Research Paper ■

Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance

Stephen E. Brossette, Alan P. Sprague, PhD, J. Michael Hardin, PhD, Ken B. Waites, MD, Warren T. Jones, PhD, Stephen A. Moser, PhD

Abstract **Objectives**: The authors consider the problem of identifying new, unexpected, and interesting patterns in hospital infection control and public health surveillance data and present a new data analysis process and system based on association rules to address this problem.

Design: The authors first illustrate the need for automated pattern discovery and data mining in hospital infection control and public health surveillance. Next, they define association rules, explain how those rules can be used in surveillance, and present a novel process and system—the Data Mining Surveillance System (DMSS)—that utilize association rules to identify new and interesting patterns in surveillance data.

Results: Experimental results were obtained using DMSS to analyze *Pseudomonas aeruginosa* infection control data collected over one year (1996) at University of Alabama at Birmingham Hospital. Experiments using one-, three-, and six-month time partitions yielded 34, 57, and 28 statistically significant events, respectively. Although not all statistically significant events are clinically significant, a subset of events generated in each analysis indicated potentially significant shifts in the occurrence of infection or antimicrobial resistance patterns of *P. aeruginosa*.

Conclusion: The new process and system are efficient and effective in identifying new, unexpected, and interesting patterns in surveillance data. The clinical relevance and utility of this process await the results of prospective studies currently in progress.

■ **JAMIA.** 1998;5:373–381.

Surveillance systems are essential in detecting new and re-emerging threats of infectious agents in public health and hospital settings.[1–3] The effectiveness of these systems is determined by their ability to rapidly analyze time-series data to detect unusual disease clusters.[1,3]

Affiliation of the authors: University of Alabama at Birmingham, Birmingham, Alabama.

This work was supported in part by cooperative agreement U47-CCU411451 with the Centers for Disease Control and Prevention (SAM), grant 1688 from the Paralyzed Veterans of America, Spinal Cord Injury Program (KBW), and predoctoral research fellowship LM-00057 from the National Library of Medicine (SEB).

Correspondence and reprints: Stephen Moser, PhD, Department of Pathology P246, 619 19th Street South, Birmingham, AL 35233-7331. e-mail: ⟨moser@uab.edu⟩.

Received for publication: 10/3/97; accepted for publication: 3/3/98.

In a chapter on computerized public health surveillance systems in *Principles and Practice of Public Health Surveillance*, Dean et al. describe an ideal public health surveillance system.[2] In doing so, they give a hypothetic example in which the user, and epidemiologist, uses the ideal system to compare data recently collected with similar data from the past. Specifying no other constraints, the user "asks the system to produce a series of maps for all conditions with unusual patterns." Identifying those patterns that are most interesting, the investigator then employs traditional database query techniques and statistical analysis to investigate them further.

In addition to the political and administrative barriers that the authors identify as obstacles to practically realizing such a system, they correctly identify the following challenge: "Several kinds of mental shifts, as well as corresponding technical developments, will be necessary before a computerized system can be used

to examine automatically a 'time slice' of disease and injury records that originate in clinics and hospitals."[2] It is this challenge that we address.

The process described in the example just given is analogous to data mining. Data mining, also known as Knowledge Discovery in Databases, is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[4] Therefore, when Dean et al. describe an ideal surveillance system as one that must "produce a series of maps for all conditions with unusual patterns," they describe a system that does data mining.

A number of statistical strategies have been developed for automatically detecting temporal patterns in surveillance data. Recently, Hutwagner et al.[1] have employed cumulative sums to detect unusual temporal patterns in *Salmonella* surveillance data. They show that cumulative sums are effective in detecting serotype-specific *Salmonella* disease outbreaks at the state level. Their analysis, however, is limited to disease incidence over two-variable state-serotype combinations. Farrington et al.[5] use a log-linear regression model to automatically detect temporal clusters of disease by organism type using the weekly infectious disease reports of England and Wales. They look for national disease outbreaks by organism and, therefore, consider disease incidence over one variable in time.

Until now, all automatic surveillance strategies have assumed that the user has a predefined outcome or case (e.g., *Salmonella* infections in a specific location) whose incidence is to be monitored for outbreaks in time. This has limited surveillance analysis to low-dimensional one-, two-, or possibly three-variable outcomes. If significant changes in incidence occur in outcomes not identified before analysis begins, these changes go undetected.

We address the problem of automatically identifying new, unexpected, and interesting patterns in surveillance data. To this end, we propose a novel data mining surveillance process that is not constrained to looking for outbreaks within user-defined outcomes. In our process, arbitrarily complex outcomes are represented by association rules, and their incidences are captured in the confidences of those rules over time. While deviations in outcome incidence can be detected using a number of different techniques, including cumulative sums and regression, we currently employ a simple chi-square–based test for this purpose.

The Data Mining Surveillance System (DMSS) is a computer data analysis system that we are developing based on the process just mentioned. In this paper we describe DMSS and experimental results obtained by using DMSS to analyze one year of *Pseudomonas aeruginosa* data from University of Alabama at Birmingham (UAB) Hospital. Hospital infection control data, which contain elements of time, place, and person, are analogous to public health surveillance data, which contain similar data elements.[6] Our process and system, therefore, apply to both public health and hospital infection control surveillance.

Background

Association rules are the subject of many papers in the data mining literature.[7–13] Many of these papers address the problem of efficiently generating association rules from large data sets.[7,9,10,12,13] This problem can be restated as one of efficiently discovering frequent sets in data.

A **frequent set** B is a set of zero or more items found together in at least T records in a data set, where T is a user-defined frequent set support threshold. Frequent set B has **support** $S(B) \geq T$, where $S(B)$ is the number of records in which B is found. The set of zero items, called *EmptySet*, has support N, the number of records in the data set.

The following example illustrates the concept of frequent set. In a supermarket database, where each record contains the names of the items in a single basket at the checkout, the frequent set (Bread, Milk, Cheese) is likely to exist in the records from a single day because bread, milk, and cheese are found together in many baskets. If the frequent set (Bread, Milk, Cheese) exists, then so do the frequent sets (Bread), (Milk), (Cheese), (Bread, Milk), (Milk, Cheese), and (Bread, Cheese).

The **association rule** $A \Rightarrow B$, where A and B are frequent sets and A intersect B is null, is a statement of conditional probability. The **confidence** of $A \Rightarrow B$ is the conditional probability of B given A, which is equal to $S(A \cup B)/S(A)$. Alternatively, an association rule is a statement about the incidence of B in A. Specifically, the confidence of $A \Rightarrow B$ is the incidence of B in A. In the supermarket setting, the confidence of association rule (Milk, Cheese) $\Rightarrow$ (Bread) is S((Milk, Cheese, Bread))/S((Milk, Cheese)), which is the probability that bread is found in the same basket as milk and cheese or, equivalently, the incidence of bread in baskets with milk and cheese. The confidence of the association rule *EmptySet* $\Rightarrow A$ is $S(A)/N$, which should be interpreted as the unconditional probability of A or the incidence of A in all observations. The **precondition support** of association rule $A \Rightarrow B$ is $S(A)$. Association rules that have relatively high precondition support are more meaningful than rules

with relatively low precondition support because they are statements about the incidence of B in non-trivial populations A. From now on, association rules that have high precondition support will be called *high-support association rules*.

Traditional association rule data mining applications focus on discovering high-support, high-confidence association rules because these rules can be used for classification.[4,7] For example, a high-support, high-confidence rule that says that young men who purchase items a, b, and c also purchase item d 65 percent of the time can be used to construct a marketing strategy in which all four items are placed contiguously on a shelf.

While high-support, high-confidence rules will be useful in our surveillance paradigm, high-support, low-confidence rules will often be more useful. The reason is simple: If B occurs every time A occurs, and A occurs frequently, then we maintain that the rule $A \Rightarrow B$ will probably be known or trivial and therefore uninteresting. However, if B occurs infrequently with A and A occurs relatively frequently, then $A \Rightarrow B$ is a low-confidence association rule, and changes in the confidence of $A \Rightarrow B$ are likely to go undetected by traditional methods.

Methods

Association Rules for Public Health Surveillance

Since, as mentioned in the previous section, the association rules we seek are commonly not of high confidence, we propose the following process for analyzing surveillance data:

- For each time-slice or partition of data, discover all high-support association rules.
- For each rule discovered in the current partition, compare the confidence of the rule from the current partition to the confidences of the rule in previous partitions.
- If the confidence of the rule has increased significantly from a previous partition, or previous partitions, to the current partition, report this finding as an *event*.

Because a large number of high-support association rules can be found even in small data sets, a successful implementation of this process depends on efficient algorithms and on data selection and preprocessing strategies that reduce the number of association rules discovered. Efficient algorithms are described elsewhere.[7,9,13] In this paper, we focus on the process just outlined, the system based on that process, and experimental results obtained from using the system to analyze real infection control data.

Notably different from other data mining research efforts are the sizes of the data sets that we use. Whereas traditional data mining work focuses on very large data sets that are megabytes to terabytes in size, we experiment with time-slices of data that are kilobytes in size. For public health and infection control data sets, however, a great number of interesting and unusual patterns may exist in kilobyte-size data sets. Although we have developed and implemented strategies that will allow us to analyze larger data sets, those strategies were not required for the results obtained here. Therefore, we will describe them in a later paper.

Processing a Time-slice of Data

The Data Mining Surveillance System is based on the data analysis process described earlier. It processes one partition or time-slice of data at a time, discovering all high-support association rules. Each partition is composed of records that may be sparse (incomplete) or of varying lengths. The only constraint is that each record contains only items from discrete or categoric attributes. The system does not yet handle items from continuous attributes. For each high-support association rule $A \Rightarrow B$ found in the current partition, P_C, the system updates a data structure called the history with $S(A)$ and $S(A \cup B)$ for the rule in P_C. The values for $S(A)$ and $S(A \cup B)$ are used in computing the confidence of $A \Rightarrow B$. The history also contains $S(A)$ and $S(A \cup B)$ for all previous partitions in which $A \Rightarrow B$ was a high-support association rule.

Searching for Interesting Patterns

After at least two partitions of data are processed, DMSS can search for association rules whose confidence has increased significantly from some previous partition to the current partition, P_C. It does this by identifying all association rules in the history that have $S(A)$ and $S(A \cup B)$ for P_C, i.e., those association rules that were discovered in the current partition. Of these rules, those that were discovered in previous time-slices and therefore have $S(A)$ and $S(A \cup B)$ for earlier partitions are analyzed to determine whether the confidence of the rule has increased significantly over time. The analysis is accomplished as follows. For each association rule R discovered in P_C, the confidence of R in P_C, Conf(R, P_C), is compared with the confidence of R in the last partition in which R was found prior to P_C, Conf(R, P_{C-1}). The comparison of confidences is done using a chi-square/Fisher-exact–based comparison of two proportions.[14,15] If Conf(R,

P_C) is greater than Conf(R, P_{C-1}) and the probability that the difference between the proportions occurred by chance is less than 5 percent, then this finding is presented to the user. If the difference is not significant at the 5-percent level, then Conf(R, P_C) is compared with Conf(R, P_{C-2}) if Conf(R, P_{C-2}) exists. This continues until a significant difference between confidences is found or no Conf(R, P_{C-n}) exists. Since the assumptions of statistical inference are not met in this type of exploratory analysis, no strong statistical statements can be made about underlying populations. Strong statements are not our objective, however, as we are trying only to identify potentially interesting patterns. The ultimate value of each pattern is dependent on its real-world interpretation by domain experts.

The strategy used to identify temporal patterns in DMSS is an interchangeable part. If other strategies (e.g., cumulative sums) are more effective for certain classes of problems, they can be used instead. The simple chi-square–based strategy, however, is effective.

Experiments with Infection Control Data

Data Source and Data Preprocessing

One year's (1996) *Pseudomonas aeruginosa* infection control data were acquired by extracting antimicrobial susceptibility results and related patient demographics from UAB Hospital's laboratory information system. Each record describes a single *P. aeruginosa* isolate and is composed of the following items: date reported, source of isolate (e.g., sputum, blood), location of patient in hospital, patient's home zip code, and a resistant (R), intermediate resistance (I), or susceptible (S) test result for piperacillin, ticarcillin/clavulanate, ceftazidime, imipenem, amikacin, gentamicin, tobramycin, and ciprofloxacin according to minimal inhibitory concentration breakpoints of the National Committee for Clinical Laboratory Standards.[16] These antimicrobials were selected because they were used in UAB Hospital for treating *P. aeruginosa* infections in 1996 and were routinely reported by the microbiology laboratory for this organism.

Duplicate records were removed so that each patient had no more than one isolate per month.[17] The resulting data set contained approximately 80 non-duplicate records per month.

In addition to removing duplicate records, items of the form S~*Antimicrobial* were removed from each record so that only I~*Antimicrobial* and R~*Antimicrobial* items remained. This was done because an isolate is more likely to be susceptible to each of the antimicrobials than it is resistant or intermediate to it. Consequently, most of R/S/I~*Antimicrobial* items will be of the S~*Antimicrobial* type. Therefore, removing S~*Antimicrobial* items significantly decreases the average number of items in a record. Because the number of frequent sets and, consequently, the number of association rules grow quickly with the average number of frequent items, this strategy significantly reduces the computational burden of generating association rules. In addition, of the association rules that contain *Antimicrobial* items, we are interested only in those that pertain to population-specific or location-specific increases in antimicrobial resistance (which include Intermediate items). The corresponding decreases in susceptibility can be inferred. Therefore, we are looking for association rules whose confidences increase significantly over time. Patterns of increasing antimicrobial resistance fit this model and will be detected. Conversely, patterns of increasing antimicrobial susceptibility (and the corresponding decreases in resistance) are defined here as not interesting and do not fit the model.

In summary, the data selection/preprocessing step significantly decreases the number of association rules generated without sacrificing any interesting information. This is important because it decreases the computational burden on the system and the number of patterns, or events, that the user must evaluate.

Experimental Design

Three separate analyses of the data were conducted, each using a different size of data partition (time-slice). Different sizes were used to determine whether short-lived interesting patterns are discovered only when the time-slice size is relatively small (e.g., one month) while other interesting patterns are discovered only when the time-slice size is relatively large (e.g., six months). For the three experiments A, B, and C, we chose partition sizes of one month, three months, and six months respectively. Therefore, in experiment A, 12 one-month partitions of data were used; in B, 4 three-month partitions were used; and in C, 2 six-month partitions were used. Within each experiment, the partitions were non-overlapping.

A frequent-set support threshold of 2 and a rule-support threshold of 10 were used for all experiments. This means that a set of items A must be found in at least two records in a partition to be a frequent set and that for each association rule $B \Rightarrow C$, S(B) must be greater than or equal to 10 for $B \Rightarrow C$ to be a high-support association rule.

The Data Mining Surveillance System was constructed in C++, and all experiments were conducted on a Silicon Graphics Indy workstation with 32 Mb of RAM.

Results

The Data Mining Surveillance System tracks many association rules for statistically significant changes in their confidences over time. In experiment A, DMSS discovered and actively monitored more than 2,000 association rules. In experiment B, it monitored more than 12,000 rules, and in experiment C, more than 20,000 rules.

System efficiency depends on how fast the system can process data and how easily the user can evaluate the discovered events. Experiments A, B, and C required a total 31, 52, and 63 seconds of machine time respectively to process all data and to search for events. Therefore, to process one year of *P. aeruginosa* data in three separate analyses required less than $2^1/_2$ minutes of computer time. From the thousands of rules discovered in each experiment, a relatively small number of events were discovered and presented to the user. In experiment A, 34 events were discovered for the entire year. In B, 57 events were discovered, and in C 28 events were discovered.

Analysis of all events reveals that many of them were discovered in only one experiment. Seventy-two percent of the events discovered in experiment A were not found in experiment B or C; 74 percent of the events discovered in experiment B were not found in A or C; and 60 percent of the events discovered in experiment C were not found in A or B. This is because of the dynamics of the organism and patient populations themselves and constraints in processing. Events discovered in experiment A and not in experiments B or C generally occur over short time frames, i.e., within three months. These inter-quarter changes are invisible in the three-month and six-month analyses of experiments B and C. Similar logic holds for events found in experiment B but not in C. Events found in experiment C but not in A or B are generally composed of low-frequency item sets whose supports become substantial (frequent) only when counted over six-month intervals.

Because of the dynamic nature of the environment and hence the data collected from it, multiple partition sizes were used to maximize the sensitivity of the experiments. Maximizing sensitivity in general comes at the expense of decreasing specificity, thereby increasing the chance of overwhelming the user with patterns and events. In this case, however, the use of three partition sizes yielded a manageable number of events. In other cases, this will not be true. Since the purpose of our process is exploratory, however, as long as the number of patterns is manageable, we would rather err by being too sensitive while sacrificing positive predictive value. If there are too many patterns, a simple way to decrease their number is to decrease α from 0.05 to 0.01 in the chi-square and Fisher exact tests. We are also investigating ways of incorporating domain knowledge into the process to make it more specific.

Discussion

Until now, surveillance strategies have required outcomes of interest be known in advance before monitoring activities begin. Therefore, interesting activity outcomes that are not known to be of interest in advance will go undetected. In addition, the lack of systems that ''recognize and promptly report significant changes and trends'' has been cited as a barrier to successfully detecting, preventing, and controlling disease outbreaks.[18]

Elder and Pregibon[19] have added:

> With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention are required. The existence of such tools can free one up to, for instance, posit a wider range of candidate data features and basis functions (building blocks) than one would wish to deal with, if one were specifying a model structure 'by hand'.

Our objectives were threefold: First, to define a new surveillance process based on association rules that is capable of identifying patterns worthy of further investigation; second, to demonstrate that a system based on this process can automatically discover interesting and unexpected patterns in real-world data sets; and third, to show that the system can accomplish the second objective efficiently. The DMSS process and system have been described above. We address the second and third objectives by analyzing the performance and results of DMSS on real-world *P. aeruginosa* data.

Pseudomonas aeruginosa was selected for these experiments because it is a clinically important bacterium exhibiting considerable variability in incidence and antimicrobial susceptibility over time and space. In Tables 1, 2, and 3, we summarize some events discovered in experiments A, B, and C and discuss them briefly in the following paragraphs.

Some of the events discovered in each experiment are relatively simple and have a rule $EmptySet \Rightarrow A$ where A contains one item. Rules of the form $EmptySet \Rightarrow A$ where A is R~*Antimicrobial* are manually evaluated on a yearly basis and could be evaluated over shorter time intervals. The same manual analysis could be performed where A is a single hospital location or source of isolate. Because of time and resource con-

Table 1 ■

Events Including Ticarcillin/Clavulanate (Ticar/Clav), Piperacillin, and Ceftazidime (Experiment B)

Event			Findings	Action Indicated
EmptySet	→	R~Ticar/Clav R~Ceftazidime R~Piperacillin	An increase from 4% (4/104) in Oct to 8% (7/86) in Nov to 11% (8/73) in Dec in the probability that a *Pa* isolate is resistant to ticar/clav, ceftazidime, and piperacillin.	A review of the use of third-generation cephalosporins to determine whether a formulary change is necessary.
R~Ceftazidime R~Piperacillin	→	SourceSP R~Ticar/Clav	An increase from 8% (2/24) in Q3 to 32% (8/25) in Q4 in the probability that a *Pa* isolate is from sputum and resistant to ticar/clav given that it is resistant to ceftazidime and piperacillin.	Supports recommendation.
R~Piperacillin	→	SourceSP R~Ticar/Clav R~Ceftazidime	An increase from 6% (2/33) in Q3 to 26% (8/31) in Q4 in the probability that a *Pa* isolate is from sputum and resistant to ticar/clav and cerfazidime given that it is resistant to piperacillin.	Supports recommendation.
R~Ticar/Clav	→	SourceSP R~Ceftazidime R~Piperacillin	An increase from 7% (2/29) in Q3 to 24% (8/34) in Q4 in the probability that a *Pa* isolate is from sputum and resistant to ceftazidime and piperacillin given that it is resistant to ticar/clav.	Supports recommendation.
R~Ticar/Clav R~Ceftazidime R~Piperacillin	→	SourceSP	An increase from 12% (2/16) in Q3 to 42% (8/19) in Q4 in the probability that a *Pa* isolate is from sputum given that it is resistant to the 3 antimicrobials.	Supports recommendation.

NOTE: Data were for one year (1996) and were analyzed for three-month intervals. Q1 indicates Jan–Mar; Q2, Apr–Jun; Q3, Jul–Sep; Q4, Oct–Dec. *Pa* indicates *Pseudomonas aeruginosa*.

straints, however, it is unlikely that anyone would monitor outcomes with more than two or three variables such as those shown in Table 1.

In Table 1, results from experiment B reveal an increase from the third to the fourth quarter of 1996 in the proportion of *P. aeruginosa* resistant to the antimicrobials ticarcillin/clavulanate, piperacillin, and ceftazidime. This same pattern was also detected in experiment A. In addition, several other events indicate that these isolates were found more frequently in sputum (Table 1). Since ticarcillin/clavulanate, piperacillin, and ceftazidime are among the drugs most frequently used in treating gram-negative infections in the hospital and because *P. aeruginosa* infections often manifest themselves as nosocomial pneumonia, increased resistance to ticarcillin/clavulanate, piperacillin, and ceftazidime is a cause for concern. A review of the usage of the three drugs and the incidence of true infections with *P. aeruginosa* is appropriate.

Results from experiments B and C indicate that the proportion of *P. aeruginosa* isolates resistant to imipenem increased in 1996 (Table 2). From the annual hospital drug utilization report, we know that imipenem utilization increased 70 percent from the first quarter to the fourth quarter and 35 percent from the first six months to the second six months of 1996. These findings suggest that the Pharmacy and Therapeutics Committee should initiate a review of imipenem utilization.

Experiment C also revealed several source-specific *P. aeruginosa* events. Events pertaining to urine isolates intermediate and resistant to imipenem, resistant to ceftazidime, and resistant to piperacillin are shown in Table 3.

In the previous examples, there are no events in specific locations within the hospital. However, as shown in Table 4, the probability that *P. aeruginosa* was from a patient in the neurologic intensive care unit and that it was from sputum increased significantly from the first half to the second half of 1996. Taken together with the results of the other analyses—e.g., increased proportion resistant to multiple antimicrobials—an investigation to explain this occurrence and determine its cause should be undertaken.

It is important to note that the patterns identified by the process are only potentially interesting and that

Table 2 ■

Events Including Imipenem (Experiments B and C)

Event			Findings	Action Indicated
Experiment B:				
EmptySet	→	R~Imipenem	An increase from 2% (6/292) in Q3 to 8% (21/263) in Q4 in the probability that a *Pa* isolate is resistant to imipenem.	Utilization review of imipenem.
R~Ciprofloxacin	→	R~Imipenem	An increase from 7% (3/43) in Q3 to 24% (8/33) in Q4 in the probability that a *Pa* isolate is resistant to imipenem given that it is resistant to ciprofloxacin.	Review of antimicrobial selection in the treatment of patients with *Pa* infection.
R~Piperacillin	→	R~Imipenem	An increase from 6% (2/33) in Q3 to 26% (8/31) in Q4 in the probability that a *Pa* isolate is resistant to imipenem given that it is resistant to piperacillin.	Review of antimicrobial selection in the treatment of patients with *Pa* infection.
Experiment C:				
EmptySet	→	I~Imipenem	An increase from 2% (7/409) in S1 to 7% (39/555) in S2 in the probability that a *Pa* isolate is intermediate to imipenem.	Supports both recommendations.
EmptySet	→	R~Imipenem	An increase from 2% (8/409) in S1 to 5% (27/255) in S2 in the probability that a *Pa* isolate is resistant to imipenem.	Supports both recommendations.
R~Ciprofloxacin	→	I~Imipenem	An increase from 6% (4/66) in S1 to 15% (11/72) in S2 in the probability that a *Pa* isolate is intermediate to imipenem given that it is resistant to ciprofloxacin.	Supports both recommendations.

NOTE: Data were for one year (1996) and were analyzed for three-month intervals (experiment B) or six-month intervals (experiment C). Q1 indicates Jan–Mar; Q2, Apr–Jun; Q3, Jul–Sep; Q4, Oct–Dec; S1, Jan–Jun; S2, Jul–Dec. *Pa* indicates *Pseudomonas aeruginosa*.

Table 3 ■

Sample Source-specific Events (Experiment C)

Event			Findings	Action Indicated
SourceUrine	→	I~Imipenem	An increase from 2% (2/96) in S1 to 9% (12/142) in S2 in the probability that a *Pa* isolate is intermediate to imipenem given that it is from urine.	Supports need for utilization review of imipenem indicated in Table 2.
EmptySet	→	SourceUrine I~Imipenem	An increase from 0.5% (2/409) in S1 to 2% (12/555) in S2 in the probability that a *Pa* isolate is from urine and is intermediate to imipenem.	Supports the above recommendation.
EmptySet	→	SourceUrine R~Ceftazidime	An increase from 1% (5/409) in S1 to 3% (16/555) in S2 in the probability that a *Pa* isolate is resistant to ceftazidime and is isolated from urine.	Combined with the next event, indicates need for review of use of third-generation cephalosporins.
SourceUrine R~Piperacillin	→	R~Ceftazidime	An increase from 50% (5/10) in S1 to 85% (11/13) in S2 in the probability that a *Pa* isolate is resistant to ceftazidime given that it is isolated from urine and that it is resistant to piperacillin.	—

NOTE: Data were for one year (1996) and were analyzed for six-month intervals. S1 indicates Jan–Jun; S2, Jul–Dec. *Pa* indicates *Pseudomonas aeruginosa*.

Table 4 ■

Sample Location-specific Events (Experiment C)

Event	Findings	Action Indicated
EmptySet ➔ LocNICU SourceSP	An increase from 0.5% (2/409) in S1 to 2% (12/555) in S2 in the probability that a *Pa* isolate is from the NICU and is isolated from sputum.	Review the intensive care procedures in the NICU and/or antimicrobial usage.
SourceSP ➔ LocNICU	An increase from 4% (2/54) in S1 to 23% (12/53) in S2 in the probability that a *Pa* isolate is from a patient in the NICU given that the isolate is from sputum.	Supports the above recommendation.

NOTE: Data were for one year (1996) and were analyzed for six-month intervals. S1 indicates Jan–June; S2, Jul–Dec. *Pa* indicates *Pseudomonas aeruginosa*.

they are not strong statements of statistical inference. Their ultimate value can be determined only in the context of expert interpretation followed by careful examination of underlying factors.

Although the experiments presented in this paper were conducted on retrospective data, we plan to use DMSS in prospective surveillance programs. In such programs, DMSS would analyze new partitions of data as they become available. Discovered events would be timely and, as such, could be interpreted, investigated, and acted on in a proactive manner.

In the future, we believe that DMSS will be particularly useful in focused intensive care unit surveillance where it could be used to identify possible disease outbreaks and their clinical and microbiologic characteristics. These functions are crucial components of nosocomial outbreak investigations.[20] In addition, DMSS could be used to assist in establishing and modifying standards for the empiric and prophylactic use of antimicrobial agents in intensive care units. While such standards are desirable,[21,22] there is little agreement on their design and implementation.[22] Events generated by DMSS indicating shifts in antimicrobial susceptibilities of resident organisms could suggest when changes in policy governing antimicrobial use should be initiated to combat emerging resistance.

Conclusions

We have defined a new exploratory data mining process for automatically identifying new, unexpected, and potentially interesting patterns in hospital infection control and public health surveillance data. This process and the system based on it, DMSS, utilize association rules to represent outcomes and association rule confidences to monitor changes in the incidence of those outcomes over time. Through experiments with *P. aeruginosa* infection control data from UAB Hospital, we have demonstrated that the DMSS process and system are effective and efficient in identifying potentially interesting and previously unknown patterns whose ultimate value depends on careful expert evaluation. Our future work will focus on experiments with public health and intensive care unit infection control data, utilizing prospective clinical studies, to determine the usefulness of DMSS in hospital infection control. In addition, evaluation of more sophisticated chi-square–based strategies for identifying outbreaks, improved event presentation to the user, and strategies for handling larger data sets are planned.

We believe that this approach to surveillance will be useful in hospital infection control programs and is a step toward the public health surveillance system described by Dean et al.[2] The clinical relevance and utility of this approach await the result of prospective studies currently in progress.

The authors thank Alan Stamm, MD, Director of the UAB Hospital Infection Control Committee, for his support and helpful suggestions.

References ■

1. Hutwagner C, Maloney EK, Bean NH, Slutsker L, Martin S. Using laboratory-based surveillance data for prevention: an algorithm for detecting salmonella outbreaks. Emerg Infect Dis. 1997;3:395–400.
2. Dean AG, Fagan RF, Panter-Conner BJ. Computerizing public health surveillance systems. In: Teutsch SM, Churchill RE (eds). Principles and Practice of Public Health Surveillance. New York: Oxford University Press, 1994:200–17.
3. Gaynes RP. Surveillance of nosocomial infections: a fundamental ingredient for quality. Infect Control Hosp Epidemiol. 1997;18:475–8.
4. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds). Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: MIT Press, 1996:1–36.
5. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A

statistical algorithm for the early detection of outbreaks of infectious disease. J R Stat Soc A. 1996;159:547–63.
6. Thacker SB. Historical development. In: Teutsch SM, Churchill RE (eds). Principles and Practice of Public Health Surveillance. New York: Oxford University Press, 1994:3–17.
7. Agrawal R, Mannila H, Srikant R, Toivonan H, Verkamo AI. Fast discovery of association rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds). Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: MIT Press, 1996:307–28.
8. Klemettinen M, Manilla H, Ronkainen P, Toivonen H, Verkamo A. Finding interesting rules from large sets of association rules. In: Proceedings of the 3rd International Conference of Information and Knowledge Management. New York: ACM Press, 1994:401–7.
9. Agrawal R, Schaffer J. Parallel mining of association rules. IEEE Trans Knowl Data Eng. 1996;8:962–9.
10. Cheung D, Ng V, Fu A. Efficient mining of association rules in distributed databases. IEEE Trans Knowl Data Eng. 1996; 8:911–22.
11. Manilla H, Toivonen H. Multiple uses of frequent sets and condensed representations. In: Simoudis E, Han J, Fayaad UM (eds). Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Menlo Park, Calif.: AAAI Press, 1996:189–94.
12. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: ACM SIGMOD 1997, New York, NY: ACM Press, 1994:255–64.
13. Zaki MJ, Parthasarathy S, Ogihara M, Li W. New algorithms for fast discovery of association rules. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press, 1997:283–6.
14. Crow EL, Davis FA, Maxfield MW (eds). Statistics Manual. New York: Dover Publishing, 1960.
15. Freeman J. Quantitative epidemiology. Infect Control Hosp Epidemiol. 1996;17:249–55.
16. National Committee for Clinical Laboratory Standards. Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically: Approved Standard. 4th ed. Wayne, Pa.: NCCLS, 1997. Document M7-AF.
17. Ngo L, Tager IB, Hadley D. Application of exponential smoothing for nosocomial infection surveillance. Am J Epidemiol. 1996;143:637–47.
18. Neu HC, Duma RJ, Jones RN, et al. Antibiotic resistance epidemiology and therapeutics. Diagn Microbiol Infect Dis. 1992;15:53s–60s.
19. Elder JF, Pregibon D. A statistical perspective on knowledge discovery in databases. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds). Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: MIT Press, 1996:83–116.
20. Doebbeling NB. Epidemics: identification and management. In: Wenzel RP (ed). Prevention and Control of Nosocomial Infections. Baltimore, Md.: Williams & Wilkins, 1993:177–206.
21. O'Hanley P, Easaw J, Rugo H, Easaw S. Infectious disease management of acute leukemic patients undergoing chemotherapy: 1982 to 1986 experience at Stanford University Hospital. Am J Med. 1989;87:605–13.
22. McGowan JE Jr. Do intensive hospital antibiotic control programs prevent the spread of antibiotic resistance? Infect Control Hosp Epidemiol. 1994;15:478–83.

Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance

Stephen E Brossette, Alan P Sprague, J Michael Hardin, Ken B Waites, Warren T Jones and Stephen A Moser

J Am Med Inform Assoc 1998 5: 373-381
doi: 10.1136/jamia.1998.0050373

Notes