 at τ.

It is important to note that kernel estimation is not in itself a technique for detecting clustering, but generates distributed maps of density which can be compared, cell-by-cell, between real point patterns and simulated spatial nulls (e.g., Kelsall and Diggle, 1995).

1.4.2 Ripley's K-function, pair correlation functions and random labelling. The K-function or *reduced second-moment function* (Ripley, 1981) describes the expected number of points within a distance h from a sample location. It is defined by

$$\lambda K(h) = E(\text{number of events within distance } h \text{ of an arbitrary event}) \qquad (3)$$

where λ is the intensity (mean number per unit area) of events. Plotting at multiple h values (or 'lags') allows point patterns to be distinguished even if their behaviour differs only at certain scales; a global index such as mean nearest-neighbour distance may fail in this regard. The observed pattern may be compared to a fully-

defined K-function describing systematic attraction or repulsion between points at each scale. The simplest example is CSR produced by a homogeneous Poisson process - i.e., simply the expected intensity of points within a circular window:

$$K(h) = \pi h^2 \qquad (4)$$

For a discussion of defined K-functions for more complex spatial processes, see Dixon (2002). In the area of epidemiology, Kingham et al. (1995) successfully combined K-function estimation with logistic regression in order to identify potential covariates in a study of childrens' respiratory health.

Edge corrections are commonly performed by comparing the current circular window to a defined boundary polygon, and weighting the results according to how much of the circumference of the circle falls within the boundary, as follows:

$$\hat{\lambda}\hat{K}(h) = \sum_i \sum_{j \neq 1} w(l_i, l_j)^{-1} \frac{I(d_{ij} < h)}{N} \qquad \text{(Ripley, 1976)} \qquad (5)$$

where d_{ij} is the distance between the *i*th and *j*th events in the area of interest, and $I(d_{ij})$ is an indicator function which is 1 if $d_{ij} \leq h$, and 0 otherwise. $w(l_i, l_j)$ is the term which corrects for reduced sampling at edges, and has the value 1 when a circle of radius h, centred at point l_i, falls entirely within the study boundary. Otherwise, $w(l_i, l_j)$ is a value less than 1 representing the proportion of that circle's circumference which falls within the study bounds.

In order to stabilize the variance of estimates of K(*h*) the expected point intensity under CSR (Equation 4) is employed to estimate Ripley's L(*h*):

$$\hat{L}(h) = [\hat{K}(h)/\pi]^{1/2} \quad (6)$$

The expected value of *L(h)* under a homogeneous random Poisson process is 0; positive values indicate clustering of points in space at this scale, and negative values indicate more regular spacing. Figure 3 illustrates the use of $\hat{L}(h)$ as a descriptor of spatial pattern for the data used in this study.

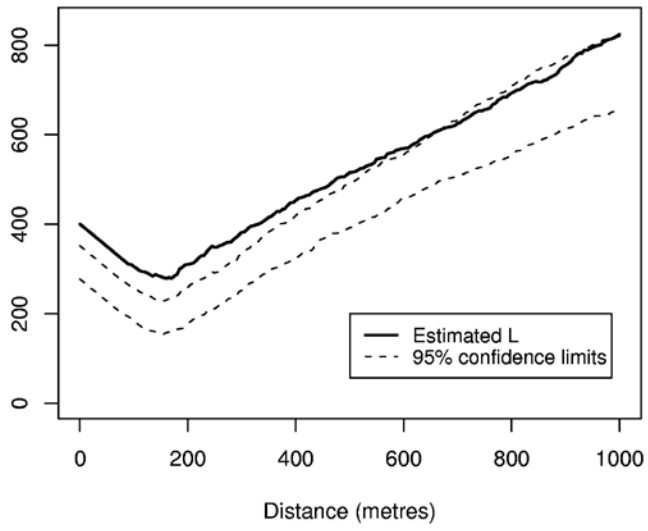

Figure 3. L-function for MRSA cases, shown against a 95% confidence envelope generated from 1000 stochastically simulated patterns based on underlying, age-stratified, population density.

Edge-corrected K-functions for all point sets were generated using the splancs 'khat' function, and transformed using equation 6 for plotting as L-functions. Randomly-simulated sets of points based on age-stratified population density at the same centroids were used to produce 95% confidence envelopes, which exhibit largely the same characteristic slopes and inflexion points as the disease data. (Envelopes were generated using the splancs function 'Kenv.csr' (Rowlingson and Diggle, 1993), and an adapted version of the

same function which read simulated point sets from file). It can be seen that MRSA cases are least clustered (i.e., most regular) at a radius of about 180 metres. This characteristic pattern stems from repulsion between OA centroids, which have an average spacing of ~200m. The MRSA cases show significant clustering above and beyond that of the underlying population, at lags between 0 and 600 metres (in contrast, the MSSA cases have a pattern which falls at the upper limit of their confidence envelope, but never significantly outside it).

It is important to note that both the K-function and the L-function are smoothed by measuring *cumulative* clustering up to distance *h.* The pair correlation function (pcf) shown below, on the other hand, considers the instantaneous clustering of a point process at each particular scale.

$$g(h) = K'(h) / 2\pi h \qquad (7)$$

where *K'(h*) is the derivative of the K-function described in (3), and *g(h)* tends to 1 at large *h*. Figures 4a and 4b show pair correlation functions for the data in this study. The oscillations in Figure 4a illustrate the structured pattern of the OA population centroids which stems from their tessellated nature. In particular, there is a downspike at around 100 m, which represents regular spacing at the average OA radius, and again around 3500 m, representing the average radius of settlements in the study area. While this pattern largely drives the spacing of the disease cases allocated to the centroids, the MRSA cases can be seen to be more clustered at very short distances (0-100 m), and are generally more clustered than the base data at distances up to 3000 m. The negative values around 3500 m correspond to the average radii of population clusters in the base data; in other words, as at 100m, neighbours are generally sparse, at this distance, for an arbitrarily-selected point within the dataset. Figure 4b compares the pcfs of MSSA and MRSA cases, showing that their patterns are broadly similar, but that MRSA is more clustered at very local scales.

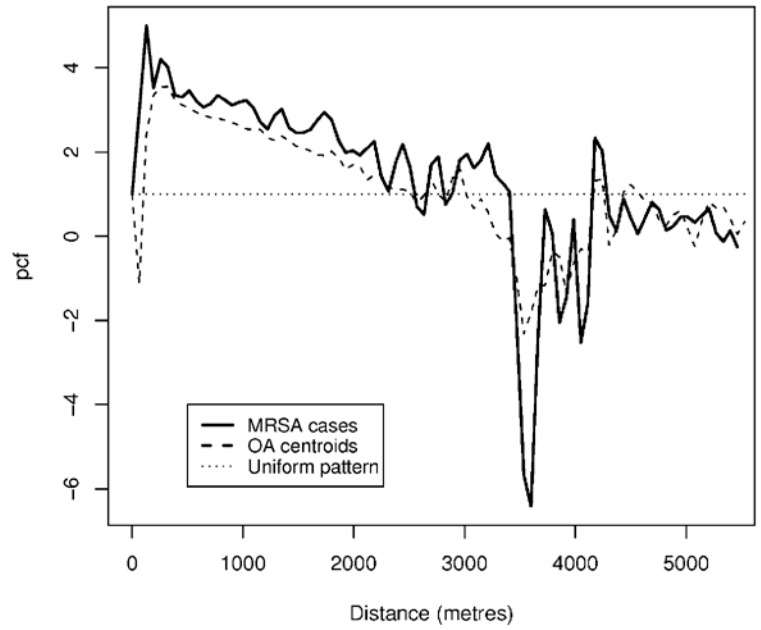

Figure 4a. Pair correlation functions for MRSA cases and the 1244 OA centroids to which data were aggregated for this study.

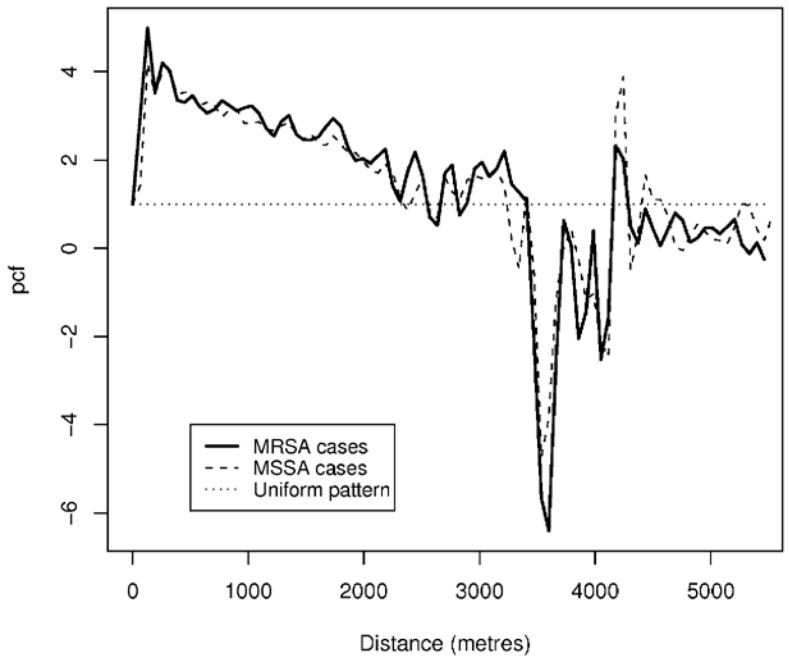

Figure 4b. Pair correlation functions for MRSA cases and MSSA cases.

K-functions are often used to model the spatial distributions of *marks* or *labels* within a single point set (for example, a set of mapped trees marked as 'infected' (cases) or 'uninfected' (controls) in a disease study). Here, the aim is to test whether events of one type tend to be surrounded by events of the same type, by a 'random labelling' of the existing points (e.g., Diggle and Chetwynd 1991). The random labelling hypothesis assumes that, among the combined group of case-control locations, the probability of a particular label at any location (in this case, a positive test for MRSA) is equal for all points, and not dependent on the label status of neighbours (MRSA / MSSA).

In a practical application of random labelling, numerous populations of labelled points are generated by randomly re-allocating the existing labels (without replacement) to the existing data locations, generating a K-function for the resulting point pattern each time. From these simulated K-functions, we can create a 95% confidence envelope, against which (Kcases – Kcontrols) can be plotted. Departure from the envelope implies that the case pattern is meaningful (i.e., more or less clustered than expected) at this scale. While this addresses scale well, it gives a global index of clustering, rather than a map of where any clustering occurs, and differs from a test of independence between the two point *patterns*, which can be assessed using repeated random toroidal shifts.

A spatial adaptation of random labelling was performed for this study by using a quartic kernel to smooth each resampled set of points, thus generating a statistical population of *surfaces*, rather than *vectors*. In order to perform a one-sided test at $p < 0.025$ for significant aggregation of points we can generate a surface, cell by cell, from the 97.5^{th} percentile of the values generated from kernel smoothing of the relabeled points. Where the surface generated from the original case data exceeds these '*pointwise tolerance contours'*, the intensity for that cell can be considered significant. This general method (Kelsall and Diggle, 1995) combines visualisation of spatial pattern with a means of testing significance, and has been used for space-time analysis of disease risk (Sabel *et al*., 2000)

1.4.3 Spatial scan statistics. The geographical analysis machine ('GAM') developed by Openshaw et al. (1988) relies on a similar 'moving window' to that used in kernel estimation. Regular grid points across the study area form the centres of variably-sized circles, within each of which the observed and expected number of cases (the latter based on local population at risk) are compared. Circles containing significantly high densities of points are drawn on the map. In the original version of GAM, circles may overlap, and contain the same case cluster, (i.e. are not statistically independent) and this was a key criticism of the technique. Methods developed to tackle this problem include spatial scan statistics, which build on the GAM approach, but constrain the geographical overlap between cluster centres and the circles around them, and which can be extended to space-time clustering by constructing volumes, rather than areas, around sets of data points. This study applied Kulldorff's spatial scan statistic (Kulldorff, 1997) which has been tested on a variety of real epidemiological datasets, and benchmarked on simulated point patterns, and which compares well to LISA indices (Hanson and Wieczorek, 2002).

2. Methods

2.1. NATURE OF THE DATA

This work uses spatial information available for *Staphylococcus aureus* isolates processed by a UK general hospital between 01/09/2004 and 31/08/2005. The samples represent a catchment area defined largely by the bounds of 3 Primary Care Trusts (see Figure 1) and cases which fell outside the boundary were discounted (82 cases of MRSA (9% of all MRSA cases) and 188 cases of MSSA (9.3% of all MSSA cases)). The population

within the bounds (based on 2001 UK Census data) was 374,883, while the total number of patients registered with GPs within the area was 398,642. This implies that 6% of the registered patients live outside the selected study area, so that the discounted proportions above are reasonable. The data available are a good representation of the study area (in other words, given the notification procedures in place in the region, the number of any isolates sent outside the study area for analysis is likely to be insignificant). The study bounds measure roughly 25 km by 25km, and the area of the selected polygon is 328.6 km^2.

For each laboratory sample, the home unit postcode provided was matched to centroid coordinates from the UK All Fields Postcode Directory (AFPD) (Office of National Statistics, 2005), and to the population centroids of 2001 Census OAs. The bounds shown in Figure 1 contained 1244 OAs, and 11,387 spatially-referenced unit postcodes, which were mapped onto 2001 Census OAs using the AFPD.

The data contained a number of inaccuracies in terms of recorded birth dates and post codes, and only some of these could be corrected. 11.5% of the original case records had incomplete unit postcodes and could not be spatially located; this was improved to 7.9% by independent telephone follow-up of the missing data to GP practices, but the outstanding cases (24 MRSA and 221 MSSA) were discounted for the purposes of this study. Inclusion of these records, which could only be geo-coded to a coarser spatial resolution, or not at all, would have severely prejudiced an assessment of spatial pattern. There was no bias towards specific postcode zones among these omitted records, so the omissions are unlikely to have introduced any statistical bias in the analysis.

Preliminary analysis identified patient age as the most important predisposing factor recorded in the dataset (see Figures 2a, 2b), and all analyses were stratified, to 3 age groups: under 65, 65-85 and over 85. The restriction of most demographic data to OA-level necessitated a tradeoff between spatial resolution and verifiable demographics, since, for most analyses stratified by the age and local density of the vulnerable population, case data were aggregated to OA population centroids.

Stratification by age, and the presence of both MRSA and MSSA data, led to 6 distinct data categories which are summarized in Table 1.

Table 1. Data categories (DCs) for the study, showing total within-bounds numbers of cases ('No. cases') and global incidence probability in the study area ('IP'). IP was calculated by dividing the number of cases in each age category by the total number of people in this age group in the 2001 Census (thus assuming a relatively static age structure).

Disease	Age category	Under 65	65 to 85	85 and over
MSSA	No. cases	1090	563	214
	IP	0.0034457	0.0109267	0.0303331
MRSA	No. cases	196	429	207
	IP	0.0006196	0.008326	0.0293409

2.2 SPATIAL ANALYSIS TECHNIQUES USED IN THIS STUDY.

2.2.1 Spatial / spatio-temporal scan. The spatial scan statistic software described in 1.4.3 is publicly available as the SaTScan software package (Kulldorff, 2006). This widely-used technique was used to identify clustering of cases against background population. SaTScan analyses were stratified by age to the six DCs shown in Table 1, by separating out MRSA and MSSA case locations, and in each case using the three age categories as a covariate in the analysis.

2.2.2 Stochastic simulation and kernel density estimation. As an alternative to spatial scan, stochastic simulation and kernel estimation were combined as recommended by Kelsall and Diggle (1995). All kernel

and k-function analysis described in this section was carried out using the R language (R, 2005), and the splancs (Rowlingson and Diggle, 1993) and spatstat (Baddeley and Turner 2005) libraries in particular.

Stochastic simulations were based on the global age-stratified Incidence Probabilities (IPs) shown in Table 1. The null spatial model (as characterized by Waller and Jacquez, 1995) is a heterogeneous underlying Poisson model with no spatial autocorrelation other than that caused by heterogeneities in the population at risk: i.e., varying population density and age structure. Monte Carlo techniques were applied to the known age structure at each OA centroid, to generate a reference distribution consisting of 1000 sets of simulated case locations, from which confidence limits can be generated. These simulated case sets can be stratified into the 3 chosen age categories, or combined to produce expected total counts for each OA.

Simulated locations were randomly generated using a Python script which produced whole numbers of expected cases for each location. Because the real case data were tied to the population centroid of the OA, simulated cases were not randomly perturbed around this centroid, but were instead assigned to the centroid, ensuring that the simulated and real datasets were more directly comparable, both before and after kernel estimation.

In order to assess significance of MRSA clusters, kernel estimation was employed as described in section 1.4.2. A quartic kernel was used to smooth each simulated point set, as well as the real data locations for MRSA, to a grid of 100m.

Selection of a bandwidth in this context is not trivial: Kelsall and Diggle (1995) note in discussing the generation of relative risk surfaces that a variety of methods may be used, and that edge-correction of the estimates for cross-validation is vital. Bandwidth selection is further complicated by the fact that discrete data such as the OA centroids used here bias the cross-validation score for smaller bandwidths (Chiu, 1991), and by the fact that, for sparse data, many ‘expected’ kernels will contain few or no data points and will again bias cross-validation estimates towards infinity (Hickson and Waller, 2003).

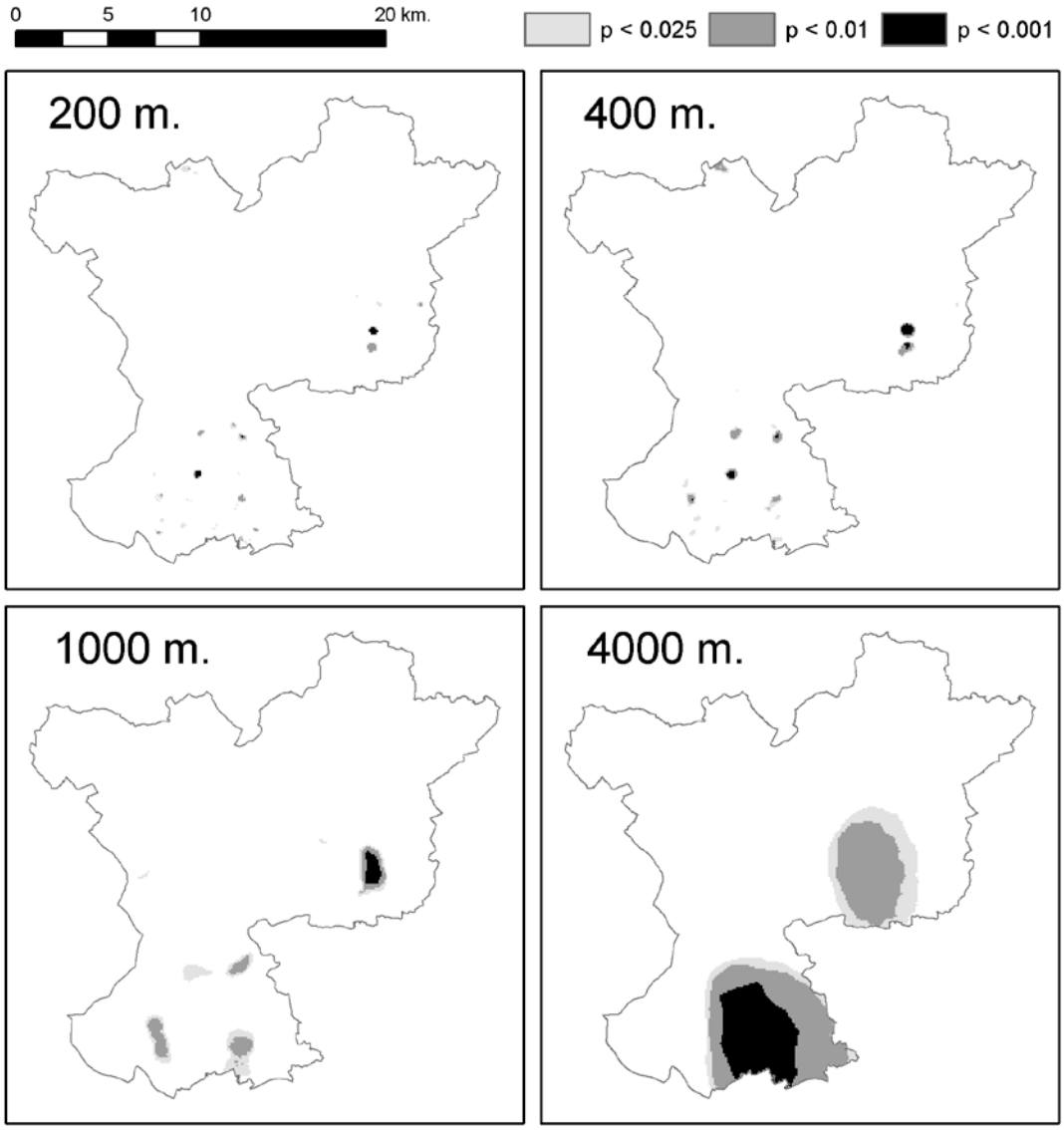

Figure 5. Kernel density estimation results for four different bandwidths: 200m, 400m, 1000m and 4000m.

For the purposes of this study, a kernel bandwidth of 200m was implied by finding the minimum from a simple plot of mean squared distance (using the splancs ‘mse2d’ function). This value was expanded to 400m in an attempt to counteract the artificial structure imposed on OA centroids, and ‘force’ neighbouring centroids to have more influence. This is largely a subjective choice, as described in Silverman (1986), but is partly based on an assessment of kernel coverage: a bandwidth of 200m left 14% of OA centroids as the only data point in a kernel (max. per kernel=8, mean=2.92) while a 400m kernel left only 3% of kernels covering

only one data point (max. per kernel=15, mean=6.8). In fact, the core clusters identified by 200m and 400m bandwidths differed very little (see Figure 5); however, an increase in bandwidth to 1km. indicates just one highly significant cluster in the East, and a further increase to 4 km. (a distance which people might be expected to travel to their place of work) moves the significant excess to the South of the study area. This bears out the contention of Waller and Gotway (2004) a variety of bandwidths, considered together, may actually give broader insight into underlying phenomena at different scales.

The chosen bandwidth of 400m was applied to random labellings of the cases and each of the 1000 simulated 'control' datasets (Kelsall and Diggle, 1995). For each grid cell in the 'real' surface, a significant value was taken to be one where at least 975 of the simulated cell values were exceeded by the real value. This extends the metaphor of the 2-dimensional confidence envelope to a 2.5D confidence surface: only where points are unusually clustered in reality do they generate local spikes which can penetrate through the 97.5^{th} percentile surface and be seen as significant. (Consideration of the 2.5^{th} percentile as a lower bound would, effectively, generate a 3-dimensional confidence volume. However, only excessive clustering above the upper bound was considered for the purposes of this analysis).

2.2.3 K-functions and kernel estimation from random labelling For random labelling of MRSA/MSSA cases, unit postcode-level locations could be used, since underlying population density was not an input. Overall sampling intensity per head of population (MRSA and MSSA combined) varied widely across the study area. There was also a natural tendency for paired results due to medical follow up; i.e., cases where an individual tested positive for MRSA and was later found to be clear, or cases where MRSA was suspected but was only identified on a subsequent test. Paired tests were found to have occurred for 15% of the MRSA cases. Where control data are available, random labelling can help to verify that clustering among cases is genuinely due to a process by which this label tends to aggregate and not simply an artefact of denser

sampling in these areas. However, given the low likelihood that an active *S. aureus* infection will fail to be notified, it is not reasonable to interpret varying MRSA density as simply a sampling artefact. Rather, sampling rates generally increase in areas where MRSA incidence is high.

K-function plots from random labelling were useful in identifying a global clustering trend, but less useful in identifying where variations might be. To spatially assess significant departures from the confidence envelopes, kernel density estimation was employed (as described in section 1.4.1). Randomly-labelled point sets were smoothed with a fixed-bandwidth 400m quartic kernel as in section 2.2.2, and the 97.5^{th} percentile of the results was used to represent the upper envelope as it exists across space. The genuine MRSA case locations were then smoothed with the same kernel, and areas where the smoothed grid exceeded the 97.5^{th} percentile grid were outlined and mapped as significance contours.

3 Results

3.1 SPATIAL AND SPATIO-TEMPORAL SCAN WITH SATSCAN

Spatial scan statistics for the MSSA data identified the entire study area as one or two clusters, and did not discriminate any local clustering. Spatial scans for MRSA with age as a categorical covariate (under 65, 65-85 and over 85) identified a number of apparently significant clusters among the MRSA cases (shown as a component of Figure 7). It is important to note that two of the highly significant clusters ($p < 0.0001$) consist of just one or two OAs where risk is strongly elevated. The same two OAs retain their importance when the analysis is extended to spatio-temporal clustering, and will be referred to in later discussion as C1 and C2 (see Figure 6).

The identification of these small areas as significant clusters was a useful outcome, and the SaTScan software scan was particularly useful in allowing an easy extension to spatio-temporal scan.

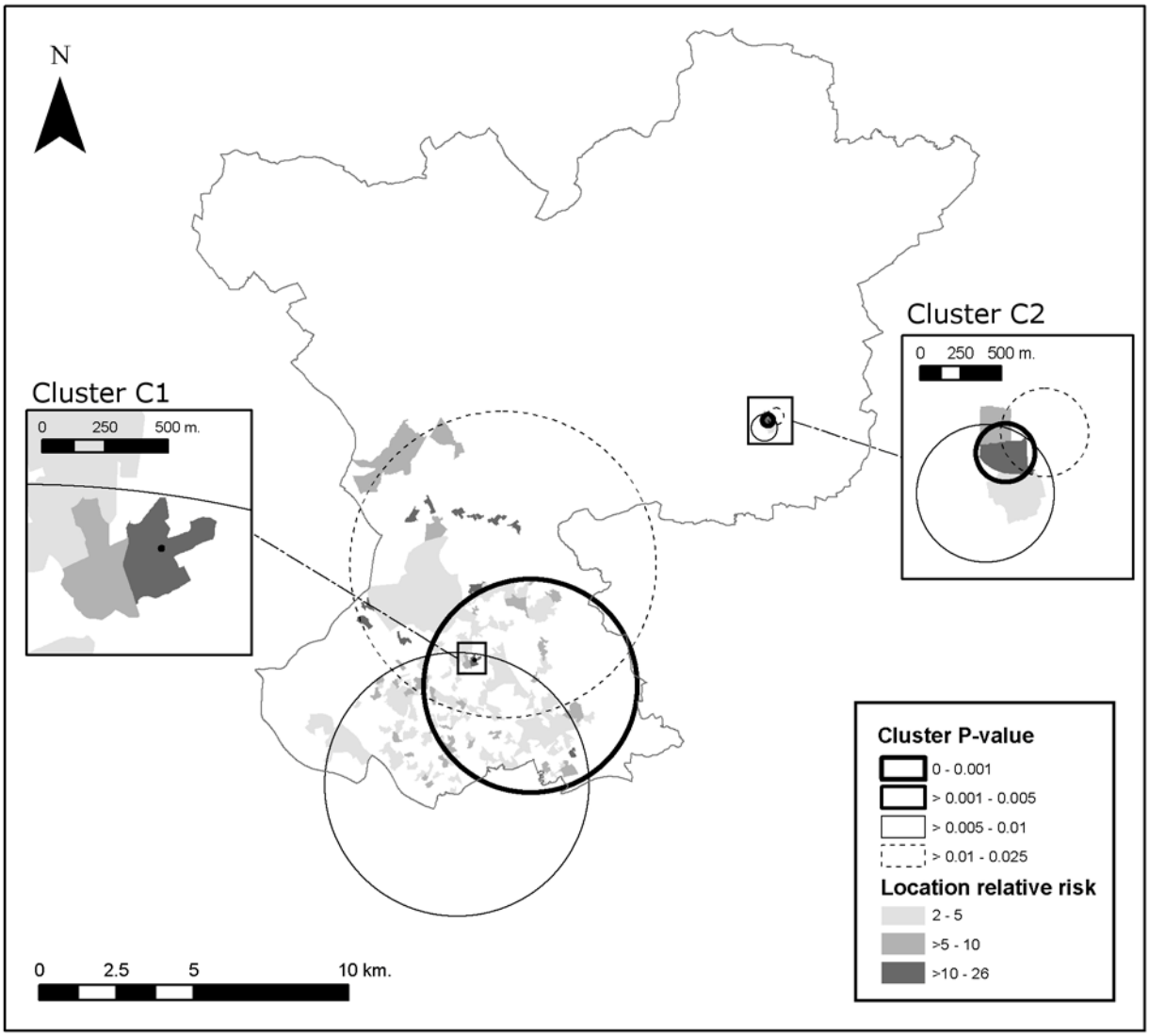

Figure 6. SaTScan-generated spatio-temporal clusters for MRSA, with 3 age categories as covariates. OA polygons whose centroids participate in significant clusters are shaded by local relative risk. NB: despite this choropleth display, the SaTScan analysis was based on point centroids rather than areas.

3.2 STOCHASTIC SIMULATION AND KERNEL DENSITY ESTIMATION

Based on kernel estimations with bandwidth of 400m, MRSA cases were found to be significantly more clustered in certain areas than the stochastic simulations based on population density (see Figure 7). The results broadly bear out the clusters identified by SaTScan. However, although both methods rely on running

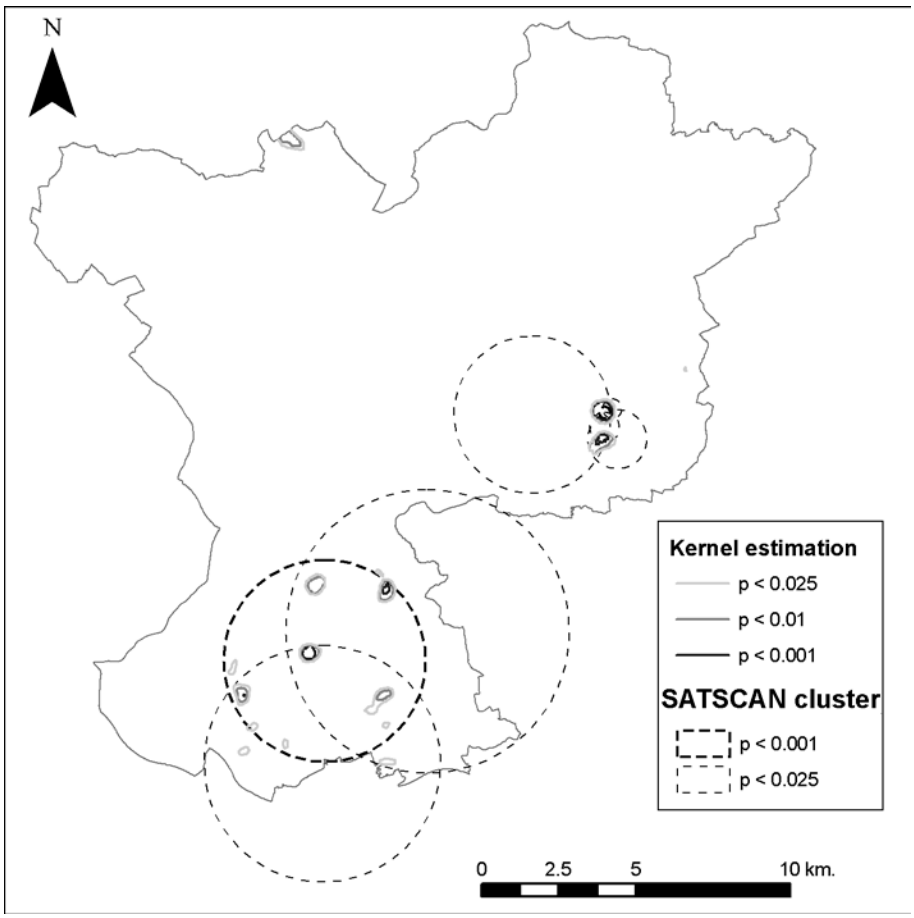

Figure 7. Significant clusters identified using pointwise tolerance contours from kernel density estimation broadly support the identification of clusters from spatial scan. The example shown here used a fixed quartic kernel with a bandwidth of 400m.

circular windows across the data, the adaptable bandwidth of the spatial scan allows us to identify ‘elevated risk’ circles of varying size, including a sizeable area in the South of the study region. The kernel estimation method, on the other hand, shows this area as a set of individual local significance peaks, unless the selected bandwidth is unrealistically large. When the pair-correlation function for all MRSA cases is plotted against 95% confidence limits generated from stochastic simulations (see Figure 8a), it can be seen that the major departure from randomness is at very short range (0-100m), reinforcing the conclusion that the major clustering effect in this data is one with virtually no spatial lag: in other words, cases co-occurring at, or at least aggregated to, the same location. The impact of this very small set of coincident points is reinforced by a

pair correlation function where the 22 MRSA cases registered to the two postcodes at the centres of clusters C1 and C2 have been removed (Figure 8b), bringing the pcf within the confidence limits for most of its range.

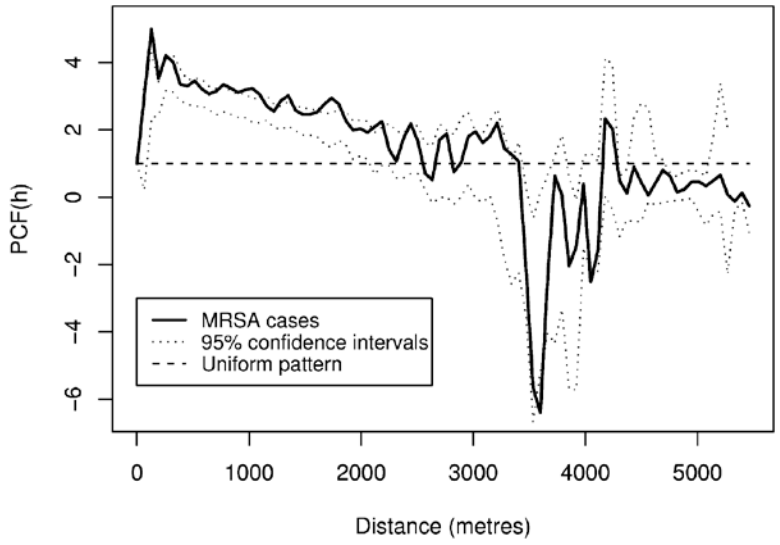

Figure 8a. Pair correlation functions for MRSA cases, and 95% confidence intervals generated from 1000 age-stratified stochastic simulations. MRSA case clustering beyond the confidence envelope is most noticeable at extremely short distances (0-100 m).

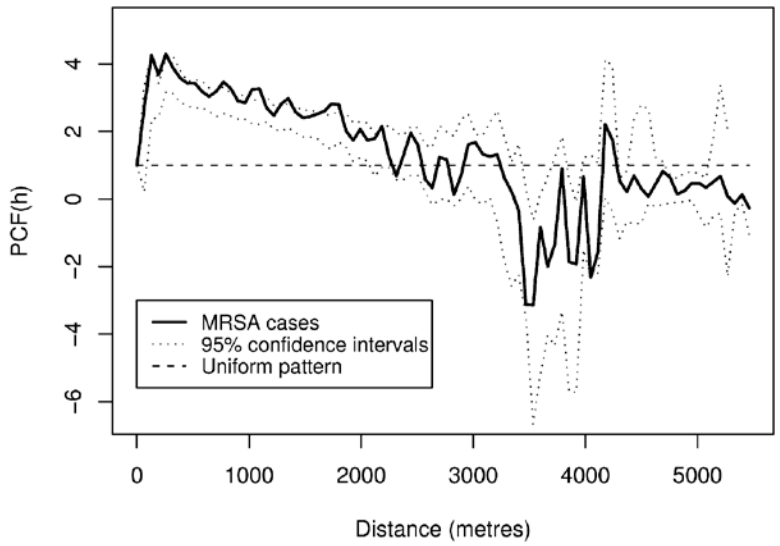

Figure 8b. Pair correlation functions for MRSA cases as above, with 22 cases from clusters C1 and C2 removed. The function now falls within the confidence limits at very short range (0-100m).

3.3 K-FUNCTIONS AND KERNEL ESTIMATES BASED ON RANDOM LABELLING

When subjected to 500 random labellings, the MRSA cases are significantly more clustered at distances of 0 to 180m than would be expected from a chance sampling of all data points. When stratified by age, however, this significant clustering is limited to the under-65 age group (Figure 9). A spatial investigation of this pattern among the under-65s was generated using the kernel method described in Section 2.2.3, and is shown in Figure 10. For cluster C2, the high density of MRSA samples in this age group is associated with an equally high density of MSSA samples; i.e., no significant excess is seen. For C1, MRSA cases are still significantly more clustered than MSSA cases among the under 65s.

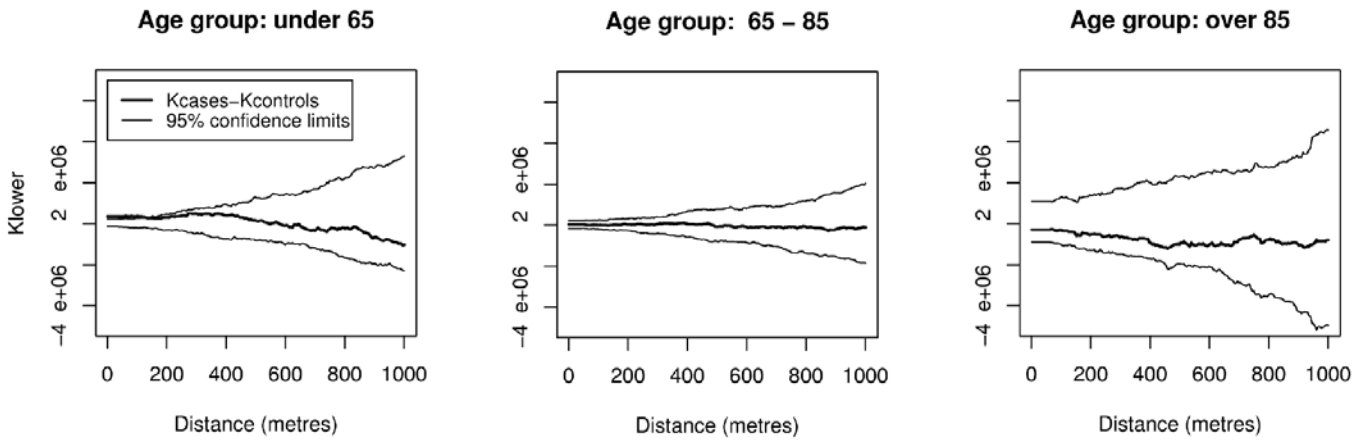

Figure 9. Stratified random-labelling analysis with MRSA as cases and MSSA as controls. Significant departure from the envelope (i.e., clustering) is seen only in the under-65 age group, at very small spatial lags (<150m).

Across the study area as a whole, this implies that sparse and dense patterns of MRSA generally coincide with correspondingly sparse/dense patterns of MSSA except in the under-65 age group, where MRSA is more clustered at certain locations. It is useful to note at this point that the soft-tissue infections characteristically caused by *S. aureus* will almost certainly be notified to a medical practitioner, and this practitioner may, in the first instance, treat with antibiotics or take a sample. If MSSA is treated successfully with a broad-spectrum antibiotic, and a sample is not taken, the case would not appear in this dataset. However, an MRSA infection

would persist to the stage where a sample must be taken and processed. The result will be a potential preferential bias towards MRSA coverage in this dataset, so that any observed difference in MSSA sampling intensity across the study area is most likely to result from a medical practitioner's tendency to treat, or to test, when the patient first presents.

It appears likely that, in general, MSSA sample density is being driven by incidence of MRSA infection, particularly among older people. Coupled with the fact that 15% of the MRSA-positive individuals had been re-tested at some point, this implies pragmatic, structured medical follow up in areas (and possibly in institutions) where perceived risk is high. For example, cluster C2, where 12 of the central 13 MRSA cases stem from a single nursing home, shows no significant excess of MRSA over MSSA. It is interesting to note the two main cases where MRSA/MSSA tracking does not apply; i.e., the locations where MRSA-MRSA clustering (in the under-65s) is most significant (see insets in Figure 10). Both of these peaks coincide with significantly high MRSA rates in general (one of them coincides with cluster C1), but they represent very different phenomena. Cluster C1 (11 cases in total) centres on 9 individuals registered at the central unit postcode of the cluster, where 2 separate medical care institutions are registered, both catering to clients under 65. This independent reinforcement of the pattern identified by the kernel estimation is encouraging. Individuals from this postcode were prioritised for further investigation by the hospital, which revealed that 7 individuals were registered at one of the nursing homes, and 2 at the other. This example highlights the necessity for data on patients' healthcare status to be attached to this type of microbiological sample. A simple flag ('resident in medical care') would enable the immediate stratification of records by healthcare status, and the assignment of different expected risks to Census OAs based on the presence and size of healthcare institutions. Simulated cases could then be based on an assumption of risk which was low for the general population, but elevated where nursing homes are known to be present. Similarly, a record of last hospital attendance would enable sample data to be, in effect, related to more realistic prior probabilities for the

assessment of significance. The pressing need for supplementary data, recorded at the time that the microbiological sample is taken, is further discussed in section 5.

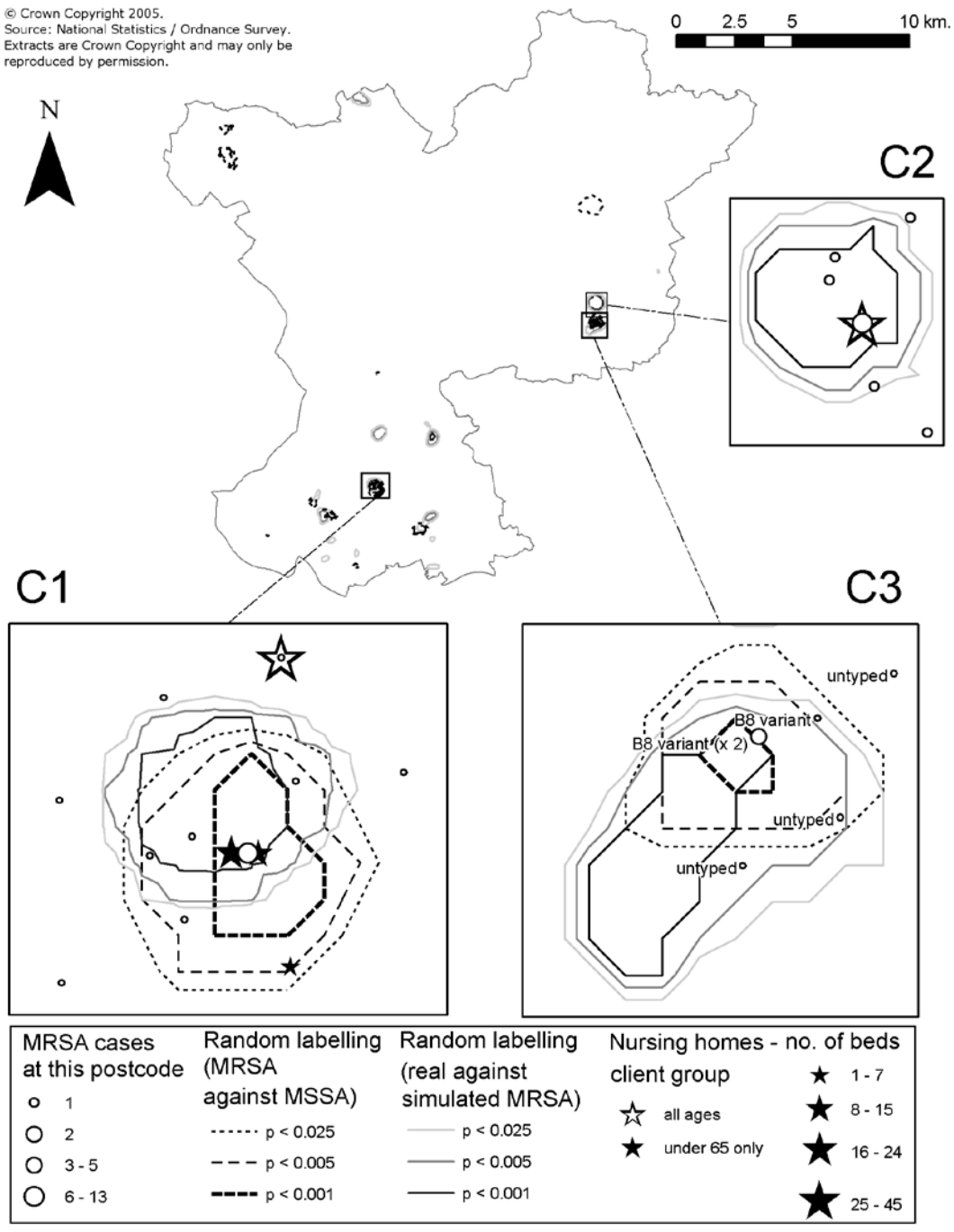

Figure 10. Kernel density analysis of randomly-labelled MRSA/MSSA data for the under-65 age group. There are two particularly significant areas (C1 and C3) where high MRSA rates overall coincide with high rates of MRSA relative to MSSA among the under 65's. NB: in cluster C1, note the cartographic displacement of the two nursing homes registered at the central postcode.

The second area of interest, which lies just to the south of cluster C2 and will be referred to as cluster C3, is not associated directly with healthcare locations, but is of particular interest because pulse-field typing data is now available for 3 of the 5 individuals in the cluster. All 3 individuals exhibited a B8 variant form of MRSA, which has a prevalence of only 10% across the whole study area. While this type of MRSA (SCCmec IV) is associated with healthcare environments, and is likely to have been contracted in a hospital, it appears that it may have been further transmitted between individuals in the community. A non-institutional context would explain why local sampling among the patients' age group (all under 65) has not risen to correspond with this increased local rate of MRSA. Again, the main outcome has been to prioritise these cases for further investigation, given the severely-limited data currently available on these cases. Of particular interest is the fact that this cluster was not identified by the SaTScan analysis.

3.4 Area-based methods: Moran's I scores. The methods described above are all restricted to a consideration of point pattern, while the population data which produce expected disease rates are tied to areas: namely, census OAs. Both centroids and areas have their disadvantages, as discussed in sections 1.3 and 1.4, and a calculation of Moran's I was undertaken on the observed/expected rates of MRSA in the region. Expected rates were based on the age-stratified probabilities summarised in Table 1, combined with population age data from the 2001 Census, and a Boolean adjacency matrix was generated for all polygons to represent connectivity. The OA constituting cluster C1 had a risk level with little relationship to those surrounding it, implying a truly isolated point cluster. C2, on the other hand, may possibly be part of a broader cluster, with high I-values in neighbouring OAs. Since cluster C2 is dominated by healthcare patients at a single nursing home, focussed molecular typing in the area will help to establish whether this location is a

reservoir for particular subtypes of MRSA, and whether these strains have moved between the healthcare environment and the local community.

4 Discussion

This study found MRSA cases (based on home unit postcode), to be significantly more clustered at certain Census OAs than would be expected based on the underlying population and its age structure. The results of a spatio-temporal scan largely concurred with a stochastic kernel estimation technique in identifying one large cluster to the south of the study area, and two very small clusters, consisting of one and two Census OAs respectively. On closer investigation, the cases within these OAs are very spatially aggregated to specific unit postcodes, at which nursing homes are also registered. In fact, of the 8 unit postcodes with the highest recorded MRSA rates, 7 had registered nursing homes. The cases highlighted in these areas are currently being investigated (a) by contacting the submitting doctors to verify whether they are healthcare-related and (b) by molecular typing to assess their similarity. Results so far indicate that, in clusters C1 and C2, the majority of patients (9 out of 11, and 12 out of 13, respectively) were registered in some form of care at the time of testing. Molecular results are still pending at the time of writing.

The ability of the spatial scan method to select the most ‘risky’ of a suite of possible circular windows (of varying size) is a particular strength when compared to the fixed-bandwidth of the kernel estimation method. Even an extension of the kernel smoothing method to an adaptive bandwidth would not explicitly address this issue. Kernel estimation as demonstrated here is computationally intensive and required a variety of bespoke programs; however, it can generate appealing and accessible visualisations, particularly in space-time contexts (e.g., Sabel et al., 2000). It also has the advantage that it readily can be used to handle multiple randomly-labelled point sets (see Figure 8), and to generate confidence ‘surfaces’ which lend themselves to 3D visualisation. All in all, the method was a useful complement to spatial scan, and the stochastically-simulated

case sets could be used to illustrate the structured nature of a random phenomenon in the context of a spatially-autocorrelated underlying factor (i.e., population density). It is interesting to note that the kernel estimation method specifically identified a cluster which appears from early molecular typing results to be a genuine case of community transmission (see the right-hand inset in Figure 10), and that this location was not specifically highlighted by SaTScan.

The identified clusters cannot necessarily be taken to show community transmission of MRSA, since it appeared that many of the cases were actually patients in medical care establishments such as hospices and nursing homes. However, although ancillary data suggest residence in an institution (and, possibly, compromised health status) at the most notable clusters, in one case elevated risk was observed in two neighbouring OAs, and this is borne out by Moran's *I* scores calculated with an adjacency matrix for OA polygons. The implication is that, while this extremely localised cluster may represent an outbreak at a medical institution, it may also be having some impact on the surrounding area. Molecular epidemiological typing is being directed towards these cases as a priority, and will be of immense value in identifying transmission events at and around these locations.

This highlights a particular gap in the data available for this study; no information was available on place or nature of employment, hobbies or other risk factors. An individual's home postcode is, in general, a very poor surrogate for this information in assessing the risk of contracting MRSA. In fact, it could be said that, due to their compromised health status, long-term residents of care homes are the people *most* likely to contract MRSA at their registered home postcode. Coupled with intermittent periods of asymptomatic colonization by MRSA (i.e., a patient could carry the bacterium for months without active infection) the reliability of such limited data imposes serious limits on tracking MRSA infection events.

The clusters of MRSA identified in this study were also found to be generally associated with high local levels of MSSA samples, such that MRSA/MRSA clustering at one of the 2 small cluster locations was not

distinguishable from MRSA/MSSA clustering under a random labelling regime. This may be another result of the medical context in which these individuals generally found themselves, and the increased surveillance in this context. However, the spatial random labelling showed several locations to have higher levels of MRSA among the under-65 age group than could have been generated by varying sampling effort alone, and these results have highlighted interesting groups of cases for further analysis.

The major influence on observed susceptibility to MRSA was age, but even an age-stratified approach failed to explain the observed local clusters, whether by stochastic simulation combined with kernel estimation or spatial scan.

The microbiological sample data presented here are actually very typical for the UK, and it is notable that, while hospital-onset cases could be removed from the analysis, there were numerous healthcare-associated cases which had not been recorded as such. The effort required to retrospectively update even a small portion of the case data with healthcare status has been substantial, and one of the major arguments of this paper is that the time and effort invested by family doctors in collecting a standard set of ancillary variables (see Table 2) will be worthwhile, and will allow a far more meaningful analysis of MRSA occurrences. It is apparent from this dataset that, if healthcare-associated cases can be labeled as such, the residual (i.e. potentially community-based) rates for small areas such as census OAs will become even smaller, with all the associated statistical problems. However, the stratification of cases on healthcare status would allow far better modeling of prior probabilities/expected rates, and open up a range of techniques suited to small-area studies, such as empirical Bayesian smoothing.

Table 2. Suggested supplementary data to be collected and attached to microbiological samples. Values marked with an asterisk should share a common coding scheme such that each hospital ward/nursing home/medical practice in the region is assigned a

unique ID. Values marked † have the potential, either alone or in combination with other information, to make patient records identifiable.

Variable	Further details	Importance
Home postcode †	Full postcode	High
Age †		High
Gender †		Medium
National Health Service number †	Unique for each patient – allows repeat tests to be automatically identified	High
Current healthcare status	Inpatient (hospital) In nursing care Not in a healthcare environment	High
Origin of sample *	Ward (if hospital inpatient) Nursing home (if resident) Medical practice (if community-based)	High
Date of microbiological sample		High
Date of last hospital visit (inpatient)		High
Ward visited *		Desirable
Postcode of regular workplace		High
Hobbies – e.g., sports		Desirable
Other postcodes regularly visited (e.g., sports clubs, family members)		Desirable

5 Conclusions

Given the international importance of MRSA, this study looks at the problem from a unique angle. It is important for effective infection control within primary care to visualise the occurrence of MRSA clusters with full consideration of the underlying population. This will facilitate effective control measures at these critical points. Given the limited nature of the patient data available, the results of this study have been extremely useful in prioritising case records for clerical follow-up and time-consuming, expensive molecular typing. The techniques assessed here were found by healthcare experts with little GIS experience to be highly suitable both for visualisation and hypothesis generation, and for assessing significance of spatial clustering in a community disease context. The work described here has also identified critical weaknesses in current epidemiological recording and reporting, and specifically identified a list of supplementary data which will be vital for any serious ongoing monitoring and tracking of MRSA in the community (see Table 2). If ancillary variables such as these can be standardised, reliably recorded and attached to bacterial isolates, the power and discrimination of cluster analysis will be greatly improved.

In the absence of occupational/lifestyle data on patients, the assumption was made that an individual's location and consequent risk of infection is adequately represented by their residential postcode. Given workplace and environmental risk, this assumption has recognised limitations, identified by Gattrell *et al*, (1996). An approach which locates cases to home postcode will be more successful in identifying clusters among people who are a) restricted in movement and b) liable to contract MRSA at their registered home address. As such, the data as they stand are suited specifically to identifying outbreaks at medical care institutions, and at this point it would appear that some of these have indeed been identified. In order to track and model patient movements and probable infection events, further information would need to be collected at the point of medical contact. A suggested list of ancillary variables, which could be presented as a paper or online form, is shown in Table 2. Even a simple categorization such as 'health care associated (inpatient)',

‘health care associated (non-inpatient),’or ‘community onset’ (as recommended by Collignon et al., 2006) would be of great value in stratifying the data addressed in this study, especially as an input to models of expected disease frequency, and for Bayesian approaches. It would be of particular value to separate out the point clustering which has practically no spatial lag (usually, individuals in medical institutions) and prioritise other cases, for example, for a more detailed assessment of risk including the workplace. The effort required to track an individual’s potential exposure to MRSA is considerable, as is the time and effort required to obtain molecular information on isolates. In both areas, this study has refined the subset of cases to be further investigated.

The ancillary variables suggested in Table 2 also have the potential to make individual patients more easily identifiable. Care should always be taken that unique patient identifiers (e.g., name, National Health number) are not included in database extracts, and that any combinations of variables which make a patient ‘potentially identifiable’ are used only according to agreed protocols.

While spatial scan is an established technique, the kernel estimation used here to visualise cluster significance is less widespread. The simulated points and surfaces are of value for visualisation, and for showing the spatial detail of point pattern summaries such as the k- and pair-correlation functions shown here. Both spatial scan and kernel estimation concurred in identifying tightly-localised clusters at single unit postcodes with no spatial lag whatsoever. However, the size and shape of the risk locations generated by the kernel estimation are bandwidth-dependent, relate to the aggregated point data rather than the unit area, and could give a false sense that risk had been predicted at unknown locations. This is an exploratory technique rather than a model of infection process, and as such is useful and interesting. Where two point patterns must be compared across space and time, kernel estimation with a well-validated bandwidth also has the arguable benefit of harmonizing data to regular, comparable grids.

Future work

Molecular typing of the isolates explored here is currently underway, and this will permit a more detailed analysis of the association of specific types of MRSA with locations in the study region.

The kernels used here for estimation were both fixed and round; it is planned to extend the kernel estimation technique adopted here to adaptive kernels, and some suggestions have been made in the literature as to alternative techniques of spatial scan which do not rely on round kernels (e.g., Tango and Takahashi, 2005).

References

BABA, T., TAKEUCHI, F., KURODA, M., YUZAWA, H., AOKI, K., OGUCHI, A., NAGAI, Y., IWAMA, N., ASANO, K., NAIMI, T., KURODA, H., CUI, L., YAMAMOTO, K. and HIRAMATSU, K., 2002, Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, **359,** pp. 1819-27.

BADDELEY, A. and TURNER, R., 2005, Spatstat: An R package for analyzing spatial point patterns. Journal of Statistical Software, 12 (6), pp. 1-42.

BAILEY, T., and GATTRELL, A., 1995, *Interactive spatial data analysis*. Longman Scientific & Technical.

BESAG, J. and NEWELL, J., 1991, The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, **154**, pp. 143-155.

BRUNSDON, C., 1995, Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. *Computers and Geosciences*, **21** (7), pp. 877-894.

CHAMBERS, H., 2001, The changing epidemiology of *Staphylococcus aureus*? *Emerging Infectious Diseases*, **7,** pp. 178-82.

CHIU, S., 1991, The effect of discretization error on bandwidth selection for kernel density estimation, *Biometrika*, **78 (2),** 436-441.

COLLIGNON, P., WILKINSON, I., GILBERT, G., GRAYSON, M. and WHITBY, R., 2006, Health care-associated *Staphylococcus aureus* bloodstream infections: a clinical quality indicator for all hospitals. *Medical Journal of Australia,* **184 (8)**, pp. 404-406.

CRESSIE, N., 1993, Spatial Point Patterns. Chapter 8 in *Statistics for Spatial Data (Revised Edition)*. (Chichester: John Wiley).

DIGGLE, P., and CHETWYND, A., 1991, Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**, pp. 1155-1163.

DIXON, P., 2002, Ripley's K-function. Pp 1796-1803 in *Encyclopedia of Environmetrics*. (Chichester: John Wiley).

GATRELL, A., BAILEY, T., DIGGLE, P. and ROWLINGSON, B., 1996, Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, **21**, pp. 256-274.

GELMAN, A, PARK, D., ANSOLABERE, S., PRICE, P., MINNITE, L., 2001, Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **164 (1)**, pp. 101-119.

GRIFFITHS, C., LAMAGNI, T., CROWCROFT, N., DUCKWORTH, G. and ROONEY, C., 2004, Trends in MRSA in England and Wales: analysis of morbidity and mortality data for 1993-2002. *Health Statistics Quarterly*, **21**, pp.15-22.

HAINING, R., 1990, *Spatial Data Analysis in the Social and Environmental Sciences*, pp. 73-74 (Cambridge: Cambridge University Press).

HANSON, C. and WIECZOREK, W., 2002, Alcohol mortality: a comparison of spatial clustering methods. *Social Science and Medicine*, **55**, pp. 791-802.

HEROLD, B., IMMERGLUCK, L., MARANAN, M, LAUDERDALE, D., GASKIN, R, BOYLE-VAVRA, S., LEITCH, C. and DAUM, R., 1998, Community-acquired methicillin-resistant *Staphylococcus aureus* in children with no identified predisposing risk *The Journal of the American Medical Association***, 279,** pp. 593-8.

HICKSON, D., and WALLER, L.A., 2003, Spatial analyses of basketball shot charts: an application to Michael Jordan's 2001-2002 NBA season. *Proceedings of the Hawaii International Conference on Statistics and Related Fields.* Available online at: http://www.hicstatistics.org/2003StatsProceedings/DeMarc%20Hickson.pdf (accessed 09/06/2006)

HOLMES, A., GANNER, M., MCGUANE, S., PITT, T., COOKSON, B. and KEARNS, A., 2005, *Staphylococcus aureus* isolates carrying Panton-Valentine leucocidin genes in England and Wales: Frequency, characterization, and association with clinical disease . *Journal of Clinical Microbiology,* **43 (5)**,

pp. 2384-2390.

ITO, T., MA, X., TAKEUCHI, F., OKUMA, K., YUZAWA, H. and HIRAMATSU, K., 2004. Novel type V staphylococcal cassette chromosome *mec* driven by a novel cassette chromosome recombinase, *ccrC*. *Antimicrobial Agents and Chemotherapy,* **48,** pp. 2637-51.

ITO, T., KATAYAMA, Y., ASADA, K., MORI, N., TSUTSUMIMOTO, K., TIENSASITORN, C. and HIRAMATSU, K., 2001. Structural comparison of three types of Staphylococcal cassette chromosome *mec* integrated in the chromosome in methicillin-resistant *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy,* **45,** pp. 1323-36.

KELSALL, J. and DIGGLE, P., 1995, Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine,* **14**, pp. 2335-2342.

KLUYTMANS, J., VAN BELKUM, A. and VERBRUGH, H., 1997, Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Review,* **10,** pp. 505-20.

KULLDORFF, M., 1997, A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, pp. 1481-1496.

KULLDORFF, M., 2006. SaTScan – software for the spatial, temporal and space-time scan statistics. *Available online at: http://www.satscan.org/.*

MA, X., ITO, T., TIENSASITORN, C., JAMKLANG, M., CHONGTRAKOOL, P., BOYLE-VAVRA, S., DAUM, R. and HIRAMATSU, K., 2002, Novel type of staphylococcal cassette chromosome *mec* identified in community-acquired methicillin-resistant *Staphylococcus aureus* strains. *Antimicrobial Agents and Chemotherapy,* **46,** pp. 1147-52.

MAINOUS, A, HUESTON, W., EVERETT, C. and DIAZ, V., 2006, Nasal carriage of *Staphylococcus aureus* and methicillin-resistant *S. aureus* in the United States, 2001-2002. *Annals of Family Medicine,* **4,** pp. 132-7

OKUMA, K. IWAKAWA, K., TURNIDGE, K., GRUBB, W., BELL, J., O'BRIEN, F., COOMBS, G., PEARMAN, J., TENOVER, F., KAPI, M., TIENSASITORN, C., ITO, T., and HIRAMATSU, K., 2002**,** Dissemination of new methicillin-resistant Staphylococcus aureus clones in the community. *Journal of Clinical Microbiology*, 40 (11), pp. 4289-4294.

ONS (OFFICE OF NATIONAL STATISTICS), 2005, All Fields Postcode Directory, November 2005. Crown Copyright 2004. Last accessed, 19th May, 2006 at http://borders.edina.ac.uk/ukborders/

OPENSHAW, S., CHARLTON, M., WYMER, C. and CRAFT, A., 1987, A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Science*, **1**, pp. 335-358.

R: A Language and Environment for Statistical Computing, (2005). R Foundation for Statistical Computing, last accessed 19th April 2006 at http://www.R-project.org

RIPLEY, B., 1981. *Spatial Statistics*. (Chichester: John Wiley).

ROWLINGSON, B. and DIGGLE, P., 1993, SPLANCS: spatial point pattern analysis code in S-Plus. *Computers and Geosciences*, **19**, pp. 627-655.

SABEL, C., GATRELL, A., LÖYTÖNEN, M., MAASILTA, P., and JOKELAINEN, M., 2000, Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. *Social Science and Medicine*, **50**, pp. 1121-1137.

SARAVOLATZ, L., POHLOD, D. and ARKING, L., 1982, Community-acquired methicillin-resistant *Staphylococcus aureus* infections: a new source for nosocomial outbreaks. *Annals of Internal Medicine,* **97**, pp. 325-9.

SILVERMAN, B., 1986, *Density estimation for Statistics and Data Analysis*, pp. 43-61. (London: Chapman & Hall).

TANGO, T., and TAKAHASHI, K., 2005, A flexibly shaped scan statistic for detecting clusters. *International Journal of Health Geographics*, **4**, pp. 11.

WALLER, L.A. and JACQUEZ, G.M., 1995, Disease models implicit in statistical tests of disease clustering. *Epidemiology*, **6**, pp. 584-590.

WALLER, L. and GOTWAY, C., (2004), Chapter 6 in *Applied Spatial Statistics for Public Health Data*. (New Jersey: John Wiley).

Acknowledgements

We would like to thank Debra Hirst for her valuable assistance with collection of data on nursing homes, and Professor Lance Waller for kind advice and encouragement with R. The advice of two anonymous reviewers was also very valuable.