Epidemiol. Infect. (2007), **135**, 100–109. © 2006 Cambridge University Press
doi:10.1017/S0950268806006650 Printed in the United Kingdom

The use of randomization tests to assess the degree of similarity in PFGE patterns of *E. coli* O157 isolates from known outbreaks and statistical space–time clusters

D. L. PEARL[1]*, M. LOUIE[2], L. CHUI[3], K. DORÉ[4], K. M. GRIMSRUD[5], 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] *Foodborne, Waterborne and Zoonotic Infections Division, 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 3 April 2006, first published online 2 June 2006*)

SUMMARY

Using isolates from reported cases of *Escherichia coli* O157 from Alberta, Canada in 2002, we applied randomization tests to determine if cases associated with an outbreak or statistical space–time cluster had more similar pulsed-field gel electrophoresis patterns, based on Dice coefficients, than expected by chance alone. Within each outbreak and space–time cluster, we assessed the mean, median, 25th percentile, 75th percentile, standard deviation, coefficient of variation, and interquartile range of the Dice coefficients of each pairwise comparison among the isolates. To assess the statistical significance of measures of location (e.g. mean) and variation (e.g. standard deviation) we created randomization distributions using all isolates or only isolates from sporadic cases. We determined that randomization tests are an appropriate tool for evaluating the similarity among isolates from cases that have been linked epidemiologically or statistically. We found little difference between using all cases or only sporadic cases when creating our randomization distributions.

INTRODUCTION

Escherichia coli O157 is a significant human pathogen that is routinely detected in public health surveillance programmes in Europe, North America, and Japan [1–4]. It is a major cause of gastroenteritis, haemorrhagic colitis, and haemolytic–uraemic syndrome in these regions [5]. Although *E. coli* O157 is a major foodborne pathogen [6], infections have also been associated with contaminated drinking and recreational water [7], direct exposure to shedding animals or humans [8, 9], and exposure to environments contaminated with this pathogen [10]. Both sporadic and outbreak cases have been reported, and the use of molecular techniques has facilitated the identification of cases associated with a common source [11].

A variety of molecular typing methods have been proposed for outbreak investigation and routine surveillance [12]. For the identification of enteric pathogens, such as *E. coli* O157, the use of pulsed-field gel electrophoresis (PFGE) has become increasingly widespread throughout North America and the world. The use of PFGE for surveillance has often

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

involved the expansion or creation of PulseNet-like molecular subtyping networks [1]. The popularity of PFGE for epidemiological studies of *E. coli* O157 is largely attributed to its high discriminatory power and reproducibility relative to other methods [13]. In retrospective evaluations, PFGE has proven highly effective at differentiating isolates associated with different outbreaks and in discriminating outbreak-associated isolates from sporadic case isolates [14–16]. In public health settings, PFGE has proven invaluable in the surveillance of *E. coli* O157 and some outbreaks have been detected based on PFGE-subtype surveillance alone [11]. The usefulness of PFGE subtyping networks of laboratories is highlighted by their ability to recognize outbreaks over multiple jurisdictions [2, 17].

Despite the success of PFGE in surveillance and epidemiological investigations, there are both technical and theoretical concerns about the use of this molecular technique [18, 19]. Users of subtyping methods based on banding patterns, like PFGE, are often required to make subjective decisions concerning pattern designation and the degree of similarity among patterns. The use of computerized systems for assisting in the assignment of bands relative to known standards helps perform these tasks but still requires manual editing and an element of subjectivity [20]. Consequently, there are efforts to produce typing techniques that provide more objective or binary results [18, 21]. Using PFGE to determine the degree of genetic similarity among isolates has also proven to be a contentious issue. Due to the lack of correlation in similarity coefficients among banding patterns produced by different restriction enzymes, it has been estimated that at least six enzymes are required to infer genetic relationships among isolates subtyped using PFGE without epidemiological data [19]. The lack of character independence among bands in PFGE patterns also limits the reliability of phylogenetic analyses based on this subtyping method [22]. However, surveillance programmes using PFGE for subtyping *E. coli* O157 rarely apply this technique without epidemiological data, and public health workers are generally not as interested in evolutionary relationships as population geneticists or evolutionary biologists.

In surveillance programmes and outbreak investigations, PFGE is used along with spatial, temporal, and other epidemiological information to make decisions concerning relationships among cases. In this context, phenotypic similarity among banding patterns is a relevant matter even if based on a single restriction enzyme. Based on epidemiological investigations, the banding patterns of isolates from outbreaks are often identical or highly similar based on the PFGE patterns produced by a single enzyme [14, 15, 23]. Recently, spatial scan statistics have been used to retrospectively identify outbreaks of *E. coli* O157 using a space–time permutation model [24]. These outbreaks were validated based on the spatio-temporal location of previously confirmed outbreaks and/or the number of isolates sharing a banding pattern. However, the presence of a few highly related isolates in a cluster of cases may be explained by chance alone depending on the size of the cluster and the typical degree of variability in banding patterns among isolates identified in a molecular surveillance programme. Consequently, the question of phenotypic similarity within a cluster of cases needs to be expressed as more than the function of the number of shared bands or the number of isolates sharing a distinct pattern. It may be more important to assess whether the degree of similarity or variation among patterns is significantly different from what would be expected by chance events.

Dice coefficients are frequently used as a measure of similarity for band-based molecular techniques [25]. Dice coefficients are calculated by multiplying the number of shared bands between two patterns by two and then dividing by the sum of the number of bands in each pattern. A variety of statistics, such as the mean or median, can be applied to Dice coefficients to measure the similarity among a cluster of isolates believed to come from a common source. However, establishing statistical significance to these measures requires some type of distribution. In theory, one could establish an empirical distribution by looking at a series of isolates collected by a surveillance programme over a defined period of time (e.g. a calendar year). By using all possible permutations of clusters of a fixed size from this database, an empirical distribution could be created. The percentile that a cluster of interest fell into relative to this distribution could be used to assess the statistical significance of the cluster. Unfortunately, enumerating all possible permutations can become computationally quite demanding. For instance, the number of possible permutations of 10 isolates from a population of 50 isolates is $1{\cdot}03 \times 10^{10}$. However, a large sample taken randomly from this population can provide an adequate representation of the complete enumeration [26, 27]. In fact, these types of randomization tests,

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press

also referred to as Monte Carlo hypothesis tests, have found a wide range of applications in biology and spatial epidemiology [27–29].

Based on a retrospective review of banding patterns from PFGE analyses performed on isolates from reported cases of *E. coli* O157, we created a randomization test to determine if cases associated with an outbreak or statistical space–time cluster were more similar and showed less variation than expected by chance alone. An outbreak was defined as two or more cases sharing an epidemiological link. A space–time cluster was defined as a statistically significant space–time cluster identified using a space–time permutation model. Dice coefficients were used to assess the similarity in PFGE patterns among the isolates of these outbreaks and space–time clusters. Our research objectives included determining: whether isolates from outbreak-related cases have statistically greater similarity and lower levels of variation than expected by chance; the impact of outbreak size on the significance level of various measures of similarity and variation; the impact of including all cases or only sporadic cases for creating a randomization distribution.

METHODS AND MATERIALS

Molecular and epidemiological data

The Alberta Provincial Laboratory for Public Health (Microbiology) is a member of PulseNet Canada and CDC-PulseNet. Like other laboratories within these networks, its members followed a standardized protocol for performing PFGE to facilitate the sharing of these patterns among different laboratories and jurisdictions [30]. They routinely performed PFGE, using the restriction enzyme *Xba*I, on all reported cases in Alberta where a microbiological sample was available. These PFGE patterns were stored electronically using BioNumerics software version 2.5 (Applied Maths, Kortrijk, Belgium).

We reviewed the laboratory's electronic database for all isolates processed in 2002. For our subsequent analyses, we only included isolates from human cases that were Alberta residents whose symptoms first appeared in 2002. In a small number of cases (<2%), multiple PFGE patterns were available from a single case due to the collection of multiple faecal samples from the same patient. In these cases, only the first isolate processed by the laboratory was used in subsequent analyses. In total, 248 isolates from sporadic and outbreak cases were included in creating randomization distributions except when the analysis only involved isolates from sporadic cases ($n=184$).

Cases were defined as sporadic or part of an outbreak based on Notifiable Disease Report (NDR) data compiled in the Communicable Disease Reporting System (CDRS) maintained by the Disease Control and Prevention Branch of Alberta Health and Wellness. Sporadic cases were defined as those that were not linked to another case by epidemiological evidence. Outbreak cases were identified in the CDRS as sharing an epidemiological link based on a unique identifier or common address. We described outbreaks where all cases shared a single address as 'household outbreaks'. In contrast, outbreaks where cases came from two or more addresses were defined as 'community outbreaks'. A more detailed description of the data editing involved in identifying outbreak cases in the CDRS has been described previously [24]. The protocol for this research was approved by the University of Guelph Research Ethics Board.

Statistics

PFGE patterns among isolates were compared using BioNumerics version 2.5 (Applied Maths). Optimization and position tolerance settings were based on empirical tests using 20 isolates with four unique laboratory-identified PFGE patterns (five isolates per pattern). Each isolate came from a different gel. Using the unweighted pair-group method using arithmetic averages (UPGMA) and testing the effect of position tolerance settings and optimization parameters over a range of 0–4% (increasing both together at 0·5% increments), we found that a setting of 0·5% for the optimization and position tolerance made the fewest errors in terms of grouping isolates with the same patterns.

A matrix of Dice coefficients used to compare the similarity in banding patterns among the study isolates was exported from BioNumerics version 2.5 as a text file. The text file was edited in Intercooled STATA 8.0 (Stata Corp., College Station, TX, USA) for Windows and this software was used to create our randomization distributions. These distributions were created for clusters ranging from 2 to 10 isolates. A program written in STATA performed the following operations 10 000 times: randomly select a fixed number of isolates without replacement; calculate the mean, median, 25th percentile, 75th percentile,

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press

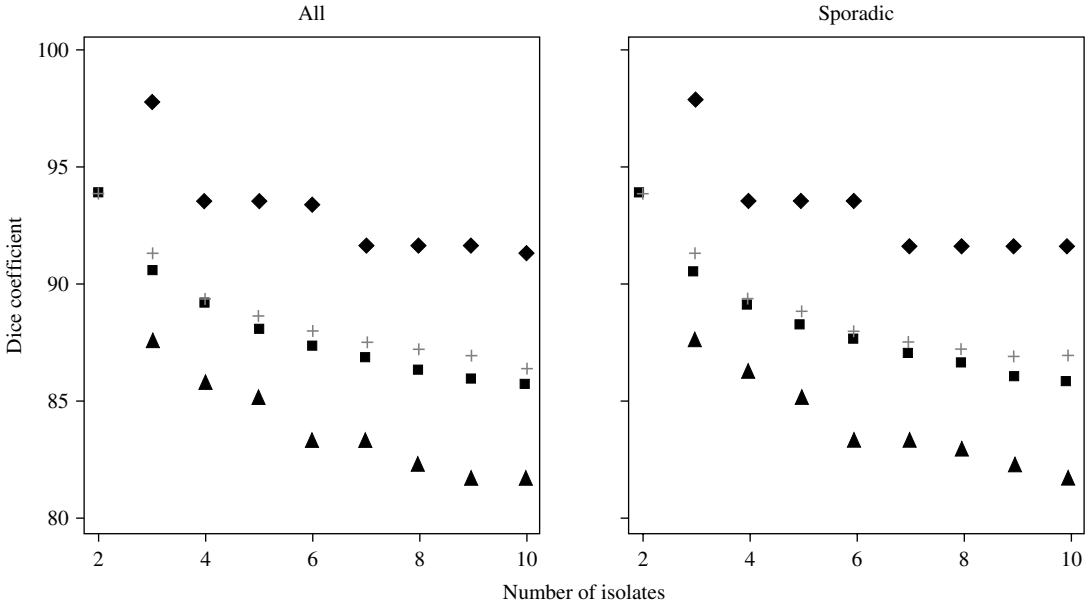

Fig. 1. The 95th percentile of the randomization distributions of the mean (■), median (+), 25th percentile (▲), and 75th percentile (◆) of Dice coefficients for clusters ranging in size from 2 to 10 isolates. These values were used to determine *P* values of $\leqslant 5\%$ for the appropriate one-tailed tests.

standard deviation, coefficient of variation and interquartile range of Dice coefficients from all possible pairwise permutations of the randomly selected isolates; and store the results of each iteration in a common file. These randomization distributions were created with all recorded isolates ($n=248$) and later with only isolates from sporadic cases ($n=184$). Using these randomization distributions for each cluster size (i.e. 2–10 isolates), we determined the Dice coefficient that marked the 2·5, 5, 10, 25, 50, 75, 90, 95, and 97·5 percentiles. These percentiles were used to determine the statistical significance of similarity and variation among isolates found in outbreaks and space–time clusters. For instance, a mean Dice coefficient that fell above the 97·5 percentile or below the 2·5 percentile of a randomization distribution would have a *P* value of $<5\%$ for a two-tailed test.

Outbreaks assessed

We assessed the statistical significance of the measures of location and variation of Dice coefficients among isolates from eight community outbreaks, 10 household outbreaks, and two space–time clusters identified in 2002 in Alberta. The space–time clusters were identified using a space–time permutation model with SatScan version 3.1.2 [24, 31]. We only included outbreaks where an isolate was analysed with PFGE in more than 75% of cases. Only 4 of 22 epidemiologically identified outbreaks failed to meet this criterion. Under our one-tailed null-hypotheses, the measures of location (e.g. mean) of Dice coefficients within an outbreak would not be greater than expected by chance and measures of variation (e.g. standard deviation) would not be less than expected by chance. We assessed these outbreaks using randomization distributions produced with all recorded isolates and with only isolates from sporadic cases. We also included the results of two-tailed tests for comparison.

RESULTS

As we increased the number of isolates in creating our randomization distributions, we found that the 95th percentile of the mean, median, 25th percentile, and 75th percentile Dice coefficient among all unique pairwise comparisons decreased (Fig. 1). In contrast, the 5th percentile of the standard deviation, coefficient of variation, and interquartile range of these Dice coefficients increased with increasing isolate number (Fig. 2). There appeared to be little difference in the pattern and values of these 'cut-off' values for assessing our one-tailed hypotheses whether we used isolates from all cases or only sporadic cases (Figs 1 and 2).

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press

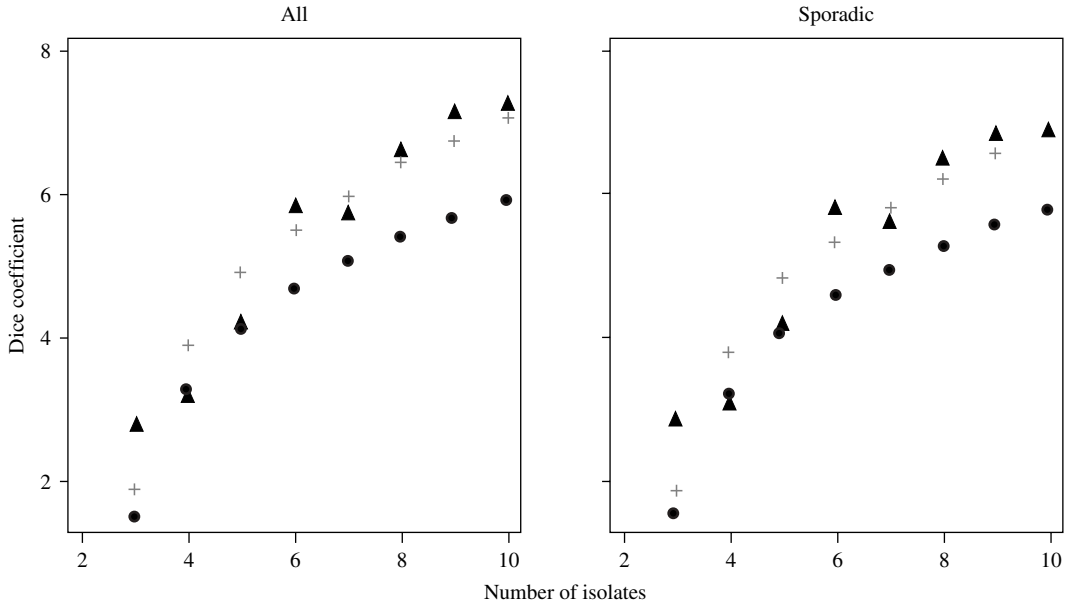

Fig. 2. The 5th percentile of the randomization distributions of the standard deviation (●), coefficient of variation (+), and interquartile range (▲) of Dice coefficients for clusters ranging in size from 2 to 10 isolates. These values were used to determine *P* values of $\leqslant 5\%$ for the appropriate one-tailed tests.

We assessed the significance of measures of location and variation of Dice coefficients for all outbreaks where appropriate. In the case of outbreaks involving two cases, there was only one pairwise comparison so there was no variation to measure. We found that in more than 60% of outbreaks, the mean, median, and 75th percentile were statistically significant using a one-tailed test whether we used all or only sporadic data to create our randomization distributions (Table 1). The 25th percentile was statistically significant in slightly less than 60% of outbreaks (Table 1). In contrast, measures of variation were rarely smaller than expected by chance alone (Table 1). The percentage of significant outcomes for each measure was similar for two-tailed tests (Table 1). In general, there was little difference between using all data or only sporadic data except in a space–time cluster (outbreak 20) where the median Dice coefficient was only significant with a two-tailed test when all the data were used (Table 2).

Eleven of the outbreaks only involved two cases (Table 2). In three of these outbreaks (outbreaks 1, 4, 5), the mean and median Dice coefficients were not statistically significant, but in two of these outbreaks (outbreaks 1 and 4) the isolates were given the same pattern number (Tables 2 and 3). A visual review of these digitized PFGE profiles suggested that differences in the placement of digital markers between

Table 1. *The percentage of outbreaks where the statistical measure was statistically significant* ($P \leqslant 0{\cdot}05$) *in a one-tailed or two-tailed test using all* (*A*) *or only sporadic* (*S*) *data*

Measure	No. of outbreaks	% Significant (one-tailed)	% Significant (two-tailed)
Mean (A)	20	70·0	65·0
Mean (S)	20	70·0	65·0
Median (A)	20	75·0	65·0
Median (S)	20	75·0	65·0
75th (A)	9	77·8	77·8
75th (S)	9	77·8	77·8
25th (A)	9	55·6	55·6
25th (S)	9	55·6	55·6
S.D. (A)	9	11·1	11·1
S.D. (S)	9	11·1	11·1
CV (A)*	9	22·2	0
CV (S)	9	22·2	0
IQR (A)	9	0	0
IQR (S)	9	0	0

Measures of location included the mean, median, 75th percentile (75th), 25th percentile (25th). Measures of variation included the standard deviation (S.D.), coefficient of variation (CV), and the interquartile range (IQR). Measures of variation were not evaluated for outbreaks involving two cases since there was only one pair for comparison.

* 11·1% of outbreaks if the one-tailed hypothesis had been in the opposite direction (i.e. greater than expected by chance alone).

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press

Table 2. *A summary of information for each outbreak*

Outbreak no.	Type	Cases	Isolates	Pairs	Patterns	Mean	Median	Percentile 25th	Percentile 75th	s.d.	CV	IQR
1	Comm	2	2	1	1	89·8	89·8	n.a.	n.a.	n.a.	n.a.	n.a.
2	Comm	2	2	1	1	94·1†	94·1†	n.a.	n.a.	n.a.	n.a.	n.a.
3	Comm	2	2	1	1	100*	100*	n.a.	n.a.	n.a.	n.a.	n.a.
4	House	2	2	1	1	89·4	89·4	n.a.	n.a.	n.a.	n.a.	n.a.
5	House	2	2	1	2	92·0	92·0	n.a.	n.a.	n.a.	n.a.	n.a.
6	House	2	2	1	2	97·9*	97·9*	n.a.	n.a.	n.a.	n.a.	n.a.
7	House	2	2	1	1	98·0*	98·0*	n.a.	n.a.	n.a.	n.a.	n.a.
8	House	2	2	1	1	100*	100*	n.a.	n.a.	n.a.	n.a.	n.a.
9	House	2	2	1	1	100*	100*	n.a.	n.a.	n.a.	n.a.	n.a.
10	House	2	2	1	1	100*	100*	n.a.	n.a.	n.a.	n.a.	n.a.
11	House	2	2	1	1	100*	100*	n.a.	n.a.	n.a.	n.a.	n.a.
12	Comm	3	3	3	2	78·3	69·4	96·0	69·4	15·4	19·6	26·6
13	House	3	3	3	1	95·8*	93·6*	100*	93·6*	3·7	3·9	6·4
14	House	3	3	3	1	92·8*	91·3†	95·7	91·3*	2·5	2·7	4·3
15	Comm	6	5	10	1	95·4*	95·8*	100*	91·7*	4·6	4·8†	8·3
16	Comm	6	5	10	3	94·8*	95·7*	100*	91·3*	4·5	4·7†	8·7
17	Stat	8	7	21	4	80·5	81·6	96·0*	69·4	14·5	18·0	26·6
18	Comm	9	8	28	2	89·9*	100*	100*	79·4	16·2	18·1	20·6
19	Comm	10	9	36	3	92·6*	93·6*	100*	87·5*	6·8	7·3	12·5
20	Stat	10	9	36	5	79·1	87·5‡	100*	80·0	26·3§	33·2	20·0

Community (Comm), household (House) or statistical (Stat) outbreak type (Type); number of cases (Cases); number of isolates where pulsed-field gel electrophoresis (PFGE) results were available (Isolates); the number of unique pairwise comparisons for Dice coefficients (Pairs); the number of different PFGE patterns (Patterns); and the mean, median, 75th percentile, 25th percentile, standard deviation (s.d.), coefficient of variation (CV), and interquartile range (IQR) of the Dice coefficients of the unique pairwise comparisons among isolates. Unless indicated, the results were not different if all the data or only sporadic data were used for the randomization distributions.

* Significant two-tailed test ($P \leqslant 0{\cdot}05$).

† Significant one-tailed test ($P \leqslant 0{\cdot}05$).

‡ Two-tailed significance if all data used but one-tailed significance if only sporadic data used to create randomization distributions.

§ Significant two-tailed test, but not in the correct direction to reject the one-tailed null-hypothesis; n.a., not applicable.

isolates with the same pattern, rather than the position of the bands on the gel, explained the relatively low Dice coefficients between identical patterns. Among epidemiologically identified outbreaks with more than three isolates, the mean, median, and 75th percentile were greater than expected by chance alone (Table 2). The isolates from the two statistically significant space–time clusters (outbreaks 17 and 20) had statistically significant higher 75th percentile Dice coefficients, and one of these clusters (outbreak 20) had a higher median Dice coefficient than expected by chance (Table 2).

In terms of variation in Dice coefficients within outbreaks, two outbreaks (outbreaks 15 and 16) with five isolates had Dice coefficients with statistically significant low coefficients of variation (Table 2). One of these outbreaks (outbreak 15) had one PFGE pattern while the other outbreak (outbreak 16) had three different patterns (Table 3). Only a statistically significant space–time cluster with nine isolates and five different patterns (outbreak 20) had Dice coefficients with a statistically significant standard deviation (Table 2). However, the standard deviation for this space–time cluster was not significant for our one-tailed null-hypothesis since the level of variation was greater than expected by chance.

DISCUSSION

Using 18 known outbreaks and two space–time clusters that were believed to capture outbreak cases from a previous study [24], we found that our randomization tests worked well at identifying isolates that were

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press

Table 3. *The national pulsed-field gel-electrophoresis (PFGE) pattern number of each isolate from each outbreak*

Outbreak no.	PFGE patterns
1	0·0765 (2X)
2	0·0665 (2X)
3	0·0508 (2X)
4	0·0483 (2X)
5	0·0013, 0·0703
6	0·0508, 0·0631
7	0·0391 (2X)
8	0·0001 (2X)
9	0·0660 (2X)
10	0·0532 (2X)
11	0·0683 (2X)
12	0·0665 (2X), 0·0664
13	0·0001 (3X)
14	0·0508 (3X)
15	0·0765 (5X)
16	0·0508 (3X), 0·0646, 0·0516
17	0·0508, 0·0355, 0·0720, 0·0722 (4X)
18	0·0368 (7X), 0·0761
19	0·0001 (7X), 0·0657, 0·0684
20	0·0001 (5X), 0·0657, 0·0654, 0·0670, 0·0684

(X) indicates the number of isolates with the same PFGE pattern number.

more similar in their PFGE patterns than expected by chance alone. In particular, the mean, median, and 75th percentile of Dice coefficients within a group of isolates appear to be most useful in identifying these close relationships rather than measures of variation (e.g. standard deviation). In the case of the statistical space–time clusters, the 75th percentile appeared to be the most useful measure for validating these statistical outbreaks. A statistical approach to outbreak identification based on a spatial scan statistic may capture a few cases that are unrelated to the outbreak. As a result, these additional cases may decrease the overall mean or median Dice coefficient, but by capturing a high number of outbreak cases, the upper quartile of pairwise comparisons continues to have high Dice coefficients. This is exemplified by comparing outbreaks 19 and 20. The space–time cluster (outbreak 20) was epidemiologically validated by sharing most of its cases with the epidemiologically identified outbreak 19 [24]. The addition of a few isolates that were not considered to be related to a daycare outbreak [8] led to a much smaller mean Dice coefficient in the statistical outbreak. Overall, the randomization tests worked well even when outbreaks were associated with multiple patterns, and the impact of using all or only sporadic cases for creating randomization distributions did not have a large impact on how we interpreted our results. However, it is important to note that all the outbreaks observed in that year involved a small number of cases. Large outbreaks may yield greater differences between randomization distributions. We predict that the inclusion of a large number of highly similar PFGE patterns in the sampled population would result in a higher degree of similarity being required for statistical significance.

Unlike cluster analyses that are typically used for studying relationships among isolates, our technique was specifically designed to test hypotheses concerning specific clusters of isolates. Rather than make qualitative or semi-quantitative judgements regarding the relative proximity of isolates on a phylogenetic tree, our method directly tests hypotheses concerning the cluster of interest with respect to the typical variation seen in the molecular surveillance system. Cluster analyses are generally used to imply evolutionary/genetic relationships, but molecular techniques such as PFGE violate basic assumptions regarding the independence of character traits. Consequently, the overall relationship of isolates among the branches of a phylogenetic tree created using PFGE should only be interpreted from a phenotypic perspective [22]. Likewise, applying our randomization tests to PFGE data should not be used to imply genotypic relationships. Tenover [32] has proposed some useful criteria for assessing the possibility that patterns are related, but these do not provide any statistical evidence, do not account for the number of isolates being compared (Figs 1 and 2), nor are they based on the typical variation seen within a surveillance system. We anticipate that our randomization tests would be most useful in confirming the phenotypic similarity of isolates from cases that are suspected to be linked based on epidemiological or statistical methods, but do not all share the same PFGE pattern. For instance, a community outbreak (outbreak 16) had multiple patterns that were highly similar, but different enough to justify unique pattern designations (Tables 2 and 3).

Recently, statistical approaches for identifying outbreaks, based on the proximity of cases in space, time, or space–time, have been tested or implemented for infectious disease surveillance [24, 33–35]. Molecular data have been used qualitatively to determine the epidemiological validity of some of these

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press

statistical outbreaks [24]. Complementing these statistical approaches to outbreak identification with our randomization tests would allow for the automation of both the search and validation procedures for outbreak identification. However, we caution that these statistical approaches may be more conservative in their overall performance than the abilities of public health workers in the field. Pearl *et al*. [24] discussed issues concerning limits to spatial resolution in some databases as well as the possibility of misclassification error when using household addresses as a proxy for the location of infection. Epidemiologists on the other hand can integrate various types of spatial information to find links among potential outbreak cases. Analogously, microbiologists would be aware of particular PFGE patterns that are new, common, or rare within their surveillance system and make appropriate decisions based on this information. The randomization tests we described are currently limited to the use of Dice coefficients and do not incorporate information concerning the prevalence of specific patterns. We suspect that the use of these statistical approaches for outbreak identification and molecular validation will be most useful for outbreaks that are more spatially and/or temporally diffuse, involve a number of isolates with different patterns, and as a consequence are more easily overlooked by public health workers. A statistically significant space–time cluster (outbreak 17) exemplifies this situation since it consisted of cases from three communities, the link among cases was only suspected as a result of a space–time scan of reported cases, and the randomization test applied to the PFGE data from this space–time cluster (Table 2) provided statistical validation that these isolates were more closely related than expected by chance alone [24].

The matrix of Dice coefficients used for these randomization tests can be impacted by a variety of factors including: the choice of settings for automated comparisons among the PFGE patterns; within and between worker variation in the placement of digital markers for the weight of particular bands; and variation in the quality of gels. While experienced laboratory workers are largely able to correct for this 'noise' when performing pattern recognition, automated procedures are more limited. Consequently, we found instances where identical patterns were given mean Dice coefficients below a level that would be statistically significant. However, it is important to emphasize that these limitations are not the result of the randomization tests themselves, but the techniques involved in automating pattern recognition. Randomization procedures can be easily adapted to quantify more objective or binary typing systems such as sequencing or multilocus variable number tandem repeat analysis. Ultimately, the benefit of this technique is that it can address hypotheses concerning the relative similarity of isolates with respect to the amount of genetic and/or phenotypic variation seen in a population of organisms. Claiming that two organisms are the same due to a certain percentage of homology based on any technique is arbitrary unless it is considered in the context of the typical amount of variation seen within the population. Defining the population of interest will inherently depend on the hypotheses being addressed, and the adoption of fixed 'rules of thumb' for significant levels of homology should be avoided.

ACKNOWLEDGEMENTS

We thank Duane Leedell, Qin Jiang, and Gene Chan for their laboratory and electronic cataloguing of PFGE patterns for the Alberta Provincial Laboratory for Public Health (Microbiology). The primary author has been supported by fellowships and awards from the Canadian Institutes of Health Research and the Ontario Veterinary College. We also acknowledge the support of the Wellcome Trust through the International Partnership Research Award in Veterinary Epidemiology.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Swaminathan B, *et al.*** PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases* 2001; **7**: 382–389.
2. **Pebody RG, *et al.*** An international outbreak of Vero cytotoxin-producing *Escherichia coli* O157 infection amongst tourists; a challenge for the European infectious disease surveillance network. *Epidemiology and Infection* 1999; **123**: 217–223.
3. **Waters JR, Sharp JC, Dev VJ.** Infection caused by *Escherichia coli* O157:H7 in Alberta, Canada, and in Scotland: a five-year review, 1987–1991. *Clinical Infectious Diseases* 1994; **19**: 834–843.

4. **Terajima J, *et al.*** Molecular epidemiological investigation of enterohaemorrhagic *Escherichia coli* isolates in Japan. *Symposium Series* (*Society for Applied Microbiology*) 2000; **29**: 99S–105S.
5. **Ochoa TJ, Cleary TG.** Epidemiology and spectrum of disease of *Escherichia coli* O157. *Current Opinion in Infectious Diseases* 2003; **16**: 259–263.
6. **MacDonald DM, *et al.*** *Escherichia coli* O157:H7 outbreak linked to salami, British Columbia, Canada, 1999. *Epidemiology and Infection* 2004; **132**: 283–289.
7. **Hrudey SE, *et al.*** A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world. *Water Science and Technology* 2003; **47**: 7–14.
8. **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. *Canada Communicable Disease Report* 2003; **29**: 21–28.
9. **Pritchard GC, *et al.*** Verocytotoxin-producing *Escherichia coli* O157 on a farm open to the public: outbreak investigation and longitudinal bacteriological study. *Veterinary Record* 2000; **147**: 259–264.
10. **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.
11. **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.
12. **Struelens MJ, De Gheldre Y, Deplano A.** Comparative and library epidemiological typing systems: outbreak investigations versus surveillance systems. *Infection Control and Hospital Epidemiology* 1998; **19**: 565–569.
13. **Heir E, *et al.*** Genomic fingerprinting of shigatoxin-producing *Escherichia coli* (STEC) strains: comparison of pulsed-field gel electrophoresis (PFGE) and fluorescent amplified-fragment-length polymorphism (FAFLP). *Epidemiology and Infection* 2000; **125**: 537–548.
14. **Allison L, Stirrat A, Thomson-Carter FM.** Genetic heterogeneity of *Escherichia coli* O157:H7 in Scotland and its utility. *European Journal of Clinical Microbiology and Infectious Diseases* 1998; **17**: 844–848.
15. **Preston MA, *et al.*** Epidemiologic subtyping of *Escherichia coli* serogroup O157 strains isolated in Ontario by phage typing and pulsed-field gel electrophoresis. *Journal of Clinical Microbiology* 2000; **38**: 2366–2368.
16. **Izumiya H, *et al.*** Molecular typing of enterohemorrhagic *Escherichia coli* O157:H7 isolates in Japan by using pulsed-field gel electrophoresis. *Journal of Clinical Microbiology* 1997; **35**: 1675–1680.
17. **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.
18. **Noller AC, *et al.*** Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates. *Journal of Clinical Microbiology* 2003; **41**: 5389–5397.
19. **Davis MA, *et al.*** Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of *Escherichia coli* O157:H7. *Journal of Clinical Microbiology* 2003; **41**: 1843–1849.
20. **Gerner-Smidt P, *et al.*** Computerized analysis of restriction fragment length polymorphism patterns: comparative evaluation of two commercial software packages. *Journal of Clinical Microbiology* 1998; **36**: 1318–1323.
21. **Zadoks R, *et al.*** Application of pulsed-field gel electrophoresis and binary typing as tools in veterinary clinical microbiology and molecular epidemiologic analysis of bovine and human *Staphylococcus aureus* isolates. *Journal of Clinical Microbiology* 2000; **38**: 1931–1939.
22. **Swofford DL, *et al.*** Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, eds. *Molecular Systematics*, 2nd edn. Sunderland, MA: Sinauer Associates, 1996, pp. 407–514.
23. **Gouveia S, *et al.*** Genomic comparisons and Shiga toxin production among *Escherichia coli* O157:H7 isolates from a day care center outbreak and sporadic cases in southeastern Wisconsin. *Journal of Clinical Microbiology* 1998; **36**: 727–733.
24. **Pearl DL, *et al.*** 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. *Epidemiology and Infection* (in press).
25. **Struelens MJ, De Ryck R, Deplano A.** Analysis of microbial genomic macrorestriction patterns by pulsed-field gel electrophoresis (PFGE) typing. In: Dijkshoorn L, Towner KJ, Struelens M, eds. *New Approaches for the Generation and Analysis of Microbial Typing Data*. New York: Elsevier, 2001, pp. 159–176.
26. **Dwass M.** Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 1957; **28**: 181–187.
27. **Manly BFJ.** *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. New York: Chapman and Hall, 1997, pp. 399.
28. **Kulldorff M.** A spatial scan statistic. *Communications in Statistics: Theory and Methods* 1997; **26**: 1481–1496.
29. **Kulldorff M, Hjalmars U.** The Knox method and other tests for space-time interaction. *Biometrics* 1999; **55**: 544–552.
30. **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.
31. **Kulldorff M, Information Management Services Inc.** SaTScan v. 3.0: software for the spatial and space-time scan statistics, 2002.
32. **Tenover FC, *et al.*** Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel

electrophoresis: criteria for bacterial strain typing. *Journal of Clinical Microbiology* 1995; **33**: 2233–2239.

33. **Kulldorff M, *et al.*** A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2005; **2**: 216–224.
34. **Mostashari F, *et al.*** Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases* 2003; **9**: 641–646.
35. **Michel P, Ellis A, Middleton D.** 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.

https://doi.org/10.1017/S0950268806006650 Published online by Cambridge University Press