

# Data Mining and Infection Control

Stephen E. Brossette, MD, PhD\*,  
Patrick A. Hymel, Jr, MD

*Cardinal Health, 400 Vestavia Parkway, Suite 310, Birmingham, AL 35216, USA*

We regard data mining as the data-driven, automated construction of descriptive and predictive statistical models. Data mining, infection control, and laboratory medicine intersect at the use of clinical laboratory data by computers to automatically construct models that describe or predict hospital epidemiology patterns of statistical and clinical significance. Of course, the main tenet of data mining is that the models and patterns contain insights that were previously unsuspected. For that reason alone, data mining is not an exercise in hypothesis-driven exploratory statistics, or hypothesis-driven statistical model building, because “hypothesis-driven” implies previously suspected. In this article, we examine data mining in laboratory medicine and infection control and describe future opportunities in the space.

Infection control is the quality control activity concerned primarily with the quantification and prevention of nosocomial infections (NIs). Its success depends on the timely identification and correction of process breakdowns that increase infection risks. It is difficult, however, for infection control to identify new risk threats, intervene, and track outcomes continuously, hospital-wide. These challenges can be mitigated by a properly designed data mining system.

Traditional collection and analysis of infection control data occur by hand. The Centers for Disease Control (CDC) recommends that NI case finding proceed by the manual application of clinical case definitions that perform poorly prospectively (sensitivity = 0.61) and retrospectively (specificity = 0.68) [1]. The inability of infection control practitioners to reliably identify NIs, much less patterns among them, is a clear limitation of the traditional CDC-endorsed system. For this reason, Brossette and colleagues [2] created the electronic Nosocomial Infection Marker (NIM,

---

\* Corresponding author.

E-mail address: [sbrossette@medmined.com](mailto:sbrossette@medmined.com) (S.E. Brossette).

patent pending, Cardinal Health). The NIM outperforms the CDC National Nosocomial Infection Surveillance system (NNIS) clinical case definitions in the ICU and the Study on the Efficacy of Nosocomial Infection Control (SENIC) case definitions house-wide [2], and is based solely on electronic clinical microbiology and electronic patient census and movement data. As a result, the NIM is reproducibly computable, thus solving a major limitation of manual case-finding methods. Data models based on the NIM, or other deterministically computable infection proxies, can specifically and reliably describe patterns of NIs, not just laboratory results, allowing for more specific and objective process improvement initiatives.

### **Predictive data mining**

Descriptive data mining should reveal and describe important, previously unknown patterns of nosocomial (and community-acquired) infections, contamination, and colonization. Predictive data mining should construct models to predict NI risk. To some extent, the NIM algorithm accomplishes as much. If an NIM is detected, it is likely associated with NI; if an NIM is not detected, an NI is likely not present [2]. The previously described Germ-Watcher system implemented culture-based definitions for NI to accomplish similar goals [3]. The NIM and GermWatcher are expert rules systems; neither generates models to predict risk. They can, however, like the NIM, be used to provide data to model-generating systems.

Predictive data mining to build models of infection risk could use any of the classifier building techniques from machine learning [4] (eg, neural networks) or even techniques used mostly for descriptive mining, such as association rules. This endeavor would require substantial research, but once developed, classifiers could be used to target proactively high-risk patients for prevention efforts. Of course, the exercise could lead to obvious conclusions, such as “neutropenic patients are at high risk for NI,” but it also may provide insights that are currently unknown or underappreciated.

### **Descriptive data mining: the Data Mining Surveillance System**

Descriptive data mining in laboratory medicine and infection control is at this time entirely represented by the Data Mining Surveillance System (DMSS) [5–8]. DMSS uses frequent set and association rule analysis to automatically construct from laboratory medicine and patient movement data patterns of statistical and clinical interest. The reason these techniques are useful for infection control is that NI risks are complex and subtle patterns of infection, colonization, contamination, and multidrug resistance often go unnoticed. This is not hard to understand; the combinatorial complexity of a simple infection event space is substantial. A hypothetical hospital with 20 common bacterial pathogens, 10 specimen sources, 10 physicians or services, and 10 in vitro antimicrobial sensitivity results (each sensitive,

intermediately resistant, resistant [S/I/R]), yields more than 100,000,000 possible events. Temporal and spatial clustering of these events compose patterns, so 100 million events with the added dimensions of time and space create a pattern space that exceeds the capacity of manual exploration. Data mining is a better way to approach these types of problems.

### *Frequent sets/association rules*

Descriptive data mining is dominated by frequent set and association rules (FS/AR) techniques. These techniques are used in DMSS and are briefly described here. A more complete discussion of association rules and other software that uses them can be found in the article by Brown, elsewhere in this issue.

Frequent sets are sets of items that commonly appear together in transaction data. For example, if motor oil, shampoo, and chocolate candy bar X are purchased together in 100 transactions in a week from a major retailer, and any set of items purchased together more than 75 times is considered frequent, then  $A = \{\text{motor oil, shampoo, chocolate candy bar X}\}$  is a frequent set. By simple closure, all subsets of  $A$  are also frequent.

Using frequent sets, association rules can be constructed. Association rules are statements of how often certain items in frequent sets are found with other items in the same set. For example, the frequent sets {bread}, {milk}, and {bread, milk} can be used to generate the association rules {bread}  $\rightarrow$  {milk} and {milk}  $\rightarrow$  {bread}. Each are statements of conditional probability of the form  $A \rightarrow B$ , read as “B given A,” where A and B are frequent sets. If A occurs 100 times and B occurs with A 63 times, then we say the rule  $A \rightarrow B$  has confidence 63/100.

Data from retail checkout systems are used to identify sets of items that people purchase together frequently. For this reason, FS/AR analysis is often referred to as market basket analysis. In market basket analysis, association rules are used to design product placement strategies and marketing campaigns. If consumers purchase items together, especially in unexpected ways, then opportunities may exist to design campaigns that use buying patterns or item attributes to increase sales.

In DMSS, the market basket contains clinical microbiology results with NIM status along with admission, location, specimen timing, and patient demographic information. Frequent sets are computed to determine which attributes occur together even at low frequency, and association rules are constructed to describe relationships between frequent sets. Once these rules are constructed for a given time-slice of data (eg, one month) their confidences are calculated historically and prospectively and are monitored for significant changes. Rather than identifying associations with a high confidence, as might be done in a simple retail market basket analysis, DMSS uses significant changes in the confidence of an association over time as an indicator of changes in the frequency of events of interest.

For example, four-drug resistant *Acinetobacter baumannii* NIMs from lower respiratory specimens from patients in the medical intensive care unit (MICU) who were in floor location X 48 hours before specimen collection is an event that may occur with a very low frequency, say once every other month among the approximately 40 patients a month who are transferred between the two locations. If this event were to occur three times in one month among 40 patients, and this was statistically significant, DMSS would generate an alert describing the change. The frequent set for this event is

{R-drug1, R-drug2, R-drug3, R-drug4, A. baumannii, NIM,  
lower respiratory, MICU, locationX-48}

Because each of the nine items can be placed on the right or left side of an association rule,  $2^9 = 512$  rules can be generated from it. DMSS generates all rules, but uses rule templates, such as “keep all resistance traits together” to prune rules that are relatively uninformative, such as  $\{R\_drug2, NIM\} \rightarrow \{R\_drug1, R\_drug3, A\ baumannii, lower\ respiratory, MICU, locationX-48\}$ . The event described above has the association rule:

$\{MICU, locationX-48\} = > \{R-drug1, R-drug2, R-drug3, R-drug4,$   
 $A. baumannii, NIM, lower\ respiratory\}$

with a confidence history something like {month1: 0/41, month2: 1/38, month3: 0/39, month4: 1/42, month5: 3/40}. When the history is broken into two parts, for example {months1-4} and {month5}, and the average confidence differs significantly between the parts, then DMSS generates an alert. The comparison of confidences is a comparison of two proportions, which can be accomplished with a Fisher’s exact test or a  $\chi^2$  test [7].

Unfortunately, data mining systems usually generate too many, often redundant, patterns and schemes must be used to reduce the pattern load on the end user. In the example above, the nine-item frequent set also has  $2^9$  frequent subsets, all of which are generated in naïve schemes, and all of which have  $2^{(their\ number\ of\ items)}$  association rules. Even after rule templates are applied, several related alerts are redundant. To address this problem, DMSS uses an alert clustering scheme to select only the most descriptive alert for presentation [7]. Conceptually, given two alerts A and B with  $A_{rule} = A_L \rightarrow A_R$  and  $B_{rule} = B_L \rightarrow B_R$  and  $B_L \subseteq A_L$  and  $B_R \subseteq A_R$ , if the data that satisfy A are removed from B, which they also satisfy, and the resultant B is not an alert, then A captures B. For example, if  $A_{rule} = \{MICU, locationX-48\} \rightarrow \{R-drug1, R-drug2, R-drug3, R-drug4, A\ baumannii, NIM, lower\ respiratory\}$  and  $A_{conf\_hist} = \{month1: 0/41, month2: 1/38, month3: 0/39, month4: 1/42, month5: 3/40\}$  and  $B_{rule} = \{MICU, locationX-48\} \rightarrow \{R-drug3, R-drug4, A\ baumannii, NIM\}$  and

$B_{conf\_hist} = \{\text{month1: } 0/41, \text{ month2: } 1/38, \text{ month3: } 0/39, \text{ month4: } 1/42, \text{ month5: } 3/40\}$ , identical to  $A_{conf\_hist}$ , then A captures B. A is more descriptive than B, and B without A is nonexistent. For cases where  $A_{conf\_hist} \neq B_{conf\_hist}$ , a “difference history” can be analyzed using a test of two proportions [7].

Once alerts are reviewed, investigative action must be taken for effects to occur. Sometimes just showing up in the right place at the right time with the right information elicits the necessary changes in process to correct the underlying problem, even if explicit defects are never discovered. Because outbreak epidemiology is complex, alerts should be viewed as windows into broken systems. Without an alert, the need to look would not exist. Additionally, process improvements reduce alert frequency, and improvement patterns can be generated. These should be used to follow and enhance compliance with improvement recommendations.

### Operational considerations

Although DMSS is a complex system and a full description is beyond the scope of this article, a few distilled operational principles and challenges are worth discussing.

#### *Data collection*

DMSS requires useable data. The “garbage-in, garbage-out” adage is applicable. Data are obtained from the laboratory information system (LIS), the admit-discharge-transfer (ADT) system, and the hospital census system. Clinical laboratory data, especially clinical microbiology data, are poorly structured, however, and contain free text and natural language.

Clinical microbiology data (including molecular testing) and infection-associated serology data from the LIS can be obtained in three ways: custom-built LIS queries, printed reports, and HL7 messages. Custom queries can be built to specification—content and presentation can be controlled—but they require programmer resources to construct and their results must be checked against gold-standard results (usually printed reports) for completeness.

Printed reports from the LIS, specifically those used to present information to clinicians, are mostly complete, but may suppress results that are selectively reported (eg, imipenem susceptibility in *Pseudomonas aeruginosa*). Suppressed results limit the ability of frequent set and association rule algorithms to detect relationships that may exist, but these limitations are usually not significant because the results are not suppressed in cases in which the information is most useful (eg, imipenem resistance in the presence of aminoglycoside and cephalosporin resistance). Printed reports can also change format with LIS upgrades, the introduction of new tests, and the removal of discontinued tests. For these reasons, structure and content must be actively monitored for change. Printed reports can be readily

obtained in file format from printer queues (usually custom queues established expressly for file retrieval), but need to be parsed to load the data into a database. Tools such as Monarch Data Pump ([www.datawatch.com](http://www.datawatch.com)) can be useful.

Clinical microbiology HL7 messages are often poorly structured but are readily available from HL7 routers in most hospitals. Their modeling and parsing, however, require considerable sophistication. Print-structured data are often simply embedded in message segments, and therefore all challenges and considerations of print report modeling apply to HL7 message modeling. Additional challenges exist, such as identifying and modeling only applicable messages. DMSS obtains LIS data by HL7 messages.

Patient movement and census data are obtained from two sources: HL7 ADT messages and electronic census reports. Although ADT messages are rich in content, near real-time, and precise, they are transaction based and occasional message omission is not uncommon. For example, if for some reason a discharge message is not generated for a patient, the patient appears to never leave the hospital. For this reason, DMSS uses census reports obtained throughout the day to reconcile ADT data errors.

### *Data cleaning/normalization*

Once data are obtained using one of the three mechanisms above, they must be loaded into a database, quality checked, and mapped. Database design and population are beyond the scope of this article (see the article by Lyman and colleagues, elsewhere in this issue for a general discussion of database design for mining), but once data files are retrieved and checked to make sure their sizes are within normal limits and their data are from the time periods expected, data can be loaded and mapped. Mapping requires, among other things, that “SA,” “S. AURIUS,” “STAPH AUREUS,” and so forth all be mapped to “Staphylococcus aureus.” Original data are also maintained. Cardinal Health DMSS databases contain data from more than 250 hospitals and have hundreds of mappings to single organisms and specimen sources. For example, there are hundreds of terms for blood specimens, including ones with misspellings of “blood.” The management of term mapping alone requires pattern recognition and quality assurance systems. After terms are mapped, data must be checked again to make sure certain common specimens, tests, and organisms exist within statistical limits.

The next step is to impart additional meaning to the data. For example, NIM criteria are applied so that NIs, community-acquired infections, and specimen contamination proxies (indicators) are computed. If this information were not imparted to the data before pattern analysis, patterns would less reliably describe nosocomial versus community-acquired infections versus colonization or contamination. Electronic proxies for these clinical and laboratory states, like the NIM, add value to the data and make data mining

more productive. Once data are annotated with these proxies, they can be analyzed.

*Frequent set and association rule analysis and alert generation*

FS/AR analyses generally work as described above, but are typically fraught with complexity for the inexperienced practitioner. Time partitioning of the data, the organization of association rules obtained from each partition, and the ability to track changes among rules need to be handled. Once rules are stored along with their confidences in time, rules whose confidences change significantly between two single or aggregate time periods compose alerts. Rules whose confidences are changing insignificantly are ignored. Alert clustering reduces alert volume by a factor of two to four and is yet another tool used to reduce pattern overload. All data mining steps from data selection to pattern presentation need to be designed with this problem in mind. Generating too many statistically significant but often meaningless or redundant patterns leads to user exhaustion and project failure.

The final step of data mining is report preparation. In DMSS reports are prepared from clustered patterns by domain experts who select patterns by their usefulness. Pattern usefulness or interestingness is a function of clinical significance and actionability, and includes an estimate of how much information the end user can use efficiently. These are largely subjective measures that are difficult to code explicitly, but through experience we know that end users do well with 5 to 10 patterns a month, about one tenth of all clustered patterns (Table 1). DMSS pattern reports are currently presented monthly.

**Data Mining Surveillance System results**

DMSS identifies new patterns of interest and detects known outbreaks in historical data [7]. Patterns can be arbitrarily complex, and can describe everything from slow changes in simple event frequency in large populations (hospital-wide, for example), to location-specific outbreaks of 10-drug resistant *A baumannii* [7], to community outbreaks of infectious diarrhea [9]. Currently, more than 225 hospitals nationwide subscribe to Cardinal Health

Table 1  
Monthly Data Mining Surveillance System statistics by hospital

|                        | Inpatient<br>admits | Specimens | Tests     | NIMs   | CIMs    | Clustered<br>patterns | Reported<br>patterns |
|------------------------|---------------------|-----------|-----------|--------|---------|-----------------------|----------------------|
| Median                 | 1498                | 2728      | 3254      | 61     | 245     | 52                    | 6                    |
| Interquartile<br>range | 809–2368            | 1367–4289 | 1604–5170 | 26–112 | 157–424 | 27.5–83               | 3–9                  |

Specimens and tests are inpatient and outpatient.  
*Abbreviations:* CIMs, community-acquired infection markers; NIMs, nosocomial infection markers.

services that include DMSS pattern analysis, and from these hospitals more than 20 DMSS-based abstracts have been presented at national conferences.

### Future directions

In its current form, DMSS provides a practical illustration of the usefulness of data mining in health care. Access to additional electronic data could extend the model-building capabilities and usefulness of DMSS. For example, additional data about the patient origin could allow models to describe or predict significant patterns from nursing homes, zip codes, counties, and so forth. Additional electronic data, such as surgical procedure, operating room, operative time, anesthesia scores, and wound class, could increase the descriptiveness of surgery-associated patterns. Antimicrobial use data or complete blood counts could increase the sensitivity and specificity of the NIM, even if for only specific subsets of patients. Any gains in pattern specificity and marker performance, however, add data acquisition costs and require additional effort for data validation and cleansing. These requirements must be matched by a corresponding increase in the clinical usefulness of alerts and reports to justify additional development.

### References

- [1] Emori TG, Edwards JR, Culver DH, et al. Accuracy of reporting nosocomial infections in intensive-care-unit patients to the National Nosocomial Infections Surveillance System: a pilot study. *Infect Control Hosp Epidemiol* 1998;19:308–16.
- [2] Brossette SE, Hacek DM, Gavin PJ, et al. A laboratory-based, hospital-wide, electronic marker for nosocomial infection. *Am J Clin Pathol* 2006;125:34–9.
- [3] Kahn MG, Steib SA, Fraser VJ, et al. An expert system for culture-based infection control surveillance. *Proc Annu Symp Comput Appl Med Care* 1993;171–5.
- [4] Mitchell Tom. *Machine learning*. McGraw Hill; 1997.
- [5] Brossette SE, Sprague AP, Hardin JM, et al. Association rules and data mining in hospital infection control and public health surveillance. *J Am Med Inform Assoc* 1998;5:373–81.
- [6] Brossette SE, Moser SA. Application of knowledge discovery and data mining to intensive care microbiologic data. *Journal of Emerging Infectious Diseases* 1999;5:454–7.
- [7] Brossette SE, Sprague AP, Jones WT, et al. A data mining system for infection control surveillance. *Methods Inf Med* 2000;39:303–10.
- [8] Peterson LR, Brossette SE. Hunting healthcare associated infections from the clinical microbiology laboratory: passive, active, and virtual surveillance. *J Clin Microbiol* 2002;40:1–4.
- [9] Peterson LR, Hacek DM, Rolland D, et al. Detection of a community infection outbreak with virtual surveillance [letter]. *Lancet* 2003;362(9395):1587–8.