Epidemiol. Infect. (2006), **134**, 699–711. © 2006 Cambridge University Press
doi:10.1017/S0950268805005741 Printed in the United Kingdom

The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli* O157 in space and time in Alberta, Canada, 2000–2002

D. L. PEARL[1]*, M. LOUIE[2], L. CHUI[3], K. DORÉ[4], K. M. GRIMSRUD[5], D. LEEDELL[3], S. W. MARTIN[1], P. MICHEL[6], L. W. SVENSON[5] AND S. A. MCEWEN[1]

[1] *Department of Population Medicine, University of Guelph, Guelph, Ontario, Canada*
[2] *Provincial Laboratory for Public Health (Microbiology), Calgary, Alberta, Canada*
[3] *Provincial Laboratory for Public Health (Microbiology), Edmonton, Alberta, Canada*
[4] *Division of Enteric, Foodborne and Waterborne Diseases, Public Health Agency of Canada, Guelph, Ontario, Canada*
[5] *Alberta Health and Wellness, Edmonton, Alberta, Canada*
[6] *Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Saint-Hyacinthe, Québec, Canada*

(*Accepted 20 October 2005, first published online 3 January 2006*)

SUMMARY

We obtained a list of all reported cases of *Escherichia coli* O157 in Alberta during the 2000–2002 period, and using scan statistics we identified yearly temporal and spatial clusters of reported cases of *E. coli* O157 during the summer and in southern Alberta. However, the location of the spatial cluster in the south was variable among years. The impact of using both outbreak and sporadic data or only sporadic data on the identification of spatial and temporal clusters was small when analysing individual years, but the difference between spatial clusters was pronounced when scanning the entire study period. We also identified space-time clusters that incorporated known outbreaks, and clusters that were suggestive of undetected outbreaks that we attempted to validate with molecular data. Our results suggest that scan statistics, based on a space-time permutation model, may have a role in outbreak investigation and surveillance programmes by identifying previously undetected outbreaks.

INTRODUCTION

Escherichia coli O157 has been recognized as a significant cause of gastroenteritis, haemorrhagic colitis, and haemolytic uraemic syndrome (HUS) in the developed world [1, 2]. The rates of infection appear to be highest in children <5 years old, and rates of HUS appear to be highest among these young children and the elderly [3, 4]. Infection with this pathogen has been associated with the consumption of contaminated meat [5], dairy products [6], fresh produce [7], drinking and recreational water [8, 9], and contact with shedding animals [10], humans [11], or an environment contaminated with this pathogen [12].

Cattle are the major reservoir for *E. coli* O157 although it has been isolated from other species including sheep, deer, rabbits, and pigs [13–15]. Ecological studies have demonstrated an association between rates of human infection in a community and the concentration of cattle, manure handling practices, and/or prevalence of shedding animals [16–18]. Rates of human infection appear to be higher in summer and early autumn [4, 16, 19, 20], and this seasonal pattern of disease is consistent with the overall pattern of shedding in cattle [14, 21–23].

* Author for correspondence: Dr D. Pearl, Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.
(Email: dpearl@uoguelph.ca)

The rate of reported cases of *E. coli* O157 in Canada ranges typically from 3–4 cases per 100 000 person-years, but in Alberta rates of disease are often more than double the national average [4]. Based on a report from the Palliser Health Authority, Galanis *et al.* [11] suggested that the rates of *E. coli* O157 in southern Alberta are among the highest in Canada. Agricultural intensity in the south of the province has also led to investigations of surface water contamination with *E. coli* O157 and *Salmonella* [24]. Most cases in the province are believed to be sporadic in nature, but small outbreaks have been identified using a combination of pulsed-field gel electrophoresis (PFGE) and standard epidemiological investigations [11]. From an epidemiological point of view, understanding the clustering of cases in space, time, or space-time, for the purpose of identifying potential risk factors or outbreak identification, requires methods that are not limited by pre-defined administrative boundaries that may cause a pre-selection bias. This is especially true when the choice of spatial and/or temporal boundaries for examining health risks have important political and economic repercussions.

A variety of statistical tests have been developed for the identification of clustering in space, time, and space-time [25–27]. These tests can be classified based on whether they detect the presence of clustering or the actual location of clusters [28]. Among these tests is the spatial scan statistic. Its popularity [29] for studying both infectious [30] and non-infectious diseases [31] is largely attributed to its ability to: identify the approximate location of clusters in space, time or space-time; limit pre-selection bias by allowing flexible scanning windows in space and time; make use of a variety of statistical models; scan retrospectively and prospectively; and use Monte Carlo simulations to adjust for multiple testing [28, 32, 33]. These features make this method ideal for exploring the spatial, temporal, and spatio-temporal clusters of human cases of *E. coli* O157 in Alberta.

The ability to differentiate outbreaks from sporadic cases is an important component of the surveillance and public health management of *E. coli* O157 and other infectious diseases. The adoption of PFGE by public health laboratories to supplement traditional epidemiological techniques is largely explained by its utility in helping identify outbreaks [34, 35]. The effectiveness of surveillance systems in differentiating outbreak from sporadic cases could have important implications for analytical studies that assume independence among cases. Scan statistics are often used in the hope that a cluster in space will help reveal a spatially stable risk factor [36, 37]. Unfortunately, the presence of an undetected outbreak may lead to the misinterpretation of an analysis intended to identify spatial clusters. The identification of a spatial cluster is often followed by the generation of hypotheses concerning relatively stable social, physical, and biological risk factors that may explain the increased risk of disease within the identified region. Outbreaks, in contrast, can be the result of exposures that only exist briefly in time or space-time.

In this paper, we present a retrospective analysis of reported cases of *E. coli* O157 from Alberta using spatial, temporal, and spatio-temporal scan statistics. Reported cases include all cases, whether independent (i.e. sporadic) or sharing an epidemiological link with another case (i.e. outbreak), that have been reported in the province's Notifiable Disease Reports (NDRs). Our research had the following objectives:

- locate and determine the statistical and biological significance of spatial and temporal clusters (i.e. areas with statistically significant increased levels of disease) among reported cases in Alberta;
- compare the results of spatial and temporal scan statistics when both sporadic and outbreak cases are included in the analysis compared to sporadic cases alone;
- determine the usefulness of a space-time permutation model in identifying outbreaks by validating the space-time clusters identified using this model with molecular and epidemiological evidence.

METHODS

Case data

Using NDR data, accessed through the Communicable Disease Reporting System, maintained by the Disease Control and Prevention Branch of Alberta Health and Wellness, we obtained a list of all reported cases of *E. coli* O157 in Alberta during the 2000–2002 period. For our analyses, we obtained, using methods that preserved patient anonymity, the following information for each case: a unique identifier (NDR number), onset date of symptoms, municipal address, health region, postal code, date of birth, sex, laboratory number, PFGE pattern number for provincial and national designation, a unique identifier used for community outbreaks (exposure indicator number), cases identified through an epidemiological link that

did not require definitive laboratory results (EPI-linked), and the NDR numbers that connected EPI-linked cases. To avoid the possible identification of cases in small communities, we coded the names of communities with fewer than 100 000 people and only identified their latitude and longitude to the nearest degree wherever the results and discussion required a specific location be identified. The protocol for this research was approved by the University of Guelph Research Ethics Board.

Geocoding

A postal code conversion file containing all valid postal codes and the names of each census subdivision (CSD) was obtained from Statistics Canada [38]. Each case had a postal code and a municipal address for each patient. The municipal address identified the city, town, or village where the patient resided at the time of their illness. Street addresses were unavailable to the senior author due to privacy legislation in the province of Alberta. We linked the case to its CSD based on postal code. Multiple CSDs could be matched to some rural postal codes so the merged files were reviewed manually using the municipal address to determine the correct CSD. In Alberta, a CSD (total = 452) may represent a city ($n = 15$), county/municipality ($n = 28$), improvement district ($n = 8$), municipal district ($n = 36$), Indian reserve ($n = 88$), regional municipality ($n = 1$), Indian settlement ($n = 4$), special area ($n = 3$), specialized municipality ($n = 2$), summer village ($n = 52$), town ($n = 110$), or village ($n = 105$) [39]. For some rural communities, the postal code and municipal address did not provide enough information to link a case to a specific CSD. However, using the 'Communities within Specialized and Rural Municipalities' list provided by Alberta Municipal Affairs, we could determine the appropriate CSD for these cases [40]. When the postal code and municipal address provided conflicting information concerning the geographical location of a case, we manually located the CSD appropriate for the municipal address since the municipal address, unlike the postal code, was always consistent with the health region where the case was recorded. The Geosuite file from Statistics Canada provided the latitude and longitude for each CSD [41].

Population data

From the 2001 Canada Census, we obtained the age and sex distribution of each of 401 CSDs that was enumerated and/or where the population was large enough for the data to be publicly released [42]. Ages were grouped into 5-year intervals beginning at 0–4 years and ending at $\geqslant 85$ years. Two Indian reserves (Ermineskin 138 and Saddle Lake 125) and one Indian settlement (Little Buffalo) were not enumerated while 48 other CSDs had populations below 40 people, therefore information concerning age and sex were suppressed to maintain confidentiality [43].

Inclusion and exclusion criteria

In total, 875 cases of *E. coli* O157 were reported in Alberta in 2000–2002. Six cases were dropped from subsequent analyses due to lack of information on gender ($n = 1$), birth date ($n = 1$), having a home address outside Alberta ($n = 3$), and/or being part of a CSD without the necessary demographic data ($n = 1$).

Household and community outbreaks

In defining outbreak cases, we recognized two types of outbreaks: household and community outbreaks. A household outbreak was defined as any series of two or more cases found exclusively in one household during the study period. To enable the identification of household outbreaks while protecting patient confidentiality, a data field anonymously coding common addresses was created within the offices of Alberta Health and Wellness. A unique identifier was given to all cases sharing a common address. Typically, the time between consecutive cases within these household outbreaks did not exceed 14 days. In one instance, two household cases were separated by more than 1 month, but their isolates had the same PFGE pattern. In two instances, household outbreaks of two patients were based on a general delivery address. However, the two cases in each of these household outbreaks had isolates of *E. coli* O157 that shared a common PFGE pattern.

The EPI-linked field and the exposure indicator number were used to identify community outbreaks. Community outbreaks were defined as any series of epidemiologically linked cases that included more than one household. While identifying household outbreaks, we found cases within these clusters that were linked to cases from other households by the EPI-linked and/or exposure indicator number fields. Consequently, the size of some community outbreaks was greater than was previously recorded in the NDR

database. In addition, a community outbreak was identified when neighbours on the same street during the same 2-week period provided faecal samples with isolates of *E. coli* O157 that shared a common PFGE pattern.

PFGE data

The Provincial Laboratory for Public Health (Microbiology) is a member of PulseNet Canada and CDC-PulseNet. Consequently, the laboratory staff make use of a standard protocol for performing PFGE to facilitate the sharing of these patterns among different laboratories and jurisdictions [44]. They routinely perform PFGE on human cases of *E. coli* O157 reported in Alberta. The NDR database has fields for these pattern numbers, but they were often not updated when PFGE information became available. Using the laboratory number field shared by the NDR database and the laboratory database we were able to determine the PFGE pattern of 88·3% of the 869 cases used in the proceeding analyses. Overall, 89% of the data from the laboratory database, that included 826 cases after correcting for multiple entries from individual patients, out of province cases, and PFGE patterns from non-human samples (e.g. food), were linked to the NDR database. PFGE was not performed on all cases, and typographical errors and missing entries in the laboratory number field from the NDR database accounted for our inability to determine the PFGE pattern for every case used in our analyses. Moreover, cases for which the PFGE data were available only through the NDR database were manually reviewed to make certain the Alberta and national PFGE pattern numbers were consistent with the nomenclature used by the laboratory.

Computer software

Database files were provided to the senior author as Microsoft Excel 2000 (Microsoft, Redmond, WA, USA) files. All database management requiring the merging of files was performed in Intercooled Stata version 8.0 (Stata Corporation, College Station, TX, USA) for Windows except for postal code conversion files where we used Microsoft Access 2000 (Microsoft). Intercooled Stata version 8.0 was also used for calculating the rates of reported cases by gender and age. All scan statistics were performed using SaTScan version 3.1.2 [29] and the software generated the standard morbidity ratios for each CSD used in these analyses. The geographical information system ArcMap 8.2 (ESRI, Redlands, CA, USA) was used for visualizing the scan statistic analyses. All maps used in the figures were provided by Statistics Canada [39].

Statistical models

In the scan statistic, the scanning window can exist in space, time or space-time [29]. The scanning window can be visualized as a series of circles in space, as a line in time, or as a cylinder in space-time with the base representing space and the height representing time. The windows begin as a point at the smallest scale defined in the study at each point in space, time or space-time. In our study, the smallest point in space is the centroid of a CSD, and in time it is the day of disease onset. The size of the window increases until it reaches the next recorded point in space, time or space-time. Each time an additional point is reached, a likelihood ratio and the relative risk is calculated to determine if the rate of disease within the window is different from outside the window based on a Bernoulli, Poisson, or space-time permutation model. In the above models, the null-hypothesis is that the rate (Poisson), proportion of cases to controls (Bernoulli), or the independence of cases in space and time (space-time permutation), is the same within and outside the scanning window. While the Bernoulli and Poisson models can be used to search for space-time clusters, only the space-time permutation model corrects for the presence of both spatial and temporal clusters within the data while it tests whether cases that are close in space are also close in time [29]. Monte Carlo simulations, generating random replications of the dataset under the appropriate null hypothesis, are used to determine the significance of these results. The P values for these tests are calculated by comparing the rank of the maximum likelihood from the real dataset with the maximum likelihoods from the random datasets with $P = \text{rank}/(1 + \text{number of simulations})$ [29]. The number of replications should be a minimum of 999 to ensure excellent power, but 9999 replications are recommended when computing time is not an issue [29]. The maximum spatial scanning window cannot exceed 50% of the population. A cluster found with a spatial scanning window that includes more than 50% of the population reflects an area with an extremely low rate outside the circle rather than an area with an extremely high rate within the circle [29]. The maximum temporal

window is recommended to be no greater than 50% of the study period although in theory it can be higher depending on the model being used [29].

Scan statistics, based on a Poisson model, were used to identify spatial clusters of cases during the study period. These models adjusted for the total number of individuals within each 5-year age class, by gender, based on the 2001 Census for each scanning window. The maximum spatial scanning window was set at the maximum allowable level (50% of the population). These scans were performed for each individual year and for the total 3-year study period, first with all cases and then with only sporadic cases. The scans for 2000, 2001, and 2002 included 324, 285, and 260 cases respectively, when all cases were included. The scans for 2000, 2001, and 2002 included 258, 236, and 190 cases respectively, when only sporadic cases were included. We used 9999 Monte Carlo replications to estimate the significance levels of these clusters. We performed similar analyses for temporal clusters; however, in this case we examined the effect of maximum temporal scanning windows of 30 and 180 days. A Bernoulli model was used to compare sporadic and outbreak cases when there were biologically relevant differences between the results when all the data or just the sporadic data were analysed. A space-time permutation model was used to determine the presence of space-time clusters, but due to computation times we reduced the number of Monte Carlo replications to 999, and the onset dates were merged into 10-day intervals. In searching for space-time clusters by year and for the entire study period, we used all the cases since the purpose was to determine the ability of this approach to find outbreaks within a dataset that did not contain additional epidemiological information.

For all analyses, the most likely (based on the size of the log-likelihood ratio), non-overlapping in space or space-time, statistically significant ($P<0{\cdot}05$) clusters are presented. SaTScan version 3.1.2 allows for the reporting of various degrees of overlap following certain reporting criteria [29]. It is noted that around any cluster a large number of less likely overlapping clusters can be found with high significance since the inclusion/exclusion of a small population may not have a large impact on the results [29]. When we conducted these analyses the software's criteria for reporting secondary clusters was set to allow some overlap as long as the secondary cluster and a previously reported cluster did not both contain each other's centroid. We report only the most likely non-overlapping clusters to simplify the presentation of results. Our reported results would have been no different if we had set the programs reporting criteria to not allow any overlap except with our space-time clusters. By allowing some overlap, we were able to identify space-time clusters that overlapped in space, but not in time. All the tests were run as one-sided tests scanning for high levels of disease since we were searching for disease clusters.

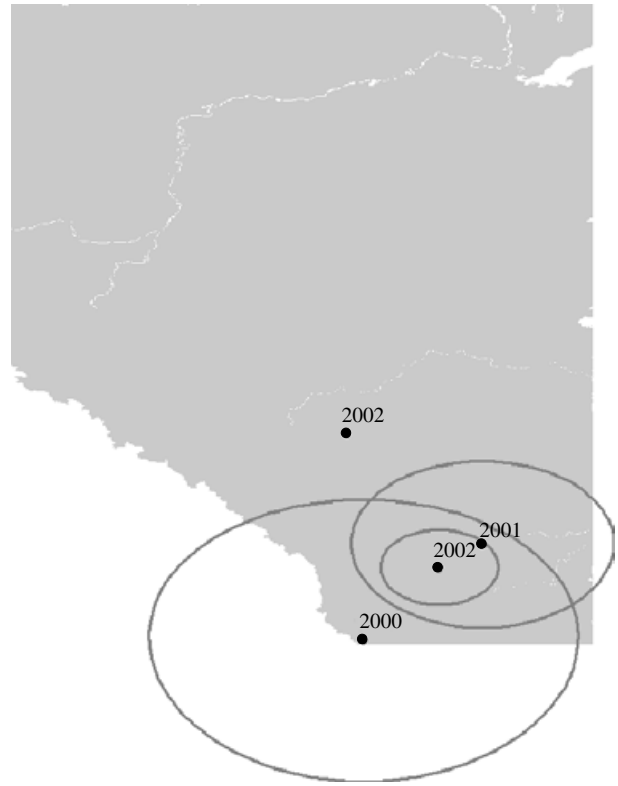

Fig. 1. The most likely non-overlapping spatial cluster(s) for each year of the study using all recorded cases. Each point and circle represents the centroid and area of the scanning window respectively, for each cluster.

RESULTS

Purely spatial scan

In each year of the study period, there was a statistically significant spatial cluster in the southern half of the province regardless of whether all cases (Fig. 1, Table 1) or only sporadic cases (Fig. 2, Table 1) were examined using a Poisson model. The location and size of these clusters varied among years, however, the within-year location and size of these clusters were similar regardless of whether all cases or only sporadic cases were used in these analyses. A spatial cluster was found in a single CSD north of these southern clusters in 2002 as a result of five cases

Table 1. *The central location (latitude and longitude in degrees), size, relative risk, and significance of the most likely non-overlapping spatial clusters of reported* E. coli *O157 cases in Alberta based on scans in 2000, 2001, 2002, and 2000–2002. These scans were performed with all cases (A) and sporadic cases (S) based on a Poisson model (P) or Bernoulli model (B)*

Year(s)	Case type	Model	Latitude (°N)	Longitude (°W)	Radius (km)	No. of cases	Relative risk	*P* value
2000	A	P	49	114	269·32	205	1·48	0·0001
2000	S	P	49	114	269·32	169	1·53	0·0001
2001	A	P	51	112	159·28	175	1·51	0·0001
2001*	S	P	51	112	147·99	136	1·42	0·0003
2002	A	P	50	113	71·73	37	5·80	0·0001
2002	S	P	50	113	105·07	34	2·63	0·0003
2002	A	P	53	114	0†	5	26·03	0·0017
2000–2002	A	P	50	113	91·01	133	2·89	0·0001
2000–2002	S	P	49	114	254·51	402	1·36	0·0001
2000–2002	A	B	50	112	77·54	50	2·03	0·0001

* Different location from the cluster including all cases in 2001 by less than 0·05 °N and 0·5 °W.
† 0 km radius indicates that the cluster is limited to one census subdivision.

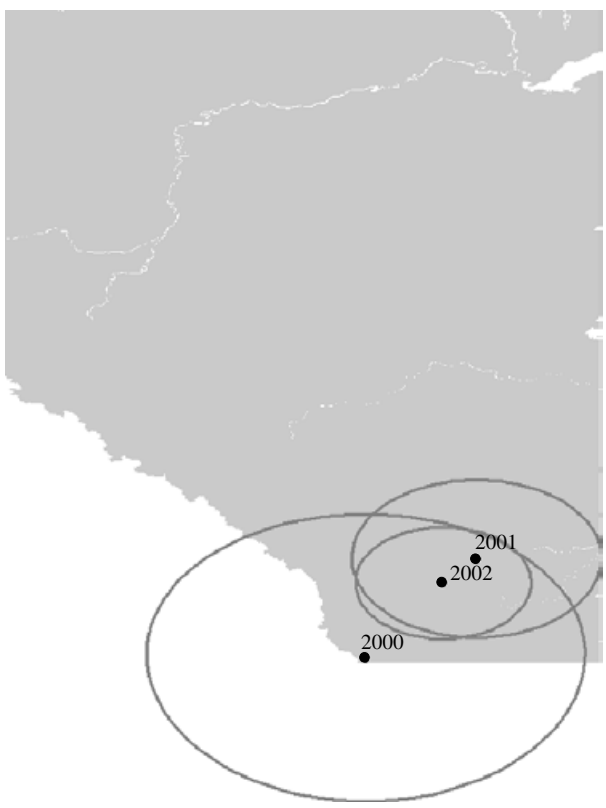

Fig. 2. The most likely non-overlapping spatial cluster for each year using only sporadic data. Each point and circle represents the centroid and the scanning window respectively, for each cluster.

occurring between March and August (Fig. 1, Table 1); three of these cases were classified as sporadic and two were part of a household outbreak. This spatial cluster was not statistically significant when the analysis was run with sporadic cases alone.

When we scanned for spatial clusters over the entire study period, the spatial cluster based on a scan of all cases was centred ~133 km northeast of the cluster found when sporadic cases were examined alone (Fig. 3). The radius of the cluster based on sporadic cases was 163·5 km longer than the cluster identified using all the data, and the increased risk of disease within the sporadic cluster (relative risk = 1·36) was much smaller than that in the spatial cluster found using all the data (relative risk = 2·89; Table 1). A Bernoulli model revealed that for the 2000–2002 period, relative to sporadic cases, outbreak cases were more heavily concentrated in a region that closely overlapped the Poisson-based spatial cluster for 2000–2002 that included all cases (Fig. 3). The 50 outbreak cases that formed the most likely non-overlapping cluster using the Bernoulli model included four household outbreaks in 2000 (13 cases), four household outbreaks (12 cases) and one community outbreak (3 cases) in 2001, and six household outbreaks (12 cases) and one community outbreak (10 cases) in 2002.

Purely temporal clusters

Temporal clusters always occurred within the late spring to early autumn period regardless of whether all cases or only sporadic cases were analysed (Table 2). The relative risk of cases occurring during this period ranged from 2·03 to 3·42 ($P = 0·0001$) depending on the size of the scanning window and

Table 2. *The date, relative risk, and significance of the most likely temporal clusters of reported* E. coli *O157 cases in Alberta based on scans in 2000, 2001, 2002, and 2000–2002. These scans were performed with all cases (A) and sporadic cases (S) based on Poisson models using 30-day and 180-day maximum scanning windows*

Year(s)	Window size (days)	Case type	Date of cluster	No. of cases	Relative risk	*P* value
2000	30	A	5 July to 3 Aug. 2000	75	2·82	0·0001
2000	30	S	8 July to 6 Aug. 2000	64	3·03	0·0001
2000	180	A	27 May to 15 Oct. 2000	255	2·03	0·0001
2000	180	S	27 May to 27 Sept. 2000	193	2·21	0·0001
2001	30	A	5 July to 3 Aug. 2001	68	2·90	0·0001
2001	30	S	5 July to 3 Aug. 2001	55	2·84	0·0001
2001	180	A	1 June to 3 Sept. 2001	176	2·37	0·0001
2001	180	S	1 June to 9 Sept. 2001	143	2·19	0·0001
2002	30	A	18 July to 16 Aug. 2002	62	2·90	0·0001
2002	30	S	5 July to 1 Aug. 2002	45	3·09	0·0001
2002	180	A	6 May to 16 Aug. 2002	163	2·22	0·0001
2002	180	S	14 May to 16 Aug. 2002	106	2·14	0·0001
2000–2002	30	A	5 July to 3 Aug. 2000	75	3·15	0·0001
2000–2002	30	S	8 July to 6 Aug. 2000	64	3·42	0·0001
2000–2002	180	A	27 May to 27 Sept. 2000	239	2·43	0·0001
2000–2002	180	S	27 May to 27 Sept. 2000	193	2·49	0·0001

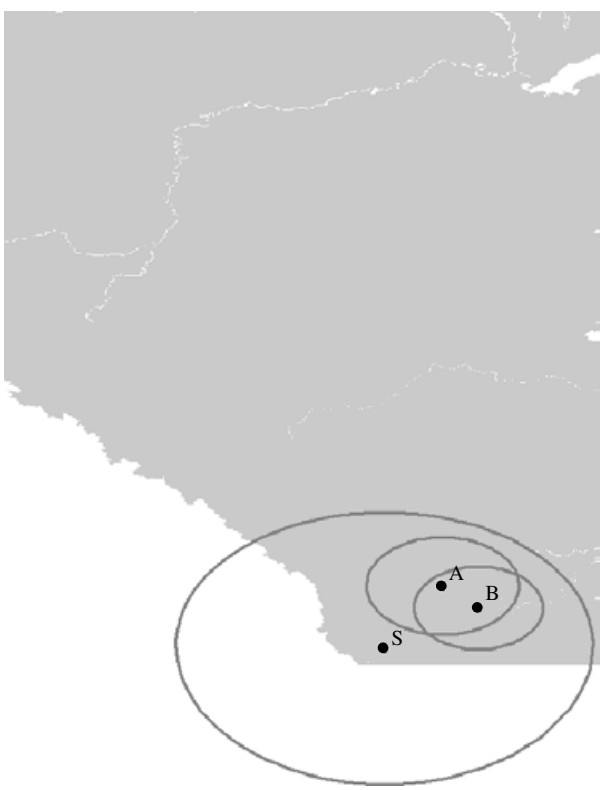

Fig. 3. The most likely non-overlapping cluster for 2000–2002 using all cases (A), sporadic cases (S), and a Bernoulli model (B). The point and circle represents the centroid and the scanning window respectively, for each cluster.

the year(s) being scanned (Table 2). The temporal clusters found when the scanning window was limited to 30 days were always nested within the temporal clusters found when the window was expanded to 180 days (Table 2). When the entire study period was included in these analyses, the summer cluster in 2000 was the most likely temporal cluster. During the study, reported case rates in the province steadily declined with crude rates of 10·9, 9·6, and 8·7 cases per 100 000 person-years in 2000, 2001, and 2002 respectively.

Space-time clusters

In reviewing the NDR database, we identified 14 community outbreaks and 55 household outbreaks that included a total of 58 and 127 cases, respectively. Using the space-time permutation model we identified statistically significant space-time clusters in each year. Unlike the spatial and temporal clusters, these space-time clusters were not limited to the southern half of the province or the late spring to early autumn period (Fig. 4, Table 3). The two space-time clusters found for the 2000–2002 period were similar in location and temporal pattern to the two space-time clusters found when the 2002 data were analysed alone (Fig. 4, Table 3). A space-time cluster was

Table 3. *The date, location (latitude and longitude in degrees), size, number of cases, relative risk, and statistical significance of non-overlapping space-time clusters of reported* E. coli *O157 cases in Alberta identified using a space-time permutation model based on scans in 2000, 2001, 2002, and 2000–2002 using all the case data*

Year(s) scanned	Date of cluster	Latitude (°N)	Longitude (°W)	Radius (km)	No. of cases	Relative risk	*P* value
2000	23 Oct. to 31 Dec. 2000	54	112	105·04	14	4·45	0·004
2001	26 May to 4 June 2001	54	115	54·14	4	25·33	0·016
2002	15 June to 24 June 2002	50	112	35·77	10	11·61	0·001
2002	13 Sept. to 12 Oct. 2002	55	118	148·04	8	9·91	0·001
2000–2002	5 June to 24 June 2002	50	112	77·70	16	13·07	0·001
2000–2002	13 Sept. to 12 Oct. 2002	55	118	148·04	8	17·42	0·001

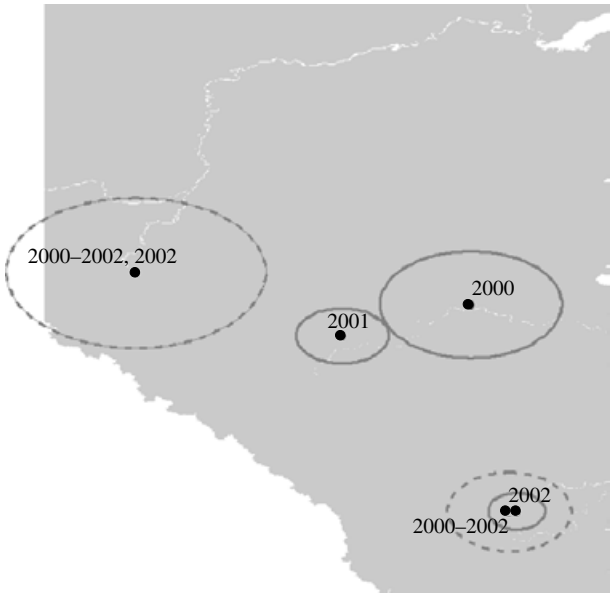

Fig. 4. The most likely non-overlapping space-time cluster(s) for each year (solid line) and the entire study period (dashed line). Each point and circle represents the centroid and the scanning window respectively, for each cluster.

found in 2000 and 2001 when data from these years were separately scanned (Table 3). Often the space-time statistical clusters contained previously identified outbreaks from the NDR database.

The 2000 space-time cluster consisted of cases that occurred over 65 days in Edmonton (nine cases) and four surrounding communities (five cases). Six of these cases were part of household outbreaks (two outbreaks in Edmonton and one in CSD 00-D). PFGE patterns were available for 11 of these cases, and eight different patterns were identified (Table 4). The 2001 space-time cluster occurred over 6 days, and consisted of four cases with one case from CSD 01-A and three cases from a household outbreak in CSD 01-B. The cases from the household outbreak all shared the same PFGE pattern (Table 4). In 2002, two statistically significant non-overlapping space-time clusters were identified. The southernmost cluster consisted of 10 cases (16 cases when the entire study period was included in the analysis), that closely matched a reported daycare outbreak in Brooks that occurred between 4 June and 25 June of that year [11]. The analysis that only included the 2002 data captured 8 of the 10 laboratory-confirmed cases while all 10 cases were found within the cluster when the scan included the entire study period. Regardless of the period included in the scan, this 2002 space-time cluster was heavily dominated by the national PFGE pattern 0.0001. The northern 2002 space-time cluster did not correspond to any outbreaks identified in the NDR database. However, four of the eight cases shared national PFGE pattern 0.0722, which only occurred five times in Alberta during the study period.

DISCUSSION

If ignored or undetected, an outbreak or series of outbreaks could bias the interpretation of a scan analysis intended to identify clustering of disease in space or time. For instance, a spatially stable risk factor, such as cattle density or a local cultural practice, may be falsely attributed to a spatial cluster that was the result of a common source of infection, such as the accidental serving of undercooked ground beef contaminated with *E. coli* O157, that only briefly existed in space-time. In a region with a relatively well developed public health system, it is unlikely that large community outbreaks would remain unidentified. However, this brings into question the potential of small outbreaks that are overlooked or remain unrecorded in public health databases to distort our

Table 4. *The date, census subdivisions (CSDs), national PFGE pattern number, and number of cases from each statistically significant non-overlapping space-time cluster. These clusters were identified using a space-time permutation model using scans for 2000, 2001, 2002, and 2000–2002*

Year(s)	Date of cluster	Cluster CSDs with cases	National PFGE pattern number(s) for each CSD	No. of cases
2000	23 Oct. to 31 Dec. 2000	00-A	0.0355	1
		Edmonton	0.0575 (3X), 0.0146, 0.0535, 0.0536, unrecorded† (3X)	9
		00-B	0.0534	1
		00-C	0.0533	1
		00-D	0.0496 (2X)	2
2001	26 May to 4 June 2001	01-A	0.0146	1
		01-B	0.0384 (3X)	3
2002	15 June to 24 June 2002	Brooks	0.0001 (6X), 0.0657, 0.0670*, 0.0684	9
		02-A-South	0.0654	1
2002	13 Sept. to 12 Oct. 2002	02-A-North	0.0720	1
		02-B-North	0.0355, 0.0722 (4X), unrecorded†	6
		02-C-North	0.0508	1
2000–2002	5 June to 24 June 2002	Brooks	0.0001 (8X), 0.0657, 0.0670*, 0.0684	11
		02-A-South	0.0654	1
		02-B-South	0.0661	1
		02-C-South	0.0660 (2X)	2
		02-D-South	0.0660	1
2000–2002	13 Sept. to 12 Oct. 2002	02-A-North	0.0720	1
		02-B-North	0.0355, 0.0722 (4X), unrecorded†	6
		02-C-North	0.0508	1

* Two PFGE patterns (1 band difference) were isolated from this patient. Only the pattern listed has a national designation.
† If a case was identified only through an epidemiological link or we were unable to link a Notifiable Disease Report with the appropriate laboratory data, the PFGE pattern was listed as 'unrecorded'.

perception of spatial or temporal clusters. In our study, we found that ignoring outbreaks in spatial scan analyses can have a large impact on the size and location of spatial clusters. Space-time clusters, identified using the space-time permutation model, appear to detect epidemiologically plausible outbreaks, but the space-time clusters found in our analyses accounted for only a small proportion of the total number of outbreak cases. Consequently, a purely analytical approach to sorting outbreak from sporadic cases may still underestimate the proportion of outbreak cases in the data.

Some studies make use of all cases in their analyses [18] while others only use cases they believe are sporadic in nature or not part of a community outbreak [45]. In particular, household outbreaks are not routinely identified in the NDR database. The availability of family names and addresses may make the recording of these events seem unnecessary, but when these databases are shared with other research groups for further analysis this information is often withheld or hidden to preserve patient privacy. Based on our study, the cases from household and small community outbreaks did not have a profound impact on the location of spatial and temporal clusters in yearly scans of the data. The increased rate of cases in the summer was identified in each year, with some variation in the start and end dates, while the spatial clusters were consistently located in the southern half of the province with some variation in their location. The relative risk of the spatial cluster identified in 2002 using both sporadic and outbreak cases was almost two times greater than the relative risk of the cluster identified using sporadic data alone which may be largely explained by an outbreak in the city of Brooks that summer [11].

In contrast, the most likely location of a spatial cluster for the entire study period was markedly different depending on whether all the data or only sporadic data were included in the analysis. When all the data were included in the analysis, a small spatial cluster whose circumference included several municipalities that contain some of Canada's largest beef cattle feedlots was apparent. Based on these results, it

would be reasonable to hypothesize that there was an association between cattle density and the rate of disease from *E. coli* O157. On the other hand, when only sporadic cases were included in this analysis, the spatial cluster had a far greater radius that included Calgary, a major Canadian city. An analysis using a Bernoulli model indicated that the difference in the distribution of outbreak and sporadic cases was not the result of the Brooks outbreak alone since excess outbreak cases were identified in each year from several community and household outbreaks. While it would be unwise to ignore the possible impact of cattle density on the high rates of disease in southern Alberta, the results of the analysis including sporadic cases alone suggest that a broader socio-ecological perspective may be required in future analytical studies.

Temporal clusters may also reveal outbreaks that are not localized in space. This type of diffuse outbreak has occurred with *E. coli* O157 when contaminated food products have been distributed over a wide geographical area [46, 47]. The interpretation of temporal clusters can be affected by the maximum scanning window used for the analysis. Temporal clusters may appear like outbreaks if the maximum scanning windows are excessively small. In our data, a 180-day maximum scanning window allowed us to differentiate seasonal peaks from diffuse outbreaks. While we may have suspected that a diffuse outbreak would occur over a shorter period, allowing a wider maximum scanning window helped to differentiate a diffuse outbreak from a seasonal effect. In addition, performing the analysis by individual year showed the consistency of the cluster pattern among years regardless of the maximum size of the scanning window.

Scan statistics may offer an additional tool for outbreak detection. The space-time permutation model, which corrects for the effect of purely spatial and temporal clusters in the data, may make it the most appropriate statistical model for outbreak detection with *E. coli* O157; a pathogen which has been associated with seasonal and spatial clustering in several studies [16, 18, 45]. In our yearly analyses, we identified four space-time clusters and three of these clusters could be easily recognized as being part of an outbreak using epidemiological and/or molecular evidence. The Brooks daycare outbreak in 2002, the largest outbreak during this period, and a household outbreak in CSD 01-B in 2001 were detected during these scans. The space-time cluster that included cases from CSD 02-B-North appears to be an undetected or unrecorded outbreak based on the disproportionate number of cases with the same PFGE pattern. The 2000 space-time cluster captured using the space-time permutation model is not readily linked to a single outbreak or dominated by a single PFGE pattern. However, it is beyond the scope of this paper to analyse the similarity among these patterns, and other outbreaks have had multiple PFGE patterns [48]. We are currently developing a test to determine the statistical significance of similarity among PFGE patterns within these space-time clusters.

Using scan statistics we were only able to identify a small proportion of the total number of recorded outbreak cases in statistically significant clusters. This may reflect the limited spatial resolution of our study, when we were searching for outbreaks using the space-time permutation model, and the potential for misclassifying the spatial location of cases by using their home address. For instance, in the Brooks outbreak, public health workers at the Palliser Health Region were able to identify the outbreak by the time of the third reported case based on the children's attendance at a common daycare facility [11]. It is evident that with the appropriate epidemiological information, public health workers should be able to identify outbreaks more efficiently. However, the scan statistic may be most appropriate for alerting public health workers to space-time clusters that encompass larger geographical areas than a single community and that may have more than one PFGE pattern due to clonal turnover or a mixed infection [11, 46–49]. The results of our study emphasize the need for testing these scan statistics prospectively under field conditions so the sensitivity and specificity of this method can be readily evaluated with an epidemiological investigation. Prospective space-time scans, using a Bernoulli model, are already showing a great deal of potential as an early warning system in the surveillance of West Nile Virus in New York City [50], and sequential mapping with spatial cluster detection, based on a Poisson model, has also been studied as a tool for identifying outbreaks of *E. coli* O157 [51]. Recently, the space-time permutation model was tested for detecting outbreaks of disease based on hospital emergency-room visits [52]. It should be noted that the space-time permutation model, like other models of space-time interaction, can be sensitive to changes in the background population so the period being scanned needs to be limited to periods when the population is relatively stable [25, 29].

The results of cluster studies always need to be tempered with knowledge of their limitations. For instance, many of these studies are based on surveillance data, therefore, reporting bias may complicate results. This bias can exist anywhere in the reporting chain from the initial tendency of a patient to seek the care of a physician to the ultimate recording of the case within the disease registry. Among jurisdictions, under-reporting may be variable so that any 'cluster' type study may only be reflecting differences in reporting among geographical regions [53, 54]. The presentation of the most likely non-overlapping cluster is usually the most efficient way to present the results of scan statistics, however, the best method for presenting these results is not without controversy [55]. There are almost always other statistically significant overlapping clusters related to the cluster being reported. Consequently, the uncertainty surrounding the exact location of the cluster is not presented. The quality of geocoding always remains an issue affected both by the resolution of the spatial information in the database and the variation in the incubation period of the disease which may result in profound uncertainty concerning the location of exposure to the agent.

Our study addresses the additional data quality issue concerning the differentiation of sporadic and outbreak cases, it also points out the potential of scan statistics to identify epidemiologically valid outbreaks that may have been overlooked by standard surveillance. Recognizing that cases within an outbreak are not independent does not mean that outbreak cases should be ignored in spatial analyses. Previously, we addressed our concern of misinterpreting a spatial cluster that is mainly composed of a single outbreak. However, it would be unwise to completely ignore the spatial clustering of cases as a result of a repeated series of outbreaks. These types of clusters may help in identifying risk factors associated with outbreaks or reveal differences among health units in reporting or detecting outbreaks.

ACKNOWLEDGEMENTS

The primary author has been supported by fellowships and awards from the Canadian Institutes of Health Research and the Ontario Veterinary College. The authors acknowledge the support of the Wellcome Trust through their International Partnership Research Award in Veterinary Epidemiology.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Ochoa TJ, Cleary TG.** Epidemiology and spectrum of disease of *Escherichia coli* O157. *Current Opinion in Infectious Diseases* 2003; **16**: 259–263.
2. **Nataro JP, Kaper JB.** Diarrheagenic *Escherichia coli*. *Clinical Microbiology Reviews* 1998; **11**: 142–201.
3. **Dundas S, Todd WT.** Clinical presentation, complications and treatment of infection with verocytotoxin-producing *Escherichia coli*. Challenges for the clinician. *Symposium Series. Society for Applied Microbiology* 2000; **29**: 24S–30S.
4. **Waters JR, *et al.*** Infection caused by *Escherichia coli* O157:H7 in Alberta, Canada, and in Scotland: a five-year review, 1987–1991. *Clinical and Infectious Diseases* 1994; **19**: 834–843.
5. **MacDonald DM, *et al.*** *Escherichia coli* O157:H7 outbreak linked to salami, British Columbia, Canada, 1999. *Epidemiology and Infection* 2004; **132**: 283–289.
6. **Gillespie IA, *et al.*** Milkborne general outbreaks of infectious intestinal disease, England and Wales, 1992–2000. *Epidemiology and Infection* 2003; **130**: 461–468.
7. **Welinder-Olsson C, *et al.*** EHEC outbreak among staff at a children's hospital – use of PCR for verocytotoxin detection and PFGE for epidemiological investigation. *Epidemiology and Infection* 2004; **132**: 43–49.
8. **Hrudey SE, *et al.*** A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world. *Water Science Technology* 2003; **47**: 7–14.
9. **Bruce MG, *et al.*** Lake-associated outbreak of *Escherichia coli* O157:H7 in Clark County, Washington, August 1999. *Archives of Pediatric and Adolescent Medicine* 2003; **157**: 1016–1021.
10. **CDC.** Outbreaks of *Escherichia coli* O157:H7 infections among children associated with farm visits – Pennsylvania and Washington, 2000. *Journal of the American Medical Association* 2001; **285**: 2320–2322.
11. **Galanis E, *et al.*** Investigation of an *E. coli* O157:H7 outbreak in Brooks, Alberta, June–July 2002: the role of occult cases in the spread of infection within a daycare setting. *Canadian Communicable Disease Report* 2003; **29**: 21–28.
12. **Howie H, *et al.*** Investigation of an outbreak of *Escherichia coli* O157 infection caused by environmental exposure at a scout camp. *Epidemiology and Infection* 2003; **131**: 1063–1069.
13. **Fischer JR, *et al.*** Experimental and field studies of *Escherichia coli* O157:H7 in white-tailed deer. *Applied and Environmental Microbiology* 2001; **67**: 1218–1224.
14. **Chapman PA, *et al.*** A 1-year study of *Escherichia coli* O157 in cattle, sheep, pigs and poultry. *Epidemiology and Infection* 1997; **119**: 245–250.

15. **Bailey JR, *et al.*** Wild rabbits – a novel vector for Vero cytotoxigenic *Escherichia coli* (VTEC) O157. *Communicable Disease and Public Health* 2002; **5**: 74–75.
16. **Michel P, *et al.*** Temporal and geographical distributions of reported cases of *Escherichia coli* O157:H7 infection in Ontario. *Epidemiology and Infection* 1999; **122**: 193–200.
17. **Valcour JE, *et al.*** Associations between indicators of livestock farming intensity and incidence. *Emerging Infectious Diseases* 2002; **8**: 252–257.
18. **Kistemann T, *et al.*** GIS-supported investigation of human EHEC and cattle VTEC O157 infections in Sweden: geographical distribution, spatial variation and possible risk factors. *Epidemiology and Infection* 2004; **132**: 495–505.
19. **Douglas AS, Kurien A.** Seasonality and other epidemiological features of haemolytic uraemic syndrome and *E. coli* O157 isolates in Scotland. *Scottish Medical Journal* 1997; **42**: 166–171.
20. **Cai Q, Olson J.** Sporadic cases of hemorrhagic colitis associated with *Escherichia coli* O157:H7 in rural Wisconsin. *Wisconsin Medical Journal* 1998; **97**: 50–53.
21. **Van Donkersgoed J, *et al.*** The prevalence of verotoxins, *Escherichia coli* O157:H7, and *Salmonella* in the feces and rumen of cattle at processing. *Canadian Veterinary Journal* 1999; **40**: 332–338.
22. **Laegreid WW, *et al.*** Prevalence of *Escherichia coli* O157:H7 in range beef calves at weaning. *Epidemiology and Infection* 1999; **123**: 291–298.
23. **Synge BA.** Recent epidemiological studies of verocytotoxin-producing *E. coli* O157 in cattle in Scotland. *Cattle Practice* 2000; **8**: 341–343.
24. **Johnson JY, *et al.*** Prevalence of *Escherichia coli* O157:H7 and *Salmonella* spp. *in surface waters. Canadian Journal of Microbiology* 2003; **49**: 326–335.
25. **Kulldorff M, Hjalmars U.** The Knox method and other tests for space-time interaction. *Biometrics* 1999; **55**: 544–552.
26. **Bell BS.** Spatial analysis of disease – applications. *Cancer Treatment Research* 2002; **113**: 151–182.
27. **Ward MP, Carpenter TE.** Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. *Preventive Veterinary Medicine* 2000; **45**: 257–284.
28. **Kulldorff M, Nagarwalla N.** Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**: 799–810.
29. **Kulldorff M, Information Management Services, Inc.** SaTScan v. 3.0: software for the spatial and space-time scan statistics, 2002.
30. **Ward MP.** Clustering of reported cases of leptospirosis among dogs in the United States and Canada. *Preventive Veterinary Medicine* 2002; **56**: 215–226.
31. **Kulldorff M, *et al.*** Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health* 1998; **88**: 1377–1380.
32. **Kulldorff M.** Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society* (*A*) 2001; **164**: 61–72.
33. **Kulldorff M.** A spatial scan statistic. *Communications in Statistics: Theory and Methods* 1997; **26**: 1481–1496.
34. **Bender JB, *et al.*** Surveillance by molecular subtype for *Escherichia coli* O157:H7 infections in Minnesota by molecular subtyping. *New England Journal of Medicine* 1997; **337**: 388–394.
35. **Swaminathan B, *et al.*** PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases* 2001; **7**: 382–389.
36. **Fang Z, *et al.*** Brain cancer mortality in the United States, 1986 to 1995: a geographic analysis. *Neuro-Oncology* 2004; **6**: 179–187.
37. **Odoi A, *et al.*** Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *International Journal of Health Geographics* 2004; **3**: 11.
38. **Statistics Canada.** Postal code conversion file, September 2002 postal codes. Ottawa: Statistics Canada, Ottawa, 2002.
39. **Statistics Canada.** Cartographic boundary files, 2001 Census: reference guide. Ottawa: Statistics Canada, 2002.
40. **Local Government Services Division Municipal Services Branch.** Communities within specialized and rural municipalities. Edmonton: Alberta Municipal Affairs, 2003.
41. **Statistics Canada.** GeoSuite, 2001 Census. Ottawa: Statistics Canada, 2002.
42. **Statistics Canada.** Profile series. Profile of age and sex, for Canada, provinces, territories, census divisions and census subdivisions, 2001 Census. Ottawa: Statistics Canada, 2002.
43. **Statistics Canada.** 2001 Census dictionary. Ottawa: Statistics Canada, 2002.
44. **Chang N, Chui L.** A standardized protocol for the rapid preparation of bacterial DNA for pulsed-field gel electrophoresis. *Diagnostic Microbiology and Infectious Disease* 1998; **31**: 275–279.
45. **Innocent GT, *et al.*** Spatial and temporal epidemiology of sporadic human cases of *Escherichia coli* O157 in Scotland (1996–1999). *Epidemiology and Infection* 2005; **133**: 1033–1041.
46. **Barrett TJ, *et al.*** Laboratory investigation of a multistate food-borne outbreak of *Escherichia coli* O157:H7 by using pulsed-field gel electrophoresis and phage typing. *Journal of Clinical Microbiology* 1994; **32**: 3013–3017.
47. **Hilborn ED, *et al.*** A multistate outbreak of *Escherichia coli* O157:H7 infections associated with consumption of mesclun lettuce. *Archives of Internal Medicine* 1999; **159**: 1758–1764.
48. **Jackson LA, *et al.*** Where's the beef? The role of cross-contamination in 4 chain restaurant-associated outbreaks of *Escherichia coli* O157:H7 in the Pacific Northwest. *Archives of Internal Medicine* 2000; **160**: 2380–2385.
49. **Faith NG, *et al.*** Prevalence and clonal nature of *Escherichia coli* O157:H7 on dairy farms in Wisconsin.

Applied and Environmental Microbiology 1996; **62**: 1519–1525.

50. **Mostashari F, *et al.*** Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases* 2003; **9**: 641–646.
51. **Michel P, *et al.*** Use of sequential mapping and cluster detection statistics for the surveillance of shiga-toxin *Escherichia coli* infection in the province of Ontario, Canada. In: Flahaut A, Toubiana L, Valleron AJ, eds. *Geography and Medicine: GEOMED'99: Proceedings of the Second International Workshop on Geomedical Systems, Paris, 22–23 November, 1999*; New York: Elsevier, 2000: pp. 49–53.
52. **Kulldorff M, *et al.*** A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2005; **2**: e59.
53. **Voetsch AC, *et al.*** Laboratory practices for stool-specimen culture for bacterial pathogens, including *Escherichia coli* O157:H7, in the FoodNet sites, 1995–2000. *Clinical and Infectious Diseases* 2004; **38** (Suppl 3): S190–S197.
54. **Flint JA, *et al.*** From stool to statistics: reporting of acute gastrointestinal illnesses in Canada. *Canadian Journal of Public Health* 2004; **95**: 309–313.
55. **Boscoe FP, *et al.*** Visualization of the spatial scan statistic using nested circles. *Health Place* 2003; **9**: 273–277.