Evaluating the utility of syndromic surveillance algorithms for screening to detect potentially clonal hospital infection outbreaks

Randy J Carnevale,1 Thomas R Talbot,2,3 William Schaffner,2,3 Karen C Bloch,2 Titus L Daniels,2 Randolph A Miller1

ABSTRACT

Objective The authors evaluated algorithms commonly used in syndromic surveillance for use as screening tools to detect potentially clonal outbreaks for review by infection control practitioners.

Design Study phase 1 applied four aberrancy detection algorithms (CUSUM, EWMA, space-time scan statistic, and WSARE) to retrospective microbiologic culture data, producing a list of past candidate outbreak clusters. In phase 2, four infectious disease physicians categorized the phase 1 algorithm-identified clusters to ascertain algorithm performance. In phase 3, project members combined the algorithms to create a unified screening system and conducted a retrospective pilot evaluation.

Measurements The study calculated recall and precision for each algorithm, and created precision-recall curves for various methods of combining the algorithms into a unified screening tool.

Results Individual algorithm recall and precision ranged from 0.21 to 0.51 and from 0.053 to 0.29, respectively. Few candidate outbreak clusters were identified by more than one algorithm. The best method of combining the algorithms yielded an area under the precision-recall curve of 0.553. The phase 3 combined system detected all infection control-confirmed outbreaks during the retrospective evaluation period.

Limitations Lack of phase 2 reviewers’ agreement indicates that subjective expert review was an imperfect gold standard. Less conservative filtering of culture results and alternate parameter selection for each algorithm might have improved algorithm performance.

Conclusion Hospital outbreak detection presents different challenges than traditional syndromic surveillance. Nevertheless, algorithms developed for syndromic surveillance have potential to form the basis of a combined system that might perform clinically useful hospital outbreak screening.

BACKGROUND

Past approaches to automated hospital outbreak detection fall into two categories: active and passive surveillance. Active surveillance approaches use decision support algorithms to automatically inform infection control staff of suspicious disease patterns that require further attention. Passive surveillance approaches provide tools that simply aggregate or display information in a more usable and manipulable electronic format for infection control staff to review on their own initiative, allowing them to better detect interesting patterns ‘manually’. Online appendix A contains a brief summary of these previous approaches to automated surveillance, with references.

Outbreaks fall into two categories: clonal and non-clonal. Non-clonal outbreaks typically occur when infection control techniques are suboptimal (eg, improper hand washing). The resulting infections involve many different bacterial species. A clonal outbreak occurs when progeny of a single organism spread to multiple patients. Non-clonal outbreaks are readily identifiable by an overall problems during the recognition and investigative processes incur delays in interventions, and with delays come increased costs and higher risks of patient morbidity and mortality.

Several recent approaches supplement older manual outbreak detection practices with automated outbreak alerting mechanisms. For more than 2 decades, various investigative groups have applied direct and straightforward algorithmic detection methods to hospital data to demonstrate improved sensitivity in inpatient outbreak alerting.2,3 Relatively few studies, however, have applied the newer algorithms developed for syndromic surveillance to single hospital inpatient surveillance. Syndromic surveillance algorithms have typically used pre-clinical data (eg, records of over-the-counter pharmaceutical purchases and of chief complaints from emergency room visits) in an attempt to detect outbreaks in outpatient settings over large geographic areas.4–6 In order to develop a screening tool that helps hospital infection control personnel to identify outbreaks in an individual hospital setting, the present study utilized microbiology culture and antibiotic sensitivity results rather than pre-clinical data as the input for algorithms initially developed for regional syndromic surveillance. The authors evaluated the algorithms’ suitability, singly and in combination, to screen culture results in a clinically useful manner.
increase in infection rates in a given hospital unit. Clonal outbreaks, however, may remain unnoticed since the increase in infections by a single rarer species may not significantly affect the overall infection rate. Genetic and molecular fingerprinting techniques remain the gold standard for determining the clonality of two bacterial isolates from different patients’ cultures of the same species. Nevertheless, it is both more efficient and more cost effective within a given institution to first screen for potential clonal outbreaks by comparing antibiotic sensitivity patterns for each bacterial species identified by cultures.7

The current exploratory study evaluated the ability of four algorithms previously applied to regional syndromic surveillance to serve as screening tools for detecting potential clonal hospital outbreaks—individually and in combination. The goal was to provide useful input to hospital infection control personnel for further review and possible additional testing. Two of these aberrancy detection algorithms originated in manufacturing quality control (CUSUM and EWMA), while the other two came from syndromic surveillance research (space-time scan statistic and WSARE).

**Statistical process control algorithms: CUSUM and EWMA**

Statistical process control originated in 1931, when Walter Shewhart of Bell Laboratories first described control chart methodologies to monitor manufacturing processes.8 Statistical process control algorithms use previous data to estimate future control limits. In statistical process control.45 The STSS algorithm incorporates spatial and temporal clustering rules for the given day and estimates the statistical significance of the cluster using Monte Carlo simulation. Many studies have employed STSS with success, including those observing commonly occurring infectious diseases,16 emerging infectious diseases,17 and cancer incidence.18 Complete details regarding the STSS algorithm appear in Kulldorff’s publications.14 15 16

As STSS addressed the growing need for incorporating spatial data, WSARE addressed the growing need for a cluster detection algorithm that could incorporate multidimensional data (eg, gender, age, and location in addition to disease status).4 5 20 WSARE first constructs a Bayesian network model based on the problem domain, but empirically, values between 0.2 and 0.5 have provided good performance in manufacturing.9 10

The typical upper and lower bounds for EWMA are similar to those used in Shewhart’s control charts, and are given by

\[ \text{EWMA}_t = \lambda Y_t + (1-\lambda)\text{EWMA}_{t-1} \]

where \( \text{EWMA}_0 \) is the historical mean, \( Y_t \) is the measurement at time \( t \), and \( \lambda \) is the decay rate of past measurements, with \( 0<\lambda \leq 1 \). At \( \lambda = 1 \), the EWMA formula matches the Shewhart control chart formula. Optimal \( \lambda \) values vary depending on the problem domain, but empirically, values between 0.2 and 0.5 have provided good performance in manufacturing.9 10

The value of \( \lambda \) affects the variance of the EWMA statistic and thus the limits, as the estimated variance is given by:

\[ s_{\text{EWMA}}^2 = \left( \frac{\lambda}{2-\lambda} \right) s^2 \]

where \( s^2 \) is the historical variance. Although more difficult to calculate, EWMA charts have the benefit of being more sensitive to small shifts in the mean than Shewhart’s control charts while still being as easy to interpret graphically.

**Syndromic surveillance algorithms: space-time scan statistic and WSARE**

Following the 2001 anthrax attacks in the USA,13 fears of bioterrorism increased interest in the nascent field of syndromic surveillance. Such systems identify infectious disease outbreaks using pre-clinical data (eg, emergency room visits, pharmaceutical purchases, etc) over a large geographic area. The current study applied two algorithms previously developed specifically for syndromic surveillance to the hospital setting: Kulldorff’s space time scan statistic (STSS) and What’s Strange About Recent Events (WSARE).

Martin Kulldorff first introduced STSS in 1997.14 At the time, most syndromic surveillance researchers used purely temporal disease cluster detection methods, including the algorithms used in statistical process control.4 5 The STSS algorithm incorporates spatial information into its detection as well to attempt to improve detection over a large geographic area. It uses a two-stage process. First, STSS searches the study area for the circular region most likely to be a disease cluster assuming the disease follows either a Bernoulli model or a Poisson model. Second, it estimates the statistical significance of the cluster using Monte Carlo simulation. Many studies have employed STSS with success, including those observing commonly occurring infectious diseases,16 emerging infectious diseases,17 and cancer incidence.18 Complete details regarding the STSS algorithm appear in Kulldorff’s publications.14 15 16

As STSS addressed the growing need for incorporating spatial data, WSARE addressed the growing need for a cluster detection algorithm that could incorporate multidimensional data (eg, gender, age, and location in addition to disease status).4 5 20 WSARE first constructs a Bayesian network model based on the problem domain’s historical data. It then uses the Bayesian network to find the single ‘best’ clustering rule for the given day and estimates a p value using Benjamini and Hochberg’s False Discovery Rate method21 to adjust for the multiple hypothesis tests.20 Because the underlying Bayesian model can include a node for each data element, WSARE easily incorporates multidimensional data. For example, if the data include gender, zip code, and influenza diagnoses, WSARE could in theory detect an increase in influenza across the study region, an increase in influenza in women region-wide, or an increase in influenza in one specific zip code. WSARE’s primary use has been in conjunction with the RODS public health surveillance system22 both for temporary short term monitoring of the 2002 Winter Olympics23 and for long-term public health surveillance of the state of Pennsylvania.24 Complete details of the WSARE algorithm appear in Wong et al.20
METHODS
Study design
This study evaluated the ability of four aberrancy detection algorithms to function as a screening tool for identifying potentially clonal outbreaks at a single site using de-identified microbiologic culture data. The four evaluated algorithms included two custom implementations (CUSUM9 and EWMA9) and two reference implementations (WSARE20 and Kullodff’s space-time scan statistic,14 SaTScan). The de-identified dataset included daily case counts for each organism taken from all microbiologic culture data collected from 2001 through 2006 from Vanderbilt University Hospital and Monroe Carell Jr. Children’s Hospital at Vanderbilt-affiliated inpatient units, outpatient clinics, and emergency rooms. It included only the first result of a given culture type (ie, organism and sensitivity pattern) for each patient on each unit to avoid giving extra weight to multiple serial cultures of the same organism from the same patient.

The study comprised three phases. Phase 1 implemented the four aberrancy detection algorithms using the hospital-derived retrospective microbiologic culture data, producing a list of potential past outbreak clusters. In phase 2, four Vanderbilt University School of Medicine Infectious Diseases faculty members who were blinded to algorithm source reviewed and categorized the suspected clusters to ascertain the performance of each phase 1 algorithm. In phase 3, project members empirically used the phase 2 results as feedback to adjust configuration parameters associated with each algorithm and investigated additional methods for combining the algorithms’ output into a single outbreak detection screening tool. The authors then carried out a 6-month retrospective evaluation of the new system. The Vanderbilt University Institutional Review Board approved the study prior to its initiation.

Phase 1: Algorithm execution
The study configured each algorithm to identify clusters of positive cultures from daily case-culture counts for each organism—both for individual hospital units and across the entire institution. The study divided the culture dataset into three parts. The first set (1 year; January 1, 2001—December 31, 2001) provided historical ‘seed’ data for each algorithm. The second set (3 years; January 1, 2002—December 31, 2004) served as a testing set for tuning the parameters of each algorithm and designing the review module before study initiation. This second set also provided additional historical baseline data for the final review. The third set (2 years; January 1, 2005—December 31, 2006) provided the testing data for the study phase 2 expert review. The study converted output from each of the four study algorithms into a common format to prevent the reviewers from identifying which algorithm had generated the cluster.

Phase 2: Expert review process
The project developed a web-based review module that collectively and serially displayed the clusters identified by the algorithms to the group of expert reviewers. Each reviewer had substantial experience as a hospital-affiliated physician-epidemiologist. Using the web-based review module, the reviewers classified each computer-generated cluster as a potential outbreak or a spurious cluster and further delineated each outbreak occurrence as ‘probable’ (likely a real outbreak), or ‘possible’ (not certain if a real outbreak). They produced their assessments based on geographic and temporal data regarding a given set of culture results comprising an algorithm-defined cluster. The reviewers could ‘drill down’ on each cluster to view narrative culture result reports and antibiotic sensitivities as needed. The reviewers also noted whether they would have conducted any further investigations had they been both aware of the cluster and responsible for hospital infection control at the time the cluster occurred. Each expert conducted an independent review while blinded to the assessments made by the other experts. As indicated in table 1, the study converted the experts’ designations into a binary classification, labeling a cluster as a ‘candidate outbreak’ if the experts identified it as a probable outbreak or a possible outbreak that merited further investigation. In an actual outbreak investigation, hospital infection control staff would conduct additional serologic or genetic testing of each candidate bacterial isolate to determine whether the cluster represented a true outbreak; no such data were available regarding the clusters the experts reviewed.

The study assigned two of the four expert reviewers to examine each algorithm-identified potential cluster independently. Discordant assessments were resolved by submitting each to a ‘tiebreaker’ reviewer randomly selected from the two reviewers who had not previously evaluated the cluster. To calibrate the reliability of the tiebreaking opinions, the study also presented the tiebreak reviewers with several randomly chosen clusters on which the first two reviewers’ determinations agreed (either as ‘candidates’ or not).

The study supplemented the list of candidate outbreaks identified by the review process (as defined above) with five infection control-investigated clusters that had been independently characterized previously by the hospital’s infection control staff. These five consisted of disease clusters subjected to genetic or serologic testing during the study time period.

Following the clinicians’ reviews, the study calculated the sensitivity and positive predictive value (recall and precision) for each cluster identification algorithm based on the ‘consensus’ classifications (by two or three reviewers, per protocol) of suspected outbreaks and infection control-investigated clusters. The study compared the individual algorithms’ performance statistics pairwise using McNemar’s test. Figure 1 summarizes the processes followed in phases 1 and 2.

Phase 3: Parameter tuning, precision-recall analysis, combined tool development, and retrospective evaluation
In study phase 3, the project empirically analyzed the effects of varying algorithm parameters on each algorithm’s ability to identify phase 2 expert-labeled candidate outbreaks. The study also explored potential methods of combining the individual algorithms with additional heuristic data to produce better candidate outbreak identification than obtained by the individual algorithms per se.

A first approach was to adjust parameters for the customizable algorithm that demonstrated better performance in phase 2 (CUSUM or EWMA) to detect as many of the candidate outbreaks as possible. For each of the expert-identified candidate outbreak clusters, the study calculated k, the minimum threshold at which the chosen algorithm would generate an alert.
for the outbreak, using varying decay rates $\lambda$ (0.05, 0.07, 0.1, 0.15, 0.2, 0.25, and 0.3). Project members recorded the number of additional alerts that would also have triggered at the given value of $k$. Based on these measurements, the study determined the optimal value of $\lambda$ and generated precision-recall curves for varying values of $k$ when using the optimized algorithm.

The study also explored methods of combining the output from the four original algorithms using various scoring metrics by which the resulting clusters could be ranked. A first step attempted to order the clusters by their previously measured value of $k$. Project members then made additional adjustments to the rank weights regarding several features identified as potentially important by the expert Infectious Disease faculty reviewers during the phase 2 review, including hospital location type (inpatient vs outpatient) and primary culture source type (urine, blood, wound, etc).

The study examined the potential for not ‘alerting’ for clusters comprised of organisms with substantially different antibiotic susceptibilities. This approach had the potential to eliminate noise due to clusters comprised of different clones from the same bacterial species. For each cluster for which sensitivity results were available for at least 50% of component cultures, project members developed an algorithm that calculated an antibiotic susceptibility difference score by summing the number of individual antibiotic sensitivity result pairwise differences and weighting the overall result by the number of cultures having each of the compared patterns. The resulting score thus represented the average number of differing antibiotic sensitivities between each pair of bacterial isolates. This filtering method, applied to the output of the individual screening algorithms, allowed the analysis to exclude clusters not meeting empirically derived uniformity limits (ie, those that appeared to be non-clonal based on varied culture sensitivities) while still allowing the system to detect potentially clonal clusters that had mutated only slightly in their antibiotic resistance over the course of the outbreak. A final best-case heuristic combination of these methods comprised the phase 3 combined detection system. With these adjustments in place, phase 3 of the study concluded by conducting a brief retrospective validation of the combined outbreak detection system’s recall. The system was run using data from January 1, 2010 to June 30, 2010 and the resulting clusters were compared to the list of confirmed outbreaks that had been previously discovered by hospital infection control staff using manual methods.

### RESULTS

#### Phase 1: Algorithm parameters

Using the first and second datasets, the authors empirically set the parameters for each algorithm. For EWMA, authors set a decay rate $\lambda=0.3$ and an alerting threshold $k=5$. For CUSUM, the authors used a V-mask for determining the alerting threshold with a daily rise of three times the standard deviation of the CUSUM statistic for each particular organism. SaTScan was executed using its purely temporal Poisson model, and WSARE with its Fisher’s exact scoring metric and 100 randomizations for each day.

#### Phase 2: Expert review results

For institution-wide microbial data covering the 2-year study period, the four outbreak detection algorithms collectively generated a total of 257 alerts (CUSUM: 114, EWMA: 66, SaTScan: 21, WSARE: 56). To present alerts to clinical expert reviewers, the study combined any computer-generated alerts with start and stop dates differing by fewer than 2 days into one single alert. As a result, six alerts detected by two algorithms and one alert detected by three algorithms were combined to form the final review list of 249 alerts.

Percent agreement on the clusters between the two assigned reviewers ranged from 79% to 88% with Cohen’s $k$ ranging from 0.11 to 0.49 (table 2). Overall, reviewers agreed on their determinations for 210 of the 249 alerts, with 17 (8.1%) deemed candidate outbreaks.

For the 39 clusters on which the pair of initial reviewer assessments disagreed, the study assigned a randomly selected third reviewer. Of the 39, the third reviewer deemed nine (23%) to be candidate outbreaks. Six randomly selected candidate outbreaks (where the two initial reviewers agreed the cluster was a potential outbreak) and six randomly selected false alarms (where the reviewers agreed the cluster was not an outbreak) were also assigned to a random third reviewer. The third reviewer agreed with the first two reviewers on all six of the false alarms. However, for the six pairwise-agreed-upon candidate outbreaks, the third expert reviewer only agreed with the initial experts’ judgment once (17%).

The hospital infection control service had previously identified five suspected outbreak clusters during the study period. Those clusters were not detected by any of the algorithms as originally configured for the phase 1 study. Of the five, two have been excluded from the study analysis. In one, the laboratory assay for the involved organism, *Clostridium difficile,* was not included in the input since the dataset only included organisms identified by microbiological culturing and thus *C. difficile* antigen could not be detected by the algorithms. In the other, the outbreak spanned several months and began prior to the beginning of the study period. The study ‘gold standard’ outbreak dataset therefore contained 29 candidate outbreaks: 17 from the initial expert consensus review, nine from the second expert conflict-resolving review, and three from the infection control archival data.

<table>
<thead>
<tr>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>% Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>86% (0.22)</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>81% (0.47)</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>88% (0.48)</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>85% (0.49)</td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td>88% (0.38)</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>79% (0.11)</td>
</tr>
</tbody>
</table>

# Figure 1 Flow of microbiologic culture data during study phases 1 and 2.
Phase 2.2: Algorithm performance
For the four evaluated algorithms, the positive predictive value relative to the study-derived gold standard ranged from 5.3% to 29%, with sensitivities ranging from 21% to 31%. Table 3 shows individual results for each algorithm. The differences in sensitivity were not sufficient to reject the null hypothesis that the algorithms had identical performance. For positive predictive value, CUSUM was significantly lower than all other algorithms \((p<0.001\) in all comparisons), and EWMA and WSARE were significantly lower than SaTScan \((p<0.001\) for each).

Stratifying the analysis by location type (hospital-wide clusters and inpatient units as inpatient; clinics and emergency rooms as outpatient) demonstrated that clusters from inpatient locations were much more likely to be considered candidate outbreaks than clusters from outpatient locations \((\text{inpatient}: 21/120\text{ clusters vs outpatient: }5/129\text{ clusters}; \chi^2=0.002)\).

Phase 3.1: Parameter adjustment
As EWMA yielded both better positive predictive value and sensitivity than CUSUM, project members adjusted EWMA’s decay rates and minimum alerting thresholds in phase 3. After the adjustments, EWMA detected up to 24 of the 29 candidate outbreaks, but its positive predictive value suffered at this sensitivity, with 629 false alarms \((3.7\%)\) at this most sensitive setting.

Phase 3.2: Scoring metrics
Using the minimum alerting threshold \(k\) as the initial ranking metric to sort the original list of 249 clusters generated by the four algorithms yielded an area under the precision-recall curve (AUC) of 0.283, where the AUC for a precision-recall curve represents the average overall precision. A linear interpolation of the expert reviewers’ performance targets of 0.5 precision at 0.9 recall and 0.75 precision at 0.25 recall gives a target AUC of 0.65. Figure 2 shows the precision-recall curve for this initial metric, with the curve for the adjusted EWMA algorithm and points for each of the individual algorithms.

To investigate whether primary culture specimen type could help to separate clinically significant clusters from less important ones, project members developed an algorithm that labeled each cluster by specimen type \(\text{(blood, urine, wound, etc)}\) if more significant ones, project members developed an algorithm that labeled each cluster by specimen type \(\text{(blood, urine, wound, etc)}\) if more significant than the study-derived gold standard. The only statistically significant relationship this analysis identified was that urine cultures were less reliable indicators of clusters than other specimen types \((2.0\% \text{ of urine vs } 13\% \text{ non-urine}; p=0.029)\). After adjusting the ranking metric downward for clusters of urine cultures, the \(k\)-sorted precision-recall AUC improved from 0.283 to 0.356. As observed in phase 2, clusters in inpatient locations were more likely to produce candidate outbreaks than clusters in outpatient units. After increasing the ranking metric for inpatient clusters, the AUC rose from 0.356 to 0.489.

Project members calculated antibiotic susceptibility difference scores for the 165 clusters that met the 50% criterion, including six of the 19 candidate outbreaks. Antibiotic susceptibility difference scores ranged from 0 to 138 in the false alarm clusters and from 0 to 2.7 in the candidate outbreaks. Based on these results, project members generated new precision-recall curves after eliminating all clusters with similarity scores greater than a conservative threshold of 5 and an aggressive threshold of 3. These adjustments increased the precision-recall AUC from 0.489 to 0.528 for the conservative threshold and to 0.553 for the aggressive threshold. Precision-recall curves for each of these adjustments are shown in figure 3.

Phase 3.3: Retrospective evaluation of combined algorithms
During the 6-month retrospective evaluation period, infection control staff identified and confirmed two single-unit outbreaks: an outbreak of vancomycin-resistant \(\text{Enterococcus}\), and an outbreak of \(\text{C difficile}\). Unlike the phase 2 dataset, in phase 3, non-culture assays were added, allowing the system to detect \(\text{C difficile}\) outbreak. The system detected a total of 41 clusters during that time period, including both of the confirmed outbreak clusters. No phase 2-type expert analyses of the other 39 clusters were conducted.

DISCUSSION
This exploratory study attempted to determine whether one or more aberrancy detection algorithms might be adapted to screening for potentially clonal hospital outbreak detection. Because each algorithm produced a list of interesting suspect clusters substantially different from the others, an ideal system in this setting would consist of multiple algorithms working together.

Cluster review
Analysis of the expert review process demonstrated the degree of subjectivity in determining which clusters were potentially interesting. The first round of reviews only managed moderate levels of inter-rater agreement as shown in table 2. Because the

### Table 3 Cluster determination by algorithm

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Candidate</th>
<th>Non-candidate</th>
<th>PPV</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUSUM</td>
<td>6</td>
<td>108</td>
<td>5.3%</td>
<td>21%</td>
</tr>
<tr>
<td>EWMA</td>
<td>9</td>
<td>57</td>
<td>14%</td>
<td>31%</td>
</tr>
<tr>
<td>SaTScan</td>
<td>6</td>
<td>15</td>
<td>29%</td>
<td>21%</td>
</tr>
<tr>
<td>WSARE</td>
<td>7</td>
<td>49</td>
<td>13%</td>
<td>24%</td>
</tr>
</tbody>
</table>

PPV, positive predictive value.
overall prevalence of true positive clusters was relatively low, measured values of Cohen’s $k$ were low despite a high percentage of agreement between reviewers. The low $k$ suggests that despite having similar training and using similar review criteria, the expert reviewers disagreed fairly often, and that constructing a true gold standard is not possible. In the second round ‘tiebreaker’ reviews, the third reviewer only agreed with the initial reviews on 17% of the ‘seed’ candidate outbreaks. By contrast, when the third reviewer examined clusters which one of the two original reviewer had designated as a candidate cluster and the other had not, the third reviewer designated the cluster as a candidate 23% of the time.

The low reviewer agreement suggests that an ideal hospital outbreak detection screening tool should favor sensitivity over positive predictive value since experts may disagree on which clusters merit further investigation. This strategy is further supported by standard infection control practice: in a prospective study, further investigation including molecular typing would have followed on each of the potentially interesting clusters to confirm clonality. Because the investigation will easily distinguish true positives from false positives, it is more important that the detection system acts as a ‘screening test’ that does not produce many false negatives.

**System performance and ranking**

The lack of consensus among alerts generated by the four algorithms and the excessive false positive rate for the parameter-adjusted EWMA algorithm suggest that none of the four algorithms evaluated can solely provide a reliable alerting mechanism. Thus, to create a functionally useful alerting system for hospital infection control purposes, some algorithmic combination technique that leverages the relative strengths of each individual algorithm will likely provide the best overall system.

Prior to the current study’s data analysis, the expert reviewers stated that performance goals for a useful outbreak screening system that they would use in practice would require a 50% positive predictive value at 0.9 sensitivity and 0.25 sensitivity at a 75% positive predictive value. Ranking the combined list of clusters using the adjusted scoring metric and eliminating clusters with dissimilar antibiotic susceptibilities allowed us to achieve a 40% positive predictive value up to a sensitivity of 0.9 and a sensitivity of approximately 0.15 at a positive predictive value of 75%. While these results did not attain the targeted performance levels, our experts found them encouraging, and further improvements may be possible.

**Limitations**

The subjectivity of the review process led to an imperfect ‘gold standard’ list of candidate outbreaks. The gold standard list could easily have missed some true outbreaks due to reviewer disagreement on what constituted a candidate cluster. Furthermore, the selection of algorithms for the study did not include the newest syndromic surveillance methods and the parameter tuning required to implement each of the four algorithms may not have been optimal, with the result that true outbreak clusters may have been omitted from the algorithms’ output lists before ever being seen by the reviewers. That none of infection control verified outbreaks during the study period appeared on the combined output list of the four algorithms suggests that suboptimal detection at the algorithmic level was a factor in our study.

The culture results dataset used to generate the alerts also contained potential methodological flaws. The study used only the first result for a given organism/patient/unit combination in the dataset. While this approach prevents spurious alerts for multiple consecutive positive cultures on the same patient, it may have been too conservative overall. For example, a patient with *Escherichia coli* cultures in January 2005 and January 2006 would only be included in 2005, although it is unlikely that the patient’s infection lasted a full year. Additional errors may also arise from the system’s lack of information about changes within the hospital over time. For example, in late 2005 (approximately halfway through the study period), the burn intensive care unit was relocated to another geographic ward, so new patient-organism-location clusters that previously would have been suppressed as duplicate cultures were not suppressed since they were reported from a ‘different’ geographic unit. In

**Figure 3** Precision-recall curves for adjusted scoring metrics.
addition, some clusters were simply a result of increased surveillance for certain organisms or an increase in a hospital unit’s size or number of patient days as the study did not adjust for increases in patient bed days.

The adjustment for antibiotic sensitivity similarity was somewhat crude. For example, if an algorithm detected a cluster made up of two distinct clones with widely differing sensitivities, the resulting average difference between the two could be large enough to eliminate the cluster from further consideration. Ideally, available antibiotic sensitivity data should be included earlier in the detection process.

Lastly, the performance of the system on retrospective datasets does not guarantee similar future performance. Because the review process was time consuming for the reviewers and the number of expected candidate outbreaks was limited, the resulting parameter adjustments have not been validated extensively. The ‘optimal’ alerting thresholds determined in the current study may be overfitted to the current data. Nevertheless, the 6-month retrospective evaluation demonstrated that the resulting system was able to detect all outbreaks confirmed by hospital infection control staff during that time period.

**CONCLUSION**

The current study explored the potential for a syndromic-surveillance-based approach to screening for potentially clonal inpatient infectious disease outbreaks. Each of the four aberrancy detection algorithms that the study examined had different performance characteristics that limited its individual applicability to the problem at hand. However, by combining the output from each algorithm and then sorting and filtering the possible clusters that the algorithms identify based on additional heuristic data that the algorithms cannot easily incorporate, the authors created a prototypic combined screening tool that demonstrated better potential to be clinically useful for hospital outbreak detection than any of the individual algorithms. Thus, while in-hospital outbreak surveillance presents different challenges than those faced by regional syndromic surveillance, the algorithms developed for syndromic surveillance may eventually be adapted to the inpatient screening setting. Further, more formal evaluation of such combined systems should occur.

**Funding** This study was funded by the National Library of Medicine, National Institutes of Health (grants T15 LM007450-08 and 5R01-LM07995-06).

**Competing interests** None.

**Ethics approval** Vanderbilt University IRB approved this study.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**REFERENCES**