Journal of Biomedical Informatics 42 (2009) 66–73

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Biosurveillance of emerging biothreats using scalable genotype clustering

Blanca Gallego [a,*], Vitali Sintchenko [a,b,c], Qinning Wang [b], Lester Hiley [d], Gwendolyn L. Gilbert [b,c], Enrico Coiera [a]

[a] Centre for Health Informatics, University of New South Wales, Coogee Campus, Sydney, NSW 2052, Australia
[b] Centre for Infectious Diseases and Microbiology, Institute of Clinical Pathology and Medical Research, Sydney West Area Health Service, Westmead, NSW 2145, Australia
[c] Western Clinical School, The University of Sydney, Sydney, NSW 2145, Australia
[d] Queensland Health Forensic & Scientific Services, Brisbane, Qld 4001, Australia

ARTICLE INFO

Article history:
Received 24 December 2007
Available online 29 July 2008

Keywords:
Biosurveillance
Molecular genotyping
Salmonellosis
Infectious disease clusters

ABSTRACT

Developments in molecular fingerprinting of pathogens with epidemic potential have offered new opportunities for improving detection and monitoring of biothreats. However, the lack of scalable definitions for infectious disease clustering presents a barrier for effective use and evaluation of new data types for early warning systems. A novel working definition of an outbreak based on temporal and spatial clustering of molecular genotypes is introduced in this paper. It provides an unambiguous way of clustering of causative pathogens and is adjustable to local disease prevalence and availability of public health resources. The performance of this definition in prospective surveillance is assessed in the context of community outbreaks of food-borne salmonellosis. Molecular fingerprinting augmented with the scalable clustering allows the detection of more than 50% of the potential outbreaks before they reach the midpoint of the cluster duration. Clustering in time by imposing restrictions on intervals between collection dates results in a smaller number of outbreaks but does not significantly affect the timeliness of detection. Clustering in space and time by imposing restrictions on the spatial and temporal distance between cases results in a further reduction in the number of outbreaks and decreases the overall efficiency of prospective detection. Innovative bacterial genotyping technologies can enhance early warning systems for public health by aiding the detection of moderate and small epidemics.

1. Introduction

Prospective infectious disease surveillance requires the ongoing collection and monitoring of infection-specific data and related information such as infectious disease counts or syndromic data. The goal of surveillance is to detect and then prevent and control outbreaks in real-time. For some infectious diseases, one or two confirmed cases are sufficient to raise an alarm (e.g. SARS, meningococcal disease). However, for many types of infections, detection requires clustering of the data based on similarity of isolates. A broad range of statistical techniques have been applied in order to improve the performance of prospective surveillance and have been extensively reviewed elsewhere [1–4]. In its simplest form, a statistical surveillance method consists of a process control algorithm for a single time-dependent variable. More complex methods involve the analysis of multi-variate spatio-temporal data sets. These early warning systems can identify large disease epidemics but there are usually significant delays and low sensitivity in detecting moderate and small outbreaks. This is due to the high level of noise in laboratory and syndromic surveillance data [4]. Better surveillance often allows the size of outbreaks to be limited as a consequence of public health interventions, as more outbreaks are detected and controlled at an earlier stage and fewer continue to a large size [5].

The molecular fingerprinting of pathogens with epidemic potential offers new opportunities for detecting and confirming clusters of community and hospital-acquired infections [6–8]. It involves rapid subtyping of isolates from infected patients for the purpose of strain discrimination. Although the discriminatory power varies according to the subtyping method, molecular genotyping is often useful to identify sources and routes of transmission [9]. However, identifying patients that share the same genotype is not enough to uniquely provide an operational definition for an outbreak. In practice, the decision to proceed with a public health intervention will depend on the severity, communicability and local epidemiology of the disease as well as on the availability of public health resources to conduct investigations and institute corrective measures [2,10]. It is therefore critical to have an outbreak definition (in the absence of epidemiological information) that optimizes the limited resources of public health practitioners while preventing further spread [11].

* Corresponding author. Fax: +61 2 9385 9006.
E-mail address: b.gallego@unsw.edu.au (B. Gallego).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.07.002

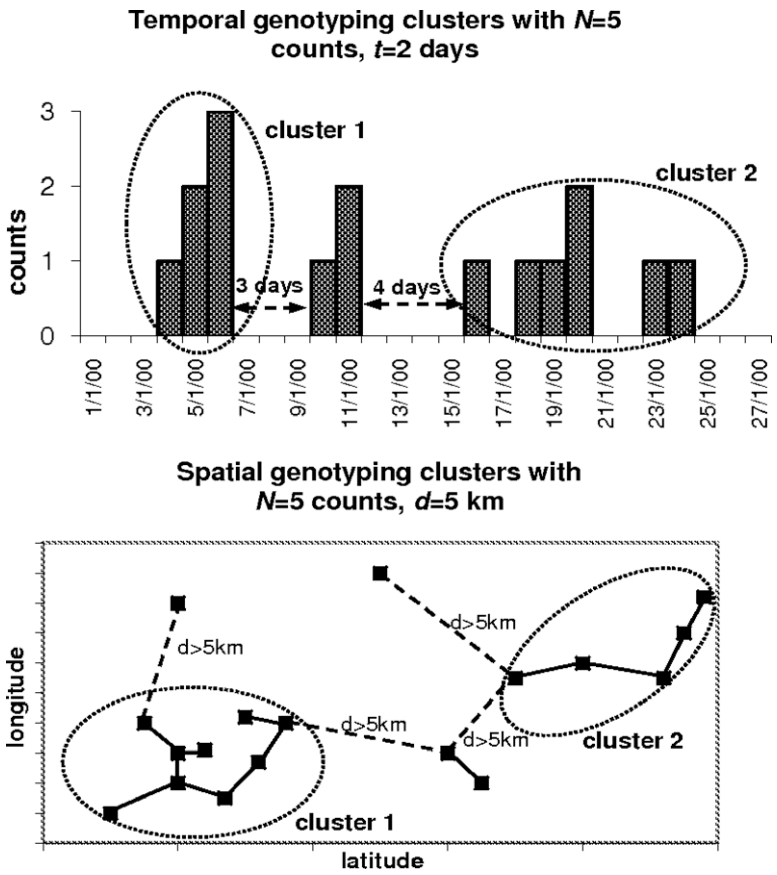

Fig. 1. Sketch depicting examples of temporal (top panel) and spatial (bottom panel) genotyping clusters. The top panel shows two temporal clusters defined as a maximal set of at least five counts with consecutive cases occurring at most 2 days from each other. The bottom panel shows two spatial clusters defined as a maximal set of at least five counts forming a spanning tree with edges at most 5 km long.

One setting that allows in-depth study of the impact of cluster definitions on prospective monitoring of bacterial genotypes is surveillance of *Salmonella enterica* serovar Typhimurium (STM) infections [12,13]. Rapid genotyping of STM has recently been widely used to characterize salmonella outbreaks. In particular, multilocus variable-number tandem repeat analysis (MLVA) of STM is a stable, easily implemented method and its results can be shared between laboratories over the Internet [14,15]. However, evidence about performance and timeliness of STM cluster detection systems remains limited [16].

To address these generic deficiencies, we introduce a working outbreak definition based upon temporal and spatial clustering of genotypes that provides unambiguous clustering of isolates and that can be tuned to accommodate the requirements and resources available for outbreak investigations. We compare this definition against statistical and epidemiologically confirmed clusters and evaluate its performance in prospective surveillance.

2. Methods and data

2.1. Genotype cluster definitions

We define a *genotyping cluster* as a *maximal set* of at least N isolates that share the same (or closely related) genotype, among a set of isolates from infected patients, each with an associated date and location (e.g. collection date and patient's address). To account for clustering in space and time, we specify:

Temporal cluster: A genotyping cluster, for which the time difference between any two consecutive cases is at most t days (see top panel in Fig. 1). The limit of $t = 0$ corresponds to clusters that last one day.

Spatial cluster: A genotyping cluster, for which locations of all cases are connectable by a spanning tree (a graph connecting a set of nodes [i.e. case locations] without any cycles) with all edges no more than d kilometers long (see bottom panel in Fig. 1). The limit of $d = 0$ indicates a cluster occurring in one location.

Spatio-temporal cluster: A combined temporal and spatial cluster characterized by parameters t and d.

These spatial and temporal cluster definitions satisfy two important properties. First, they provide a unique way of clustering cases that is independent of the order in which the isolates are considered. This property guarantees that any two cases assigned to a cluster at a given time will remain in one cluster in the presence of additional cases. This makes it possible to search, retrospectively, for clusters (for given parameters N, t and d) in historical data, compute the number of clusters and determine how early they would have been detected, prospectively. In this way, one can adjust future values of N, t and d according to prospective surveillance needs and availability of public health resources. For simplicity we have assumed that the parameters N, t and d are independent of genotype. Second, except for the limits $t = 0$ and $d = 0$, the duration and area of a cluster is not prescribed, making definitions scalable. A more naive outbreak definition as a set of at least N isolates of a given genotype occurring within a fixed duration and/or fixed area does not fulfill these properties. Furthermore, definitions with fixed duration are obviously not appropriate for prospective surveillance.

An algorithm that implements the working definitions of outbreaks described in this paper has three steps:

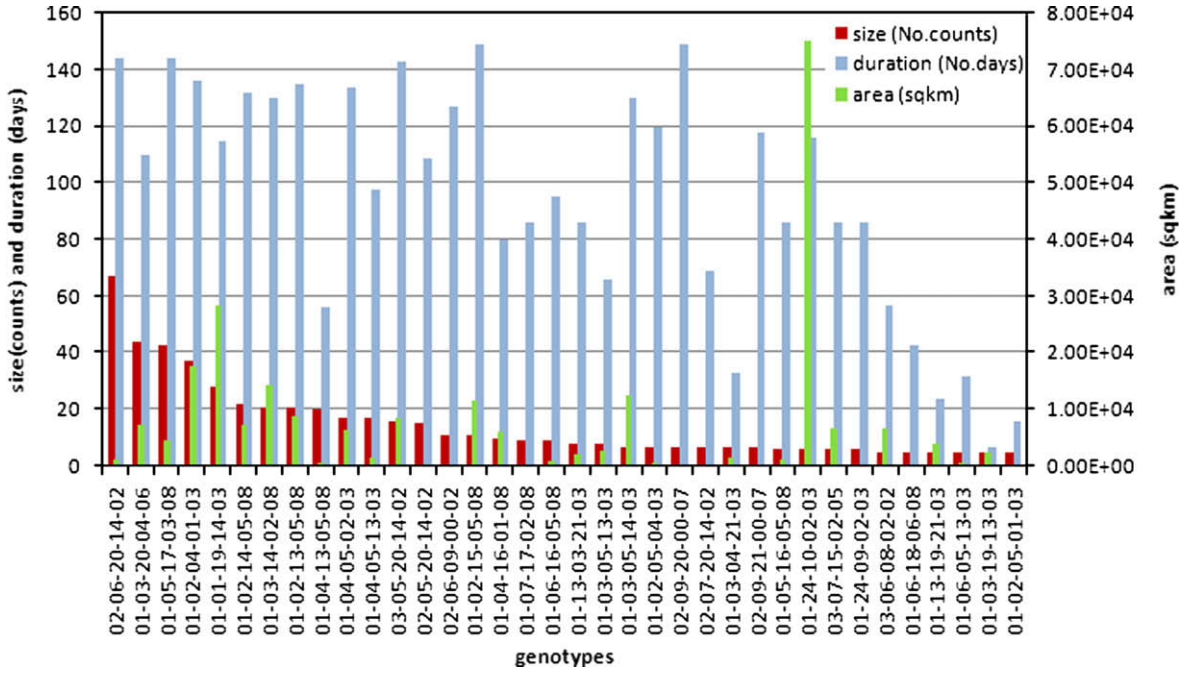

Fig. 2. Size, duration and surface area of genotyping clusters with $N = 5$. Clusters are ordered by size.

I. Compute temporal and/or spatial distance of each new isolate with existing same-genotype isolates.

II.

(a) If an existing isolate is found for which temporal and/or spatial distance is smaller or equal to t and/or d then the new isolate joins the set of this existing isolate.

(b) If more than one isolate is found for which temporal and/or spatial distance is smaller or equal to t and/or d and they belong to different sets then the sets merge into one.

(c) When no isolates have been found for which temporal and/or spatial distance is smaller or equal to t and/or d then the new isolate forms a new set.

III. A set becomes a cluster the moment it reaches N or more isolates.

The clustering algorithm is easy to implement and requires short computational times, since its order does not depend on the number of locations or dates under surveillance. Also, its accuracy is independent of the number of cases under consideration.

2.1.1. Prospective surveillance

The performance of these outbreak definitions in a prospective surveillance system can be tested by computing how long it would take to detect each cluster in real-time using a given outbreak definition. Following the algorithm described above, the detection date of an outbreak is simply the date at which a set of same-genotype isolates that fulfil the appropriate spatio-temporal restrictions reaches N or more isolates and becomes a cluster (step III).

2.2. Spatio-temporal scan statistic

An alternative and more complex way to account for time and space correlations within a set of disease counts (and potentially within a genotyping cluster) is to use a statistical method, such as the space–time permutation scan statistic [17]. In this method, a cluster is defined as the region in space and time where the probability of an incident case occurring is higher inside than outside. Expected values are estimated from existing counts (for a given day and location the expected counts are proportional to all the cases that occurred in that location multiplied by all the cases that occurred in that day), and the underlying probability function is the hypergeometric distribution. A cluster is considered to be statistically significant when its p-value[1] is smaller than a certain threshold. Many statistical methods have been suggested in the literature for the detection of disease clusters. Among the methods that consider spatio-temporal clustering, we have chosen scan statistics as our "representative" statistical method because: (a) it is one of the most popular; (b) it can be fitted into a general cumulative sum framework [18]; and (c) it has been implemented in a publicly available software package used in many surveillance publications before (see http://www.satscan.org/references.html).

2.3. MLVA genotyping

Molecular fingerprint of a given pathogen is a set of marker scores displayed by an isolate obtained from a patient which is used for assessment of epidemiological relatedness among bacterial isolates. Different techniques have been applied to obtained fingerprints of different pathogens [19].

The most common methods used for the subtyping of *S. enterica* Typhimurium (STM) are phage typing (PT), pulse-field gel electrophoresis (PFGE) and more recently multiple-locus variable-number tandem-repeats analysis (MLVA) [12,20]. MLVA is based on the detection of short sequence repeats that vary in copy number in the microbial genome at various loci. MLVA detects polymorphisms at five different sites in the genome. Four regions of detection are on the bacterial chromosome and one is located on the serotype specific plasmid *pSLT*. MLVA has high discriminatory power within clonal species and appears to be more rapid and more amenable to standardization than pulse-field gel electrophoresis for both surveillance and outbreak investigations of STM [13].

[1] $p-\text{value} = \frac{R_{up}+1}{R+1}$, R = number of randomly generated replicas and R_{up} = number of randomly generated replicas with higher maximum likelihood ratio than potential cluster.

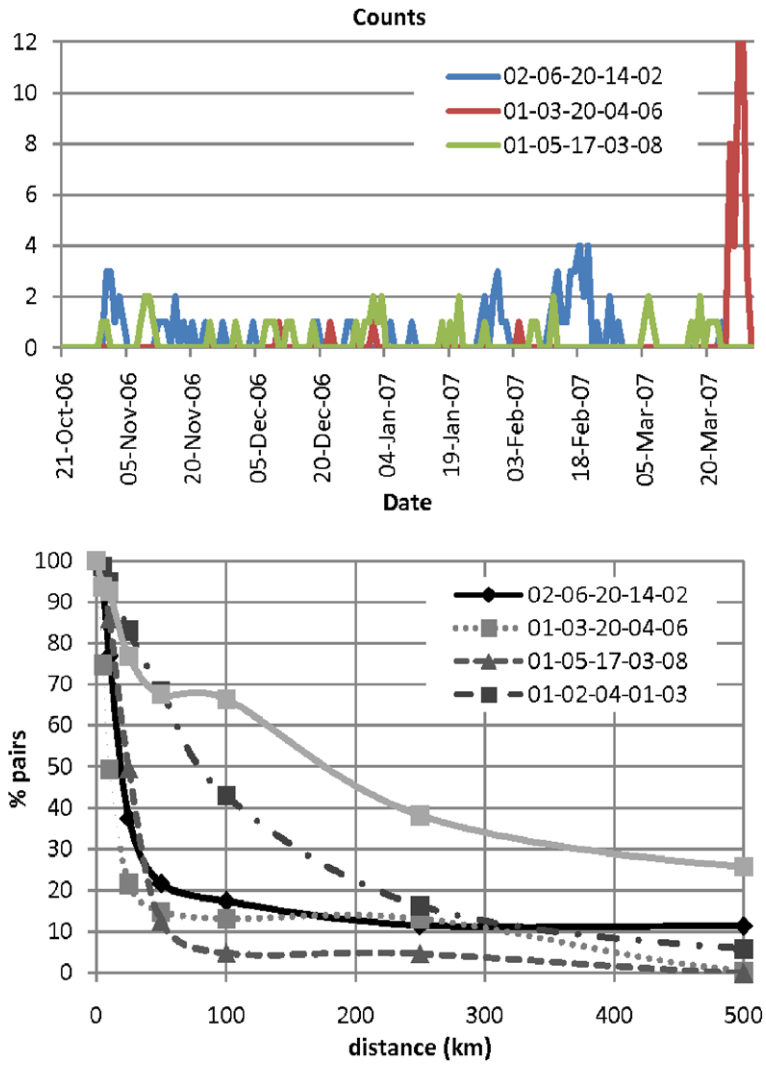

Fig. 3. Temporal and spatial characteristics of largest STM genotyping clusters. Top panel: time series of the 3 largest genotyping clusters. Bottom panel: spatial range of the 5 largest genotyping clusters measured as percentage of pairs of counts, in which patients are found at different minimum distances from each other. A point (x,y) in this graph records the percentage y of pairs of same MLVA-type isolates in which patients are found at a distance of x km or more from each other (for the whole period of sampling).

Epidemiological concordance for molecular multilocus typing patterns has been demonstrated [21].

2.3.1. Dataset

In Australia, all salmonella isolates obtained from patients with gastrointestinal illness are serotyped by state reference laboratories. The Centre for Infectious Diseases and Microbiology Laboratory Services at the Institute of Clinical Pathology and Medical Research (ICPMR), Sydney West Area Health Service and the Queensland Health Forensic & Scientific Services (QHFSS) are state reference facilities for enteric pathogens for New South Wales and Queensland, respectively. Since 2006, all confirmed STM isolates which were characterized in these two laboratories have been further subtyped using MLVA as part of state surveillance. Phage typing (PT) was provided by the Microbiological Diagnostic Unit at the University of Melbourne.

The dataset used in this paper consists of STM isolates from humans referred to the ICPMR and QHFSS between 23-October-2006 and 31-March-2007. Each isolate has an associated MLVA type, PT type, specimen collection date and postcode of patient's address. ICPMR and QHFSS at the time of testing were using different conventions for translating the number of repeats in STM MLVA loci into a MLVA genotype number.

3. Results

3.1. Salmonella Typhimurium clusters

Our genotyping cluster definitions were applied to MLVA genotyping data for STM isolates from humans referred to the two state reference laboratories during the study period. There were 816 isolates, displaying 226 different MLVA profiles, the most common of which was found in 67 or 8.2% of isolates. Each isolate had an associated specimen collection date and postcode of patient's address. Distances between cases were defined as those between the geographical centers of the patient's postcode areas. For simplicity, only identical MLVA profiles were considered as part of the same cluster[2]. We refer to cluster size as the number of isolates or cases (counts) within the cluster, cluster duration as the number of days between (and including) the first and last collection dates in the cluster and, cluster area as the sum of areas of the patients' postcodes plus those of the enclosed postcodes.

Fig. 2 shows the sizes, durations, and areas of the 36 genotyping

[2] It is possible to relax the condition of identical MLVA types in the definition of genotyping cluster to account for genotypes that are genetically close and that may have undergone a mutation during an outbreak.

clusters each containing at least five isolates ($N = 5$). The average cluster size was 15 cases and the median 9. Their durations ranged from 7 (MLVA 01-03-19-13-03) to 149 days (MLVA 01-02-15-05-08 and 02-09-20-00-07), and averaged 96. Their mean area was 6839 km^2, representing 0.27% of the total area of New South Wales and Queensland. MLVA 02-09-21-00-07 cluster occupied the smallest area (0.003% of total) while MLVA 01-24-10-02-03 spread throughout approximately 3% of the whole area. Many different MLVA clusters shared the same phage type (9 and 8 clusters shared PT135a and PT170, respectively). The majority of MLVA profiles appear "endemically" throughout the 6-month study period, with occasional clusters lasting 3–40 days. Some MLVA types also have a wide geographical spread, with significant differences in temporal and spatial distributions among genotypes (see Fig. 3).

3.1.1. Temporal clusters

In view of the temporal structure of the genotyping clusters, it makes sense to describe STM outbreaks, for prospective surveillance, using our temporal cluster definition. For instance, if an outbreak is defined as a maximal set of at least 5 isolates with the same MLVA profile, each of which collected within up to 1 day of the next ($t = 1$), there were 12 outbreaks with a mean of 12 cases each, lasting an average of 4 days and occupying on average 514 km^2 (Table 1). Variation with t in the number, size, duration and area of temporal clusters, for which $N = 5$, is shown in the top panel of Fig. 4. For sensitivity analyses, the range $t = 0$–7, is most interesting; the number of clusters varies (as a proportion of 36, the clusters detected when t was unlimited) from 4 or ~10% when $t = 0$, to 29 or 80% when $t = 7$. The mean size of clusters does not vary significantly with t, although that of larger outbreaks is less with lower t values. Both the average duration and areas of clusters fall sharply with smaller t values.

For prospective surveillance, when information about epidemiological links between patients is lacking, the value of t can be chosen according to the number of expected future outbreaks. The more outbreaks we are willing to deal with (assuming, for example, that the first 5 patients whose isolates share the same MLVA profile will be investigated) the larger the t value for a given N. This tuning of our model parameters is the equivalent of the selection of alarm thresholds used in standard prospective surveillance systems (see e.g. Reis et al. [22]). In these statistical models, the threshold is typically adjusted to allow for given average false alarm rates. In our model, the parameters controlling the number of "alarms" have a direct relationship with the size, duration and area of the outbreaks. The values of N and t could also be adjusted to account for seasonal patterns, local prevalence or differences between genotypes, but that is beyond the scope of this paper.

3.1.2. Spatio-temporal clusters

Limiting the distance between cases in the functional definition of an outbreak may also be useful, particularly for genotypes with wide geographical distribution. This is done by applying the definition of a spatio-temporal cluster. Examples of STM spatio-temporal genotyping clusters with $N = 5$, $t = 1$ and $d = 5$ are shown in Table 1. The bottom panel of Fig. 4 shows changes in the number, size, duration and area of potential outbreaks, with variations in d for spatio-temporal genotyping clusters with $N = 5$ and $t = 2$. Similar results (not shown) are observed for other values of t. Like t, d can be varied according to the number of future investigations public health officers are able to perform.

3.1.3. Scan statistic clusters

The space–time permutation scan statistic method was applied to our salmonella data set retrospectively with default values for bonds on cluster size (50% of population at risk and 50% of study period). The spatial scanning window was circular and only non-geographically overlapping clusters with p-value < 0.05 were considered. Calculations were performed using the freely available software SatScan [23]. When clustering the entire data set without distinguishing between genotypes, 4 clusters were found. Two of them correspond to genotyping clusters 01-03-20-04-06 (29 isolates out of 30) and 01-04-13-05-08 (all 14 isolates). The other two clusters are 75% (3 isolates out of 4) associated with genotype 01-03-19-13-03 and 33% (12 out of 36 isolates) associated with genotype 02-06-20-14-02. This latter cluster included 17 different genotypes. The calculation was repeated using phage types as covariates. This found 3 clusters that roughly coincide with the first 3 clusters found without phage type information, plus another cluster containing 8 isolates, which included 4 different genotypes. Computations using MLVA types as covariates or as individual data sets did not result in statistically significant clusters.

Table 1
Example of STM temporal genotyping clusters with $N = 5$ and $t = 1$ (upper section) and STM spatio-temporal genotyping clusters with $N = 5$, $t = 1$ and $d = 5$ (lower section)

	Cluster	Genotype	Size (counts)	Duration (days)	Area (km2)	Phage types	Confirmed by scan statistic[a]	Confirmed by epi. investig[b]
Temporal clusters	1	01-03-20-04-06*	39	5	184.19	9, 12	Yes	Yes
	2	02-06-20-14-02**	24	10	735.91	197	Yes	Yes
	3	01-04-13-05-08*	16	3	626.14	170	Yes	Yes
	4	02-06-20-14-02**	10	5	165.74	197		
	5	01-02-04-01-03**	8	3	795.69	U302, 186, 35		
	6	02-05-20-14-02**	8	4	81.64	197		Yes
	7	02-06-20-14-02**	7	4	228.30	197		
	8	01-05-17-03-08*	7	5	86.33	135a	Yes	
	9	01-02-04-01-03**	6	3	183.31	UNK, 35, U302, 29, RDNC		
	10	01-05-17-03-08*	6	4	184.70	135a		
	11	01-01-19-14-03**	5	2	2,792.20	RDNC, 44		
	12	01-04-05-13-03**	5	3	104.31	135a, RDNC		Yes
Spatio-temporal clusters	1	01-03-20-04-06*	29	5	98.62	9	Yes	Yes
	2	01-04-13-05-08*	14	3	314.64	170	Yes	Yes
	3	02-06-20-14-02**	7	5	37.81	197	Yes	Yes

* Identified at ICPMR (NSW).
** Identified at QHFSS (Qld).
[a] The space–time permutation scan statistic method [17] as applied retrospectively without distinguishing between genotypes. In most cases, the scan statistic clusters included fractions of genotyping clusters together with isolated with isolates with different genotypes. Computations applied to sets of same-MLVA type isolates did not result in statistically significant clusters.
[b] A total of five epidemiological investigations took place during the period under considerations.

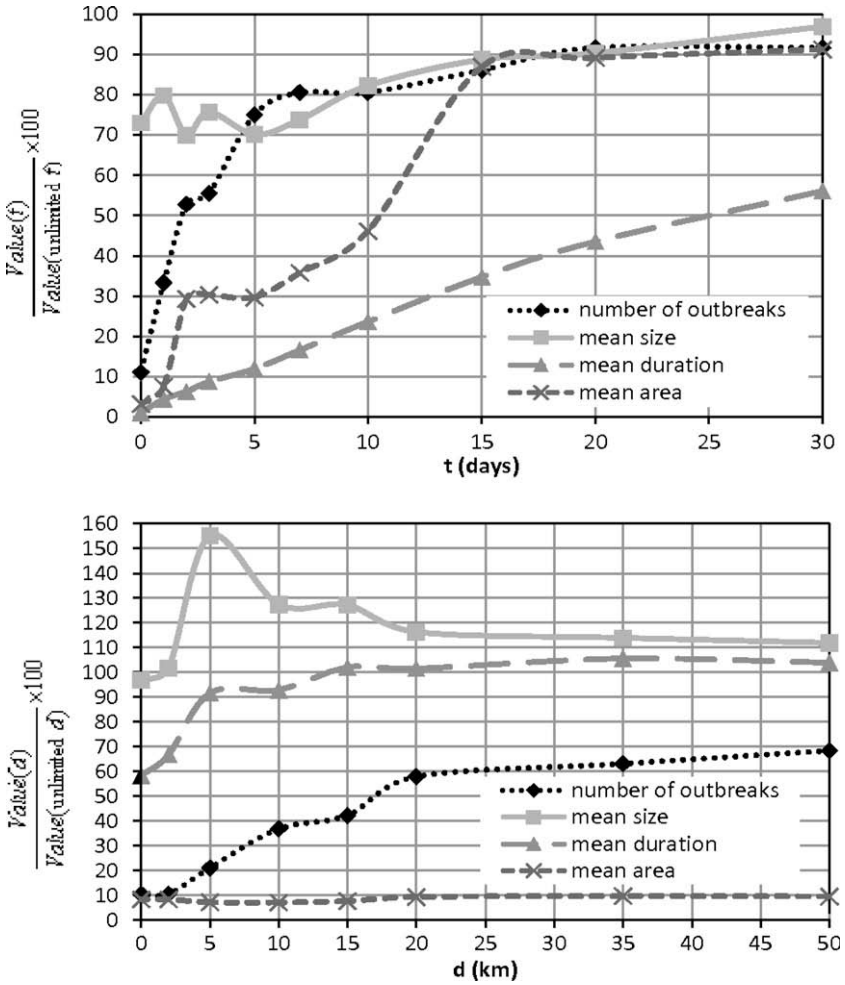

Fig. 4. Sensitivity of potential number of outbreaks, their mean size, mean duration, and mean surface area: with changes in t for STM temporal genotyping clusters with $N = 5$ (top panel) and with changes in d for STM spatio-temporal genotyping clusters with $N = 5$, $t = 2$ (bottom panel) .A 100% value corresponds to the case of unlimited t and d, respectively.

As it can be seen in Table 1, the space–time permutation scan statistics detects clusters that are similar to all of the genotyping spatio-temporal clusters with $N = 5$, $t = 1$, $d = 5$, but not all of the corresponding genotyping temporal-only clusters. This is to be expected since the space–time permutation scan statistic method will detect a cluster when there is a high proportion of excess cases in a given area with respect to the surrounding areas, during a specific period of time, while this is not necessarily the case for all of the temporal-only clusters.

3.2. Prospective surveillance of genotyping clusters

Next, we address the question of timeliness of detection by computing the proportion of genotyping clusters detected within a given time (expressed as a fraction of cluster duration) for different values of N, t and d. Timely outbreak detection was best when $N = 3$, 4 or 5, resulting in identification of 58, 44 or 36 clusters, respectively, of which more than 50% were detected before reaching the peak or midpoint of the cluster duration.

Imposing restrictions on intervals between collection dates, without restricting distances between cases, does not significantly affect the timeliness of detection, particularly for values of $N > 2$ (see upper row in Fig. 5). For instance, temporal clusters with $N = 5$ and $t = 1$ were detected, on average, 2 days after identification of the first case and had a mean duration of 4 days. Similarly, outbreaks defined using $N = 5$ and $t = 8$ were detected, on average, within less than 10 days of the first case and lasted, on average, 18 days. The differences in timeliness of detection among spatio-temporal clusters is slightly larger, particularly for $N = 2$. In general, clustering in both space and time decreased the overall efficiency of prospective detection (see lower row in Fig. 5). For example, the 3 spatio-temporal clusters with $N = 5$, $t = 1$ and $d = 5$ lasted 5, 3, and 5 days and were detected at days 4, 2, and 4, respectively.

4. Discussion

Like many other infectious diseases, salmonellosis occurs as a mixture of temporally and/or geographically clustered cases, superimposed on non-clustered endemic cases. This is also the case for each MLVA type separately, which raises the question of refining outbreak definition beyond genotyping groups. For example, Torpdahl et al. [13] defined a salmonella outbreak as at least five isolates with the same MLVA type found within a 4-week period. This definition can be ambiguous, since it does not provide a unique way of clustering cases and, furthermore, cannot be used in prospective surveillance. Other frequently used tools for identification of disease clusters rely on statistical methods that detect regions in space and/or time where disease counts are significantly higher than expected. Review and performance of such algorithms for outbreak detection in automated syndromic surveillance systems can be

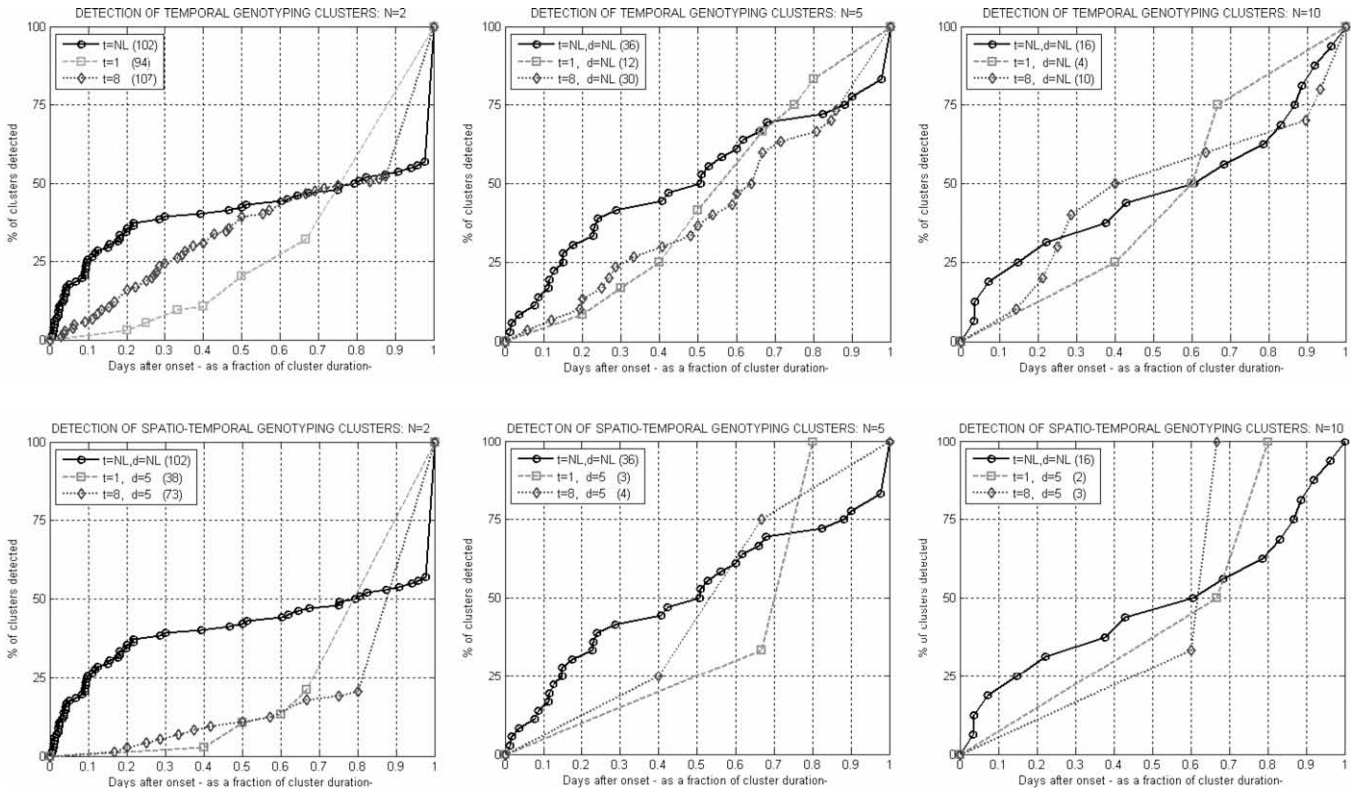

Fig. 5. Percentage of genotyping clusters as a function of their detection time (expressed as a fraction of cluster duration). Upper row: detection of temporal genotyping clusters with t = no limit, 1, 8 for N = 2(left), N = 5(centre), and N = 10(right); lower row: detection of spatio-temporal genotyping clusters with (t,d) = (no limit, no limit), $(t,d) = (1,5)$, $(t,d) = (8,5)$ for N = 2 (left), N = 5 (centre), and N = 10 (right); the total numbers of clusters for each definition appear in parenthesis within the legend.

found in [22,24]. One statistical method, which has been commonly (and often successfully) used in syndromic and population health surveillance, is scan statistic [25,26]. However, this methodology does not guarantee that any two cases assigned to a cluster at a given time will remain in the same cluster when new cases are added, since the latter may change the likelihood ratios, and it is much slower to run. More importantly, statistical methods may fail to identify clusters when applied to the highly specific sets of same-genotype isolates (as is the case with same-MLVA-type STM sets) due to low signal to noise ratios. In contrast, our temporal and spatio-temporal cluster definitions, based on molecular fingerprints of bacteria with epidemic potential, are unambiguous, scalable and generalizable. They provide a methodology by which the number of clusters identified can be varied according to resources available and are applicable to any new molecular technologies used in public health surveillance. In our reference laboratories, the average turn-around time for MLVA genotyping is between 3 and 7 days after identification of STM which enables more rapid public health investigations. Another important benefit of molecular fingerprinting of pathogens is that it can be easily standardized, allows the comparison of strains across many jurisdictions (states of Queensland and New South Wales, and potentially all Australian States and Territories, in our case) and subsequent clustering across regional and national borders and, finally, encourages international public health networks [5].

4.1. Limitations

There are limitations to surveillance capabilities of systems such as that described in this paper. Genotyped samples represent a small proportion of infectious cases in the population [27], and the date and the location associated with the specimen only approximate epidemiologically relevant parameters. Our definitions have been tested in the domain of food-borne bacterial infections and by using clusters identified by only one of many possible molecular genotyping methods. Nevertheless, these definitions are domain-independent and could be extended to other domains of prospective biosurveillance. Our observations are in line with previous experiences in molecular microbiology and the central hypothesis of public health, that a single outbreak is usually related to a single infecting strain [5]. Furthermore, turn-around time of amplification-based molecular fingerprinting techniques performing 'real-time' genotyping may now be feasible for reference laboratories. However when indistinguishable isolates are identified, appropriate public health actions must be taken. This implies that patients must be identifiable to public health officers to enable hypothesis-generating interviews; therefore, especially stringent criteria of the patient's privacy protection should be applied to such early warning systems.

4.2. Conclusions

We have used emerging genotyping techniques to introduce a working outbreak definition based on temporal and spatial clustering of genotypes. This definition provides unambiguous clustering of isolates that can be tuned to accommodate the requirements and resources available for outbreak investigations thus addressing a significant problem in current public health surveillance information needs. It allows timely recognition, source identification and capacity to apply public health action and will enable better early warning systems for new and established biothreats.

Acknowledgments

This work was supported by the Australian Research Council through the Linkage Grant LP0667531.

References

[1] Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. J R Stat Soc A 2003;166(1):5–21.
[2] Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. J Biomed Inform 2005;38:99–113.
[3] Kleinman K, Abrams A. Assessing surveillance using sensitivity, specificity and timeliness. Stat Methods Med Res 2006;15:445–64.
[4] Heisterkamp S, Deckers A, Heijne J. Automated detection of infectious disease outbreaks: hierarchical time series models. Stat Med 2006;25: 4179–96.
[5] Tauxe RV. Molecular subtyping and the transformation of public health. Foodborne Pathog. Dis. 2006;3:4–8.
[6] Sintchenko V, Iredell JR, Gilbert GL. Genomic profiling of pathogens for disease management and surveillance. Nat Microbiol Rev 2007;5:464–70.
[7] Gosselin P, Lebel G, Rivest S, Douville-Fradet M. The integrated system for public health monitoring of West Nile virus (ISPHM-WNV): a real time GIS for surveillance and decision-making. Int J Health Geog 2005:4–21.
[8] Deplano A, Denis O, Nonhoff C, Rost F, Byl B, Jacobs F, et al. Outbreak of hospital-acquired clonal complex-17 vancomycin-resistant *Enterococcus faecium* strain in a haematology unit: role of rapid typing for early control. J Antimicrob Chemother 2007;60:849–54.
[9] Gilbert GL. Molecular diagnostics in infectious diseases and public health microbiology: cottage industry to postgenomics. Trends Mol Med 2002;8(6):280–7.
[10] Buehler J, Hopkins R, Overhage J, Sosin D, Tong V. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks: Centre for Communicable Diseases. Report No. 53 (RR05). USA; 2004.
[11] Elliot P, Wakefield J. Disease clusters: should they be investigated, and if so, when and how? J R Stat Soc A 2001;184(Part 1):3–12.
[12] Lindstedt B, Vardund T, Aas L, Kapperud G. Multiple-locus variable-number tandem-repeats analysis of *Salmonella enterica* subsp. *enterica* serovar Typhimurium using PCR multiplexing and multicolor capillary electrophoresis. J Microbiol Methods 2004;59:163–72.
[13] Torpdahl M, Sorensen G, Lindstedt B, Moller Nielsen E. Tandem repeat analysis for surveillance of human *Salmonella* Typhimurium infections. Emerg Infect Dis 2007;13(3):388–95.
[14] Hopkins KL, Maguire C, Best E, Liebana E, Threlfall EJ. Stability of multiple-locus variable-number tandem repeats in *Salmonella enterica* Serovar Typhimurium. J Clin Microbiol 2007;45:3058–61.
[15] Lindstedt BA, Torpdahl M, Nielsen EM, Vardund T, Aas L, Kapperud G. Harmonization of the multiple-locus variable-number repeat tandem analysis method between Denmark and Norway for typing *Salmonella* Typhimurium isolates and closer examination of the VNTR loci. J Appl Microbiol 2007;102:728–35.
[16] Sintchenko V, Gallego B, Chung G, Coiera E. Towards bioinformatics assisted infectious disease control. In: Summit on Translational Bioinformatic, 10–12 March 2008, San Francisco, USA: American Medical Informatics Association; 2008. p. 105–9.
[17] Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space–time permutation scan statistic for disease outbreak detection. PLoS Med 2005;2(3):e59.
[18] Sonesson C. A CUSUM framework for detection of space–time disease clusters using scan statistics. Stat Med 2007;26(26):4770–89.
[19] Fournier P-E, Drancourt M, Raoult D. Bacterial genome sequencing and its use in infectious diseases. Lancet Infect Dis 2007;7:711–23.
[20] Wang Q, Kong F, Jelfs P, Gilbert GL. Extended phage locus typing of *Salmonella enterica* serovar Typhimurium, using multiplex PCR-based reverse line blot hybridization. J. Med. Microbiol 2008;57(Pt 7):827–38.
[21] Chan M, Maiden M, Spratt BG. Database-driven multi locus sequence typing MLST of bacterial pathogens. Bioinformatics 2001;17:1077–83.
[22] Reis B, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. PNAS 2003;100(4):1961–5.
[23] Kulldorff M, Information Management Services, Inc. SaTScan TM v7.0: Software for the spatial and space-time scan statistics; 2007. Available from: http://www.satscan.org/.
[24] Buckeridge DL. Outbreak detection through automated surveillance: a review of the determinants of detection. J Biomed Inform 2007;40:370–9.
[25] Weinstock M. A generalized scan statistic test for the detection of clusters. Int J Epidemiol 1980;10:289–93.
[26] Kulldorff M. A spatial scan statistic. Communications in Statistics. Theory Methods 1997;26:1481–96.
[27] Hall G, Raupach J, Yohannes K. An estimate of under-reporting of foodborne notifiable diseases: *Salmonella*, Campylobacter, Shiga toxin-producing *E. coli* (STEC). National Centre for Epidemiology Working Paper No. 52, Australian National University; 2006.