Interpretable Predictive Models for Knowledge Discovery from Home-Care Electronic Health Records

Bonnie L. Westra, PhD, RN, FAAN[1,*]; Sanjoy Dey, BSc[2]; Gang Fang, M.Sc[2]; Michael Steinbach, PhD[2]; Vipin Kumar, PhD[2]; Cristina Oancea, PhD-C[3]; Kay Savik, MS[1]; Mary Dierich, PhD-C, RN[1]

[1]University of Minnesota, School of Nursing, Minneapolis, MN 55455, USA.
[2]University of Minnesota, Department of Computer Science and Engineering, Minneapolis, MN 55455, USA.
[3]University of Minnesota, School of Public Health, Division of Environmental Health Sciences, Minneapolis, MN 55455, USA.

Submitted June 2010. Accepted for publication December 2010.

ABSTRACT

The purpose of this methodological study was to compare methods of developing predictive rules that are parsimonious and clinically interpretable from electronic health record (EHR) home visit data, contrasting logistic regression with three data mining classification models. We address three problems commonly encountered in EHRs: the value of including clinically important variables with little variance, handling imbalanced datasets, and ease of interpretation of the resulting predictive models. Logistic regression and three classification models using Ripper, decision trees, and Support Vector Machines were applied to a case study for one outcome of improvement in oral medication management. Predictive rules for logistic regression, Ripper, and decision trees are reported and results compared using F-measures for data mining models and area under the receiver-operating characteristic curve for all models. The rules generated by the three classification models provide potentially novel insights into mining EHRs beyond those provided by standard logistic regression, and suggest steps for further study.

Keywords: electronic health records, oral medication management, data mining, home care, rules classification

*Corresponding author: Professor Bonnie L. Westra, University of Minnesota, School of Nursing, 5–140 Weaver Densford Hall, 308 Harvard St SE, Minneapolis, MN 55455, USA. *westr006@umn.edu*. Other authors:*sanjoy@cs.umn.edu*, *gangfang@cs.umn.edu*, *steinbac@cs.umn.edu*, *kumar@cs.umn.edu*, *sco@cccs.umn.edu*, *savik001@umn.edu*, *dieri001@umn.edu*

1. INTRODUCTION

The terabytes or even petabytes of health data available in electronic health records (EHRs) present new opportunities and challenges for research that aims to effectively use these data to discover new knowledge to improve healthcare. Under the American Recovery and Reinvestment Act of 2009, $19.2 billion in grants have been awarded to help achieve the goal that 90% of physicians and providers, along with 70% of hospitals, fully implement EHRs by 2019 [1]. Providers and hospitals will receive incentive payments starting in 2011 for meaningful use of EHRs requiring abstraction of EHR data to report quality measures. This is a laudable first step, but the more important opportunity is the reuse of EHR data to discover new knowledge for predicting outcomes and testing new methods to improve outcomes.

Challenges exist in the reuse of EHR data for research, because it is intended for clinical rather than research purposes, namely, to describe patient needs, care provided, and results of care. First, branching questions are often used to streamline documentation, but for research, branching questions result in missing data for dependent questions. An example of a branching question is whether the patient had a primary caregiver. If not, then dependent questions about the type and frequency of care giving are skipped. A second issue is that some variables are clinically important but have little variance in terms of their values in the dataset. Interventions are an example; some interventions are performed infrequently yet are highly relevant to the outcome of interest.

Well-known methods for developing predictive models, such as logistic regression, are available for use with healthcare outcomes. However, with the massive amount of data now available from EHRs, the application of data mining techniques can aid in knowledge discovery of the factors contributing to the quality of care and outcomes. Knowledge discovery is a multistep process that extracts potentially useful and previously unknown information from raw data [2, 3]. Data mining, which is one step in the process of knowledge discovery, has been defined as "the semiautomatic exploration via a computer driven program, and analysis of large quantities of data in order to discover meaningful patterns and rules [4] (p.5)." It combines methods from statistics, machine learning (artificial intelligence), and pattern recognition to extract new, interesting, and nonobvious knowledge from the data. The multiple steps for discovering patterns to inductively build profiles or models include data selection and abstraction, preprocessing, data mining (analysis), and validation. Although data mining has been used in many other fields, its application to analyzing health information is relatively new. In healthcare, it has been useful in examining patient factors and agency characteristics associated with the length of home health care services [5], predicting admissions to acute care facilities [6], severe-trauma management [7], and identification of medication errors or near misses [8, 9]. In this paper, we explore methods for knowledge discovery from EHRs and share insights gained from comparing methods of analyzing EHR data.

1.1. Purpose of Study

The purpose of this methodological study was to compare methods for developing predictive rules from EHR visit data that are parsimonious and clinically interpretable

using logistic regression and three data mining classification models. To that end, we compared a linear predictive model, logistic regression, which is a standard analysis technique used in the healthcare and biostatistics domains, with selected data mining techniques applied to a case study for one outcome of improvement in oral medication management. Three well-established and widely used methods of classification were used: rules (Ripper), decision trees (DT), and Support Vector Machines (SVM). Classification is the discovery and validation of models, such as a set of rules, which uses the patient-specific values of selected variables to predict classes of patients (for example, those patients who are likely to improve or not improve with respect to their oral medication management).

2. METHODS

This study is a secondary analysis of a limited set of home care EHR data from 15 home care agencies that used 1 of 2 different EHR software systems. The software vendors were selected because they used the only known systems that included and could abstract the required data. After IRB approval, the software vendors selected agencies that they considered effective users of their software for at least six months prior to 2004. The investigators were blinded to the identity of the home health care agencies. The software vendors signed an agreement to participate in the study, which included contacting their home health care customers, explaining the study and obtaining a written consent from the agencies to participate in the study, abstracting the data from the agencies' EHRs, and securely providing it to the investigators.

The selected EHR records included all nonmaternity patients, 18 years of age and older, with **O**utcome and **AS**sessment **I**nformation **S**et (OASIS) data representing a discharge from home care to the community for patients who at admission had problems with managing oral medications. Patients were included if they received services any time in 2004 and had a start or resumption of care OASIS assessment followed by a discharge assessment.

2.1. Data Source

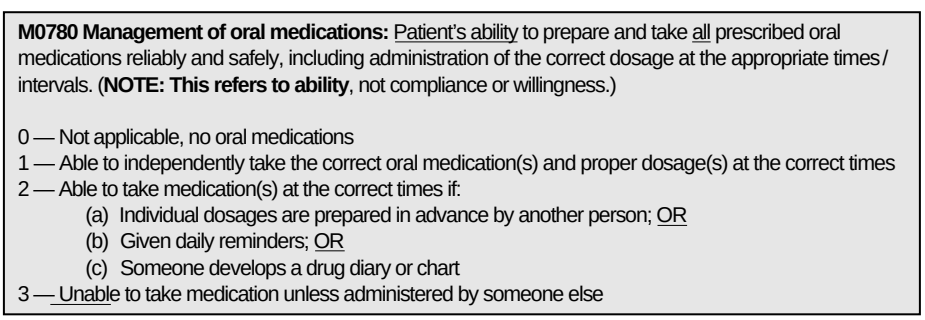
Three types of data were used in this study: OASIS patient assessments, Omaha System interventions, and numbers of medications per patient. OASIS (Version B) includes 79 items that are completed within five days of the start of home health care or within two days of resumption of care and again within five days of discharge from home care to the community [10]. The OASIS data contain demographic and patient history information, living arrangements and social support, health status, functional status, medication and equipment management, and the need for therapy. OASIS is the "gold standard" for assessment and outcome measurement in home care based on 15 years of development [11] and is required for all Medicare and Medicaid patients. The OASIS data are highly audited for accuracy because they are used for quality improvement, public reporting, and calculation of Medicare's prospective payment rate.

Interventions by clinicians during home visits were documented using the Omaha System, which is a structured terminology used to describe the processes of care for environmental, psychosocial, physiological, and other health-related areas of concern

for nursing. An intervention includes a combination of three elements: a problem focus, a type of action, and a more specific target or focus of care. The Omaha System (Version 1) includes 42 problems, 4 categories of action (teaching, guidance, and counseling; providing treatments and procedures; case management; and surveillance), and 63 more specific targets or foci for an intervention. An example intervention is "Pain: providing treatment - medication administration." The Omaha System is a valid and reliable terminology developed over 18 years with federal funding and is nationally recognized by the American Nurses Association [12]. Multiple studies have demonstrated the reliability, validity, and usefulness of the Omaha System for home care, public health, and other practices [13].

Medications were abstracted from the medication record and included all over-the-counter and prescribed medications. No information exists to support the reliability or validity of these records, but they are assumed to be valid because they are part of the legal chart and validated on physician orders.

The variable for calculating the outcome of improvement in oral medication management is shown in Figure 1 and is measured at admission and discharge using the OASIS question M0780: Management of Oral Medications [10]. The variable was recoded such that NA = 0 resulting in a scale from 0–3. The outcome of improvement in oral medication management was evaluated as a change in status from admission to home care to the time of discharge. Improvement was a change from a higher score to a lower score; otherwise, patients who remained the same or got worse were coded as not improving. The final dataset in this study included 1,759 cases and included all patients who had oral medications at admission. This dataset is different than that used by the Centers for Medicare and Medicaid Services (CMS); only patients who could improve in oral medication management are included by CMS in the denominator for the outcome calculation.


M0780 Management of oral medications: Patient's ability to prepare and take all prescribed oral medications reliably and safely, including administration of the correct dosage at the appropriate times / intervals. (**NOTE: This refers to ability**, not compliance or willingness.)

0 — Not applicable, no oral medications
1 — Able to independently take the correct oral medication(s) and proper dosage(s) at the correct times
2 — Able to take medication(s) at the correct times if:
(a) Individual dosages are prepared in advance by another person; OR
(b) Given daily reminders; OR
(c) Someone develops a drug diary or chart
3 — Unable to take medication unless administered by someone else

Figure 1. Assessment of management of oral medications.

2.2. Preprocessing for Data Analysis

Prior to analysis, the data were preprocessed and transformed by several steps: removing records with duplicate assessments and considerable missing data, matching assessments into episodes of care, transforming variables using various data-reduction

strategies, and calculating the outcome of improvement in oral medication management. An episode of care is defined as a continuous number of days during which a patient intermittently receives home visits. A start or resumption of care assessment was matched with a discharge assessment to represent an episode of care. Patients can have more than one episode of care, and each episode was treated as a separate case in this study. The initial dataset contained 2,065 episodes of care that ended with a discharge to the community. The number of episodes of care ranged from 5 to 405 per agency.

Due to the large number of OASIS variables compared with the number of episodes, the following steps were used to reduce the OASIS data. The primary diagnoses were transformed into clinically meaningful diagnoses groups using the Clinical Classification Software [14] and then further reduced by clinical experts into 51 groups of primary diagnoses [15]. A unique listing of 17 possible fields for secondary diagnoses, the reason for an inpatient stay, change of treatment, or payment were transformed into the Charlson Index of Comorbidity to create a summative score for severity of illness (0–37) [16]. By summing related OASIS items, we created scales with the range of possible scores in parentheses: prognosis (0–2), pain (0–4), and respiratory status (0–7). Interventions were clustered into 23 meaningful groups following the Agency for Healthcare Quality Research process for grouping procedures [17]. The complete list of how interventions were grouped is available online [18].

Data preparation included transforming variables to meet the assumptions and methods of analysis. For easy interpretation, some values were collapsed before creating models. For example, originally the patient's ability to transfer had 5 values, but values 0–1 were collapsed and interpreted as meaning that the patient was independent or needed minimal assist; otherwise, the patient was unable to transfer independently. The final data dictionary is available online and includes the variable name, description, and format [19].

2.3. Predictive Modeling

Logistic regression is commonly used to analyze health care data and develop predictive models. We accomplished this by manually doing a stepwise logistic regression with PROC GENMOD in SAS v9.2 that takes into account repeated observations within agencies. One advantage of logistic regression is that it is robust to unbalanced data. PROC GENMOD addresses this in the specification of the binomial distribution for the outcome variable, and the logit as the link function. We used chi-square for evaluating candidate variables and retained variables with a significance level of their association with the outcome at $p \leq .1$. Whereas logistic regression is well accepted in the healthcare literature, one disadvantage is that it does not produce rules, per se, but rather generates models that show the probability of an outcome based on a set of unique variables. Because we were interested in the added value of the interaction of these variables in creating predictive rules, we used data-mining classification techniques that produce such results.

Classification techniques from the data-mining domain are increasingly being used in healthcare to extract information from EHRs. However, each classification technique

has different advantages and disadvantages, and the choice of the appropriate method depends on the challenges arising from the domain and the goal of the study [20]. For the task of inferring new knowledge about home care oral medication management, we used well-established and widely used methods of data-mining classification: Ripper, DT, and SVM with polynomial kernel. We used Weka, which is an open-source machine-learning software package. One of the advantages of these classifiers is the ability to combine multiple variables to create rules; independently, some of these variables may not be significant, but when variables are combined, they may be significant as part of a rule. One of the disadvantages of data mining classifiers is their difficulty in handling imbalanced classes.

Rule-based classifiers, such as Ripper, develop a set of *if-then* rules directly from the dataset by comparing each predictor variable with the outcome class. The "if" portion of the rule contains a combination of predictor variables and the "then" part of the statement is the outcome class (improvement vs. no improvement in oral medication management). The set of rules are exhaustive in the sense that each record is covered by at least one of the rules. Moreover, the rules can be ordered in decreasing order of their priority based on different measures (e.g., accuracy, coverage, or order in which they are generated).

DT produces a set of nodes, organized into a tree, that are either decision points (interior nodes) or nodes that assign a class (leaf nodes). Any particular record follows a path down the tree through various decision nodes until it reaches a leaf node, where it is classified. The path followed depends on the attribute values of the record. DT can be transformed into sets of rules.

SVM views the data records as points in a multidimensional space and classifies records by finding a hyperplane in the data space that produces the widest gap for separating the classes of the records. The idea is that the points with the same class will form a group that is separate from the points that belong to another class. In practice, this simple idea requires a number of sophisticated techniques, such as using kernel functions to map the points to high-dimensional space and optimizing an objective function to obtain the best separating hyperplane. Further descriptions of SVM and the other techniques can be found in an introductory data mining text by Tan, Steinbach, and Kumar [3].

Ripper and DT are useful for creating clinically interpretable models, namely, rules or DT (which can also be turned into rules), while SVM produces a "black box" model with less transparency. However, SVM is known for its robust classification performance, and thus provides a benchmark to measure how much, if any, classification performance is lost by the decision to use classification techniques with more easily interpretable models. For this reason, we analyzed Ripper and DT models for finding clinically interpretable rules and compared the performance of those models with that of SVM.

2.4. Determining Reliability and Validity of Results

The reliability and validity of the results were established and compared in several ways. For logistic regression, the final model included variables associated with the

outcome at $p \leq .05$. For data mining, we used 10-fold cross validation, partitioning the data into 10 parts or folds. A training dataset representing nine folds was used to create the model which was then tested on the 10^{th} fold. This procedure was repeated until each of the folds had been used as a testing set and the results were averaged over the 10 iterations. We compared the data-mining final models for classification using an F-measure, which combines precision and recall measures for predicting both improvement and no improvement of oral medication management. However, the F-measure will vary depending on the relative frequencies of the different classes (i.e., improvement or no improvement). Discrimination (the ability to distinguish patients likely to improve or not improve in oral medication management) was quantified using the area under the receiver-operating characteristic curve (ROC) [21], which is insensitive to class frequency. Values for the F-measure and the area under the ROC (AUC) are reported as a fraction with values ranging from 0–1, with higher accuracy indicated by a closer value to 1. The AUC under the ROC generated by the logistic-regression model was used for comparison with the models derived from data mining. We also compared the logistic-regression model with final Ripper and DT models in terms of the parsimony of the model, ease of analysis, and its clinical interpretability.

3. RESULTS

The initial dataset contained 2,065 episodes of care representing 1,938 patients (6.2% had more than one admission that ended with a discharge to the community), and ranged from 5 to 405 episodes per agency. From the initial dataset, 306 episodes of care were eliminated either due to missing data or no oral medication, leaving 1,759 episodes of care to be analyzed. Of these, 282 (16.0%) improved in managing oral medications by discharge, whereas 1,477 (84.0%) did not. Patients were predominantly Caucasian (98.8%) and older adults (71.5% age 65 or older), with a higher proportion of females (63.5%), and most were recently discharged from an inpatient setting in the previous two weeks (82.6%). The primary payor for home health care services was Medicare (83.2%). The majority of patients (88.3%) were discharged in less than 60 days. Patients had numerous medications to manage, ranging from 1 to 41 medications; almost half (47.3%) had 10 or more medications. The most frequent primary diagnoses groups were: (a) central nervous system, vision, and hearing (26%), (b) orthopedic (20%), and (c) cardiovascular (17%). Patients with central nervous system diagnoses largely represented patients with gait disorders, requiring physical therapy or after-care following hospitalization (82.8%). The most frequent orthopedic conditions were after-care for therapy, rehabilitation, or osteoarthritis (35%), and the most frequent cardiovascular primary diagnoses were atrial fibrillation (7.4%), congestive heart failure (21.1%), and cerebrovascular accidents (10.8%).

3.1. Logistic Regression Results

The dataset used for logistic regression had 1,668 records with complete data, and all variables were included in the final model for creating Ripper and DT rules. The relationships of all variables with the outcome were assessed as candidates for the

logistic model using Chi-Square with $p < .1$, resulting in 35 variables being retained. These variables included the following:

- demographics (age, Medicare payor, no inpatient stay, and living alone);
- health status (prognosis, life expectancy, vision, pressure ulcer, stasis ulcer, surgical wound, respiratory status, cognitive function, time of day confusion occurs, depressive mood, gastrointestinal primary diagnoses);
- functional status (grooming, dressing upper body, toileting, transferring, ambulation/ locomotion, feeding, preparing light meals, housekeeping, using the telephone);
- medication management (oral, inhalant, and injectable medications); and
- interventions of teaching, providing or monitoring respiratory and circulation; teaching and monitoring medications; teaching about disease treatments, providing "other" therapies, and monitoring injury prevention.

Variables significantly associated with the final model of improvement in oral medication management as determined by stepwise logistic regression are shown in Table 1. The AUC for the ROC generated by the model was 0.85 as shown in Figure 2.

Table 1. Results of logistic regression

Predictor Variable	Odds Ratio	95% Confidence Intervals
No prior inpatient stay previous 14 days	0.32	0.20–0.51
Prepare light meals	0.61	0.48–0.78
Oral-medication management at admission	8.50	6.27–11.52
Teaching medications	1.55	1.12–2.14

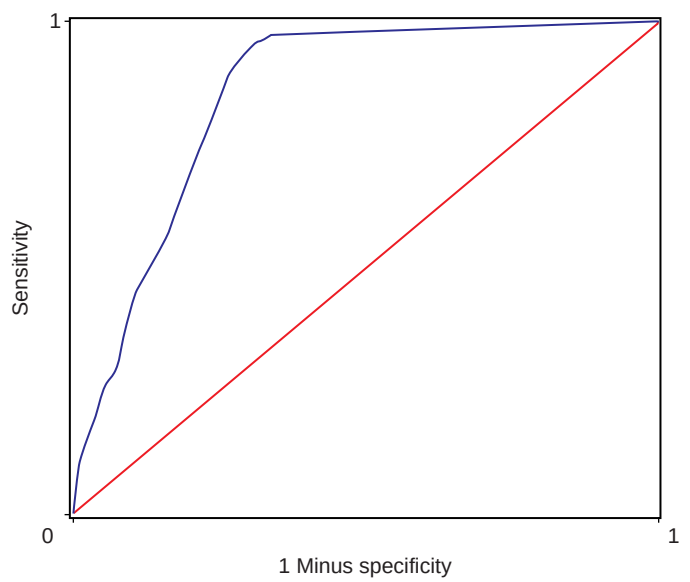


Figure 2. ROC curve for logistic regression model.

3.2. Data Mining Experiments and Results

One issue of the dataset, data with little variance, was handled in data mining by assessing 4 experiments and comparing models as shown in Table 2. Variables with little variance included the following: race/ethnicity; prior indwelling/suprapubic catheter; prior disruptive or socially inappropriate behavior, parenteral nutrition; high risk factors of alcoholism or drugs; depressive symptoms of sense of failure, hopelessness, recurrent thoughts of death, or thoughts of suicide; behaviors that are verbally disruptive, physically aggressive, socially inappropriate or delusional; frequency of behavior problems; and variables that occur within a branching question for wounds. Experiments 1 and 3 included intervention variables, whereas experiments 2 and 4 dropped intervention variables and tested models for the presence or absence of variables with little variance.

The most challenging problem for building classifiers using data-mining classification compared with logistic regression was how to handle the imbalanced classes. The majority class of no improvement contained 1,477 (84.0%) of the cases, while the minority class contained 282 (16.0%) of the cases for improvement (for the second and fourth experiment). To build a classifier on such an imbalanced dataset would yield results that are biased toward the majority class (the largest number of records are for no improvement). If every case is classified as no improvement, then the classifier will accurately classify 84.0% of cases. This is 100% correct for subjects who showed no improvement, but 0% correct for those who did show improvement. Although various methods can be used for managing the problem of imbalanced classes [3], for this study, we chose the straightforward approaches of doing upsampling or downsampling. Upsampling increases the number of the records for the minority class (by using multiple instances of the records of the minority class), whereas downsampling decreases the number of records in the majority class. Each of the four experiments in Table 2 was conducted using both upsampling and downsampling. However, both upsampling and downsampling need to be performed with caution when conducted with cross-validation (CV) framework. If sampling is performed first and then used in the standard k-fold CV, some samples can be present in both the training and testing set because of the possible duplicate entries in the sampling process. To avoid such bias, we designed the following special setups.

Table 2. Dataset and size for experiments for improvement in oral medication management

Experiment	Interventions Variables	"Little Variance" Variables	Sample Size	Number of Variables
1	Yes	Yes	1,668	118
2	No	Yes	1,759	95
3	Yes	No	1,668	99
4	No	No	1,759	76

3.2.1. Upsampling

The upsampling approach duplicates the minority class (improvement) using the ratio of the majority class to minority class (4.91 for experiments 1 and 3, and 5.24 for experiments 2 and 4, respectively). To overcome the problem of common samples in both training and testing classes, both classes were divided into k-folds first and then one sample from each class was used to form the test set. Duplication of the minority class in the test set was not performed in order to better evaluate performance of the classifier in the real world. To build the training set, the left-out samples from the majority class and minority class were used, and then the minority class was duplicated to make both of the classes roughly equal in number. As shown in Figure 3, all duplicate samples belong to the training class only and, hence, the classification model will not be biased. Classification models are then built on the training set of each fold and the model is tested on the test set.

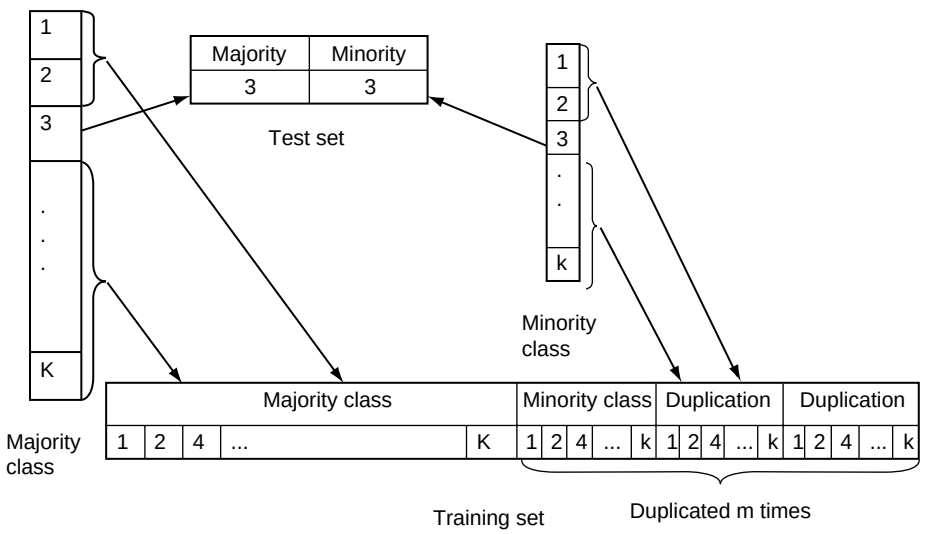


Figure 3. The process of upsampling for the third dataset of CV; m is the ratio of the majority class to minority class.

3.2.2. Downsampling

Similar to upsampling, downsampling was also performed inside the CV loop. First, we divided the majority and minority data into 10-folds for CV, and a random sample of cases was selected without replacement from the majority class to equal the sample size of the minority class (282 in our case), resulting in an equal number of samples for both the majority and minority classes. However, the dataset was greatly reduced (564 total records) in comparison with the original sample size (1,759). Thus, to obtain a better estimate of classification performance, we performed 5 samplings of the majority class for each CV fold. The mean and the standard deviation of F-measures and AUCs for

each random sampling were calculated. As with upsampling, the test dataset was left imbalanced to estimate the performance of the model in the real world.

3.2.3. Results of the Four Data-Mining Experiments

The Weka data-mining tool (version 3.6) was used for all classification models. J48, Ripper, and Sequential Minimal Optimization classifiers from the Weka tools were used for the three chosen classification models: DT, rule-based classifier, and SVM, respectively. SVM used the Sequential Minimal Optimization classifier with polynomial kernel along with the default parameter settings provided by Weka (other kernels implemented in Weka were tried, and polynomial kernel provided the best result). However, we estimated the pruning parameters of each of the three models by a separate CV framework to avoid possible overfitting of the model. Hence, overall, we had two nested CV frameworks: the outer CV used for estimating the performances of the classifiers, and the innermost CV used to estimate the optimum parameter for the training dataset obtained from the outer CV loop.

More specifically, we divided each training data (obtained from the outer CV loop for both upsampling and downsampling) into another k-fold CV framework for creating separate validation sets to tune the parameters for the best performance result. We optimized the model pruning parameters (-C option for J48, -F option for Ripper, and -C option for SVM in Weka) to maximize the AUC. This approach reduces classification bias, although it adds complexity to the process, specifically introducing another level of CV inside the original CV framework developed for upsampling and downsampling. However, the results obtained from the optimized parameters justify this added complexity.

The F-measures and the AUC were the performance metrics used for evaluation of models, as they are considered appropriate metrics for imbalanced classes [3]. We compared the four experiments for all three methods of modeling (Ripper, DT, and SVM) using upsampling and downsampling. F-measures for improvement in oral medication management across all four experiments are shown in Table 3.

We observed very little difference across all four experiments. The F-measures for no improvement in oral medication management across all four experiments are shown in Table 4. The F-measures for all experiments for no improvement were high across all three classification methods. However, the F-measures for any models of improvement in oral medication management were not good. This outcome is a result of the imbalanced nature of the test datasets, which leads to poor precision and ultimately a poor F-measure for the minority class. This result is illustrated using the confusion matrix shown in Table 5. True positives (TP) are records for improvement that are correctly classified, whereas false negatives (FN) are the number of records for improvement that are incorrectly classified as no improvement. False positives (FP) are records for no improvement incorrectly classified as improvement, whereas true negatives are no improvement records correctly classified. Precision is defined as TP/(TP + FP) and recall is defined as TP/(TP + FN). For improvement in oral medication management, precision = 17/(17 + 30) = 0.36 and recall = 17/(17 + 9) = 0.65. With respect to recall, the classifier performed reasonably well for classifying the

Table 3. F-measures for improvement in oral medication management

Experiment	Intervention Variables	"Little Variance" Variables	F-measure Downsampling			F-measure Upsampling		
			Ripper	DT	SVM	Ripper	DT	SVM
1	Yes	Yes	0.52	0.51	0.53	0.49	0.49	0.53
2	No	Yes	0.51	0.51	0.54	0.51	0.51	0.54
3	Yes	No	0.51	0.52	0.53	0.52	0.47	0.55
4	No	No	0.51	0.51	0.54	0.52	0.50	0.55

Table 4. F-measures for no improvement in oral medication management

Experiment	Intervention Variables	"Little Variance" Variables	F-measure Downsampling			F-measure Upsampling		
			Ripper	DT	SVM	Ripper	DT	SVM
1	Yes	Yes	0.78	0.80	0.84	0.82	0.83	0.89
2	No	Yes	0.78	0.79	0.84	0.81	0.80	0.88
3	Yes	No	0.79	0.80	0.84	0.81	0.84	0.89
4	No	No	0.78	0.80	0.84	0.81	0.80	0.88

Table 5. Confusion matrixes for a downsampling result with precision 0.362 and recall 0.654

	Classified as Class No Improvement	Classified as Class Improvement
Class 0 – No Improvement (140)	110	30
Class 1 – Improvement (26)	9	17

improvement, but had poor precision because of the large number of false positives (30) for improvement. The number of records for no improvement is relatively large compared with the number of records for improvement; therefore, even a small error rate for classifying a no improvement record as an improvement record can add a relatively large number of false positives to the improvement class.

This example illustrates two aspects of the F-measure that are important in practice. First, for imbalanced datasets, it is hard to avoid poor precision and a poor F-measure for the minority class unless the false-positive rate is very low. Thus, if the classification models for this study were to be used in practice, and if the class distribution was approximately the same in the general population as in the analyzed dataset, many patients recommended for treatment (roughly two-thirds) would not benefit.

Second, for a fixed-classification model, the F-measure will change as the class proportion changes. For example, the classification models shown above would have a far higher F-measure for the improvement class if the classes were balanced in number. For this reason, many prefer to evaluate classifiers, particularly when comparing different classifiers, by using the AUC, which is insensitive to variations in class distribution. However, that does not change the fact that in a practical setting involving imbalanced classes, the number of FP may exceed the number of TP, sometimes by a large margin, even for the best classifier that can be found.

The AUCs computed for the different classification models are shown in Table 6.

Table 6. AUCs for oral medication management

Experiment	Intervention Variables	"Little Variance" Variables	AUC Downsampling			AUC Upsampling		
			Ripper	DT	SVM	Ripper	DT	SVM
1	Yes	Yes	0.81	0.80	0.79	0.76	0.74	0.74
2	No	Yes	0.81	0.80	0.81	0.80	0.77	0.76
3	Yes	No	0.81	0.81	0.78	0.79	0.67	0.76
4	No	No	0.81	0.80	0.80	0.81	0.74	0.76

In this study, the SVM model was used strictly for estimating how well the Ripper and DT models performed, because SVM results are not easily interpreted [3]. In terms of the F-measures and the AUC performance measures, the rule-based classifiers obtained by Ripper and DT performed reasonably as compared with the SVM. For this reason, only Ripper and DT were explored for rule-based models generated by Weka in the next sections.

As a result of these experiments, we selected experiment 3 with downsampling for subsequent analysis. The interventions were clinically important, and the performance of the models before and after dropping variables with little variance was almost the same. Therefore, we decided to use experiment 3, which included interventions but dropped variables with little variance. No consensus exists on how to choose between upsampling and downsampling. We conducted five random samples with 10-fold CV for experiment 3 to get an overall idea about variations for the different sampling performance metrics.

3.2.4. Ripper Classifier Results

The average F-measures for improvement and no improvement for rules from Ripper experiment 3 with downsampling were 0.51 and 0.79, respectively, while the AUC was 0.81. These values were estimated using 10-fold CV, where estimates for five random samples without replacement were calculated for each fold. The standard deviations for these F-measures and the AUC were 0.06, 0.07, and 0.05, respectively. Two models were created using the default setting in Ripper (-F3) and also with pruning. Pruning

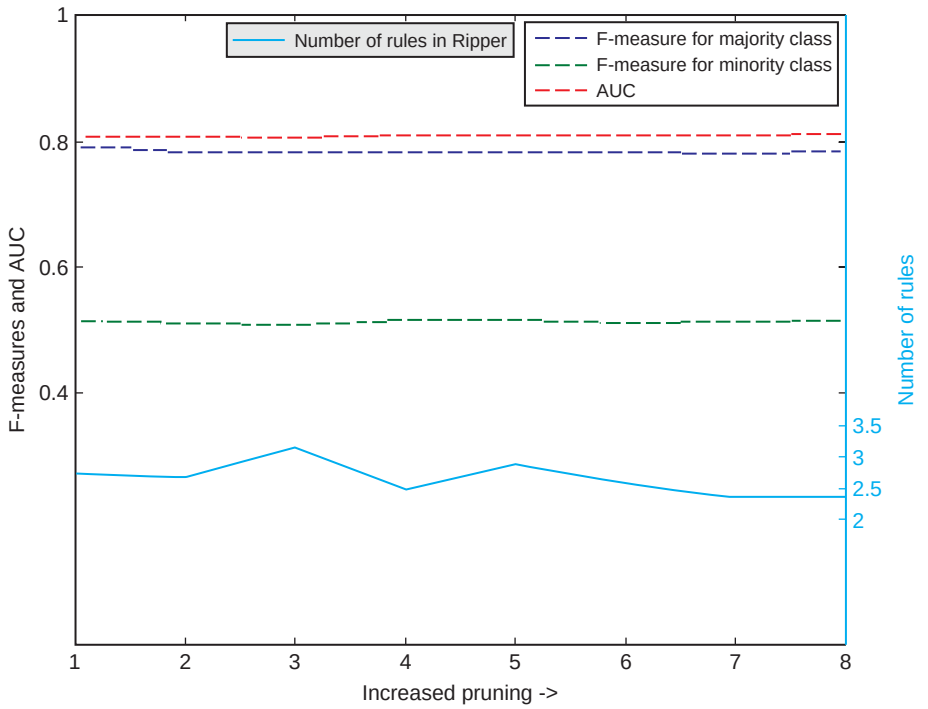


Figure 4. The pruning effect on Ripper with different parameters using downsampling.

had almost no effect on the performance of the Ripper models or the number of rules in those models (Figure 4). The x-axis represents increased pruning for the F option in Weka, which lies in the range of (2, 25). The left-side y-axis is for the F-measure and the AUC, whereas the right side y-axis measures the number of rules for the model.

Models containing both six and eight rules were examined. The model resulting in six rules was not only more parsimonious, but also more clinically interpretable. Figure 5 shows the rules, outcome predicted, and the records correctly and incorrectly classified for the downsampling model with six rules. For each rule predicting improvement, the first component of rules 1–5 included the patient needs assistance with oral medication management.

3.2.5. Decision Tree Results

A DT with downsampling was considered using the Weka tool with the default parameter for our pruning (-C 0.1), but this model produced a large tree with a few hundred nodes that precluded clinical interpretation of the model. To obtain a smaller size DT with more interpretability, the DT was pruned with different parameter settings

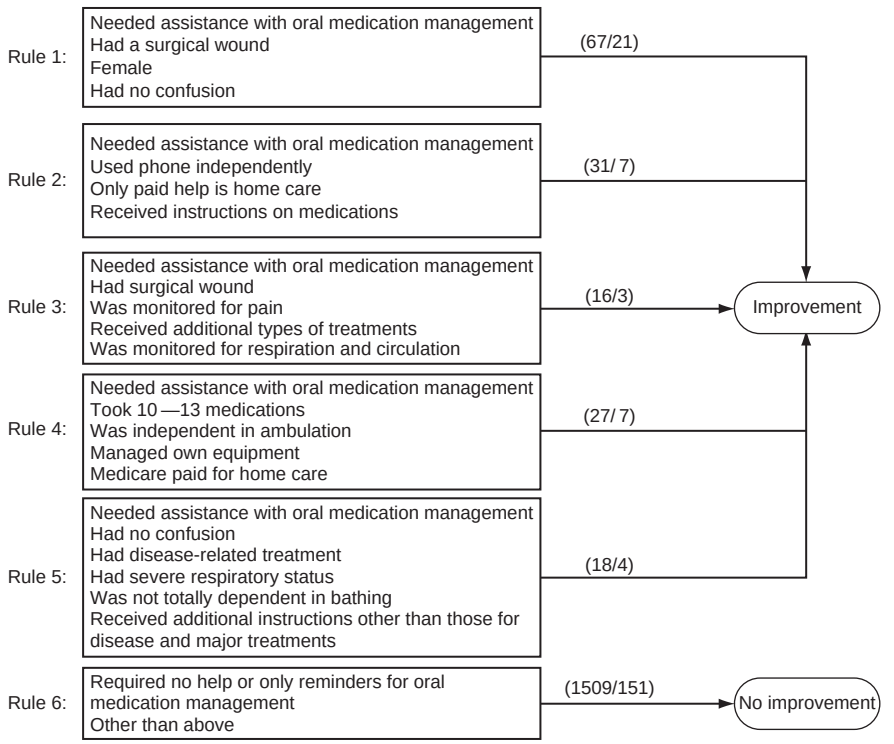


Figure 5. Ripper downsampling model for predicting oral medication management improvement. Numbers in parentheses show records correctly and incorrectly classified. (Parameters: -F 4 -N 2.0 -O 2 -S 1).

ranging from 0.25 to 0.005 in Weka. As shown in Figure 6, there is a slight change for the F-measures; the F-measure for minority class increases and for the majority class there is a decrease as pruning increases, whereas the AUC increases with pruning. The DT model became interpretable for parameter values -C < 0.03. The left-side y-axis shows the F-measures and the AUC, whereas the right-side y-axis measures the number of nodes in DT.

We selected the most pruned tree with the parameter settings with -C = 0.03 for the DT model for downsampling. The mean and standard deviation of the F-measure for the majority class were 0.80 and 0.06, respectively, whereas those for the minority class were 0.52 and 0.07. The mean and standard deviation of the AUC were 0.81 and 0.04, respectively. Each successful branching included the rule for a specific line, as well as any higher rule within the same branch. Figure 7 shows the rules created by the DT.

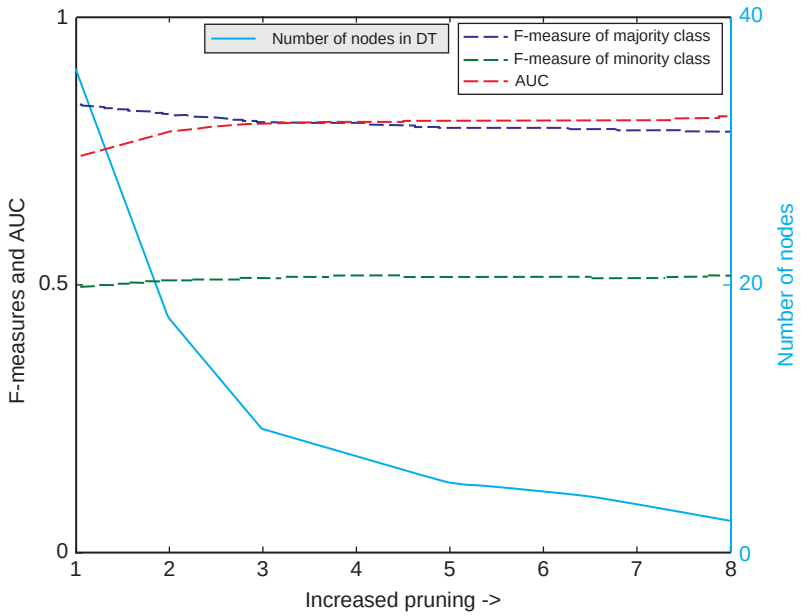


Figure 6. The effect of pruning on the decision tree for the parameter of -C in Weka over the range (0.25, 0.005).

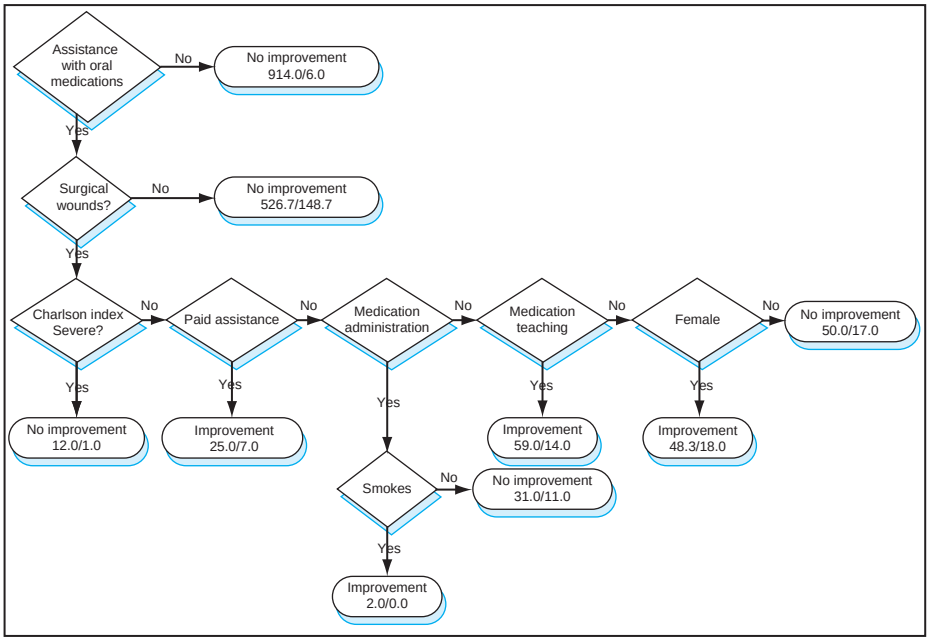


Figure 7. Decision tree rules with downsampling for experiment 3. The numbers for each decision rule represent the number of records correctly classified followed by the number of records incorrectly classified by the rule.

4. DISCUSSION

The purpose of this methodological study was to compare methods of developing predictive rules from EHR visit data that are parsimonious and clinically interpretable using logistic regression and three data-mining classification models. We addressed two major issues: variables with little variance and managing imbalanced classes. We used EHR data to conduct four experiments with different classification methods and compared different data-mining techniques to determine which model(s) can provide the more useful and accurate results. We found that Ripper and DT models performed better compared with the SVM model with respect to both the F-measures and the AUC, as well as with the logistic regression. The data-mining models (AUC .78 – .81) were consistent with the logistic regression model (AUC = .85).

It is not unusual to have imbalanced classes for outcomes in healthcare, particularly in home care where services are provided for a predominantly chronically ill elderly population [22]. In this methodological study, steps were outlined to reduce bias in sampling procedures in order to balance classes for data-mining algorithms. Downsampling was selected as the better procedure in terms of the AUC, but this may be a data-dependent result and could be different in other studies. One of the advantages of stepwise logistic regression using PROC GENMOD is that it does not require balancing classes of outcomes and results in a higher number of records for analysis. In logistic regression, the distribution of outcomes may be as extreme as 90/10 and still provide valid models. However, logistic regression does not result in a set of rules, but results can be easily turned into a model that predicts the probability that the outcome will improve.

Another area of interest in comparing logistic regression and data-mining classification methods was ease of use. Stepwise logistic regression using PROC GENMOD requires manual entry of variables and comparison of each resulting model to determine "stepwise" effects on the overall model. Ripper and DT were fully automatic through CV, and manual selection of variables was not required. However, manual pruning was required for DT, as was additional creation of samples for testing and training sets due to imbalanced classes.

We were able to produce a parsimonious set of clinically interpretable rules with all methods. Logistic regression produced a 4-variable set of predictors easily interpretable for improvement in oral medication management. Ripper produced a set of six rules using 15 variables. The rules are mutually exclusive and covered all cases for improvement or no improvement in oral medication management. We were unable to produce a parsimonious set of rules with DT without further pruning. However, pruning produced eight decision rules that combined 1–4 decisions in a branch. We shared results of this study with four home care and geriatric clinicians/researchers. They were able to grasp the rules, and indicated that Ripper rules in Figure 5 were more easily interpreted. Both Ripper and DT rules are complex to implement without creating a computerized algorithm; however, they do reflect the complexity of decision making by clinicians. Compared with logistic regression, rule-based models are more complex, but they elucidate more information for clinical decision making, which can be easily integrated into current clinical system through a computer-based algorithm.

We found some similarities, but many differences, in the predictors for improvement in oral medication management. The level of dependency at admission was the strongest predictor of improvement across all three models. Receiving instruction on

medication management was also a strong predictor. Patients who were more dependent at admission for oral medication management and those who received instruction on medication management were more likely to improve compared with patients needing little or no help. Although this seems intuitive, it is interesting to note that, at discharge, only 16.0% of patients actually improved in oral medication management in this study. One possible explanation is that a ceiling effect exists in the measurement of oral medication management. No other studies were found addressing this issue. Another common predictor of improvement was whether a patient received home care for an acute or chronic health problem. In all models, patients with more acute conditions (indicated by the presence of a surgical wound, services covered by Medicare, or whether they were admitted from an inpatient facility) were more likely to improve in managing oral medications. All models included at least one measure of functional status; however, the specific functional dependency differed, perhaps due to associations between functional status variables. Future research should consider creating functional-status scales combining the variables and evaluating overall functional-status abilities associated with outcomes.

To our knowledge, this study is the only one that compares the standard logistic regression model to data mining created rules that combine multiple variables for predicting the outcome of improvement in oral medication management for home care patients. We did, however, find studies with similar patient characteristics and interventions associated with managing medications for community-dwelling older adults [23–25]. We found that interventions were relevant for development of rules across all methods. In a previous study, we investigated the best way to cluster interventions for predictive modeling, as the way interventions are clustered could influence outcomes [26]. We are continuing to develop methods for comparing the clustering of interventions on prediction of home care outcomes. A final issue to consider is the appropriate level of accuracy for models to support clinical decision making. What is not yet clear in healthcare is what rate of accuracy needs to be attained before deciding that applying a rule is worthwhile. It may be that approximately 80% or higher is sufficient for clinical decision-support in an EHR, as decision rules are intended as alerts and not prescriptions for practice. Continued investigation is needed to understand how to create the best rules to support clinical decisions.

4.1. Limitations and Future Research

This is the first study for discovering rules from practical data-mining methods applied to improvement in oral medication management. Further validation is needed with new datasets. Additional research is also needed with new datasets to generalize results to other clinical populations. Because we included all patients with oral medications, not just those who can improve, this was valuable in demonstrating that our models correctly predicted those patients who could not improve and were predicted not to improve, providing confidence in our findings. However, because CMS only includes patients who can improve in the denominator for the outcome of improvement in oral medication management, this study should be repeated with this more limited sample. Another limitation in this study was the amount of missing data, which can influence

results. Because this study was a secondary analysis of EHR data, no inter-rater reliability for documentation of intervention was possible, and future studies would benefit from standardizing documentation first. This study is also limited by a small sample size and should be repeated with a larger dataset. Future studies should include some recent work on visualizing SVM models and compare these results against obtained sets of rules in terms of clinical usability.

5. CONCLUSIONS

In this methodological study, logistic regression and data-mining rules classification models were compared for knowledge discovery from home care EHR data to develop a parsimonious set of clinically interpretable rules. The F-measures obtained by Ripper, DT, and logistic regression were similar to that of SVM, and AUC performance measures were comparable with logistic regression. We demonstrated methods for managing two common problems encountered in EHR data: evaluating clinically important variables with little variance and demonstrating ways for managing imbalanced datasets. Logistic regression produced a more parsimonious clinically interpretable model, while classification rules better reflected the complex decision making. Further development of data-mining algorithms is needed to automate some of the challenging works of balancing classes. This study provides a foundation for building classification rules in the future for clinical decision-support.

ACKNOWLEDGEMENTS

This study was funded by the University of Minnesota Grant-in-Aid program and research assistantship support by the Department of Computer Science, University of Minnesota.

REFERENCES

[1] Steinbrook, R., Health Care and the American Recovery and Reinvestment Act, *N. Engl. J. Med.,* 2009, 360(11), 1057–1060.

[2] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining,* AAI Press/ The MIT Press, Menlo Park, CA, 1996.

[3] Tan, P., Steinbach, M., and Kumar, V., *Introduction to Data Mining,* 1st ed., Pearson Addison-Wesley, Boston, 2005.

[4] Berry, M.J.A., and Linhoff, G., *Data Mining Techniques for Marketing* Wiley, New York, 1997.

[5] Madigan, E. A., and Curet, O. L., A Data Mining Approach in Home Healthcare: Outcomes and Service Use, *BMC Health Serv. Res.,* 2006, 6:18.

[6] Abbott, P. A., Quirolgico, S., Candidate, D., Manchand, R., Canfield, K., and Adya, M., Can the US Minimum Data Set Be Used for Predicting Admissions to Acute Care Facilities? *Medinfo,* 1998, 91318-1321.

[7] Demsar, J., Zupan, B., Aoki, N., Wall, M. J., Granchi, T. H., and Robert Beck, J., Feature Mining and Predictive Model Construction from Severe Trauma Patient's Data, *Int. J. Med. Inform.,* 2001, 63(1–2), 41–50.

[8] Rudman, W. J., Brown, C. A., Hewitt, C. R., Carpenter, W. O., Campbell, B., Tubb, T., and Noble, S. L., The Use of Data Mining Tools in Identifying Medication Error Near Misses and Adverse Drug Events, *Top Health Inform. Manag.,* 2002, 23(2), 94–103.

[9] Hauben, M., and Reich, L., Potential Utility of Data-Mining Algorithms for Early Detection of Potentially Fatal/Disabling Adverse Drug Reactions: A Retrospective Evaluation. *J. Clin. Pharmacol.,* 2005, 45(4), 378–384.

[10] Centers for Medicare and Medicaid Services, CMS Home Health Quality Initiatives: Overview, *http://www.cms.hhs.gov/HomeHealthQualityInits/*, 2010, (accessed January 1 2011).

[11] Rosati, R. J., The History of Quality Measurement in Home Health Care, *Clin. Geriatr. Med.,* 2009, 25(1), 121–134.

[12] Martin, K.S., *The Omaha System : A Key to Practice, Documentation, and Information Management,* Elsevier Saunders, St. Louis, 2005.

[13] Bowles, K.H., Use of the Omaha System in Research, in: Martin, K.S. ed., The Omaha System: A Key To Practice, Documentation, And Information Management, 2nd ed., Elsevier Saunders, St. Louis, 2005, 105–133.

[14] AHRQ, Clinical Classifications Software (CCS) for ICD-9-CM, *http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp*, 2009, (accessed February 16 2009).

[15] Westra, B. L., Clinical Classification Software Diagnoses Groupings, *http://www.nursing.umn.edu/ICNP/OtherProjects/index.htm*, 2009, (accessed January 1 2011).

[16] Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R., A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *J Chronic Dis.,* 1987, 40(5), 373–383.

[17] HCUP CCS, Healthcare Cost and Utilization Project (HCUP), *http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp*, 2009, (accessed May 31 2009).

[18] Westra, B. L., Omaha System Intervention Expert Groups, *http://www.nursing.umn.edu/ICNP/OtherProjects/index.htm*, 2009, (accessed January 1 2011).

[19] Westra, B. L., Oral Medication Data Dictionary for Data Mining, *http://www.nursing.umn.edu/ICNP/OtherProjects/index.htm*, 2009, (accessed January 1 2011).

[20] Bellazzi, R., and Zupan, B., Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines, *Int. J. Med. Inf.,* 2008, 77(2), 81–97.

[21] Harrell, F. E., Lee, K. L., and Mark, P., Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors, *Stat. Med.,* 1996, 15361–387.

[22] Centers for Medicare & Medicaid Services, OASIS - Based Home Health Agency Patient Outcome and Case Mix Reports, *http://www.cms.hhs.gov/OASIS/09a_hhareports.asp*, 2008, (accessed December 13 2008).

[23] Marek, K.D., and Antle, L., Medication Management of the Community Dwelling Older Adult. in: Hughes, R.G. ed., *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. AHRQ Publication No. 08-0043. Agency for Healthcare Research and Quality, Rockville, MD; *http://www.ahrq.gov/qual/nurseshdbk/*, 2008, (accessed January 1 2011).

[24] Ellenbecker, C. H., Frazier, S. C., and Verney, S., Nurse's Observations and Experiences of Problems and Adverse Effects of Medication Management in Home Care. *Geriatr. Nurs.,* 2004, 25(3),164–170.

[25] Shearer, J., Improving Oral Medication Management in Home Health Agencies [Corrected] [Published Erratum Appears in HOME HEALTHC NURSE 2009 May;27(5):270], *Home Healthc. Nurse,* 2009, 27(3),184–192.

[26] Monsen, K., Westra, B. L., Yu, F., Ramadoss, V. K., and Kerr, M. J., Data Management for Intervention Effectiveness Research: Comparing Deductive and Inductive Approaches, *Res. Nurs. Health,* 2009, 32(6),647–656.