

## Knowledge Discovery: Detecting Elderly Patients with Impaired Mobility

Der-Fa Lu<sup>a</sup>, William Nick Street<sup>a,b</sup>, Connie Delaney<sup>a</sup>

<sup>a</sup> College of Nursing, The University of Iowa, Iowa City, Iowa, USA

<sup>b</sup> Tippie College of Business, The University of Iowa, Iowa City, Iowa, USA

### Abstract

*Immobility is an important health concern for the elderly patients and healthcare providers who care for the elderly. The purpose of this study was to test a knowledge discovery method to detect elderly patients with impaired mobility in a large clinical dataset. The research method applied an exploratory design and a data mining classification method (cost sensitive Decision Tree J48 from WEKA) to classify patients. Important factors were identified by the Feature Selection method. The Decision Tree algorithm classified patients in the dataset with 65% sensitivity and 72% specificity for a reduced model. The results were evaluated by 10-fold cross validation. Examples of decision rules were also extracted. The study can be applied to classify different health problems in different populations and serves as a foundation for the development of healthcare decision support systems.*

**Keywords:** Knowledge Discovery, Elderly Patients, Impaired Mobility

### Introduction

By the year of 2030, more than 70 million (20%) of the United States population will be aged 65 and older (1). Physical mobility will take part in a key role in managing health problems for the elderly population, e.g., depression, and other chronic diseases (2). Healthcare providers can benefit from knowledge discovered from clinical patient care data regarding impaired mobility for the elderly patients. Clinical patient care datasets represent complex patient care information.

Traditional statistical methods have limitations in performing data analysis for large healthcare datasets with complex variables. Knowledge discovery in databases and data mining can be used to detect novel and useful patterns of information. The purpose of this study was to test a knowledge discovery method to detect elderly patients with impaired mobility problems in a large clinical dataset. The specific aim of this study was to detect patterns of information and important attributes in elderly patients with impaired mobility.

### Methods

A clinical patient care dataset from a community hospital in the Midwest was used for this study. The dataset

contained information regarding patient demographic data and patient care data. Knowledge discovery in databases (KDD) processes were used. The KDD processes include 5 steps (3,4): 1) dataset selection; 2) data preprocessing; 3) data transformation; 4) data mining; and 5) data interpretation.

### Data Selection

One of the investigators has more than 10 years of collaboration with the research hospital, and provided the access to the patient care dataset for the study. The research hospital has used nursing diagnosis in their patient care documentation system for more than 20 years. The hospital provides educational programs on standardized nursing languages training to both new and existing nursing staffs to ensure the reliabilities of data in the patient care dataset. The hospital also has an on-going audit system to monitor the validity of the data entered by clinicians. After identifying the research question, a subset of data was extracted, based upon patient age (greater or equal to 65 years old) and inpatient care. Impaired mobility nursing diagnosis was the most critical health problem among these elderly patients. The study dataset had 8,259 elderly patient records, with 20% of their records with nursing diagnosis of Impaired Mobility, and 80% of patient records without Impaired Mobility. The patient services distributions are as follows: 33% (n=2,726) were Medicine patients, 26.9% (n=2,221) were Surgical, 15.5% (n=1,281) were Transitional-skilled unit, and 6.5% (n=535) were Neurology patients. Eight independent variables based upon literature and the researchers' clinical knowledge (Table 1) were chosen for the model.

The International Classification of Disease Clinical Modification Version 9 (ICD-9-CM) was used in the original dataset for coding diseases. There were more than 800 disease codes found in the extracted elderly patient dataset.

### Data Preprocessing

Computer programming PERL (Practical Extraction and Report Language) was used in parsing the dataset and generating recode statements for transforming disease codes. The original ICD-9-CM disease codes in the dataset were not feasible for data analysis due to small number of sample in each disease category. The Clinical Classifications Software (CCS) for ICD-9-CM conversion from the Agency of Healthcare Research and Quality (AHRQ) (5) was used to recode disease into 250

categories. The data set consisted of 8 independent variables (Gender, Race, Service, Primary Insurance, Marital Status, Religion, Disease code, and Age), and 1 dependent variable (Impaired Mobility). The dataset was transformed into an Attribute Relationship File Format (ARFF) in preparation of using WEKA software.

**Data Transformation**

No data transformation was needed for this study.

**Data Mining**

Waikato Environment for Knowledge Analysis (WEKA) is data mining software which provides both supervised and unsupervised machine learning methods (6); in this case, supervised learning methods were applied to predict impaired mobility.

A cost sensitive classifier was used to increase the weight of Class 1 cases. After several experiments, the cost sensitive matrix at 3 for the cost of false negative (when Class 1 patients with Impaired Mobility were classified as Class 0 without Impaired Mobility), and 1 for the cost of false positive (Class 0 patients without Impaired

Mobility) were classified as Class 1 with Impaired Mobility) performed the best. Among different methods of classifications, the Decision Tree (J48 algorithm) produced the best results. The Decision Tree (J48) is based on the Information Grain theory (split on the node with the most information gain), and generates decision rules that are easy to interpret (7). Ten-fold cross validation method was used to evaluate the performance. In ten-fold cross validation, 10% of data was used for testing, and 90% of data for training (7). This process was repeated 10 times, so that every data was used once for testing

The Feature selection to identify important factors was performed using Wrapper Subset Evaluator and Naïve Bayes classifier. The above data mining procedure was repeated on these selected features. Feature selection identified 5 variables with the strongest association with impaired mobility, they are: Gender, Service, Marital Status, Disease code, and Age. Three variables (Race, Primary Insurance and Religion) were not included from this procedure, due to lower predictability.

Table 1. Variables and descriptions

| Variable names        | Descriptions   |
|-----------------------|--|
| Independent Variables |  |
| Gender                | Female, male or missing  |
| Race                  | 7 categories for ethnicity   |
| Service               | 12 major types of patient care services  |
| Primary Insurance     | Medicare, Medicaid, and commercial insurances  |
| Marital Status        | 6 categories to describe marital status  |
| Religion              | 30 categories to describe different religious preferences Convert from ICD-9-CM to CCS (250) |
| Disease code          | 9-CM to CCS (250)  |
| Age                   | Continuous variable ( $\geq 65$ )  |
| Dependent Variable    |  |
| Impaired Mobility     | Class 0 = “without Impaired Mobility”, Class 1 = “Impaired Mobility”                         |

**Data interpretation**

Results from the data mining were validated by a group of nurses who have experiences in caring for elderly patients. Among these nurse experts, 2 qualified with PhD’s in gerontological nursing, and 1 as a PhD candidate working as an Informatician at the hospital where the clinical data originated.

independent variables. The examples of decision rules from data mining methods are illustrated in Table 2. The model with 5 variables generated sensitivity of 65%, and specificity of 72%. The sensitivity from the reduced model is complementary to results from the full model with 4% difference.

**Results**

The cost sensitive Decision Tree classifier was able to detect elderly patients of impaired mobility with a 69% of sensitivity and 70% of specificity using all 8

The elderly patients in the chosen hospital are primarily Caucasian, Christians, and Medicare Medicaid recipients. Lack of variation in race, primary insurance, and religion accounted for their exclusion from the model.

Table 2. Examples of Decision Rules from Data Mining

| Decision Rules (1 = patient with impaired mobility)   |
|---|
| If Service = Rehabilitation, then 1.  |
| If Service = Transitional-skilled unit, then 1.   |
| If Service = Surgery, and disease = chest pain, then 1.   |
| If Service = Surgery, and disease = Dysrhythmia, and marital status = single or widow or unknown, then 1. |
| If Service = Surgery, and disease = Acute CVD, and gender = female, then 1.                               |

If Service = Surgery, and disease = pre-cerebral occlusion, then 1.  
 If Service = Surgery, and disease = other circulation problem, and gender = female, then 1.  
 If Service = Surgery, and disease = other disease of veins and lymphatic system, then 1.  
 If Service = Surgery, and disease = acute bronchitis, then 1.  
 If Service = Surgery, and disease = COPD, and gender = female, then 1.  
 If Service = Surgery, and disease = COPD, and gender = male, and insurance = Medicare, then 1.  
 If Service = Surgery, and disease = Cancer of Colon, and age greater than 77.36 years, then 1.  
 If Service = Surgery, and disease = skin and subcutaneous tissue infection, and gender = female, then 1.  
 If Service = Surgery, and disease = skin and subcutaneous tissue infection, and gender = male, and age greater than 79.49, then 1.

## Discussion

This study was able to classify elderly patients with impaired mobility nursing diagnosis with an acceptable range of sensitivity around 69% for the whole model of 8 variables and 65% for the reduced model of 5 variables. Different decision rules with patient characteristics were extracted. These various conditions identified from decision rules of the study are useful to influence the decision process for clinicians who care for elderly patients. For example, more than 40% (6/14) of identified rules are associated with diseases in circulatory system. Older male patients (greater than 79.49) are at higher risk for impaired mobility than their counterparts when they have skin tissue infection and surgical procedures.

To the knowledge of the investigators, there is no previous published study which used data mining method to classify elderly patients with impaired mobility. This study provides information about risk factors for elderly patients with impaired mobility, which can be used in the design of quality improvement interventions, and outcome studies. However, the study only demonstrated association and could not prove causal relationships between the independent and dependent variables. The study procedure can also be applied to other nursing diagnoses in different patient populations. Furthermore, the collection of decision rules can be used to as the foundation in development of clinical decision support systems.

## References

- [1] Center and Disease Control and Prevention, & Merck Institute of Aging & Health, 2005. *The State of Aging and Health in American*, 2004. Washington DC: Merck Company Foundation.
- [2] Center and Disease Control and Prevention, & Merck Institute of Aging & Health, 2005. *Spotlight: physical activity and older Americans*. In *The State of Aging and Health in American*, 2004. Washington DC: Merck Company Foundation.
- [3] Dunham, M. H. *Data mining: introductory and advanced topics*. New Jersey: Pearson Education., 2003
- [4] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). *Advances in knowledge discovery and data mining*. CA: AAAI/MIT Press., 1996
- [5] Agency on Healthcare Research and Quality. Retrieved from the Web, March 14, 2004. <http://www.hcup-s.ahrq.gov/toolsoftware/ccs/ccs.jsp>
- [6] WEKA 3-4, (2004). *Waikato Environment for Knowledge Analysis 3-4*. New Zealand: University of Waikato.
- [7] Witten, I., & Frank, E. *Data mining: practical machine learning tools and techniques with JAVA implementations*. CA: Morgan Kaufmann Publishers., 2000