Journal of Korean Academy of Nursing (2006) Vol. 36, No. 4, 652–661

Knowledge Discovery in Nursing Minimum Data Set Using Data Mining

Myonghwa Park, RN, PhD[1], Jeong Sook Park, RN, PhD[2], Chong Nam Kim, RN, PhD[3], Kyung Min Park, RN, PhD[4], Young Sook Kwon, RN, MSN[5]

Purpose. The purposes of this study were to apply data mining tool to nursing specific knowledge discovery process and to identify the utilization of data mining skill for clinical decision making.

Methods. Data mining based on rough set model was conducted on a large clinical data set containing NMDS elements. Randomized 1000 patient data were selected from year 1998 database which had at least one of the five most frequently used nursing diagnoses. Patient characteristics and care service characteristics including nursing diagnoses, interventions and outcomes were analyzed to derive the meaningful decision rules.

Results. Number of comorbidity, marital status, nursing diagnosis related to risk for infection and nursing intervention related to infection protection, and discharge status were the predictors that could determine the length of stay. Four variables (age, impaired skin integrity, pain, and discharge status) were identified as valuable predictors for nursing outcome, relived pain. Five variables (age, pain, potential for infection, marital status, and primary disease) were identified as important predictors for mortality.

Conclusions. This study demonstrated the utilization of data mining method through a large data set with standardized language format to identify the contribution of nursing care to patient's health.

Key Words : Nursing minimum data set, Knowledge discovery, Data mining

INTRODUCTION

Background and significance

Interest in patient outcomes in response to intervention of health care has grown rapidly in the last decade (Delaney, Ruiz, Clarke, & Srinivasan, 2000). There is widespread concern about the quality and effectiveness of health care, the relationship of health care workers' knowledge and quality outcome, the high cost of health care, and the access to health care. Without collecting information systematically, having standardized classification system, and analyzing collected clinical data, nurses could not demonstrate their contribution to patient health. One of the nursing efforts to determine nursing contribution to patient outcome is the development of the Nursing Minimum Data Set (NMDS) (Werley & Lang, 1988). The NMDS provides a structure to promote the standardization of data to facilitate its retrieval and analysis from large clinical databases and to describe and compare nursing practices (Blewitt, & Jones, 1996; Brossette, Sprague, Hardin, Waites, Jones, & Moser, 1998).

Transforming nursing data and other patients' data included in the Nursing Minimum Data Set to knowledge such as relationship between nursing intervention and

1. Assistant Professor, College of Nursing, Keimyung University
2. Professor, College of Nursing, Keimyung University
3. Professor, College of Nursing, Keimyung University
4. Associate Professor, College of Nursing, Keimyung University
5. Associate Professor, College of Nursing, Keimyung University
This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (A050909)
Corresponding author: Myonghwa Park, RN, PhD, College of Nursing, Keimyung University
194 Dongsan-dong, Daegu 700-712, Korea
Tel: 82-53-250-7552 Fax: 82-53-250-6614 E-mail: mhpark1@kmu.ac.kr
Received April 17, 2006 ; Accepted June 10, 2006

nursing outcome will demonstrate nursing contribution to patients' care. An ability to link and analyze the linkage of nursing intervention and nursing sensitive outcomes will demonstrate the effectiveness of nursing treatment in responding to patients' problems. Analysis of patient demographic characteristics will provide information on the type of nursing treatments and outcomes regarding the variation of patient characteristics (Goossen, Epping, Feuth, van den Heuvel, Hasman, & Dassen, 2001).

The development of high powered computers and storages technology over the last decade has made possible the collection and storage of extremely large volumes of patient and nursing data. New techniques and tools have been developed to manage these large amount of data into useful information and knowledge. One of these innovative approaches is data mining, or knowledge discovery in databases (KDD) (Delaney et al., 2000). Applying data mining techniques into data management will improve nurses' ability to understand important links between massive volume of collected data and the outcome of patient care, and to predict the desired outcomes (Harris, Graves, Solbrig, Elkin, & Chute, 2000).

Knowledge discovery and data mining

A growing number of publications in various fields (e.g. business, engineering, and medicine) have devoted to knowledge discovery and data mining. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions (Goodwin & Iannacchione, 2002). The term Knowledge Discovery from databases incoporates data mining as a step but also covers the full process from initial data cleansing and preprocessing to interpretation of the induced patterns. (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; McDonald, Brossette, & Moser, 1998; Windle, 2004). Nurses have just begun to explore this tool for nursing knowledge discovery in the last few years and only a few published nursing studies utilizing data mining strategy exist. Eriksen, Turley, Denton, and Manning (1997) applied data mining tool, based on a cluster analysis technique, to pilot project studying the staffing levels according to the patient acuity and volume. The investigators concluded that the data mining technique assisted in understanding the complex relationships which were embodied in the data.

A collaborative research project proposed by nurses in the U.S. and Austria studied differences in patient positioning on the patient outcomes of adult respiratory distress syndrome (ARDS) (Goodwin et al., 1997). Clinical data from Austria's large database were used for the study and data mining techniques were applied to this project to find clinically relevant patterns in predicting risks for ARDS. The multidisciplinary research team at Duke University used an extensive clinical database of obstetrical patients to identify factors contributing to perinatal outcomes (Goodwin et al., 1998). Recently, data mining is increasingly used to predict clinical outcomes, for example, nursing practice domain in spinal cord injury clinical database (Kraft, 2003), osteoarthritic knee rehabilitation outcomes using a predicton model (Tam, Cheing, & Hui-Chan, 2004), and birth outcome prediction (Goodwin, & Iannacchione, 2002).

Rough set model which was introduced by Zdzislaw Pawlak in Poland in the early 1980's (Ohrn & Rowland, 2000) is one of the data mining methods. Rough set is a set theory that classifies objects into sets based on attributes of the objects. The classification can then be used to access objects or sets of objects where objects are roughly equal or roughly overlap. Rough set is one of the non-statistical methodologies for data analysis and concerns the classificatory analysis if imprecise, uncertain, or incomplete information expressed in terms of data acquired from experience. Basically, rough set theory deals with the approximation of sets that are difficult to describe with the available information. Rough set evaluates the importance of conditional attributes, reduces redundant attributes, determines minimum subsets of attributes ensuring satisfactory classification of objects, creates models for objects, and results in the form close to natural language (decision rules) allowing researchers understand and interpret the outcome more easliy. One of the main advantages of rough set theory is that it does not need any preliminary or additional information or statistical assumptions about data compared with other statistical methods. Therefore, rough set thoery can be used easily for healthcare data which have a variety of attributes, missing data, and incompleted data (Kusiak, 2000).

A set of interests could be the set of patients with a certain disease or outcome (Ohrn & Rowland, 2000). Podraza and Podraza (1999) applied a rough set approach to a data set consisting of clinical and laboratory examination of children with acute lymphoblastic leukemia. This approach led to the conclusion that intensive, high dose central nervous system prophylactic irradiation seemed to be better prevention against Central

Nervous System relapse. Woolery and Grzmala-Busse (1994) applied a machine learning program based on a rough set to analyze three large datasets. They concluded that machine learning could generate expert system rules for prediction of preterm delivery.

Therefore, applying data mining techniques to the NMDS which contained a large volume of patients' data would be able to demonstrate outcome effectiveness of different interventions over a period of care in this study. Moreover, the pattern of the linkage among elements (e. g. relationship between length of stay and comorbidity) in the NMDS would be assisted by utilizing this technique.

Purpose of study

The purposes of this study were to apply rough set to nursing specific knowledge discovery process and to identify the utilization of rough set as a data mining skill for clinical decision making.

Objective of study

1. To describe the knowledge discovery process.
2. To apply the data mining technique based on rough set model to nursing knowledge discovery.
3. To identify the methodology of data preparing process.
4. To understand patterns generated from data mining.
5. To identify the utilization of rough set model for clinical decision making.

METHODS

Conceptual framework

The conceptual framework used to guide this study was based on the Knowledge Discovery in Databases (KDD) Process Model by Fayyad et al. (1996), which explained the management and processing of nursing data, information, and knowledge to support the practice and the delivery of nursing care (see Figure 1).

Setting

The target setting was a 500-bed community hospital in the Midwest of U.S. The nursing information system linked the elements of the NMDS and used the North American Nursing Diagnosis Association classification (NANDA, 1996), Nursing Intervention Classification (NIC) (Iowa Intervention Project, 1996), and the Nursing Outcomes Classification (NOC) (Iowa Outcome Project, 1996). Within the automated on-line care planning system, the nurses were provided sets of standardized care plans, while at the same time allowing nurses to select the nursing interventions for each patient. At discharge, nursing clinical data from the care plan were coded and manually keyed into the hospital information system for electronic archiving and retrieval.

Population and sample

All adult and elderly inpatients discharged between January 1, 1998 and December 31, 1998 from one of the 8 inpatient units were used to determine the effectiveness of nursing interventions for the most frequently occurring nursing diagnoses targeted to the most frequently occurring DRGs. One thousand patients' records were retrieved from the original sample.

Data preparing process

1) Selection of five most frequently used nursing diagnoses: Five nursing diagnoses were selected based on frequencies. Each visit (record) had at lease one of these di-

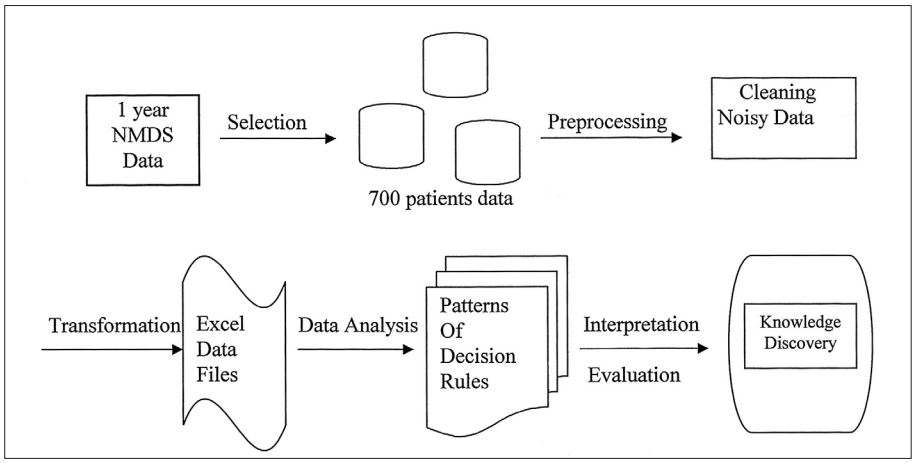

Figure 1. Steps of the KDD process

agnoses. In the database, the diagnoses were represented by five dichotomous variables, coded as 1 if that diagnosis appeared in the record and 0 if not. These variables were Pain, Potential for infection, Potential for injury, Immobility, and Impaired skin integrity.

2) Identifying the subsample: Randomized 1000 patient data were selected from the year 1998 database. The data set was randomly split into a training set with 700 cases and a test set with 300 cases. All these patient data in the training set consisted of 60 attributes. Each record had at least one of the five nursing diagnoses above.

3) Identifying 16 most frequently used nursing interventions: Sixteen most frequently used nursing interventions were identified for specific five nursing diagnoses. These were coded as 1 if the specific NIC appeared and 0 if not.

4) Identifying 17 most frequently used nursing outcomes: Seventeen most frequently used nursing outcomes were identified for specific five nursing diagnoses. These were coded as 1 if the specific NOC appeared and 0 if not.

5) Data cleaning and preprocessing: Original data kept in access database were transformed to Excel spreadsheet. Some redundancy data variables (columns) were excluded. Real values such as ICD 9 were categorized into groups. All missing values were designed for specific values to fill up the field. Finally, a table composed of 700 rows corresponding to patients (objects) and 60 columns corresponding to attributes (variables).

6) Data mining with the training set: The training set of data was selected to extract the rules and analyzed with the data mining program based on rough set theory (Kusiak, 2000). The data mining approach used in the study was that a group of patients with a unique set of characteristics (e.g., nursing diagnoses and nursing interventions) shared the same outcome. Data mining algorithms dynamically analyzed large databases and identified unique characteristics of the groups of patients, rather than analyzing the entire population. Unique

Table 1. Demographic Characteristics of the Study Sample (N= 700)

Variables	Range	Mean	SD
Age in year	25−102	70.31	14.63
Number of Co-morbidity	0−4	3.5	0.76
Length of Stay in day	1−63	5.4	5.27
Variables	**Category**	**Number**	**%**
Gender	Male	271	38.7
	Female	429	61.3
Marital Status	Married	378	54.1
	Single	139	19.8
	Separated	32	4.6
	Widowed	139	19.8
	Unknown	12	1.8
LOS	LOS ≤ 3	385	55.5
	LOS > 3	315	45.5
Discharge Disposition	Home	430	61.6
	Home with Home health Care Services	80	11.6
	Skilled Care/Rehabilitation Facility	43	6.2
	Transferred to other hospitals	15	2.2
	Left the Hospital Against Medical Advice	10	1.4
	Expired	122	17.6
Service	Medical	251	35.9
	Surgical	218	31.1
	Obstetrics	118	16.9
	Oncology	36	5.1
	Others	77	11.0
Primary Disease	Myocardiac Infarction	73	10.4
	Angina	52	7.4
	Anemia	44	6.3
	Arrhythmia	16	2.4
	Heart Failure	15	2.1
	Others	500	71.4

Table 2. Data Attributes

PT-ID: Patient ID

GENDER: Patient gender
M = Male, F = Female

Comorbidity: Number of comorbidity
Comorbidity ≤ 2 = 1
Comorbidity > 2 = 2

Age
1 = below 30
2 = 31–65
3 = above 65

Race
A = Caucasian/White
B = African American/Black
C = American Indian/Eskimo
D = Chinese/Taiwanese
E = Japanese
F = Filipino
G = Hawaiian
H = Korean
I = Asian Indian/Pakistani
J = Vietnamese
K = Laotian
L = Patient Refused to Provide
M = Other
N = Unknown
O = Hispanic

ZIPCODE

LOS = Length of stay
LOS ≤ 3 = 1
LOS > 3 = 2

SERVICE: department at discharge
Med = meidicine
Neu = neurology
OBI = obstetric
OBV = observation
ONC = cancer
OPS = outpatient surgery
PUL = pulmonary
REH = rehabilitation
SUR = surgery
TSU = ransfer to skilled unit

DISSTAT: Discharge status
1 = Home
2 = Other hospitals
3 = Skilled
4 = Home Care Services
5 = Rehabilitation Services
6 = Expired
7 = Against Medical Advice

Marital status
D = divorce
M = married
S = single
W = widow
X = legally separated

RELIG: Religion
1 = Plymouth Brethren Assemblies
2 = Congregational
3 = Unitarian Universalist
4 = United Church of Christ
5 = Evangelical Free
6 = Evangelical Covenant
7 = Protestant
8 = Christian Science
9 = Bahi
A = Apostolic
B = Baptist
C = Catholic
D = Disciples of Christ
E = Epsicopal
F = Friends
G = Assembly of God
H = Church of Christ
I = Church of God
J = Jewish
K = Christian & Missionary Alliance
L = Lutheran
M = Methodist
N = None
O = Other
P = Presbyterian
Q = Christian

NINSUR1: Primary Insurance

NINSUR2: Secondary Insurance

NINSUR3: Third Insurance

CCS1a: Primary medical diagnosis from ICD9 coding system

CCS2a: Secondary medical diagnosis from ICD9 coding system

CCS3a: Third medical diagnosis from ICD9 coding system

CCS4a: Fourth medical diagnosis from ICD9 coding system

PROCED1: Medical treatment

PROCED2: Medical treatment

DRG: Medical diagnosis from DRG coding system

Nursing Diagnoses including:
PAIN = Pain
INFEC = Infection
INJURY = Injury
IMMOBILE = Immobility
SKIN = Impaired skin integrity

NIC: Nursing interventions, 1 is having this intervention, 0 is not having this intervention
NIC1 = pain management
NIC2 = exercise therapy : ambulation
NIC3 = Environmental management: comfort
NIC4 = infection protection
NIC5 = skin surveillance
NIC6 = medication management
NIC7: = analgesic administration
NIC8 = safety (environmental management)
NIC9 = bedrest care
NIC10 = perineal care
NIC11 = fall prevention
NIC12 = infection control
NIC13 = exercise therapy: joint mobility
NIC14 = surveillance:safety
NIC15 = wound care
NIC16 = positioning

NOC: patient outcomes, 1 is having this outcomes, 0 is not having this outcomes
NOC1 = remain free of infection
NOC2 = states realistic level of comfort
NOC3 = maintain a febrile state
NOC4 = states relieved pain
NOC5 = wound incision appear clean and free of purulent drainage
NOC6 = identify intervention to prevent /reduce risk of infection
NOC7 = demonstrate pain relief measures
NOC8 = identify causative factors
NOC9 = maintains or improves tissue oxygenation
NOC10 = requests assistance for mobility
NOC11 = follows activity restrictions
NOC12 = Utilizes safety measures
NOC13 = transfer with assistant/independently
NOC14 = requests assistance for toileting
NOC15 = patient/significant other states awareness of environmental risk factors (specify)/patient
NOC16 = reposition with assist/independently
NOC17 = patient will have intact dry skin

NONIC: No Nursing Intervention
0 = No, Patient receives some interventions
1 = Yes, Patient doesn't receive any interventions

NONOC: No Outcomes
0 = No, Patient has some outcomes
1 = Yes, Patient doesn't have any outcomes

characteristics were not specified in advance. The sets of unique characteristics were determined for a group of patients by the data-mining algorithm itself.

7) Rule testing with the testing set: The rules were evaluated with the testing set data (300 patients) for their predictability and ROC (Receiver Operating Characteristcs) was drived to see the discminatory performance of rules.

RESULTS

Demographic characteristics

Demographic characteristcs of the study sample are shown in Table 1. The age ranged from 25 to 102 and the mean age was 70.31. The mean number of comorbidity was 3.5 and average length of stay (LOS) was 5.4 days. The proportion of women was 61.3% (n = 429). Nearly 62% (n = 430) of the patients were discharged to home without any additional services and 17.6% were expired at the hospital. About 36% of patients received the care service at medical units and 31.1% did at surgical units. The most common primary disease was myocardiac infarction (10.4%), followed by angina (7.4%).

Descision table with data attributes

Data were discretized and represented with decision table in which each row represented a patient (an object) and each column represented a variable (an attribute) that could be measured for each object. All variables have been transformed into categorical ones, even though some of them were inherently numerical (see table 2).

Table 3. Sixteen Most Frequent Nursing Interventions (N=700)

Nursing Diagnoses	Frequency	Percent
Pain management	324	46.3
Infection protection	312	44.6
Analgesic administration	227	32.4
Skin surveillance	203	29.0
Medication management	109	15.6
Fall prevention	97	13.9
Infection control	80	11.4
Surveillance: safety	79	11.3
Wound care	66	9.4
Positioning	45	6.4
Perineal care	44	6.3
Exercise therapy : ambulation	41	5.9
Environmental management: comfort	41	5.9
Exercise therapy: joint mobility	37	5.3
Safety (environmental management)	25	3.6
Bedrest care	18	2.6

Characteristics of nursing care

The five most common nursing diagnosis was potential for infection (n = 593, 84.7%), followed by pain (n = 504, 72.0%), potential for injury (n = 202, 28.9%), impaired skin integrity (n = 159, 22.7%), and immobility (n = 100, 14.3%).

The 16 most frequently used nursing interventions are listed in Table 3.

The intervention most freqeuntly used was pain management (n = 324, 46.3%). Other frequently used interventions were infection protection (n = 312, 44.6%), analgesic administration (n = 227, 32.4%), and skin surveilance (n = 203, 29.0%).

The 17 most frequently used nursing outcomes are listed in Table 4.

The nursing outcome most freqeuntly used was remain free of infection (n = 384, 54.9%), followed by states realistic level of comfort (n = 289, 41.3%), maintain a febrile state (n = 206, 29.4%), and states relieved pain (n = 195, 27.9%).

Example rules based on rough set theory

Decision: Length of Stay

Data mining found 135 rules and the 135 rules were pruned down to 10 rules. Five variables such as the number of comorbidity, marital status, risk for infection (NANDA), infection protection (NIC), discharge status were identified as valuable predictors for length of stay.

Table 4. Seventeen Most Frequent Nursing Outcomes (N= 700)

Nursing Diagnoses	Frequency	Percent
Remain free of infection	384	54.9
States realistic level of comfort	289	41.3
Maintain a febrile state	206	29.4
States relieved pain	195	27.9
Wound incision appear clean and free of purulent drainage	177	25.3
Identify intervention to prevent / reduce risk of infection	163	23.3
Demonstrate pain relief measures	135	19.3
Identify causative factors	87	12.4
Maintains or improves tissue oxygenation	100	14.3
Follows activity restrictions	73	10.4
Requests assistance for mobility	65	9.3
Patient will have intact dry skin	64	9.1
Transfer with assistant/independently	57	8.1
Requests assistance for toileting	51	7.3
Patient/significant other states awareness of environmental risk factors (specify)/patient	42	6.0
Reposition with assist/independently	36	5.1
Utilizes safety measures	32	4.6

Rule 1.

(Comorbidity < 2) & (Marital Status = M) & (INFEC =1) & (NIC4=1) & (DISSAT=1) => (LOS<3); [92, 23.4%, 87.2%]

If	Comorbidity < 2 and
	Marital Status = M and
	INFEC =1 and
	NIC4=1 and
	DISSAT=1
Then	LOS < 3
With	Coverage 23.4%
	Accuracy 87.2%

Figure 2. Decision Rule Example One

Rule 2.

(Age = 3) & (SKIN = 1) & (PAIN=1) & (DISSAT = 2 or 3) => (NOC2=0) [90, 32.4%, 78.0%]

If	Age = 3 and
	SKIN = 1 and
	PAIN = 1 and
	DISSAT = 2 or 3
Then	NOC2 = 0
With	Coverage 32.4%
	Accuracy 78.0%

Figure 3. Decision Rule Example Two

Rule 3.

(Age = 3) & (PAIN=1) & (Potential for Infection) & (Marital Status = W or X) & (CCS1a = 411) => (DISSAT = 6) [33, 27.3%, 80.0%]

If	Age = 3 and
	PAIN=1 and
	Potential for Infection = 1 and
	Marital Status = W or X and
	CCS1a=411
Then	DISSAT=6
With	Coverage 27.3%
	Accuracy 80.0%

Figure 4. Decision Rule Example

Applying the rules to the test set resulted in the area of under ROC curve of 89.0% with standard error of 2.5% showing the high predictability. Figure 2 shows an example rule with high accuracy.

This decision rule indicates that if patients had less than 2 comorbidity, were married, had a risk for infection as nursing diagnosis and received infection protection as nursing intervention, and discharged to home, 92 patients had less than 3 days of length of stay. It describes exactly 42 patients, which represents 23.4 % of pateints with LOS < 3.

Decision: Nursing Outcome

Data mining found 88 rules for nursing outcome, Relieved Pain, and the 88 rules were pruned down to 7 rules. Four variables (age, impaired skin integrity, pain, and discharge status) were identified as valuable predictors. Applying the rules to the test set resulted in the area of under ROC curve of 90.1% with standard error of 2.3% showing the high predictability. Figure 3 shows an example rule with high accuracy.

This decision rule indicates that if patients older than 65 had impaired skin integrity and pain as nursing diagnoses, and discharged to home with home health care services or skilled care, the patient did not show any relieved pain as nursing outcome; 90 (32.4%) patients met this decision rule with 78.0% of accuracy.

Decision: Mortality

Data mining found 155 rules for being expired at hospital and the 155 rules were pruned down to 15 rules. Five variables (age, pain, potential for infection, marital status, and primary disease) were identified as valuable predictors for mortality. Applying the rules to the test set resulted in the area of under ROC curve of 91.1% with standard error of 1.8% showing the high predictability. Figure 4 shows an example rule with high accuracy.

This decision rule indicates that if patients were older than 65, had pain and potential for infection as nursing diagnoses, were widowed or seperated, and had myocardiac infarction as a primary medical diagnosis, the patients had high probability of expiration at hospital; 33 (27.3%) patients met this decision rule with 80.0% of accuracy.

DISCUSSION

Data mining techniques have been applied actively in many studies outside nursing, showing the value for knowledge discovery in clinical data set. Even though nurses are the primary users of clinical data, the importance and utilization of large clinical data set have been neglected in nursing. Data mining skill can be one of the essential methods for knowledge discovery to enhance decision making for quality assessment and improvement. This study explored the potential for utilization of data mining tool to discover the information and knowledge in the NMDS.

In this study, the most freqently used nursing disgnoses for inpatient care in an acute care setting were identified as potential for infection, pain, potential for injury, impaired skin integrity, and immobility, which showed similar pattern with the results of other studies (Kraft, 2003; Windle, 2004; Park, Delaney, Maas, & Reed, 2004). The most frequently identified primary diseases were related to cardiovascular dysfunction such as myocardiac infarction, angina, anemia, cardiac dysrhythmias, and heart faliure. This pattern of primary diseases of the sample population is likely reason why nurses used nursing diagnoses related to acute cardiac dysfunction most frequently such as infection, pain, injury, skin intergrity, and immobility.

The rules identified by data mining revealed that number of comorbidity, marital status, nursing diagnosis related to risk for infection and nursing intervention related to infection protection, and discharge status were the predictors that determined the length of stay. This is a similar finding to other studies which explored the predictors for length of hospital stay (Lee, 1998; Newhouse, Johantgen, Pronovost, & Jonhnson, 2005). Patients who had more comorbidity tended to require longer care and early detection of risk of infection while intervention for infection protection shortened the length of hospital stay.

One of the decision rules for nursing outcomes indicated that if patient older than 65 had impaired skin integrity and pain as nursing diagnoses, and discharged to home with home health care services or skilled care, the patient would not show any relieved pain as nursing outcome. This rule suggested that impaired skin intergrity inhibited pain relieving for older patients and eventually needed further care after discharge such as home care or long-term care services (Goossen, Epping, Feuth, van den Heuvel, Hasman, & Dassen, 2001).

Five variables (age, pain, potential for infection, marital status, and primary disease) were identified as valuable predictors for mortality in this study. Older patients have a higher risk of death in the hospital than do younger patients. Pain and risk for infection were associated with mortality of old patients with myocardiac infarction and needed more attention as mentioned in other studies (Berkman, Leo-Summers, & Horwitz, 1995; Gliksman, Lazarus, Wilson, & Leeders, 1995). Patients with MI who were separated or widowed had a higher risk of death in hospital than those who were married. One reason might be the limited sources of support, because they lived alone. Social support has been found to be an important variable in adapting to MI (Morse,

Jones, Brosette, 1999).

The effectiveness and potentials of rough set model as a data mining tool have been proven through this study. The patterns of decision rules were readily interpretable and allowed possible new insights into how various attributes interact and serve as baseline data for decision making. Researchers may benefit from the use of data mining based on rough set model to identify factors that might be missed by using traditional statistical methods such as regression , and analysis of variance, etc. It is notable, however, that more than 50% of the data mining process was devoted to the data selection, cleaning and reduction. This implies that the most important prerequisites of effective data mining are the exploration and selection of the right data set. Clear and in-depth understanding of the selected data sets should be followed. It also should be noted that cleaning and preparing data sets are one of the critical processes of data mining because these works enhance the reliability and validity of data, resulting in quality improvement of the data. Prior knowledge of the pattern of data can also significantly reduce the data mining step as well as all the other steps in the KDD process.

A potential limitation of this study would be the quality of the nursing records. The study hospital is known for maintaining consistent quality in its computerized nursing documentation system, including standardized nursing diagnoses and nursing interventions and outcomes. The clinical contents in the information system mirror the paper chart contents exactly mirror the contents of the paper charts, which are maintained as a part of the permanent patient records. In addition, the standardized language used in the NMDS incorporated into the site's clinical ladder system and consequently is an explicit criterion in annual performance appraisals. With all these facts, the reliability and validity of the data should still be considered.

CONCLUSION

It is the beginning step to use the method of knowledge discovery and data mining in manipulating nursing data set as is tried in this study. Currently, few published studies looked at computerized nursing practice data from longitudinal perspectives using data mining techniques for evidence of nursing contribution to patient outcomes or for diagnostic and intervention patterns. Therefore, more studies are necessary to explore the potential of applying knowledge discovery and data mining to various types of nursing data. In addition, nurses need to demonstrate the effectiveness of nursing intervention through the large data set with standardized language format to unify data from different settings.

This study represents an initial step in the determination of nursing care and patient characteristics that contribute to predicting the outcomes of hospitalized patients using data mining tool based on rough set model. A better understanding of patient and nursing pattern enables this kind of nursing information to be used for care planning and resource allocation.

Based on the overall findings of this study, indications exist to warrant further investigation in several specific areas. First, it is recommended that the study be replicated using a larger sample from a variety of acute care settings representing different areas and different population of patients, especially from Korean settings where the elecronic health record systems are increasing very rapidly. Second, it is needed to compare the data mining method based on rough set theory with other data ming methods or traditional statistical methods such as logistic regression to verify the predictability of rough set model. Finally, since this study used the data prepared based on five frequently used nursing diagnoses, future study is needed to explore the nursing services and patient characteristics related to a variety of specific medical diagnoses or DRGs (Diagnosis Related Groups).

The value of nursing database research is directly related to the value placed on the contribution of nursing practice to research and knwoledge development. Nursing data from the real clinical settings can produce meaningful information that can contribute to improving the quality of nursing care. Such information then can further empower nurses in their clinical decision-making and problem solving.

References

Berkman, L. F., Leo-Summers, L., & Horwitz, R. I. (1992). Emotional support and survival after myocardial infarction. *Annals of Internal Medicine*, *117*, 1003-1009.

Blewitt, D. K., & Jones, K. R. (1996). Using elements of the nursing minimum data set for determining outcomes. *J Nurs Adm*, *26*(6), 48-56.

Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., & Moser, S. A. (1998). Association rules and data mining in hospital infection control and public health surveillance, *J Am Med Inform Assoc*, *5*(4), 372-381.

Burn-Thornton, K. E., & Edenbrandt, L. (1998). Myocardial infarction-pinpointing the key indicators in the 12-Lead ECG using

data mining. *Comput Biomed Res*, *31*, 293-303.

Delaney, C., Ruiz, M. E., Clarke, M., & Srinivasan, P. (2000). Knowledge discovery in databases: data mining NMDS. Proceedings of the 7th IMIA International Conference on Nursing Use of Computers and Information Science. *Aukland*, New Zealand, 61-65.

Eriksen, L. R., Turley, J. P., Denton, D., & Manning, S. (1997). Data mining: A strategy for knowledge development and structure in nursing practice. In U. Gerdin, M. Tallberg, & Wainwright. Nursing informatics: the impact of nursing knowledge on health care informatics.. proceedings of NI˙97, Sixth Triennial International Congress of IMIA-NI, Nursing Informatics of International Medical Informatics. *Amsterdam*, Netherlands: IOS Press.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Ed.). *Advances in knowledge discovery and data mining*. (pp. 1-31). Menlo Park, CA: AAAI Press/MIT Press.

Gliksman, M. D., Lazarus, R., Wilson, A., & Leeder, S. R. (1995). Social support, marital status and living arrangement correlates of cardiovascular disease risk factors in the elderly. *Social Science & Medicine*, *40*(6), 811-814.

Goossen, W. T. F., Eppng, P. J. M., Feuth, T., van den Heuvel, W. J. A., Hasman, A., & Dassen, T. W. N. (2001). Using the nursing minimum data set for the Netherlands (NMDSN) to illustrate differences in patient populations and variations in nursing activities. *Int J Nurs Stud*, *38*(3), 243-257.

Goodwin, L., Saville, J., Jasion, B., Turner, B., Prather, J., Dobousek, T., & Egger, S. (1997). A collaborative international nursing informatics research project: Predicting ARDS risk in critically ill patients. In U. Gerdin, M. Tallberg, & Wainwright. *Nursing informatics: the impact of nursing knowledge on health care informatics. proceedings of NI˙97*, Sixth Triennial International Congress of IMIA-NI, Nursing Informatics of International Medical Informatics. Amsterdam, Netherlands: IOS Press.

Goodwin, L., & Iannacchione, M. A. (2002). Data mining methods for improving birth outcomes prediction. *Outcomes Manage*, *6*(2), 80-85.

Goodwin, L., Prather, J., Schlitz, K., Iannacchione, M. A., Hage, M., Hammond, W. E., & Grzymala-Busse, J. (1998). Data mining issues for improved birth outcomes. *Biomed Sci Instrum*, *34*, 291-298.

Harris, M. R., Graves, J. R. Solbrig, H. R., Elkin, P. L., & Chute, C. G. (2000). Embedded structures and representation of nursing knowledge. *J Am Med Inform Assoc*, 7(6), 539-549.

Kraft, M.R. (2003). *Mining a spinal cord injury clinical database for nursing information: A source of nursing knowledge*. Unpublished Doctoral Dissertation, Loyola Univerity of Chicago.

Kusiak, A. (2000). *Computational Intelligence in Design and Manufacturing*. New York: John Wiley.

Lee, T. (1998). *An Analysis of the Relationship among Patient Profile Variables in Predicting Home Care Resource Utilization and Outcomes*. Unpublished Doctoral Dissertation, University of Maryland, Baltimore.

McCloskey, J. C., & Bulechek, G. M. (1996). *Nursing Interventions Classification (NIC)* (2nd ed). St. Louis, MO: Mosby Year Book.

McDonald, J. M., Brossette, S., & Moser, S. A. (1998). Pathology information systems: Data mining leads to knowledge discovery. *Archive of Pathology Laboratory Medicine*, *122*, 409-411.

Moser, S. A., Jones, W. T., & Brossette, S. E. (1999). Application of data mining to intensive care unit microbiologic data. *Emerging Infection Disease*, *5*(3), 454-457.

Newhouse, R. P., Johantgen, M., Pronovost, P. J., & Jonhnson, E. (2005). Perioperative nurses and patient outcomes: Mortality, complications, and length of stay. *AORN J*, *81*(3), 508-518.

North American Nursing Diagnosis Association. (2000). *Nursing diagnoses: Definitions & classification*. Philadelphia, PA: Authors.

Ohrn, A., & Rowland, T. (2000). Rough sets: A knowledge discovery technique for multifactorial medical outcomes. *Am J Phys Med Rehabil*, *79*, 100-108.

Park, M., Delaney, C., Maas, M., & Reed, D. (2004). Using a nursing minimum data set with older patients with dementia in an acute care setting. *J Advan Nurs*, *47*(3). 329-339.

Ryan, P., & Delaney, C. (1995). Nursing Minimum Data Set. In J.J. Fitzpatrick & J.S. Stevenson (Ed.), *Annual review of nursing research*, (pp. 169-194). New York: Springer Publishing Company.

Podraza, W., & Podraza, H. (1999). Childhood leukemia relapse risk factors. A rough sets approach. *Medical Informatics, 24*(2), 91-108.

Tam, S., Cheing, G. L. Y., & Hui-Chan, C. W. Y. (2004). Predicting osteoarthritic knee rehabilitation outcome by using a prediction model devleoped by data mining techiniques. *Int J Rehabil Re*, *27*(1), 65-69.

Werley, H. H., & Lang, N. M. (1988). *Identification of the nursing minimum data set*. New York: Springer.

Windle, P. E. (2004). Data mining: An excellent research tool. *J Peri Anesth Nurs*, *19*(5), 355-356.

Woolery, L. K., & Grzymla-Busse, J. (1994). Machine learning for an expert system to predict preterm birth risk. *J Am Med Inform Assoc*, *1*(6), 439-445.