MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Can the US Minimum Data Set Be Used for Predicting Admissions to Acute Care Facilities?

Patricia A. Abbott, MS, RNC[a-b], Stephen Quirolgico, MS, Doctoral Candidate[b], Roopak Manchand[a], MSc, Kip Canfield, PhD[d], Monica Adya, PhD[b]

[a]*University of Maryland School of Nursing,*
[b]*University of Maryland Baltimore County,*
[c]*University of Maryland School of Medicine,*
[d]*Laboratory for Healthcare Informatics - University of Maryland Baltimore County*

Abstract

This paper is intended to give an overview of Knowledge Discovery in Large Datasets (KDD) and data mining applications in healthcare particularly as related to the Minimum Data Set, a resident assessment tool which is used in US long-term care facilities. The US Health Care Finance Administration, which mandates the use of this tool, has accumulated massive warehouses of MDS data. The pressure in healthcare to increase efficiency and effectiveness while improving patient outcomes requires that we find new ways to harness these vast resources. The intent of this preliminary study design paper is to discuss the development of an approach which utilizes the MDS, in conjunction with KDD and classification algorithms, in an attempt to predict admission from a long-term care facility to an acute care facility. The use of acute care services by long term care residents is a negative outcome, potentially avoidable, and expensive. The value of the MDS warehouse can be realized by the use of the stored data in ways that can improve patient outcomes and avoid the use of expensive acute care services. This study, when completed, will test whether the MDS warehouse can be used to describe patient outcomes and possibly be of predictive value.

Keywords

Knowledge Discovery in Large Databases; Classification; Minimum Data Set; Nursing Informatics.

Introduction

Advances in database technology in concert with the push for computerized patient records have enabled healthcare organizations to collect and store data at rates unimaginable decades ago. Unfortunately, this headlong rush into the capture and storage of patient related data has led to massive data warehouses that have become unwieldy for purposes of analysis. This conundrum is not specific to healthcare; in reality, many other fields such as banking, industry, finance and manufacturing are faced with identical problems. Fayyad [1] in describing the process of capturing data for analysis from the torrents of data entering warehouses likened it to "drinking from a fire hose". Indeed, the issue at hand is how to integrate, harness and analyze the massive amounts of data that are entering and being stored in large data repositories.

KDD Process

The techniques used in KDD and data mining are not new, having been discovered in the field of Artificial Intelligence (AI) research in the 1980's. However, these techniques are just now gaining popularity in the analysis of large datasets in healthcare. Several factors have contributed to convergence, such as the widespread availability of scaleable information technology, the dawning awareness of the value of information, and widespread deployment of high volume integrated health care information systems. Prior to this confluence, the vast majority of health related data was stored in disparate data silos with (in many instances) limited capacity.

The move towards integrated health systems and the tracking of data from "cradle to grave" has highlighted the need for a method by which the vast amounts of data can be analyzed and visualized. KDD can be used to sift through large repositories to "discover" trends, predictive patterns, or correlation's in data; confirm hypotheses; and highlight exceptions. This functionality becomes critical in the light of intense competition for scarce healthcare resources. KDD and data mining allows organizations to begin to make sense of the data being collected. Failure to do so increases the risk of surrendering the value of the data that enormous amounts of resources are being devoted to capture.

Fayyad, Piatetsky-Shapiro, and Smyth [2] define KDD as the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." According to Goodwin [3] data mining and KDD in healthcare uses a "combination of artificial intelligence and computer science techniques to help build knowledge in complex health care domains." In essence, KDD can be viewed as extracting high-level knowledge from low level data [1]. KDD uses discovery-based approaches in which pattern recognition and matching, classification or clustering schemas, and other algorithms are used to detect key relationships in the data. The ability to "discover" facts separate from the original hypothesis is viewed as a value-added activity.

KDD in Healthcare

In practical terms related to healthcare, the use of KDD and data mining holds much promise. Studies have shown that the overload of data provided to healthcare providers can inhibit the decision making process [4]. With the advent and increasing frequency of computer based patient record initiatives, providers are being faced with the potential for drowning in patient data. While it is commendable that the computerization of health data is moving forward, the inability to organize, access, and analyze large datasets is troubling. Providers, administrators, and organizations require methods such as KDD to be able to filter the "essence from the bulk". In other words, the results of the classification resulting from KDD and data mining can be incorporated into clinical systems to assist clinicians in transforming voluminous amounts of data into comprehendible information.

Of particular interest to these authors is the Health Care Finance Administration (HCFA) resident assessment instrument (RAI), based on the Minimum Data Set (MDS), for patients in long-term care (LTC) facilities. The MDS is a patient assessment tool with over three hundred data points per assessment period which is collected by all LTC facilities in the US that participate in Medicare/Medicaid programs. The MDS is administered quarterly, and more often if marked change in the patient's status is evidenced. Two studies of particular importance focused upon the use of the MDS in US long term care facilities [5]. The results of these two studies demonstrated a positive impact in three separate areas; resident outcomes, processes of care, and administrative perspectives. The MDS was shown to be an excellent outcome measurement tool in which data can be collected, aggregated and analyzed to examine processes of care.

MDS

The Omnibus Budget Reconciliation Act (OBRA) of 1987 contained Nursing Home Reform Amendments. The Amendments imposed new rules for nursing home care, requiring that long-term care facilities which participate in Medicare and Medicaid programs perform periodic assessments of the functional capacity of individual residents. OBRA required the establishment of a minimum data set by which the functional capacity of residents could be measured.

This minimum data set (MDS) was to be composed of a set of core elements and common definitions for use by the LTC facilities. In addition, guidelines were established for use of the MDS and an assessment instrument was designated. In 1992, HCFA proposed as a requirement for participation in the Medicare and Medicaid programs that facilities encode the MDS in a machine-readable form capable of reporting MDS data to HCFA on a routine basis. [6] The original deadline for compliance with this mandate was October 1, 1994. HCFA anticipated that a national MDS collection would:

1. foster the growth of a management information systems in LTC's which would contribute to an improvement in the cost-effectiveness of LTC;
2. identify effective care patterns which would facilitate trends, highlight the use of resources, assist in the evaluation of policy and program options in long-term care delivery and financing, and promulgate national and regional analyses. In particular, the interests included: case-mix payment, patterns, organizational structure, and physician participation.
3. encourage quality monitoring by focusing on targeted reviews which would increase the chances of detecting problem facilities. The collection of large amounts of data from the MDS would also facilitate the development of nationwide patient norms by which all LTC residents could be measured. Once aggregated by facility, the use of the data sets could be used for aggregate analysis and comparisons at state, regional and national levels.

The use of the MDS has met with limited success in LTC facilities. A noted failure of the MDS is the lack of longitudinal and cross-sectional feedback to clinicians. This contributes to the perception of the MDS as additional administrative burden rather than a clinical tool for improving or measuring patient outcomes. As noted previously, the MDS is an excellent assessment tool in which the effect or processes of care can be examined both longitudinally and cross-sectionally. The benefits to be gained by benchmarking the MDS both within and across facilities, regions, and healthcare systems are tremendous. The three major challenges to be overcome include: the mining of data required for research from the MDS warehouse, the ability to properly classify or predict clinical trends, and the need to return the information *essence* back to clinicians and administrators in LTC facilities.

Acute Care and LTC Residents

The use of acute care services by LTC residents is a negative outcome, potentially avoidable, and expensive. Previous research in elder patient care has shown that admission to an acute care facility from a LTC results in a 59% increase in complication rates [7], a three-fold increase in hospital morbidity and mortality [8], and a general worsening of psychiatric status in 40% of patients [9]. Although many admissions to an acute care facility cannot be avoided, early detection and interventions aimed at minimizing hospitalization are of critical importance in this frail population. The value of the MDS warehouse can be realized by the use of the stored data in ways that can improve patient outcomes and possibly avoid the use of expensive acute care services.

In order to effect the rate of hospitalization, we must first be able to measure and predict risk. Measuring and predicting risk can be used as a clinical decision support resource to impact upon provider behaviors and care patterns. The prediction of risk of admission to an acute care facility from a LTC may be based upon a comprehensive patient assessment that addresses clinical, functional, and interventional components. In addition, factors specific to the LTC such as ownership, bed size, types of special units, and category/numbers of staff require investigation in relation to their contribution to admission risk.

It is premised that with the use of Knowledge Discovery in Large Datasets (KDD) procedures and connectionist machine learning techniques, the data in MDS warehouses can be inves-

tigated and the assessment of risk of admission to an acute care facility from a LTC can be performed. The assessment and subsequent alerts to providers of admission risk can be used to signal that processes of care may need attention, with the potential for improvement of patient outcomes.

Due to the enormity of the HCFA MDS data warehouse, the effort to set the parameters, extract, and clean data is daunting. The goals (as set forth by HCFA) are contingent upon the ready access and analysis of data stored by HCFA in the MDS files. The immensity of the warehouse is a challenge, requiring advanced mining, analysis and classification work to complete the expectations and alter user perceptions as mentioned in the previous section.

Current Study

The selection of the target data from the massive MDS warehouse required the setting of several inclusion factors. The inclusion criteria included:

1. the presence of three of more MDS assessments in the specified time period; assessments for the 1993-1995 time period.
2. facilities with a bed size of 500 or larger and located in a similar geographic region.

These criteria were necessary to reduce the data set to manageable and comparable levels. These criteria resulted in the selection of six large facilities in the New York area and further extraction of assessments that met the additional inclusion criteria as noted above. All facility and resident identifying information were stripped prior to dataset delivery to this researcher.

The data that is being used in this study includes the MDS data files from 1993-1995, which included 8,103 residents for a total of 44,733 individual assessments. Each assessment has 772 variables. The MDS file is part of the HCFA Multi-State Nursing Home Case-Mix and Quality (NHCMQ) demonstration project. The HCFA Inpatient Claims Datafile, which contains 8,524 claims with 51 variables each (matching the residents and assessments included in the MDS file), was extracted and will be used for classifier training. Twenty-eight OSCAR certification surveys specific to the selected facilities (272 fields per survey) will be used for the "facility factor" variables.

The inclusion of the Inpatient Claims File is based upon the need for the outcome of "admission" to validate the classification scheme derived from the data mining experiment. Presence of inpatient claims data can be said to represent an acute care admission. The Inpatient Claims File includes artificially constructed patient identifiers for matching purposes, admission and discharge dates, and diagnosis codes. This file will be used to validate the results of the classifier. The variables to be used as classifier input will require reduction via a Principle Components Analysis. Several data mining algorithmic approaches such as association, clustering, classification, sequence-based analysis, and estimation, can be used to extract the relevant relationships in the data. The most common approach (and that chosen for this study) is classification algorithms, based on neural networks.

The classification being utilized involves using a set of pre-classified examples to develop a model, which can then be used to classify (via a standard back propagation neural network) other data in the same domain. In this case, the training, testing, and implementation of the classifier will involve subsetting a fraction of the MDS dataset and running the neural net on a training subset comprised of pre-classified cases (those who have been admitted to an acute care facility). The neural net will then be used to determine the set of parameters and weights required to properly discriminate the outcome of "admission". The performance of the classifier will then evaluated by comparing the scheme results to actual outcomes of "admission" contained within the data set. Implementation of the classifier in the predictive mode will then used to classify the remaining cases (those without the claims file attached).

The importance of this technique is that the classifier, if validated on this particular standardized health dataset, can be implemented in the predictive mode to support decision-making activities in similar LTC settings. This can have important implications for the design of automated MDS systems for LTC. Immediate feedback in the form of alerts to clinicians about the risk of admission to an acute care facility based on the input to the MDS may alter the plan of care. This in turn has the potential to lead to improved patient outcomes and a reduction in cost shifting from LTC to acute care facilities.

Conclusion

This design and preliminary study description paper presents work which is currently underway as part of a doctoral dissertation. The intent of this study is to investigate and test the potential of the MDS assessment as a tool for predicting admission to acute care from long-term care facilities. The use of KDD and classification algorithms were presented and described as ways to perform this study. The process of set extraction and cleaning is underway, which has resulted in a dataset comprised of 44,733 assessments. The development and testing of the classifier on this large MDS file is anticipated in the near future.

References

[1] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, Oct. 1996: 20-25.

[2] Fayyad, U., Piatetesky-Shapiro, G., Smyth, P. From Data Mining to Knowledge Discovery. In Fayyad. U., Piatetesky-Shapiro, G., Smyth, P (Eds) Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA; 1996: 1-36.

[3] Goodwin, L. Data Mining for Improved Patient Outcomes. In press; Connections, 1997.

[4] Kushniruk, A., Patel, V., Fleizer, D. Analysis of Medical Decision Making: A Cognitive Perspective on Medical Informatics. In; Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care. Hanley & Belfus: 1995, 193-197.

[5] Phillips, C., Hawes, C., Mor, V., Fries, B., & Morris, J.

Evaluation of the Nursing Home Resident Assessment Instrument; Executive Summary. HCFA Contract #88-500-0055. 1996.

[6] http:/www.rti.org/publications/RAI_gerontologist.html

[7] Duxbury, (1996). Geriatrics:Unmasking polypharmacy problems and adverse drug effects. Consultant, April, pp.762-776.

[8] Charleson, M. (1987). Morbidity during hospitalization: Can we predict it? Journal of Chronic Disease, (40), 705.

[9] Gillick, M. (1982). Adverse consequences of hospitalization in the elderly. Society of Science in Medicine, (16), 1033.

Address for correspondence

Patricia A. Abbott, MS, RNC
University of Maryland School of Nursing
655 West Lombard Street
Department of EAHPI
Baltimore, MD 21201 USA
abbott@gl.umbc.edu