Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 36 (2003) 379–388
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Data mining issues and opportunities for building nursing knowledge

Linda Goodwin,[a,*] Michele VanDyne,[b] Simon Lin,[a] and Steven Talbert[a]

[a] *Duke University, Durham, NC 27710, USA*
[b] *IntelliDyne, Inc., Kansas City, MO, USA*

Received 29 August 2003

Abstract

Health care information systems tend to capture data for nursing tasks, and have little basis in nursing knowledge. Opportunity lies in an important issue where the knowledge used by expert nurses (nursing knowledge workers) in caring for patients is undervalued in the health care system. The complexity of nursing's knowledge base remains poorly articulated and inadequately represented in contemporary information systems. There is opportunity for data mining methods to assist with discovering important linkages between clinical data, nursing interventions, and patient outcomes. Following a brief overview of relevant data mining techniques, a preterm risk prediction case study illustrates the opportunities and describes typical data mining issues in the nontrivial task of building knowledge. Building knowledge in nursing, using data mining or any other method, will make progress only if important data that capture expert nurses' contributions are available in clinical information systems configurations.

Keywords: Data mining; Data mining issues; KDD; Knowledge discovery in data; Knowledge base; Clinical information systems; Nursing knowledge; Expert nurses; Nurse knowledge workers

1. Introduction

Expert nursing knowledge workers hold an important key to both high quality patient care and cost effectiveness in health care, but their extensive knowledge base has not been clearly articulated and remains undervalued in the health care system. Nursing faces two opportunities to build important knowledge for clinical practice: (1) valuing and explicating the knowledge of expert nurses, and (2) using data mining methods for building nursing knowledge. Each of these opportunities involves multiple issues that create enormous challenges for the profession. As nursing better identifies, structures, and standardizes nursing data, the clinical information systems that collect those data will provide a foundation for data mining analyses that have the potential to build nursing knowledge regarding the relationships between data, nursing interventions, and patient outcomes.

In this paper, we first describe two opportunities related to building nursing knowledge for clinical practice. Second, we provide a brief overview of selected data mining techniques as background to a case study that utilized the chosen methods. Third, using a preterm risk prediction case study, we describe typical data mining issues.

2. Issue and opportunity #1: valuing what expert nurses know

Problems in nursing data are generally twofold in nature. First, the data may not be collected in any permanent form. Nursing's data are often transferred in the verbal transactions nurses exchange with the health care team and is not recorded, either in paper or computerized format. Much of nursing's data remain in paper format and is not captured in electronic form or not stored in such a way that it can be easily retrieved. Worse yet, it may exist in computerized form but data entry quality includes large numbers of errors and/or missing values.

Many, perhaps most, health care systems still focus on what nurses DO and few value what nurses KNOW. In an era of unprecedented health care reform, and

* Corresponding author. Fax: 1-919-681-8899.
E-mail address: lkgoodwin@yahoo.com (L. Goodwin).

1532-0464/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2003.09.020

while the world deals with an information revolution and demand for knowledge workers, nursing knowledge remains poorly articulated and undervalued. Nursing's knowledge base covers both breadth and depth that includes foundations in science (biology, psychology, sociology, math, chemistry, anatomy, physiology, microbiology, genetics, and more) as well as complex concepts of health and illness that emerge from nursing theory, nursing science, nursing process, patient–environment interaction, and human responses to actual and potential diseases. An important fact that goes undervalued in health care is that nurses monitor medical conditions in their patients, and are often the first to detect and diagnose serious medical problems. Confusion arises in that a nurse's scope of practice is not authorized for medical diagnosis, yet they need to have at least a modicum of medical knowledge if they are to provide safe and quality care to their patients. Adding to the complexity of nursing's knowledge base, nurses' knowledge of complex systems is critical to coordinating all elements of the patient's care which includes consideration of the patient's family and community, and coordinating an interdisciplinary and often poorly organized care environment within the context of organizational, legal, ethical, social, political, religious, spiritual, economic, and technology forces that impact the patient's care. Given the breadth, depth, and complexity of nursing's' knowledge base, it is not surprising that nurses with advanced education are typically more expert than others in applying this extensive knowledge to patient care.

Expert nursing knowledge workers hold an important key to both excellent quality and cost effectiveness in health care. For example, Brooten et al. [1] reported results of a randomized trial where high-risk pregnant women received half of their prenatal care via home visits from a master's prepared clinical nurse specialist. There were no differences between groups for race, marital status, education, and public health insurance. The intervention group had lower infant mortality, fewer preterm births, more twin pregnancies carried to term, and significantly fewer hospital days with an estimated savings of $2,880,000. Advanced practice nurses clearly made a significant difference in patient outcomes in Brooten's study.

Advanced practice nurses have been undervalued and underfunded in health care for more than two decades, although they typically have a master's degree and many have expertise and a knowledge base that is not found anywhere else in the health care system. Nursing and society would be well served through development of goals and strategies that articulate the knowledge used by expert nurses in caring for patients. But expert care is founded on complex and embedded knowledge that the nursing profession has been unable to capture in any meaningful way to communicate its value to the larger health care system. The difficulty encountered in studying nurse experts, or any domain experts, lays within the experts themselves, and the process by which the person becomes an expert. Experts have two kinds of knowledge: the knowledge they use to explain the task or problem, and the knowledge they use that actually performs the task. Researchers found that the method the expert tells you they used for solving the problem (called a reconstructed method of problem-solving) is usually different from the experiential knowledge that was used during the actual method of solving the problem [2]. Johnson [3] called this the 'paradox of expertise' in that the very knowledge we wish to capture is the knowledge the expert can least discuss.

The problem becomes one of extracting knowledge from an expert who can only tell you what they use to *explain* the task (not the knowledge they use to actually *perform* the task). In nursing, this discrepancy is a fundamental problem for articulating nursing's knowledge bases. And without a clearly articulated and valued knowledge base, nursing will continue to struggle with repetitive historical issues. Informatics research methods that include techniques for data mining and knowledge discovery in data (KDD) offer new tools and opportunities for knowledge development in nursing and other domains.

3. Issue and opportunity #2: knowledge discovery in (patient) databases

Knowledge discovery in data or databases (KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from raw data [4]. Knowledge discovery uses data mining and machine learning techniques that have evolved through a synergy in artificial intelligence, computer science, statistics, and other related fields [5]. Although there are technical differences, the terms 'machine learning,' 'data mining,' and 'KDD' are often used interchangeably. Data mining is a powerful methodology that can assist in building knowledge directly from clinical practice data for decision-support and evidence-based practice in nursing. As data mining studies in nursing proliferate, we will learn more about improving data quality and defining nursing data that builds nursing knowledge. The simplicity of Fig. 1 belies the difficulty of work that will be required to build knowledge in nursing. One of many challenges before us is to find ways to use data mining tools and methods to develop knowledge bases

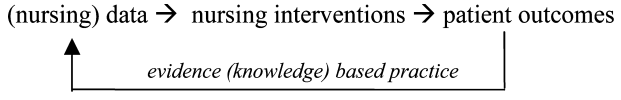

Fig. 1. Opportunity for (nursing) data to build knowledge.

that clearly identify patterns and important links between data, nursing interventions, and patient outcomes. As we better understand these important links, nurses may be able to use this knowledge to improve quality of care and patient outcomes.

Health care now collects data in gigabytes per hour volume. Data mining can help with data reduction, exploration, and hypothesis formulation to find new patterns and information in data that surpass human information processing limitations. Review of the literature finds a proliferation of articles that apply data mining and KDD to a wide variety of health care problems and clinical domains and includes diverse projects related to cardiology [6], cancer [7,8], diabetes [9], finding medication errors [10], and many others.

Over the past two decades, it is clear that we have been able to develop systems that collect massive amounts of data in nursing and health care, but now what do we do with it? Data mining methods use powerful computer software tools and large clinical databases, sometimes in the form of data repositories and data warehouses, to detect patterns in data. Within data mining methodologies, one may select from an extensive array of tools that include, among many others, neural networks, decision trees, and rule-based (if–then) systems. Application and utilization of data mining and KDD in nursing requires an understanding of the methods available for knowledge base development.

4. Data mining overview

To provide context for data mining opportunities, an overview of selected data mining methods is provided. Following this overview, a case study in building knowledge will demonstrate application of data mining methods for a preterm risk problem domain. Important issues that impact data mining methods are described in the context of the preterm case study.

Carbonell [11] identified four major machine learning paradigms as (1) *inductive* learning, (2) *analytic* learning, (3) *genetic* algorithms, and (4) *connectionist* learning methods. He also provided a definition of machine learning. ''Learning can be defined operationally to mean the ability to perform new tasks that could not be performed before or perform old tasks better (faster, more accurately, etc.) as a result of changes produced by the learning process.''

Carbonell [11] identified the *inductive* paradigm as one which generally works from a set of data where the classification of an example (patient) is known, and the system learns how to discriminate between different classifications given the data values associated with various patterns in the data. Classification and regression trees (CART) are just one of many approaches in the inductive paradigm, where relationships amongst data are visually graphed in decision-tree fashion. The ''tree'' output is reasonably easy to understand and interpret and helps clinical experts make sense of the data. Fig. 2 illustrates one form of inductive paradigm output, which is a tree structure.

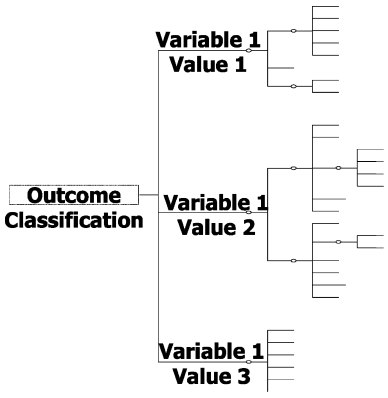

Fig. 2. Inductive paradigm tree representation.

According to Carbonell [11], *analytic* learning occurs from a few exemplars using a rich underlying domain theory. The inference method is deductive rather than inductive, and the focus is on improving the efficiency of the system without sacrificing accuracy or generality.

The *genetic* paradigm [12] is inspired by genetics and evolution via natural selection. Concept descriptions are represented as individuals in a population, and induction occurs through the recombination (reproduction) of these individuals. A rule discovery process in classifier systems is sometimes implemented using genetic algorithms [11]. The genetic algorithm selects high strength classifiers as parents and forms offspring from them by recombining the components. The advantages of genetic algorithms are their capacity to deal with mutually contradictory, partially confirmed rules which allows the system to deal with noisy data.

A *connectionist* paradigm frequently uses programs called ''neural networks'' that were inspired by the way densely interconnected, parallel structures of the mammalian brain process information. Neural networks purport emulation of neural pathway development and adaptive learning in biological nervous systems. The analogy is more symbolic than real, as neural networks are actually collections of mathematical formulas. Hinton [13] described the realm of *connectionist* learning procedures in which the goal is to discover efficient learning techniques (Fig. 3).

Support Vector Machine (SVM) techniques emerged in the 1990s, based on statistical learning theories developed by Vapnik [14]. SVM separates samples with a hyperplane. SVM is a robust classifier with the capacity to handle noisy and high dimensional data. In most of the cases, the performance of SVM either matches or outperforms other machine learning approaches (Fig. 4).

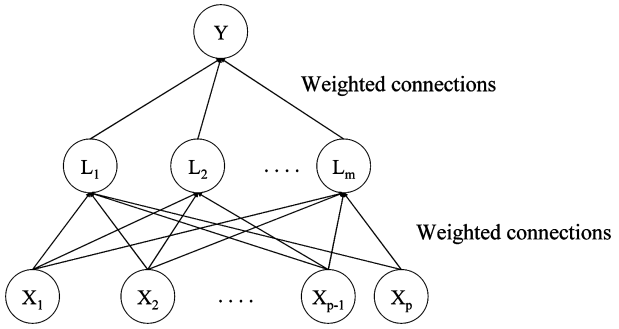

Fig. 3. Connectionist paradigm neural network representation. Bottom layer (X) nodes are inputs that are weighted and computed in a hidden middle layer (L). The output node (Y) is at top. The neural network can be trained to do either regression (Y as continuous variable) or classification (Y as logical variable).

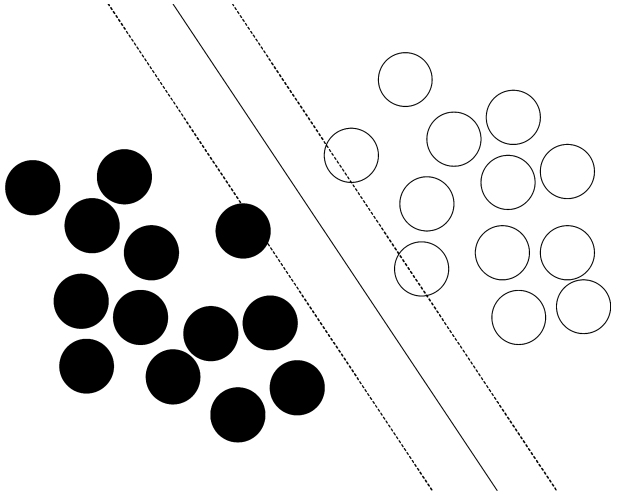

Fig. 4. Support vector machine representation. A separating hyperplane (solid decision line) in two dimensions with maximal margin (dashed line) classifies the dots into two groups. The data points on the dashed lines are the support vectors.

5. Application of data mining methods

The case study that follows experimented with traditional statistical analyses in combination with inductive (CART, rule generation) and connectionist (neural network) learning methods, as well as Support Vector Machines and a genetic (bucket brigade) algorithm to increase predictive accuracy.

5.1. Preterm knowledge discovery case study

Research interest in preterm birth prevention began with questions that arose during the primary author's clinical nurse specialist role in the early 1980s. An early project built a prototype expert system that used both the clinical nurse specialist's knowledge and items proposed by Creasy and Herron [15] but the prototype system failed miserably in trying to predict preterm birth risk in real patients. This outcome was surprising and discouraging, since the system used prevailing clinical wisdom. It became apparent that traditional methods for extracting knowledge from experts would not work for preterm birth prediction, and after more than a decade of use in clinical practice, paper-based preterm risk scoring tools were found unreliable and invalid [16,17]. Machine learning techniques emerged in the 1980s and offered a data-driven approach to this complex preterm birth problem (Table 1).

Fig. 5 provides a model for building knowledge with data mining/KDD methods. Foundational to all research, including KDD studies, is *data*. Preprocessing the data includes multiple steps to assure the highest possible data quality, thus efforts are made to detect and remove errors, resolve data redundancies, and (in this case study) to remove patient identifiers. Data are analyzed using both statistical and data mining methods to produce *information*; output formats will vary depending upon the method used. Predictive modeling efforts are iterative, thus statistical and data mining results are repeated with different permutations until the best results (metrics) are obtained.

It quickly becomes apparent that building *knowledge* in complex domains is a nontrivial task. In spite of nearly a decade of research, additional studies are needed to improve preterm predictive accuracy before reliable and valid models are available for clinical practice. As seen in Fig. 5, the methods for building knowledge in nursing use both the information derived from statistical and data mining analyses of the data, combined with iterative analyses that optimize performance metrics. Only those models that are validated by experts are retained in the knowledge base for system testing and verification. Future work will test decision support systems that are embedded in real-time clinical information systems and used by nurses, and other care providers, in caring for pregnant women. Until we can better predict who is at risk for preterm birth, it is difficult to tailor appropriate interventions for pregnant women. Over time, the goal is to improve patient outcomes by prolonging gestation and preventing preterm birth. The knowledge development approach in Fig. 5 is a data-driven process that will build both decision support and an evidence base for clinical practice.

Data. Duke's TMR perinatal database provided data for more recent studies, and is the only known clinical database that electronically collected data on pregnant women for more than two decades. The final research data set included 1622 variables and 19,970 patients after cleaning and filtering procedures were completed for data extraction of 71,753 records and approximately 4000 potential variables per patient.

Information. Data analyses focused on 1232 variables (of 1622) collected between 10 and 20 weeks of pregnancy since this time frame would offer the most

Table 1
Knowledge discovery program of research in preterm birth prevention

Type of study	Funding agency	Data sources	Information (KDD method)	New knowledge
Dissertation 1990–1992	None	$n = 2436$ with 77 variables. Race/ethnicity not recorded.	ID3 inductive learning	Machine learning methods were validated as accurate. 88 production rules were not validated.
Small Business Innovation Research (SBIR) 1992–1993	National Institutes of Health/National Institute of Nursing Research (NIH/NINR)	3 Databases: • $n = 2436$ with 77 variables • $n = 3186$ with 52 variables • $n = 13,216$ with 129 variables Race/ethnicity not consistently recorded—primarily Caucasian	(LERS = [Inductive] Learning from Examples using Rough Sets)	520 expert verified production rules yielded 53–89% predictive accuracy
Extramural Research Project (RO1) 1997–2003	National Institutes of Health/National Library of Medicine (NIH/NLM)	Duke's TMR patient record database • $n = 19,970$ with 1622 variables 55% Black; 3% Hispanic; 1% Asian; 1% Native American; 39% Caucasian; 2% Unknown	Logistic regression, neural networks, classification and regression trees (CART), and experimental machine learning software methods yielded similar results	7 demographic variables yielded 0.72 area under the curve (ROC analyses) where 1.0 is perfect prediction. Adding more than a thousand additional variables added only 0.3 area under the curve.
Under review	National Institutes of Health/National Institute of Nursing Research (NIH/NINR)	• Prospective data collected by pregnant women • Duke's OB TraceVue electronic patient records	Multiple methods to analyze currently missing variables	Building knowledge for improved predictive accuracy

opportunity for intervention to prolong gestation and prevent preterm birth. Receiver Operator Characteristics (ROC) analyses are particularly useful in a paradigm that deals with prediction, where the probability of detection, or the true positive rate and probability of false alarm, or the false positive rate can be graphed. Fig. 6 shows that ROC curves plot sensitivity (Y axis) against 1 minus the specificity (X axis) providing a clear visualization of area under the curve (AUC). The greater the accuracy, the greater the AUC (1.0 is perfect prediction).

Results of a plethora of data analyses using multiple data mining methods produced similar information for Duke's TMR perinatal data, where all ROC curves overlaid each other. This finding was a surprise, since we anticipated that different methods would yield different results. But our findings confirmed that building knowledge in our preterm risk domain is data-driven, and not a function of any particular method used to analyze the data. This method-independent result is significant in that it may lessen the requirement for analyzing data with multiple KDD methods, and permit efficiency in the analysis process using one or two techniques.

Knowledge. Clinical experts validated the 0.72 area under the curve model (see Fig. 6) as valid for inclusion in the preterm risk knowledge base. This finding supports ongoing concerns and questions about socio-demographic factors associated with race, poverty, and disparity in preterm risk [18]. Demographic variables are non-invasive and are low cost, since they should be collected on all patients anyway. Where seven demographic variables produced results of 0.72 area under the curve (ROC), the addition of more than a thousand variables added only 0.03 area under the curve (ROC). While 0.72 is a very respectable result for preliminary data mining analyses, current and future studies seek to increase the area under the curve for improved predictive accuracy, by adding relevant (and currently missing) variables to our predictive models.

Important issues need careful consideration when using data mining methods for building knowledge in nursing and other domains. Issues include dealing with patient privacy, data quality, (non)standardized language, missing values, and other typical problems in any research study. Additional issues that are somewhat unique to data mining methods include dimensionality reduction issues and metrics issues where overfitting data can be a problem.

5.2. Data mining issue: patient privacy

Most data mining research is interested in aggregate data that can be accomplished without patient identifiers. However, in spite of careful de-identification procedures in our data cleaning processes, a very young

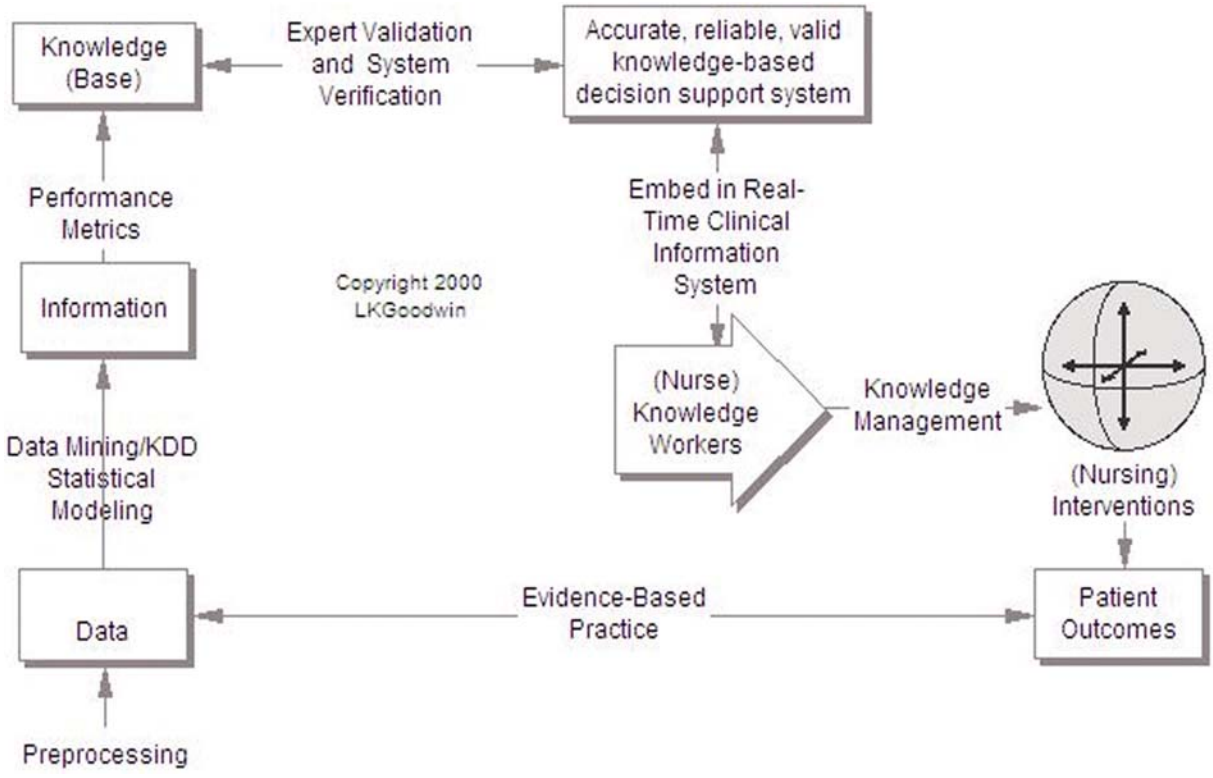

Fig. 5. Data mining methods for nursing knowledge development.

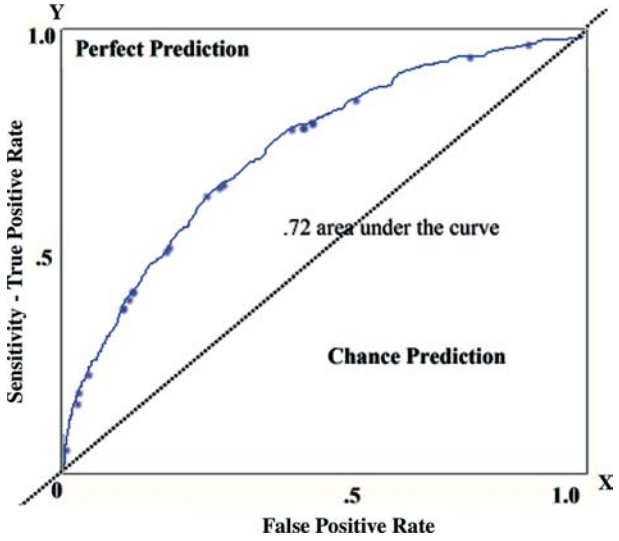

Fig. 6. Seven demographic variables produced 0.72 area under the curve.

pregnant girl was still individually identifiable by some providers because of her age, and we are reminded that privacy issues may persist even in de-identified large clinical databases. The young girl's record was removed from our research data. Patients and health care consumers are increasingly concerned about the privacy of their personal health information. In a spirit of entrusted stewardship, our data mining research carefully attempted to create completely anonymous data before analyses were begun. Comparing previously used "anonymization" methods with emerging HIPAA de-identification standards found that earlier procedures required minimal modification to be HIPAA compliant [19].

5.3. Data mining issue: data quality

Problems with clinical data quality and cleaning have been described in all previous studies and persist with current work. Anticipating that data will be 100% complete and error free is unrealistic when working with patient data that are collected in complex health care systems. Cleaning the data proved a nontrivial and tedious task that consumed approximately 16 months to complete the research data warehouse.

As might be expected, both anticipated and unanticipated problems and issues arose where data extractions contained 71,753 patient records and approximately 4000 potential variables per patient. Data error identification was both an automated and a manual process, and required an iterative procedure that drew upon expertise from the clinical experts as well as statistical experts and the data warehouse engineer. Errors that detected out-of-range values (for example, a systolic blood pressure of 700) were identified by the clinical

experts and eliminated from the research data sets. Errors where a variable included inconsistently recorded text required an iterative extraction and programming solution; clinical experts reviewed the text extraction and provided guidelines for converting data for consistency, coding, or deleting the variable if data conversion was not possible.

The 800 MB volume of data in our Microsoft SQL research data repository is relatively small when compared to many data warehouses. Still, the data volume required vast amounts of computer processing time that consumed multiple machines for extensive periods of time. The data volume yielded what is called 'noisy' data. According to Quinlan [20], it is important to eliminate noise affecting classification in the training set, but it is not worthwhile to expend effort in eliminating noise from the training set if there will be noise in the testing set. It may be better to simply dispose of noisy, less important attributes since the payoff in noise reduction increases with the importance of the attribute. Using a data volume of 1,232 variables resulted in noise-to-signal problems that we believe masks the effect of other potentially informative variables, and provides the basis for ongoing work.

5.4. Data mining issue: standardized language (or the lack thereof)

Articulating knowledge from nursing's data is obstructed by a "Tower of Babel" phenomenon that results from lack of standardization in how health care, in general, and nursing, in particular, collect and label data variables and values. Even where high quality computerized data does exist, lack of standardized data and terminology interferes with our ability to mine the data for patterns and patient outcomes.

Our preterm studies provide abundant examples of the problems that are generated when data and terminologies are not standardized. Remember the Duke TMR data included 1622 variables. When comparing specific data items for prenatal database fields between TMR and three other databases used in prior studies, only 10 variables used consistent data types and terminology between all four databases. And yet all the data are collected for a prenatal population with fairly well understood clinical parameters. Plans to merge the databases were abandoned, since the data were so disparate that merging databases would yield an enormous volume of missing values.

Clearly, there is both a challenge and an opportunity to standardize data and terminology in nursing. Numerous efforts have been made toward this goal and those that meet rigorous development criteria are recognized by the American Nurses' Association (ANA) Council for Nursing Practice Information Infrastructure (CNPII) [21]. Before data mining and KDD methods can be used effectively in nursing, appropriate, structured, and standardized nursing data elements must be captured in clinical information systems. The currently ANA recognized nursing data sets and vocabularies provide a necessary but not yet sufficient foundation for advanced clinical data mining to yield knowledge of the value of nursing's impact on patient outcomes. It is important to emphasize that we must standardize more than simple task language. Capturing simple nursing tasks fails to document the complexity of nursing's knowledge base and also fails to provide the rich data set needed to build knowledge for nursing practice and improved patient outcomes.

5.5. Data mining issue: repeated measures

The number of (1622) variables appear large, but are actually seven demographic variables plus approximately 400 clinical variables measured at each prenatal visit. Prior work rolled each repeated measure out as a new variable (column) in the data set. However, the TMR data volume met with early hardware and software failures. Thus a summary and data reduction decision was made; within each 10-week prenatal block, repeated measures of interval data were collapsed into a mean, median, mode, standard deviation, range, and frequency count for how many times the variable was measured. The trade-off for data loss resulted in manageable data sets that the machines could process in hours rather than days.

5.6. Data mining issue: missing values

Researchers anticipate and develop strategies for dealing with missing values in any large data set. Common strategies for dealing with missing values include: (1) elimination of the records with missing values, (2) substitution of the variable mean, and (3) substitution of a Bayesian frequency count for the most common value of a particular variable. Our research finds no statistically significant differences in any of the missing values treatments with prenatal data sets used in our data mining studies [22–24].

5.7. Data mining issue: missing variables

In spite of ever-increasing and extensive electronic patient records in prenatal populations, important variables are missing. For example, most prenatal information systems do not collect variables for nutritional status, fatigue, stress, depression, oral hygiene, social support, intimate partner violence, or sexual practices and yet research studies report these variables associated with preterm birth outcomes [25]. No matter how large the data set (our earlier studies included $n = 18{,}890 - \text{SBIR}, 19{,}970 - \text{RO1}$) results will not yield

the desired knowledge base if important variables needed for predictive modeling are missing. Participation by clinical experts is absolutely essential for effective clinical data mining, to identify and capture missing variables for building knowledge. This is true in a preterm problem domain as well as any number of other problem domains where nurses play an important role in patient outcomes.

5.8. Data mining issue: overfitting data

Data mining problems are often created due to overfitting a model to a specific data set. Data are divided into training versus testing sets in order to manage the overfitting problem. The larger training set (usually 75–90% of the data) is used to train the models while the remaining (testing) data (10–25%) are set aside for final evaluation of the model. Training sets often achieve optimistic and positive results that are not replicated in the testing data. Since the true value of data mining lies in its ability to predict "unseen" data, every effort should be taken to prevent overfitting. Two methods are commonly used to improve predictive model performance on test data. First, the data set is randomly divided into training and testing data sets. In addition to randomly choosing and isolating the testing data set, a process using cross validation methods with the training data helps avoid overfitting and provides good estimates of overall model accuracy [26].

5.9. Data mining issue: variable/feature selection

Hardware and software limitations have historically forced data analysts to "pre-select" those variables they believed were best/most relevant. Data mining researchers tend to use experts for dimensionality reduction through a process called "feature selection" to reduce the number of attributes (variables) for analysis. Feature selection is frequently a daunting task when deciding which attributes are important. Dalal et al. [27] described that human pre-selection bias for Challenger space shuttle data ultimately ended in disaster, and that when analyses with *all* available data were conducted, the O-ring failure was accurately predicted. Health care data mining will continue to struggle with this issue since there are no easy answers and solutions are often context and domain dependant.

5.10. Data mining issue: dimensionality reduction

Data mining research seeks to reduce redundancy in data and thereby reduce its complexity and dimensionality; the larger the data set, the more complex is its dimensionality. Once the data warehouse was extracted and anonymized, the next step was to remove infant data from the research data set, since most infant data was recorded after birth, and would not be helpful for pregnancy prediction models. Retained infant data included information recorded during pregnancy for multiple gestations and sometimes for infant sex. Eliminating infant data reduced the number of records by nearly half, and also diminished the number of variables by eliminating many variables that were infant-specific. Multiple tables in the data warehouse contained redundant entries for the same (or a similar) variable. Data redundancy issues were managed using experts who had 10–15 years experience in working with the TMR system and reviewed equivalent data elements to identify the 'best' data source for redundant variables.

Dimensionality reduction for our research began with decisions to avoid human pre-selection of features (variables). Based on descriptive statistics, all binary variables that had mean gestation periods less than 37 weeks and all continuous variables that accounted for more than 1% of the variance in the output were included for subsequent analyses. The outcome variable (GEST_DEL = weeks gestation at delivery) was originally categorized as follows:

IF GEST_DEL < 20 then DELETE; IF GEST_DEL < 37 THEN PRETERM $= 1$; ELSE PRETERM $= 0$;

Subsequent work analyzed 3-group classifications to distinguish between very preterm (20–31.9 weeks), preterm (32–37 weeks), and full term (37+ weeks).

Dimensionality reduction was highly successful since 1232 variables were reduced to seven demographic variables with a 0.72 area under the curve (ROC, see Fig. 6). However, using such a large data volume of variables resulted in noise-to-signal problems that we believe masks the effect of other potentially informative variables. Ongoing research continues to apply new algorithms to search for other informative variables that improve the predictive value of the models.

5.11. Data mining issue: metrics

Data mining typically uses an accuracy metric to analyze output (number of accurately predicted divided by the total number of cases). However, where approximately 90% of women deliver full term, this skews accuracy ratings and provides a false sense of highly accurate predictions, while it often fails to detect women at risk for preterm birth.

In our preterm risk problem domain, sensitivity and positive predictive value (PPV) are the performance metrics of greatest interest, as they indicate our ability to accurately detect women at risk for preterm birth (TP). We are much less concerned with a predictive error where a pregnant woman is predicted for preterm delivery and actually delivers full term (FP); false positive predictions have minimal risk for the infant, family, and care provider and associated costs of extra prenatal care are minimal in comparison with care and risks

associated with false negative (FN) predictions. False negatives (FN) occur when full term delivery is predicted but the woman actually delivers preterm, which may result in potential mortality and morbidity impacts on the baby and family, and a liability risk for the care provider. The problem of increasing sensitivity while avoiding false negative predictions will be critical for developing decision support systems for clinical practice.

5.12. Data mining relationship with statistics

Skeptics sometimes argue that data mining is a fishing expedition, rather than a scientific method. This attitude stems from Selvin and Stuart's [28] description of "unfettered exploration of data" as data dredging or fishing. Thirty years later, many statisticians have adopted Tukey 's [29] philosophy of exploratory data analysis and acknowledge that model search is an important step in the modeling process.

Some have argued that data mining does not need to bother with statistical assumptions and sample-to-population inferences, because the samples are large enough to be considered the population itself. Interpreting an individual p value is easy; however, data mining often generates thousands of p values which are difficult to interpret, especially when considering that thousands of hypotheses can also be generated for testing in large data sets. Given typically large sample sizes in data mining, statistical significance is often achieved but expert interpretation finds the significance has no clinical merit or validation. Statistical significance does not adequately address whether the results in a given study will replicate [30]. Thus, in data mining, replicability of results is frequently of greater interest than statistical significance. Procedures for splitting data into training versus testing sets that use sub-sampling and cross-validation methods are used to analyze replicability of results.

6. Discussion

It is important for nurses, and all health care providers who document in the patient's record, to think about the long-term implications of clinical data mining. As the profession moves forward in defining and capturing an improved nursing data set in health care information systems, every effort should be made to minimize data errors and missing values. Administrators must consider the importance of providing their employees with adequate technical, clinical, and psychological support that keeps users motivated to perform quality data entry into databases that feed clinical data repositories and warehouses. Data that were not entered or were entered inaccurately could have a negative impact when statistical and data mining output is used for both administrative and clinical decision-making. The old 'garbage-in... garbage-out' cliché still holds true!

Increasing use of computers and clinical information systems creates explosive growth in patient data. Nurses, and others in the health care system, are inundated with an overload of data and information that can interfere with decision-making. A seminal work by Grier [31] found that overwhelming volumes of data interfered with clinical judgment and decision-making in nursing.

Data mining methods offer solutions to help manage data and information overload and build knowledge for information systems and decision support in nursing and health care. As with any method for dealing with complex problem domains, data mining deals with typical research issues as well as a few that are unique to data mining methods, but careful planning and rigorous attention to managing those issues will yield results for nursing knowledge development.

7. Conclusions

Nursing has opportunities to explicate what expert nurses KNOW and work toward identifying and structuring nursing data, as well as developing standardized nursing terminologies. As nursing better identifies, structures, and standardizes nursing data, the clinical information systems that collect those data will provide an opportunity for data mining methods to assist with building nursing knowledge. Data mining methods and a model (Fig. 5) for nursing knowledge base development have been valuable for building knowledge in a preterm birth risk domain. But building knowledge is a nontrivial and tedious task with inherent issues in the data mining process. Building knowledge in nursing, using data mining or any other method will make significant progress only if important data that incorporate expert nurses' *knowledge* are made available in clinical information systems configurations.

References

[1] Brooten D, Youngblut JM, Brown L, Finkler SA, Neff DF, Madigan E. A randomized trial of nurse specialist home care for women with high-risk pregnancies: outcomes and costs. Am J Manag Care 2001;7(8):793–803.
[2] Ericsson K, Simon H. Protocol analysis: verbal reports as data. Cambridge, MA: MIT Press; 1984.
[3] Johnson P. What kind of expert should a system be? J Med Philos 1983:77–97.
[4] Fayyad U. Advances in knowledge discovery and data mining. Cambridge, MA: MIT Press; 1996.
[5] Mitchell TM. Machine learning. New York: McGraw-Hill Science/Engineering/Math; 1997.
[6] Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods

for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. J Invest Med 1995;43(5):468–76.
[7] Penson D, Albertsen P. Lessons learnt about early prostate cancer from large scale databases: population-based pearls of wisdom. Surg Oncol 2002;11(1–2):3–11.
[8] Stephan C, Cammann H, Semjonow A, et al. Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies. Clin Chem 2002;48(8):1279–87.
[9] Breault J, Goodall C, Fos P. Data mining a diabetic data warehouse. Artif Intell Med 2002;26(1):37–54.
[10] Rudman W, Brown C, Hewitt C, et al. The use of data mining tools in identifying medication error near misses and adverse drug events. Top Health Inf Manag 2002;23(2):94–103.
[11] Carbonell J. Introduction: paradigms for machine learning. In: Carbonell J, editor. Machine learning: paradigms and methods. Cambridge, MA: MIT Press; 1989. p. 1–9.
[12] Booker L, Goldberg D, Holland J. Classifier systems and genetic algorithms. In: Carbonell J, editor. Machine learning: paradigms and methods. Cambridge, MA: MIT Press; 1989. p. 235–82.
[13] Hinton G. Connectionist learning procedures. In: Carbonell J, editor. Machine learning: paradigms and methods. Cambridge, MA: MIT Press; 1989. p. 185–234.
[14] Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.
[15] Creasy R, Herron M. Prevention of preterm birth. Semin Perinatol 1981;5(3):295–302.
[16] Creasy RK. Preterm birth prevention: where are we. Am J Obstet Gynecol 1993;168(4):1223–30.
[17] Edenfield S, Thomas S, Thompson W, Marcotte J. Validity of the Creasy risk appraisal instrument for prediction of preterm labor. Nurs Res 1995;44(2):76–81.
[18] Goodwin LK, Iannacchione MA, Hammond WE, Crockett PW, Maher S, Schlitz K. Data mining methods find demographic predictors of preterm birth. Nurs Res 2001;50(6):340–5.
[19] Goodwin LK, Prather JC. Protecting patient privacy in clinical data mining: a comparison of HIPAA de-identification with careful data filtering procedures. J Health Inf Manag 2002;16(4):62–7.
[20] Quinlan J. The effect of noise on concept learning. In: Michalski RS, Carbonell J, Mitchell T, editors. Machine learning: an artificial intelligence approach, vol. II. Los Altos, CA: Morgan Kaufmann; 1986. p. 149–66.
[21] Coenen A, McNeil B, Bakken S, Bickford C, Warren J. Toward comparable nursing data: American Nurses Association criteria for data sets, classification systems, and nomenclatures. Comput Nurs 2001;19(6):240–6.
[22] Grzymala-Busse J, Woolery LK. Improving prediction of preterm birth using a new classification scheme and rule induction. In: Proceedings of the The Annual Symposium on Computer Applications in Medical Care; 1994. p. 730–4.
[23] Grzymala-Busse J, Goodwin LK. Predicting preterm birth risk using machine learning from data with missing values. Bull Int Rough Set Soc 1997;1:17–21.
[24] Grzymala-Busse J, Gryzmala-Busse WJ, Goodwin LK. A comparison of three closest fit approaches to missing attribute values in preterm birth data. Int J Intell Syst 2002;17(2): 125–34.
[25] Goodwin L. Building nursing knowledge to reduce preterm disparities. Grant proposal under review; January 2, 2003.
[26] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence; 1995.
[27] Dalal SR, Fowlkes EB, Hoadley B. Risk analysis of the space shuttle: pre-challenger prediction of failure. J Am Stat Assoc 1999;82:112–22.
[28] Selvin H, Stuart A. Data dredging procedures in survey analysis. Am Stat 1966;20(3):20–3.
[29] Tukey J. Exploratory data analysis. Reading, MA: Addison-Wesley; 1977.
[30] Carver R. The case against statistical significance testing. Harv Educ Rev 1978;48:378–99.
[31] Grier M. Information processing in nursing practice. Annual review of nursing research. New York: Springer; 1985. p. 265–87.