IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4144

Knowledge Discovery In Clinical Databases: Issues and Processes

M.R. Kraft
VA Hines Hospital-Hines, Illinois, Niehoff School of Nursing, Loyola University, Chicago, Margaret.Kraft@med.va.gov

K. C. Desouza*
Dept. of Information & Decision Sciences, University of Illinois at Chicago, Chicago, Illinois
Tel: +1 312 829 8447, Fax: +1 312 413 0385, kdesou1@uic.edu

I. Androwich
Niehoff School of Nursing, Loyola University, Chicago, iandro@luc.edu

ABSTRACT

The healthcare field is facing strong pressures for cost reduction along with demands for increased quality of service. As a strategy to address these issues, healthcare information systems are being utilized in the field to help in decision support and knowledge management. The knowledge discovery process consists of several segments: data gathering, data cleansing, information generation, information mining, knowledge generation, knowledge storage and distribution. Knowledge Discovery in Clinical Databases: Issues and Process examines the processes of data acquisition and data cleaning as they pertain to nursing information management within a spinal cord injury (SCI) clinical database. A long-term goal of clinical database analysis is the achievement of greater efficiency and effectiveness in the use of resources for nursing care delivery. The focus of the paper is on pertinent data cleaning issues faced while gathering data for a data-mining project. Areas of future research are also discussed.

INTRODUCTION

The healthcare industry faces contradictory pressures of lowering cost and increasing quality of service, both of which require efficient decision-making. Healthcare facilities have at their disposal vast amounts of data. However the challenge is to extract relevant information from this data and act upon it in a timely manner (Desouza, 2002). Generating information and knowledge calls for organizing data into a useful form. Efficient decision-making is a by-product of thorough analysis of available data on a given problem.

Nurses need large amounts of health data and are confronted daily with constantly changing information needed to manage care. As nursing practice data are computerized, the ability to capture, store, retrieve, organize, and analyze the information of nursing practice can provide information for nursing decision support, enhancement of documentation, and identification of nursing care trends and costs with the ultimate goal of improved patient care.

Exploration of nursing data elements within a spinal cord injury (SCI) database is proposed as a mechanism to help in the identification of major phenomena basic to SCI nursing care. Utilization of information in SCI databases may be a means of bringing more focused and appropriate care to SCI individuals who, as consumers of significant costly care resources, are 'outliers' in the healthcare system (Lincoln & Builder, 1999). Patients identified as outliers are those whose annual care costs far exceed normally expected healthcare costs. In our rapidly changing health care system, it is important to know aggregate costs of SCI to ensure that adequate funds are allotted for care of the SCI population. Although SCI occurs much less frequently than other types of injury and debilitating disease, the cost of SCI to individuals and to society is staggering. Berkowitz, O'Leary, Kruse, & Harvey (1998) estimate that SCI costs the nation more than 9.7 billion dollars per year. Direct care costs within the first year of injury average $223,261 with an additional annual cost for SCI care of at least $26,000. Equipment, supplies, medications, and environmental modification costs increase both figures. Indirect costs related to loss of income and productivity is more difficult to compute with consideration given to age at injury and earning potential but indirect cost estimates can be projected as significant. The aggregate annual direct and indirect costs of new cases of SCI may be between 7.2 and 9.5 billion dollars (Berkowitz et al., 1998).

Information Generation

For purposes of this paper and within the framework of systems theory, information is defined as organized and processed data that can be communicated, and/or received, and when received, is meaningful and useful to the recipient. This database analysis will use data mining techniques to determine if there are patterns of patient needs, nursing diagnoses, nursing interventions, and patient outcomes that can contribute information that can improve the efficiency and effectiveness of the delivery of SCI nursing care. Analysis may demonstrate that information patterns related to the presence of specific nursing diagnoses and the choice of specific nursing interventions that promote desired outcomes can be used to allocate resources for SCI care delivery. The application of the data mining process to this SCI clinical database may determine that this research method can lead to a better understanding of how to use data to improve SCI nursing practice.

Healthcare Information

Three dimensions in health information outlined by Sorthon, Braithewite, and Lorenzi (1997) include management information, professional information, and patient information. Overlap and commonalties are identified but fundamental differences exist in the types of information required for each dimension, the way the information is used, and the way standards are maintained. The achievement of a comprehensive and integrated data structure that can serve the multiple needs of each of these three dimensions is a goal in most health care information system development.

Patient health records are a rich source of data for research. Such data is generally accessible, accurate, and relatively inexpensive. Traditionally, the paper record is documented in a 'diary' style (Gabrieli, 1990) and includes documentation that is for the most part legally and medically accurate and reliable but the expectation is that such records are read only by colleagues and are not for public use. There are disadvantages to using health care records as a data source. Too often the data are so disconnected that information is not useful. Concerns related to such data are that data are collected as a by-product of some other processes; data are probably collected and entered by many people without any quality check; data may have different structure even within the same database, and missing data may be common (Lange & Jacox, 1993). Other disadvantages are related to the non-research purpose of the record, the presence of selective information,

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

the need for interpretation of certain information in the record, and the difficulty with data verification (Krowchuk, Moore, & Richardson, 1995).

Threats to validity inherent in large databases include sampling and measurement errors. Sampling errors are the result of the selection of cases and measurement errors develop as the result of problems with operational definitions of concepts. Because data bases exist over long periods of time, reliability threats are created by such things as clerical error, subtle changes in data collection techniques with improved diagnostic skills, and the instrumentation used to collect data. Appraisal of data includes consideration of accuracy, representativeness, authorship and authenticity (Reed, 1992). The validity of conclusions in research depends partly on the completeness and accuracy of the data. Incomplete data is meaningless. Both random and systematic errors can occur in data collection and management. Such errors may be identified with measures of frequency, central tendency, range, and dispersion. Knowing the data well can help in the identification and resolution of potential errors (Roberts, Anthony, Madigan & Chen, 1997). Rather than accepting data at face value, the researcher must consider all potential limitations.

The cost of information is usually not stated or, indeed, even known but information represents a large percentage of the healthcare cost structure. About 1/3 of the cost of health care in the United States, some 300 billion dollars—represents the cost of capturing, storing, and processing such information as patient's records, physicians' notes, test results, and insurance claims (Evans & Wurster, 1997). It has been estimated that 25% of hospital cost is spent on information handling primarily as a means of communication. There are indications that the most frequent problem with healthcare information is a lack of availability. Too much information, information in the wrong place, incomplete, inaccurate, inconsistent, illegible or difficult to understand information is also noted. There is little evidence that nurse researchers seek aggregate patient data that might reveal trends and patterns among patients with similar situations or treatments. Such information can be useful in understanding patterns and in predicting patients' responses to conditions and interventions. Information systems can be designed to aggregate such data and present it in a variety of formats.

Information technology has encouraged the accumulation of an unlimited quantity of health care data but has also created a resurgence of controversy in the issues of privacy and confidentiality (Rittman & Gorman, 1992). Styffe (1997) addressed the unique meaning of privacy, confidentiality, and security as related to patient data in clinical information systems. Individuals are concerned with privacy as their right to determine when, how, where, and to what extent their information is transmitted. Confidentiality is the concern of health care providers and organizations and "is the trust placed that information shared will be respected and used only for the purpose disclosed." It is based on the relationship between the person disclosing and the person receiving information. Security is built into clinical information systems and addresses the levels of authorization necessary for access to data and information. Computer security involves the protection of data against accidental or intentional disclosure to unauthorized persons. Permission to use data beyond the original intent has rarely been obtained explicitly. Human subject committees will determine whether use of data represents a threat to confidentiality. If risk is high, subjects may have to be re-contacted to get permission to use. Confidentiality is an emerging problem in computerized clinical data sets.

Knowledge Discovery in Databases

Data mining and knowledge discovery in databases (KDD) relate to the process of extracting valid, previously unknown and potentially useful patterns and information from raw data in large databases. "The analogy of "mining" suggests the sifting through of large amounts of low grade ore (data) to find something valuable. It is a multi-step, iterative inductive process. It includes such tasks as problem analysis, data extraction, data preparation and cleaning, data reduction, rule development, output analysis and review. Because data mining involves retrospective analyses of data, experimental design is outside the scope of data mining. Generally, data mining and KDD are treated as synonyms and refer to the whole process in moving from data to knowledge. The objective of data mining is to extract valuable information from data with the ultimate objective of knowledge discovery. However, a small number of published studies address the value of data mining within the healthcare industry.

STUDY SETTING

The setting for this study is a large tertiary care Veteran's Health Administration (VHA) Hospital located on a 62-acre campus within the metropolitan Chicago area. The Veterans Administration (VA) is involved in the full continuum of SCI care and has the largest single network of SCI care in the nation (DVA, 2000). This hospital has two acute rehabilitation / continuing care inpatient SCI units with a total of 68 beds, a hospital-based SCI home care program, and a 30 bed residential SCI unit. The hospital uses the national VA hospital information system (HIS) known as the Veterans Health Information Systems and Technology Architecture (VistA). VistA, one of the most extensive hospital information systems in the world, is an internally developed comprehensive integrated system that provides for both administrative and clinical support and documentation of care. Over a 20-year period, VistA has evolved to include over 70 applications as well as numerous links to commercial products. VA software is written in MUMPS (Mass General Utility Multi-Programming System), an ANSI (American National Standards Institute) programming language now call "M" (Kolodner, 1997).

The modular design of the VA nursing software within VistA allows computerization of data for clinical, administrative, research, and educational purposes as well as quality improvement (Vance, Gillian-Storm, Kraft, Lang, & Mead, 1997; Vance, Kraft, & Lang, 1998). The data collection system of the VA nursing software incorporates the elements of the nursing minimum data set (NMDS) as defined by Werley and others. The NMDS standardizes the items of essential core nursing data for collection, storage, and retrieval. It includes 16 elements categorized into three broad groups: nursing care, client demographics, and service (Werley & Leshe, 1991). These elements represent data used on a regular basis by nurses in any setting where nursing care is provided and are considered necessary for the analysis of nursing practice and its impact on outcomes and cost effective care. The goal of the NMDS is to provide for comparability of nursing data across clinical populations, settings, and geographic areas. The specific nursing care elements in the NMDS are nursing diagnoses, nursing interventions, outcomes and intensity of nursing care. Most computerized nursing information systems (NISs) now utilize the NMDS as the framework for data capture.

The VA nursing database for patient health problems is built on the North American Nursing Diagnosis (NANDA) taxonomy and the care planning process of diagnosis, intervention, and outcome reflects the nursing process. Nursing diagnoses provide a common language within the profession which can enhance communication between nursing clinicians, improve continuity of care, help formulate expected outcomes, assist in addressing cost-effectiveness of care, and allow emphasis on clinical nursing research. Nursing diagnoses have been recognized as the nursing equivalent of Diagnostic Related Groups (DRGs). The use of nursing diagnosis increases the possibility of giving comprehensive care by identification, validation, and documentation of response to specific health concerns. Nursing diagnosis allows clinicians to describe nursing practice within a shared framework.

Permission to use the VHA SCI database for this study was obtained from the facility's institutional review board (IRB) which includes the Human Studies Subcommittee (HSS) of the Research and Development (R&D) Committee and the R&D Committee itself. Since there were no interventions and no direct contact with patients, the facility IRB gave an expedited review approval. IRB approval for the study was also obtained from the Institutional Review Board of Loyola

University. Confidentiality for this study was maintained by using the internal VA patient coding to download data. This data was immediately re-recoded by the investigator to remove all possibility of patient identification.

PROCESS AND ISSUES

Data Gathering

The 525 patients with 1107 admissions to the study unit between July, 1989 and June, 2000 became the study sample. The list of admissions to the study unit was downloaded from an ORACLE mainframe database built through nightly data extracts from VistA. After identification of the study population, nursing diagnoses and interventions selected for these patient encounters were identified using an identification and ranking query that is part of the VistA nursing software. Since the nursing data elements of interest in this study are not included in the VA national data warehouse, this data was downloaded directly from the operational database to a P.C. Data related to age, date of injury and level of injury was obtained directly from the mainframe SCI Registry database that is another VistA software package.

Data Cleaning

It is estimated that 80% of the time spent in a data mining project is spent in data preparation and cleaning (Desouza, 2001; Gerber, 1998). Data preparation includes data selection (identification and extraction of data); data preprocessing (sampling and quality testing); and data transformation (conversion into an analytical model) (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998). Goodwin et al. (1997) identify the issues obstructing progress in data mining for improved health outcomes as "data quality, data redundancy, data inconsistency, repeated measures, temporal (time-contextual) measures, and data volume" (p.291). Computerization of data does not make up for bad data but once data has been cleaned, the analysis of vast amounts of data may identify potentially important relationships that do not emerge from sparse data. The analyst must formulate a query to extract data from a database, extract the aggregated data, visualize the results in a graphical way, and analyze the results. Invariably, routinely collected data is full of errors and incompleteness. Much of the data collected from this computerized database was found to be non-standardized and at a nominal level of measurement. As a result, data were visually inspected, structured, and checked for accuracy, reliability, and redundancy. Data "noise" included redundant, insignificant, erroneous, and missing data. Differences in punctuation and case or changes in word sequence were recognized by the computer software as new terms, new labels, or new variables. This required the researcher to make a visual inspection of all diagnostic and interventional labels and create a structure of labels that represent label clusters with a common or shared meaning. Data visualization is an invaluable counterpart to data mining. Visualization includes displays of trends, clusters, and differences. The visual review of all eleven years of data in this study took approximately 300 hours of time.

Data Categorization

There were 4750 different diagnostic labels in the cumulative eleven-year database that after visual inspection were determined to represent 161 unique nursing diagnoses. Through further inspection, these were clustered into 20 diagnostic categories. Two domain experts with significant SCI knowledge and experience reviewed the categories to reach a consensus on the labels for the diagnostic categories. The selected diagnostic categories for the cumulative data were: Skin Care; Elimination; Self Care Deficit; Infection Prevention/Control; Mobility; Respiratory Function; Psychosocial Adaptation; Pain Management; Knowledge Deficit; Nutrition; Fluid Volume Maintenance; Acute Problem Management; Safety/Prevention of Injury; Activity/Rest; Cognitive Functioning; Temperature Control; Sexual Health; Communication, and Miscellaneous. Any diagnostic label within the cumulative database that did not appear at least eleven times during the eleven-year study period was assigned to the category of "Miscellaneous." A map of the annual diagnostic rankings for each of the eleven years in the study was developed to determine if there were significant changes in nursing diagnosis over the study time frame (See Table 2).

ONGOING AND FUTURE RESEARCH

The potential for original research with pre-collected data is tremendous. The better acquainted the researcher is with the database, the greater the potential for creative new research. Secondary analysis is described as "extremely versatile in that it can be applied to studies designed to understand the present and the past; to understand change; and to examine phenomena comparatively" (Kiecolt & Nathan, 1985, p47). Advantages of such data analysis include larger samples, elimination of instrument development, sample selection, and data collection (Abel & Sherman, 1991). Aggregated data may provide insight and information that is useful in patient care delivery, program planning, and policy development. Lange and Jacox (l993) have identified interest in using clinical and administrative health care databases for health policy research because of national concern about the quality, cost, and outcomes of health care.

The data set is currently being examined using data mining methodology. The use of data mining, by definition, excludes the possibility

Table 1: Ranking of nursing diagnoses clusters

Years	89-90	90-91	91-92	92-93	93-94	94-95	95-96	96-97	97-98	98-99	99-00
Skin Care	1	1	1	1	1	1	1	1	1	1	1
Elimination	3	2	2	2	2	2	2	2	2	2	2
Self Care Deficit	2	3	4	3	3	4	3	4	4	5	5
Infection Prevention	5	4	3	4	4	3	4	3	3	3	3
Mobility	7	5	10	6	5	5	5	5	7	6	6
Psychosocial Adapt.	4	6	6	5	7	7	6	6	5	7	7
Respiratory Function	6	7	5	7	6	6	8	7	6	4	4
Comm. Reintegration	N/A	8	11	9	N/A	8	10	8	8	9	9
Pain Mgmt.	8	9	9	8	8	9	9	10	12	8	8
Knowledge Deficit	9	10	7	12	10	10	7	9	9	10	10
Fluid Volume Maint.	11	11	12	13	14	14	14	13	13	14	16
Nutrition	10	12	8	10	9	13	12	11	14	15	15
Miscellaneous	13	13	13	15	17	16	17	16	17	11	13.
Acute Medical Mgmt.	12	14	15	14	11	11	13	14	15	16.	14
Activity/Rest	13	15	14	11	12	15	16	15	11	13	11
Prevention of Injury	N/A	16	16	16	13	12	11	12	10	12	12
Temperature Control	N/A	17	19	19	N/A	N/A	N/A	N/A	N/A	N/A	17
Cognitive Functioning	N/A	18	18	17	16	17	15	17	16	17	18
Sexual Health	N/A	19	17	18	15	19	19	18	N/A	N/A	N/A
Sensory/Perceptual Deficit	N/A	N/A	N/A	N/A	N/A	18	18	N/A	N/A	N/A	N/A

of testing preconceived hypotheses. Data miners do not pose a question, as much as ask the system to discover data patterns that may be predictive. The process of data mining may result in the identification of hypotheses for future research. Of specific interest is a predictive model using artificial neural networks for hospital length of stay based on nurse diagnosis. Care is being taken in the evaluation and analysis of data sets since data set variables may not adequately reflect the secondary analyst's concepts of interest. The task of designing a study using available data can be challenging.

CONCLUSION

The identification of patient outcomes sensitive to nursing care is a priority for nursing research. The need to capture outcomes has been recognized by providers, payers, and policy makers. Knowledge discovery in clinical databases is a step toward outcome identification. Outcomes may be classified as 'generic' or pertinent to all health care consumers or 'condition-related' and pertinent to sub-populations of patients with specific diseases or conditions. In addition, time becomes a dimension of outcome measurement. Outcome related data might come from multiple sources such as the patient, families and caregivers, health care professionals, and biomedical instrumentation (Zielstorff, 1995). Assessment of effectiveness of care, according to Ozbolt (1996) requires standardized data aggregated in databases for comparison across times, conditions, and institutions. To analyze outcome data, it is critical that data are stored in a retrievable format according to standards that will allow for data sharing and data queries while patient privacy and confidentiality is protected. There must also be a way to link outcome data to all influencing factors such as comorbidities, procedures, treatments, interventions, patient demographics, etc.

Research related to SCI nursing diagnoses and SCI nursing interventions may demonstrate under which particular circumstances specific interventions promote the most effective outcomes for SCI patients.

REFERENCES

Abel, E. & Sherman, J. (1991) Use of national data sets to teach graduate students research skills. Western Journal of Nursing Research. 13 (6): 794-797.

Berkowitz, M., O'Leary, P., Kruse, D., & Harvey, C. (1998) Spinal cord injury: An analysis of medical and social costs. New York: Demos Medical Publishing, Inc.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998) Discovering data mining: From concept to implementation. Upper Saddle River, NJ: Prentice Hall, Inc.

Department of Veterans Affairs (2000) VA expands spinal cord injury care. Washington, D.C.

Desouza, K. (2001) Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands: Institute for Healthcare Technology Management.

Desouza, K. (2002) Knowledge management with artificial intelligence. Westport, CT: Quorum Books, In Press.

Evans, P. & Wurster, T. (1997) Strategy and the new economics of information. Harvard Business Review September-October: 71-82.

Gabrieli, E. (1990) Electronic healthcare records: A discourse. Journal of Clinical Computing XVIII (5&6): 130-143.

Gerber, C. (1998) Excavate your data. http://www.PlugIn/workvench/datamine/exacv.htm.

Goodwin, L., Prather, J., Schlitz, K., Iannacchione, M., Hage, M., Hammond, W., & Grzymala-Busse, J. (1997) Data mining issues for improved birth outcomes. Biomedical Sciences Instrumentation 34:291-296

Kiecolt, K. & Nathan, L. (1985) Secondary analysis of survey data. Thousand Oaks, Calif.: Sage Publications.

Kolodner, R. (1997) (Ed.) Computerizing large integrated health networks: The VA success. New York: Springer Verlag.

Krowchuk, H., Moore, M., & Richardson, L. (1995) Using health care records as sources of data for research. Journal of Nursing Measurement 3 (1): 3-12.

Lange, L. & Jacox, A. (1993) Using large data bases in nursing and health policy research. Journal of Professional Nursing 9 (4): 204-211.

Ozbolt, J. (1996) From minimum data to maximum impact: Using clinical data to strengthen patient care. Advanced Practice Nursing Quarterly 1 (4): 62-69.

Reed, J. (1992) Secondary data in nursing research. Journal of Advanced Nursing 17: 877-883.

Rittman, M. & Gorman, R. (1992) Computerized databases: Privacy issues in the development of the nursing minimum data set. Computers in Nursing 10 (1): 14-17.

Roberts, B., Anthony, M., Madigan, E., & Chen, Y. (1997) Data management: Cleaning and checking. Nursing Research 46 (6): 350-352.

Sorthon, F., Braithwaite, J., & Lorenzi, N. (1997) Strategic constraints in health informatics: Are expectations realistic? International Journal of Health Planning and Management12: 3-13.

Styffe, E. (1997) Privacy, confidentiality, and security in clinical information systems: Dilemmas and opportunities for the nurse executive. Nursing Administration Quarterly Spring: 21-28.

Vance, B., Gilleran-Strom, J., Kraft, M., Lang, B., & Mead, M. (1997) Nursing use of systems. In Kolodner, R (Ed) Computerizing large integrated health networks. New York: Springer-Verlag.

Vance, B., Kraft, M. R., & Lang, B. (1998) Nursing software development and implementation: An integral aspect of the Veterans Health Administration information system infrastructure. In Moorhead, S. & Delaney, C. (Eds.) Information systems innovations for nursing: New visions and ventures. Thousand Oaks, Calif.: Sage Publications.

Werley, H. & Leshe, J. (1991) Standardized comparable, essential data available through the nursing minimum data set. In Turley, J. & Newbold, S. (Eds) Nursing Informatics 91: Pre-conference proceedings. Heidelberg-Berlin: Springer-Verlag.

Zielstorff, R. (1995) Capturing and using clinical outcome data: Implications for information systems design. Journal of the American Medical Informatics Association. 2: 191-196.