International Journal of Nursing Studies 52 (2015) 102–111

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

Exploring factors associated with pressure ulcers: A data mining approach

Dheeraj Raju [a,*], Xiaogang Su [b], Patricia A. Patrician [a], Lori A. Loan [c], Mary S. McCarthy [c]

[a] School of Nursing, University of Alabama at Birmingham, United States
[b] University of Texas at El Paso, United States
[c] Center for Nursing Science & Clinical Inquiry, Madigan Army Medical Center, Tacoma, United States

ARTICLE INFO

Article history:
Received 29 January 2014
Received in revised form 2 August 2014
Accepted 7 August 2014

Keywords:
Data mining
Predictive modeling
Pressure ulcers
Braden scale

ABSTRACT

Background: Pressure ulcers are associated with a nearly three-fold increase in in-hospital mortality. It is essential to investigate how other factors besides the Braden scale could enhance the prediction of pressure ulcers. Data mining modeling techniques can be beneficial to conduct this type of analysis. Data mining techniques have been applied extensively in health care, but are not widely used in nursing research.
Purpose: To remedy this methodological gap, this paper will review, explain, and compare several data mining models to examine patient level factors associated with pressure ulcers based on a four year study from military hospitals in the United States.
Methods: The variables included in the analysis are easily accessible demographic information and medical measurements. Logistic regression, decision trees, random forests, and multivariate adaptive regression splines were compared based on their performance and interpretability.
Results: The random forests model had the highest accuracy (C-statistic) with the following variables, in order of importance, ranked highest in predicting pressure ulcers: days in the hospital, serum albumin, age, blood urea nitrogen, and total Braden score.
Conclusion: Data mining, particularly, random forests are useful in predictive modeling. It is important for hospitals and health care systems to use their own data over time for pressure ulcer risk prediction, to develop risk models based upon more than the total Braden score, and specific to their patient population.

What is already known about the topic?

- The Braden scale is one of the widely used tool for assessing pressure ulcer risk.
- Data mining techniques have been applied extensively in health care, but are not widely used in nursing research.
- The nursing research studies use one or another technique but do not compare them.

What the paper adds?

- This paper adds to our knowledge of how other factors enhance assessing the probability of developing pressure ulcers when combined with the Braden scale.
- The paper extends the knowledge of data mining to the nursing statistical toolbox.

* Corresponding author at: The Center for Nursing Research, 1720 2nd Avenue South, NB 1019F, Birmingham, AL 35294, United States. Tel.: +1 205 975 5336.
E-mail address: seeth001@uab.edu (D. Raju).

http://dx.doi.org/10.1016/j.ijnurstu.2014.08.002
0020-7489/

1. Introduction

Pressure ulcers (PU) are a substantial burden for patients and for the health care system in general. The National Patient Care Safety Monitoring Study (Lyder et al., 2012) of over 51,000 patients found that 4.5% of Medicare beneficiaries developed a pressure ulcer during their hospital stay and 5.8% had a pressure ulcer on admission. Pressure ulcers regardless of whether they were present on admission were associated with a nearly three-fold increase in in-hospital mortality, 69% increase in 30-day mortality, and an increased length of stay of 6.4 days (Lyder et al., 2012). As of 2008, hospital acquired stage III and IV pressure ulcers are no longer reimbursed by the U.S. Centers for Medicare and Medicaid Services (CMS), leaving the hospitals themselves to absorb the cost of care for patients, which is estimated at $43,180 per patient (Armstrong et al., 2008). Thus it is imperative to discover factors associated with both community and hospital acquired pressure ulcers and institute additional care measures to prevent their occurrence.

Because the causative factors for pressure ulcers are "multifactorial and not well understood" (Benoit and Mion, 2012, p. 341), it is critical for hospitals, nursing homes, and home care agencies to systematically monitor patients for pressure ulcer rates, assess risk, and enhance prevention efforts. Although not all pressure ulcers are avoidable (Black et al., 2011), frequent monitoring may lead to better risk predictions and more thoughtful application of resources (i.e., evidence-based nursing preventive interventions such as turning and repositioning) to those who need it most. As more hospitals adopt electronic medical records, the large clinical data repositories could help improve clinical care through the study of their own best practices and lessons learned directly from their patients. Analyzing clinical data collected from discharge abstracts or directly from clinical records and comparing those who developed or did not develop a pressure ulcer can inform problem identification in quality improvement. Data mining modeling techniques can be beneficial to conduct this analysis. The purpose of this paper is to build and compare data mining models for pressure ulcer prevalence (both community and hospital acquired) and determine the variables that are associated with pressure ulcers based on a four year study database collected from 12 military hospitals. The variables included in the analysis are easily accessible demographical information and medical measurements. We carefully selected a group of data mining techniques that not only supply high predictive accuracy but also allow for meaningful interpretations. The Braden scale developed by Bergstrom et al. (1987) is one currently available tool for assessing pressure ulcer risk. Given the multifaceted nature of pressure ulcers, it is of keen interest to see whether and how other factors could enhance the performance of predicting pressure ulcers when combined with the Braden scale. The eventual wide scale use of electronic medical records will enable hospitals to apply these data mining techniques to their own patient level data to determine factors associated with pressure ulcers.

2. Background

Identifying patients who may have a pressure ulcer on admission to a health care facility or who may develop one during hospitalization is the starting point for primary, secondary and tertiary preventive activities aimed at reducing this costly and debilitating complication. A recent systematic review of 54 pressure ulcer studies (Coleman et al., 2013) identified three primary risk factors: mobility level, perfusion, and skin status. Other secondary risk factors that emerged from this literature were skin moisture, age, basic serum metabolic measures, nutrition and general health status.

Risk assessment scales have been used for many years to forecast the patients who are at high risk for developing pressure ulcers; however, their sensitivity and specificity are far from ideal (Pancorbo-Hidalgo et al., 2006; Schoonhoven et al., 2006). It is important that risk assessment scales predict the likelihood of getting a pressure ulcer, so that scarce resources can be applied in an evidence-based manner to the highest risk patients. The Braden Scale for Predicting Pressure Sore Risk (Bergstrom et al., 1987) is the most widely used and most widely researched risk prediction scale for pressure ulcers (Balzer et al., 2007; Pancorbo-Hidalgo et al., 2006). According to a meta-analysis of risk assessment scales, the Braden has the "best sensitivity and specificity balance" (weighted means were 57.1% and 67.5% respectively) and was the best at predicting risk (Odds Ratio of 4.08, 95% CI = 2.56–6.48) compared to Norton Scale, Waterlow Scale, and Nurses' clinical judgment (Pancorbo-Hidalgo et al., 2006).

The Braden scale consists of six subscales derived from two main components of a patients' skin, pressure and tissue tolerance. The subscales mobility, activity, and sensory perception are associated with pressure, whereas the subscales moisture, friction/shear, and nutritional status are associated with the level of tissue tolerance for pressure. These six subscales are rated on scales of 1–4, with the exception of the friction/shear scale which is rated 1–3. When added together, the scores range from 6 to 23, with lower numbers indicating higher risk for pressure ulcer development. Scores of 18–23 indicate no risk; 15–18 for low risk; 13–14 moderate risk; 10–12 high risk, and 6–9 very high risk (Cremasco et al., 2012).

Although the Braden scale is widely used for predicting patients at risk, it is not infallible. In the Coleman et al. (2013) review, mobility subscale scores were found more predictive than total risk assessment scores among all risk assessment scales. Furthermore, the Braden scale does not evaluate perfusion and skin status as defined in the Coleman study. There are other factors important to consider in predicting patients at risk. For example, in 7 of the 11 studies Coleman et al. (2013) reviewed that evaluated serum albumin, it was noted to be statistically associated with pressure ulcer development, such that lower albumin was associated with pressure ulcers (odds ratios of 0.4–0.8). Serum albumin is associated with nutritional status, which is of practical importance in wound healing. Because of the difficulty in capturing all aspects important to predict pressure ulcers, we used data mining techniques to provide a comprehensive assessment

in determining the factors associated with pressure ulcers in our population.

2.1. Data mining

Data mining is the digging for or "mining" of prior unknown useful information from data. Defined as "the exploration and analysis, by automatic or semiautomatic means of large quantities of data in order to discover meaningful patterns and rules" (Berry and Linoff, 2004), data mining applies diverse algorithms for finding patterns in data. Data mining predictive models are non-parametric in nature. Therefore, unlike classical statistical techniques, most of the data mining has minimal prior assumptions for model building. Data mining is useful for the following purposes:

1. Exploratory data analysis – examining the dataset with graphical pictures and basic descriptive statistics;
2. Descriptive modeling – partitioning the data into groups;
3. Predictive modeling – building statistical models to predict the target variable;
4. Discovering patterns and rules – discover items that occur frequently in databases; and
5. Retrieval by content – finding patterns in a new dataset using the procedures from a prior analysis (Hand et al., 2001).

Data mining techniques have been applied extensively in health care, but are not widely used in nursing research. In a pressure ulcer study, Lahmann and Kottner (2011) used Chi-square Automated Interaction Detection (CHAID) algorithm, a data mining technique to explore empirical relationships between friction forces and category II pressure ulcers and between pressure forces and categories III and IV pressure ulcers. A few recent studies in nursing data mining application used either decision trees or logistic regression models (Lahmann and Kottner, 2011; Lahmann et al., 2011; Almasalha et al., 2013; Kottner et al., 2014). Typically in nursing research when data mining techniques are used, only one or two types of models are explained (Cheng et al., 2005; Lee et al., 2011; Vincent et al., 2010). To remedy this methodological gap, this paper will review, explain, and test several data mining models to examine patient level factors associated with pressure ulcers.

3. Methods

3.1. Data mining models

There are numerous methods and procedures available for exploring factors associated with binary outcome (i.e., prevalence of pressure ulcers), but of course, the model choices depend on the research aims. Because the research objectives were to accurately predict pressure ulcer prevalence and identify clinically relevant factors that are associated with pressure ulcers, we cautiously selected four predictive modeling methods: logistic regression, decision trees, random forests (RF), and multivariate adaptive regression splines (MARS). These four data mining models were selected because of their high predictive performance and meaningful interpretations they supply.

3.1.1. Logistic regression

Logistic regression (see, e.g., Hosmer and Lemeshow, 2000) is the standard statistical Generalized Linear Model (GLM) approach for modeling binary outcomes, i.e., whether or not a pressure ulcer is present. In this approach, the logit of the conditional probability of having pressure ulcer is formulated as a linear function of covariates.

The slope parameters in a logistic model can be interpreted as log of odds ratio. The advantages of logistic regression include simple linear structure, widely available fitting software, some flexibility to deal with categorical variables and model interaction terms. Its disadvantages mainly stem from linearity as well. The linear functional form may not provide satisfactory fit when strong nonlinearity and complex interactions are present. The idea is to approximate nonlinear curve with broken lines (first-order spline functions) with thresholds. These terms are data driven and found by automated greedy search procedures.

3.1.2. Decision trees

Decision trees fit piecewise constant models by recursively partitioning the predictor spaces. They are helpful in identifying sub-populations with high/low pressure ulcer incidence rates via easily interpreted grouping rules. A rule is induced by a binary split on covariates with questions such as "Is age less than 40?" or "Is subject male or female?" According to some criterion, the algorithm searches for the best split among all possible splits and the data is partitioned accordingly. The procedure is repeated till the data set is split into a number of mutually exclusive groups. To address the tree model selection and other issues, Breiman et al. (1984) proposed the Classification and Regression Trees (CART) procedure, which has made tree models widely popular in various application fields. Su et al. (2011) introduced different decision tree methods to nursing research. Advantages of decision trees include efficiency in handling categorical variables, invariance to monotone transformations on predictors, ease of understanding, handling missing data, and ability to deal with complex interactions. One of the drawbacks is that the tree models are highly data-adaptive and unstable, meaning that minor alterations in the sample data may cause dramatic changes in the tree model structure. Also predictions from a single tree analysis are often unsatisfactory.

3.1.3. Random forests

The random forests (RF) are among the techniques that help to address the weakness of a single decision tree, by borrowing strength from the instability of tree models. Trees are instable in the sense that predictions from each single tree tend to have small bias but large variance. As a model ensemble method, random forests reduce variance by averaging the predicted values from a number of tree models. The main idea is model ensemble: build up a large number of tree models by perturbation (e.g., bootstrapping)

and then combine the predictive power from all tree models. RF achieves high prediction accuracy and can handle a large number of predictors of different types and missing data. The Gini index is used to calculate the importance rank of predictors (Brieman, 2001). The drawback of random forests is that it does not supply an explicit functional form (i.e. an equation) for the predictive model and the model interpretation is not so easy compared to a single tree model. To remedy this, random forests implements two ways of extracting interpretation: variable importance ranking helps sort variables in terms of their predictive power and partial dependence plot depicts the functional relationship between each predictor and the response after adjusting for other predictors.

3.1.4. Multivariate adaptive regression splines

Multivariate adaptive regression splines (Friedman, 1991) adaptively fits piecewise linear models with truncated power (often of first order for the sake of feasibility) spline basis functions. The final multivariate adaptive regression splines (MARS) model can be written as a model form. MARS is similar to logistic regression in retaining a model form, however adds more flexibility in handling categorical variables and nonlinear patterns and interactions and little requirement on data preparation and variable selection. Its drawback is that it does not provide as good a fit as random forest technique.

3.2. Software

R programming language was used as the data mining software for all the analyses. The following R packages facilitated building the different data mining models:

(1) Logistic regression model – inbuilt R function *glm* with *logit* family option. To select variables, nonconvex penalty Smooth Clipped Absolute Deviation (SCAD) was used as implemented in R package *ncvreg* (Breheny and Huang, 2011).
(2) Decision tree model – package *rpart* by Therneau and Atkinson (2012).
(3) Random forests model and imputation – package *randomForest* by Brieman and Cutler (2011).
(4) MARS model – package *polspine* by Kooperberg (2013).

3.3. Dataset

We extracted data from the Military Nursing Outcomes Database (MilNOD), a nurse staffing and adverse event database that was compiled over four years (2003–2006) (Patrician et al., 2010). The pressure ulcer data used for this analysis was obtained by annual prevalence studies in the participating hospitals' medical–surgical, critical care, and step-down units ($N = 1653$ patients). In addition to documenting pressure ulcer presence and stage, the patients' age, gender, Braden scale and subscale scores, nurse-assessed risk, metabolic test data (blood urea nitrogen, creatinine, and serum albumin), admission source, and days in hospital were collected. The full dataset had many less severely ill patients that were less likely to receive metabolic screening for blood urea nitrogen (BUN), and creatinine, and highly unlikely to undergo serum albumin laboratory tests, therefore these lab values were missing on a large number of patients (54% was missing). For that reason, patients without serum albumin values were removed. This dataset with patients without serum albumin values was the actual data that were used for building the data mining models ($N = 680$). The missing data across all variables for the dataset ranged from 0 to 5%. The data were imputed using random forests technique in the first step and then the data mining models were built (see Fig. 1). This imputation method uses similar algorithm as the random forest *modeling* technique discussed earlier (Brieman, 2001). This non-parametric random forest imputation technique can handle high dimensional data (continuous or discrete) with higher computational efficiency (Stekhoven and Bühlmann, 2012). Random Forest imputation method averages many regression trees which constitutes a multiple imputation scheme. Shah et al. (2014) noted that random forest imputation may be useful for imputing complex epidemiologic datasets in which some patients have missing data. Therefore the random forests method was not only used as modeling strategy but was also used as imputation method for all data mining models. Random forests provides high classification accuracy, ranks variable importance, models complex interactions, and its algorithm for imputation decreases bias in the imputed values by preserving variability in the data (Cutler et al., 2007).

To resolve the problem of highly correlated variables, e.g., Braden score and its subscales, we applied a nonconvex penalty Smoothly Clipped Absolute Deviation (SCAD) to handle the variable selection in logistic regression model only (because decision trees, RF, and MARS methods make automatic variable selection in a recursive manner which accounts for correlated variables).

All models were fit with all predictors as indicated in Fig. 1. All models were cross-validated using a 10-fold cross-validation method of sampling to ensure that each observation was represented at least once in both training and validation data. A 10-fold cross-validation method uses a stratification method of sampling; the entire dataset is divided 10 times (10 folds) and nine out of ten datasets are used as training data and the left over dataset is used as validation data. The same process is repeated 10 different times so that each observation in the dataset has a predictive probability for pressure ulcer prevalence when the observation belongs to a test sample. The C-index was computed using the predicted probability. The C-statistic calculated from receiver operating characteristics curve was used to select the best predictive model (Nisbet et al., 2009). In summary, patients with missing lab values for serum albumin were deleted and then the dataset was imputed using Random forests methods. Next, we used nonconvex SCAD to address the variable selection for logistic regression. The model building and selection process with all other three methods were automated via built-in cross-validation or bootstrap. Afterwards to compare these four data mining models, we used 10-fold cross validation to compute the predicted probabilities and accordingly obtain C index. Finally, the best model was selected based on the C index. Fig. 1 summarizes the model building process (see Fig. 1).

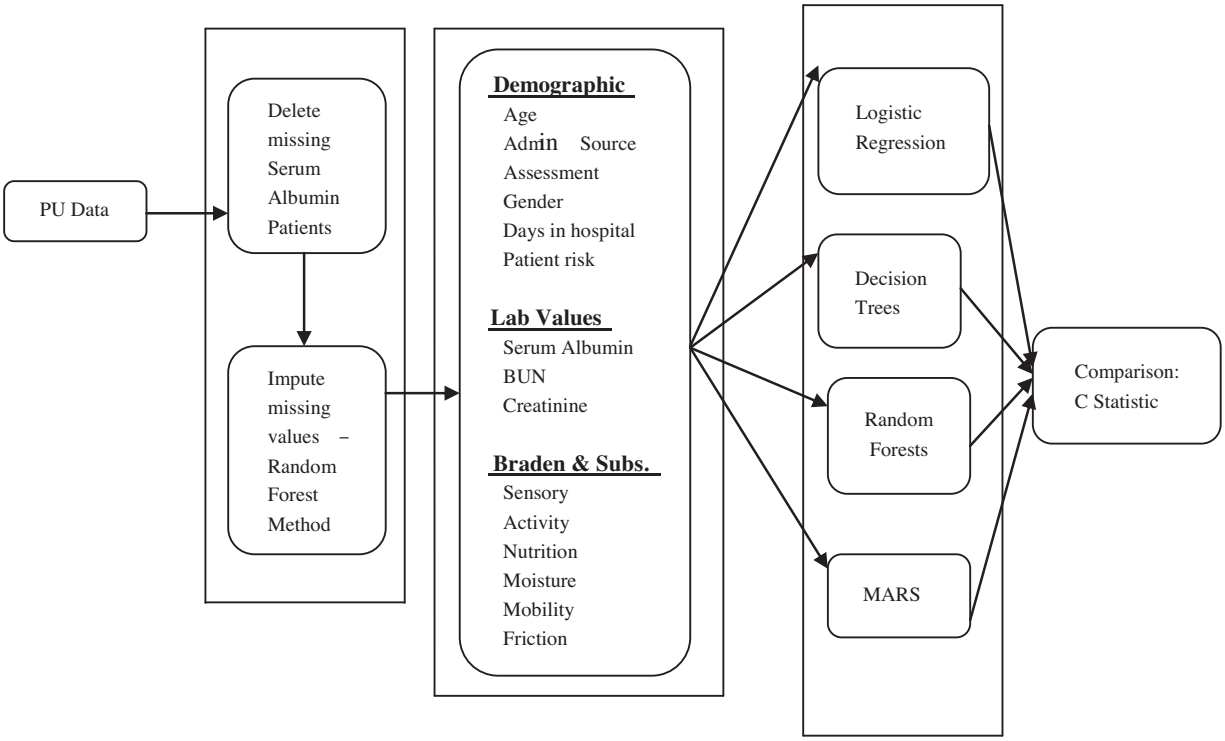

Fig. 1. Data mining model building process.

4. Results

4.1. Exploratory data analysis

The full dataset contained 1653 patients, among which 333 (20%) had a pressure ulcer of any stage. Table 1 shows the demographic and summary statistics of variables included in the analysis for the full dataset. Serum Albumin had the highest percent missing (753/1653 = .5) followed by Creatinine (.3) and BUN (.2). The overall average patient was 54 years old (SD 21.5; range 18–93) and was in the hospital nearly 11 days (SD 23.5; range 1–468). The Braden score for patients without pressure ulcers had a mean of 19.0 (SD 3.5) whereas patients with pressure ulcer had a mean of 15.7 (SD 4.0). The serum albumin lab values for patients with and without pressure ulcer were 2.8 and 3.4, respectively. The BUN level for patients without pressure ulcer was lower (18.4) compared to patients with pressure ulcer (25.2). The Wilcoxon rank-sum test indicated there was no difference ($p < .05$) in creatinine level.

Column *N* shows the total number of observations for all levels of discrete variables (see Table 2). The pressure ulcer prevalence for female patients constitutes around 17.5% compared to 22.2% for males. Of patients admitted from home ($N = 1219$) 16.7% had a pressure ulcer and 30.1% patients admitted from other acute facilities ($N = 196$) had a pressure ulcer. The highest proportion of pressure ulcers was seen in the population presenting to the hospital from skilled nursing facilities (45.2%) or from rehabilitation centers (62.5%). Almost all patients were assessed for risk of pressure ulcer within 24 h of admission ($N = 1432$). Patients who were assessed for risk of pressure ulcer within 24 h had a pressure ulcer prevalence of 21.2% compared to 16.1% for patients who were not assessed. Of 239 patients assessed as at risk for a pressure ulcer, 34.3% had a pressure ulcer. The final dataset (after deleting subjects without serum albumin) that was used to build all for data mining models contained 680 patients with 240 subjects identified as having a pressure ulcer (35.3%).

4.2. Data mining models

4.2.1. Logistic regression

Based on $p < .05$, the logistic regression model indicated age, days in hospital, serum albumin, BUN, and mobility as statistically significant factors associated with pressure ulcers (Table 3).

4.2.2. Decision trees

The decision tree model showed that data was split based on mobility subscale value of 2.5. The left side of the branch contains cases with mobility greater than or equal to 2.5. This node was further divided by the number of days in hospital. There were 66 patients with pressure ulcer who had mobility greater than or equal to 2.5 and spent less than 11.5 days in hospital. Patients with days in hospital greater than or equal to 11.5 were further divided based on the level of BUN. There were 46 cases of pressure ulcer for patients with mobility greater than or equal to 2.5 with days in hospital greater than 11.5 and level of BUN greater than or equal to 11.5 (see Fig. 2).

Table 1
Summary statistics of interval variables for complete data ($N = 1653$).

	Total	PU – no			PU – yes			Wilcoxon rank sum
		n	Mean	SD	*n*	Mean	SD	*p*
Age	1639	1306	52.4	20.9	333	59.3	22.7	<.01*
Days in hosp.	1643	1309	7.8	14.1	334	22.3	40.1	<.01*
Serum albumin	753	499	3.4	0.9	254	2.8	0.8	<.01*
BUN	1080	767	18.4	13.9	313	25.2	21.2	<.01*
Creatinine	1088	776	1.4	5.1	312	1.6	2.7	0.1
Braden	1342	1039	19.0	3.5	303	15.7	4.0	<.01*
Sensory	1343	1040	3.6	0.7	303	3.0	0.9	<.01*
Activity	1342	1039	2.8	1.2	303	2.1	1.1	<.01*
Nutrition	1342	1039	3.0	0.8	303	2.5	0.9	<.01*
Moisture	1342	1039	3.6	0.7	303	3.2	0.9	<.01*
Mobility	1342	1039	3.2	0.8	303	2.5	0.9	<.01*
Friction	1342	1039	2.7	0.5	303	2.3	0.7	<.01*

* $p < .05$.

The second branch on the top right for missing serum albumin data contains cases with mobility less than 2.5. This node was further partitioned with the level of serum albumin. There were 117 patients with pressure ulcer and only 58 patients without pressure ulcer when mobility was less than 2.5 and the level of serum albumin less than 3.8 (see Fig. 2).

4.2.3. Random forests

The random forests model outputs the importance rank of predictors that explains the importance by the size of vertical bars (see Fig. 3). Random forest model ranked days in hospital, serum albumin, age, Braden score, BUN, creatinine respectively as variables that are associated with pressure ulcers (see Fig. 3).

The partial dependency plot facilitates in visualizing the predictor effects in the generated Random Forests model for all the variables. For illustrative purposes the partial dependency plot for serum albumin is shown (see Fig. 4). The plot indicates that patients with lower levels of serum albumin have higher pressure ulcer prevalence.

4.2.4. Multivariate adaptive regression splines model

The MARS model identified serum albumin, mobility, BUN, and days in hospital as significant variables that are associated with pressure ulcers (see Eq. (1)).

$$PU = 0.52 + 0.06(\text{serum albumin}) - 0.03(\text{BUN}) - 0.01(\text{day in hospital})_+ - 0.003(\text{admission source}) + 0.13(\text{mobility}) \quad (1)$$

$(x - a)_+ = x - a$ if $x \geq a$; and 0 otherwise.

Table 2
Summary statistics of discrete variables for complete data ($N = 1653$).

Variables	Characteristics	*N*	PU – yes	%
Gender	Male	973	216	22.2
	Female	669	117	17.5
Admission source	Home	1219	203	16.7
	Home care	22	7	31.8
	Skilled nursing	31	14	45.2
	Board/care	5	1	20.0
	Other acute	196	59	30.1
	Rehab	8	5	62.5
	Other	116	40	34.5
PU assessment – 24 h	No	155	25	16.1
	Yes	1432	304	21.2
Patient @ risk	No	897	143	15.9
	Yes	239	82	34.3
	Not assessed	517	109	21.1

4.2.5. Model comparison

Table 4 shows the C-statistics for each of the models. The random forest model and logistic regression using nonconvex SCAD technique had 83% and 82% C-statistic values respectively. Overall the random forests model had the highest C-statistic value followed by logistic regression with SCAD and multivariate adaptive regression spline models. The decision tree model had the lowest C-statistic.

5. Discussion

The prevalence of all stages of pressure ulcers among this sample for the entire dataset, regardless of whether hospital-acquired or not, was 20.3%, a much higher rate than nationally reported in the recent literature. During our data collection period, a 14–17% prevalence in acute care settings was reported (Whittington and Briones, 2004). There is evidence that pressure ulcers have decreased over time with national rates of 13.5 and 12.3 in 2008 and 2009, respectively (Vangilder et al., 2009). During the period of data collection for this

Table 3
Logistic regression model.

	Estimate	SE	*p*-Value
Logistic regression model with SCAD (N = 680)			
Intercept	1.69	0.66	.01*
Age	−0.01	0.00	<.01*
Days in hospital	0.02	0.00	<.01*
Serum albumin	−0.50	0.12	<.01*
BUN	0.01	0.00	<.01*
Mobility	−0.65	0.11	<.01*
Moisture	−0.29	0.15	.05

* Indicates significance at $\alpha = .05$.

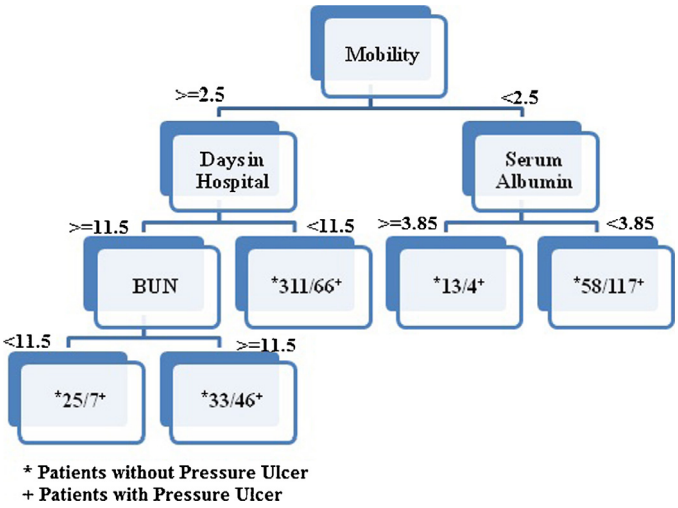

Fig. 2. Decision Tree model.

study (2003–2006), the war in Iraq and Afghanistan was ongoing and many patients seen in our study hospitals were the younger, active duty military population who were returning to the US following injuries sustained in the war (Crumble and Kane, 2010). The age range in our sample was 18–93, with an average age of 59.3 (SD 22.7 years) among those who had pressure ulcers; the age range of patients who had pressure ulcers in the National Medicare Patient Safety Monitoring study (Lyder et al., 2012) was statistically significantly older (78 years average with a lower SD of 11.2), representative of the Medicare beneficiary population upon which the study was based.

The random forest model had highest accuracy for exploring factors associated with pressure ulcers in this sample, i.e. it produced the best C-statistics (.8). Furthermore to compare the results obtained with list-wise deletion of missing values a post analysis was performed. Since random forest had the highest C-statistic, a random forest model was fit to a list-wise deleted dataset and a completely imputed dataset. Whole sample ($N = 1653$) was used and missing values were imputed. In addition a three level variable was created indicating lab test status: 0 – patient had no lab test for BUN, creatinine, or serum albumin; 1 – patient had BUN and/or creatinine value, but not serum albumin; 2 – patient had serum albumin level documented in the medical record. This new variable, serum albumin availability, served as a surrogate for lab test. The results from the random forest model for the list-wise deleted dataset and completely imputed dataset produced the exact same results indicating following variables in order of importance: days in the hospital, serum albumin, age, total Braden score, BUN, creatinine, and the Braden mobility subscale, less important associations of pressure ulcers were: admission source, the other four Braden subscales, gender, and nurse-assessed risk. This post analysis validated the results obtained from building models with imputed dataset.

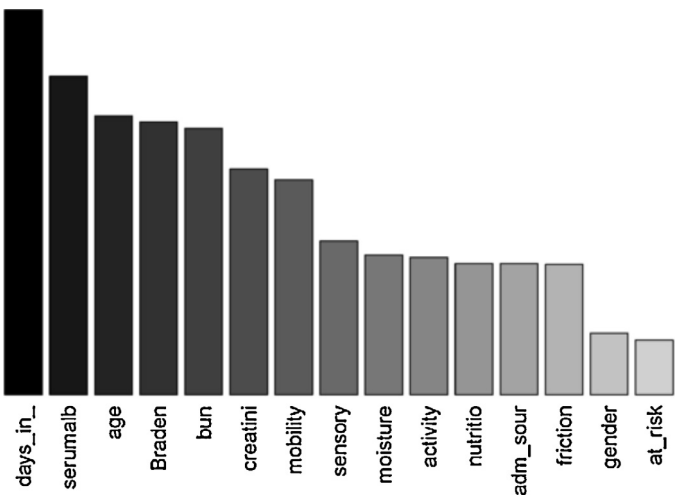

Fig. 3. Random forest model importance rank of predictors.

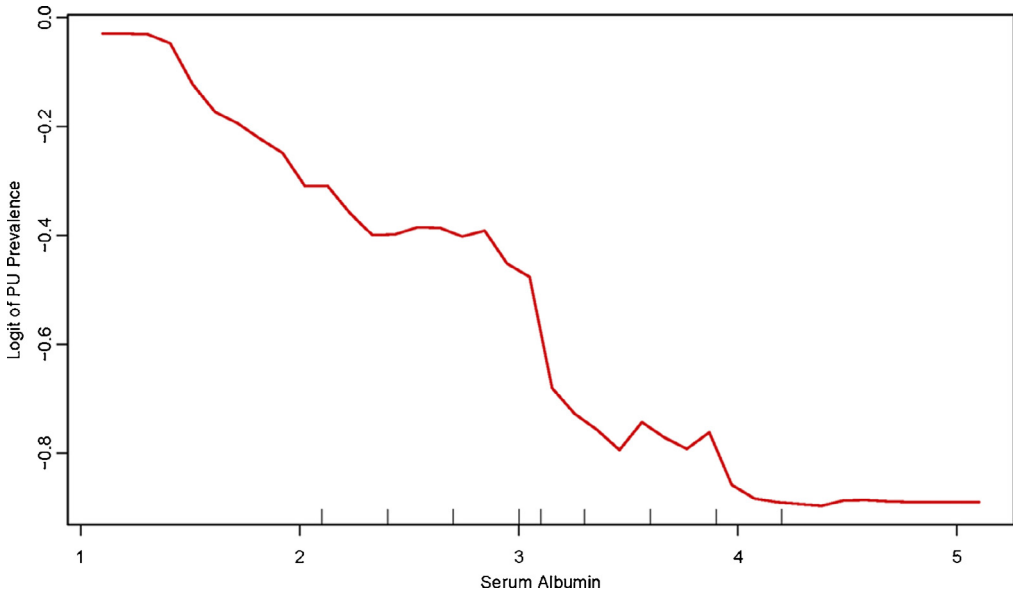

Fig. 4. Partial dependency plot for serum albumin.

The tree models had mobility and serum albumin as the strongest predictors. Generally, logistic regression and decision trees with CART algorithm tend to produce similar results. All models (logistic regression, decision trees, random forests, and MARS) were in agreement that using the Braden score alone is not adequate to explore factors that are associated with pressure ulcers, as has been discussed recently by the instrument developer (Braden, 2012). Out of all the data mining models, only the RF model selected a total Braden score as a significant predictor (by virtue of it being the fourth most important). All data mining models selected the following as important factors that are associated with pressure ulcers: days in hospital, serum albumin, and the mobility subscale. When comparing our findings to that of the extant literature, it was apparent that pressure ulcer studies are not consistent in the variables that are measured, or even in the definitions of prevalence and incidence (Baharestani et al., 2009). However several of our findings are similar to those reported in the literature. The National Medicare Patient Safety Monitoring study (Lyder et al., 2012) found that risk-adjusted length of stay for those with pressure ulcers was 6.4 days longer than for patients without pressure ulcers. Length of stay was also found to be an important predictor of pressure ulcers in a study by Cox (2011). Using logistic regression analysis, she also found that age, the mobility subscale and several other factors we did not measure (i.e., vasopressor infusion and cardiovascular disease) in addition to the length of stay, explained a large part of the variance in pressure ulcers.

Table 4
C-statistic values for all models.

Data mining models	C-statistic
Random forest	0.83
Log regression – nonconvex SCAD	0.82
MARS	0.78
Decision trees	0.63

Our finding that a low albumin level was predictive of pressure ulcers has some support in the literature. Coleman et al. (2013) reported in her systematic review of pressure ulcer risk factors that 7 out of 11 studies reported an association between low albumin levels and pressure ulcer development. Serra et al. (2012) found that hypoalbuminemia on admission to a critical care unit was an independent risk factor for the development, and the severity of pressure ulcers in intensive care. As described by Cox (2012), nutritional deficiencies lead to low protein states and protein-calorie malnutrition which can alter the ability of the skin to tolerate prolonged pressure, thus increasing the risk for pressure ulceration. However, the physiologic and immunologic derangements in any acutely ill or critically ill patient render biochemical analysis of protein stores unreliable as albumin is an acute phase reactant. This may explain why no studies examining Braden subscales found the Nutrition subscale to be an important predictor of pressure ulcer development in ICU patients. Older age was predictive of pressure ulcers in several studies (Coleman et al., 2013; Cox, 2011; Schoonhoven et al., 2006), but whether it is an independent risk factor was equivocal. Elevated creatinine, along with BUN, was found to be predictive of pressure ulcer development in a study (Serpa and Santos, 2007), but not in another (Okuwa et al., 2006). It is also important to note the cross-sectional nature of the study which implies no causation.

Overall, the Braden score alone does not do a thorough job in forecasting the risk of pressure ulcers. Its predictive ability can be enhanced by other variables, namely serum albumin, age, and days in the hospital. The mobility subscale was found to be extremely important in predicting pressure ulcer prevalence and it should be carefully considered and measured daily. Coleman et al. (2013) found mobility to be one of the three primary risk factors,

along with perfusion and skin/pressure status. Early progressive mobility protocols, if not contradicted, can be important to preventing pressure ulcers. Indeed, immobility for even short periods of time causes tissue and muscle breakdown. Because serum albumin is a strong predictor, care must be taken to address patients' nutritional needs on admission and re-evaluate with each shift assessment. Keeping hospitalized patients without food and fluids (as is often the case for diagnostic test preparation) is detrimental to nutritional status and may compromise skin integrity. As with any secondary analysis, the study is limited by the variables that were collected in the original data set and there were missing serum albumin lab values for patients who were not severely ill. Additionally the organizational and cultural characteristics at military hospitals are dissimilar compared to civilian hospitals.

Hence it is unreasonable to generalize the study to all hospitals in the United States. Nevertheless, the data mining methodology used in this research can be applicable to other hospitals in analyzing their respective pressure ulcer data to accurately find variables that are highly associated with pressure ulcers in their respective setting, basing their problem identification on actual data. Then hospital leaders can plan pressure ulcer reduction strategies to target these particular variables.

6. Conclusion

Data mining is a useful method when one has a dataset of many variables that are potentially associated with an outcome. In particular, the Random Forest model was most predictive of pressure ulcers in this sample. The final predictive model included the following in order of importance to predicting pressure ulcers: days in the hospital, serum albumin, age, BUN, and total Braden score. It is important for hospitals and health care systems to use their own data over time for pressure ulcer risk prediction, to develop risk models based upon more than the total Braden score, and specific to their patient population. An important next step would be to validate this model using another pressure ulcer prevalence data repository.

Conflict of interest: None declared.

Funding: This research is sponsored by the TriService Nursing Research Program, Uniformed Services University of the Health Sciences (Grant # N10-C01); however, the information or content and conclusions do not necessarily represent the official position or policy of, nor should any official endorsement be inferred by, the TriService Nursing Research Program, Uniformed Services University of the Health Sciences, the Department of Defense, or the U.S. Government.

Ethical approval: None.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijnurstu.2014.08.002.

References

Almasalha, F., Xu, D., Keenan, G., Khokhar, A., Yao, Y., Chen, Y., Wilkie, D., 2013. Data mining nursing care plans of end-of-life patients: a study to improve healthcare decision making. Int. J. Nurs. Knowl. 24 (1), 15–24.

Armstrong, D.G., Ayello, E.A., Capitulo, K.L., Fowler, E., Krasner, D.L., Levine, J.M., Smith, A.P., 2008. New opportunities to improve pressure ulcer prevention and treatment. J. Wound. Ostomy Continence Nurs. 35, 485–492.

Baharestani, M.M., Black, J.M., Carville, K., Clark, M., Cuddigan, J.E., Dealey, C., Sanada, H., 2009. Dilemmas in measuring and using pressure ulcer prevalence and incidence: an international consensus. Int. Would J. 6 (2), 97–104.

Balzer, K., Pohl, C., Dassen, T., Halfens, R., 2007. The Norton, Waterlow, Braden, and Care Dependency Scales. Comparing their validity when identifying patients' presure sore risk. J. Wound. Ostomy Continence Nurs. 34 (4), 389–398.

Benoit, R., Mion, L., 2012. Risk factors for pressure ulcer development in critically ill patients: a conceptual model to guide research. Res. Nurs. Health 35, 340–362.

Bergstrom, N., Braden, B.J., Laguzza, A., Holman, V., 1987. The Braden scale for predicting pressure sore risk. Nurs. Res. 36, 205–210.

Berry, M., Linoff, G., 2004. Data Mining Techniques for Marketing, Sales, and Customet Relationship Management. John Wiley & Sons, New York.

Black, J.M., Edsberg, L.E., Baharestani, M.M., Langemo, D., Goldberg, M., McNichol, L., Panel, N.P., 2011. Pressure ulcers: avoidable or unavoidable? Results of the National Pressure Ulcer Advisory Panel Consensus Conference. Ostomy Wound Manage. 57 (2), 24–37.

Braden, B., 2012. The Braden Scale for Predicting Pressure Sore Risk: reflections after 25 years. Adv. Skin Would Care 25 (2), 61.

Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat. 5 (1), 232–253.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.I., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Brieman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brieman, L., Cutler, A., 2011. Package 'Randomforest'. Random forests for classification and regression. CRAN-R. http://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

Cheng, B.-W., Chang, C.-L., Liu, I.-S., 2005. Enhancing care services quality of nursing homes using data mining. Total Qual. Manag. 16 (5), 575–596.

Coleman, C., Gorecki, C., Nelson, E., Closs, S.J., Defloor, T., Halfens, R., Nixon, J., 2013. Patient risk factors for pressure ulcer development: systematic review. Int. J. Nurs. Stud., http://dx.doi.org/10.1016/j.ijnurstu.2012.11.019.

Cox, J., 2011. Predictors of pressure ulcers in adult critical care patients. Am. J. Crit. Care 20 (5), 364–375.

Cremasco, M.F., Wenzel, F., Zanei, S., Whitaker, I.Y., 2012. Pressure ulcers in the intensive care unit: the relationship between nusing workload, illness severity and pressure ulcer risk. J. Clin. Nurs. 1–9.

Crumble, D.R., Kane, M.A., 2010. Development of an evidence-based pressure ulcer program at the National Naval Medical Center: Nurses' role in risk factor assessment, prevention, and intervention among young service members returning from OIF/OEF. Nurs. Clin. North Am. 45 (2), 153–168.

Cutler, R., Edwards Jr., T., Beard, K., Cutler, A., Hess, K., Gibson, J., Lawler, J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Friedman, J., 1991. Multivariage adapte regression splines. Ann. Stat. 19, 1–141.

Hand, D., Mannila, H., Smyth, P., 2001. Principles in Data Mining. MIT Press, Massachusetts/London, England.

Hosmer, D., Lemeshow, S., 2000. Applied Logistic Regression. Wiley, New York.

Kooperberg, C., 2013. Package 'polspline'. *Polynomial Spline Routines*. CRAN-R, In: http://cran.r-project.org/web/packages/polspline/polspline.pdf.

Kottner, J., Blume-Peytavi, U., Lohrmann, C., Halfens, R., 2014. Associations between individual characteristics and incontinence-associated dermatitis: a secondary data analysis of a multi-centre prevalence study. Int. J. Nurs. Stud. http://www.journalofnursingstudies.com/article/S0020-7489(14)00039-X/pdf.

Lahmann, N., Kottner, J., 2011. Relation between pressure, friction and pressure ulcer categories: a secondary data analysis of hospital patients using CHAID methods. Int. J. Nurs. Stud. 48 (12), 1487–1494.

Lahmann, N., Tannen, A., Kottner, J., 2011. Friction and shear highly associated with pressure ulcers of residents in long-term care – Classification Tree Analysis (CHAID) of Braden items. J. Eval. Clin. Pract. 17 (1), 168–173.

Lee, T.-T., Liu, C.-Y., Kuo, Y.-H., Mills, M.E., Fong, J.-G., Hung, C., 2011. Application of data mining to identification of critical factors in patient falls using a web-based reporting system. Int. J. Med. Inf. 80, 141–150.

Lyder, C.H., Wang, Y., Metersky, M., Curry, M., Kliman, R., Verzier, N.R., Hunt, D.R., 2012. Hospital-acquired pressure ulcers: results from the national Medicare Patient Safety Monitoring System study. J. Am. Geriatr. Soc. 60 (9), 1603–1608.

Nisbet, R., Elder, J., Miner, G., 2009. Handbook of Statistical Analysis & Data Mining Application. Academic Press, San Diego.

Okuwa, M., Sanada, H., Sugama, J., Inagaki, M., Konya, C., Kitagawa, A., Tabata, K., 2006. A prospective cohort study of lower-extremity pressure ulcer risk among bedfast older adults. Adv. Skin Wound Care 19 (7), 391–397.

Patrician, P., Loan, L., McCarthy, M., Brosch, L., Davey, K.S., 2010. Towards evidence-based management: creating an informative database of Nursing-Sensitive Indicators. J. Nurs. Scholarsh. 42 (4), 358–366.

Pancorbo-Hidalgo, P.L., Garcia-Fernandez, F.P., Lopez-Medina, I.M., Alvarez-Nieto, C., 2006. Risk assessment scales for pressure ulcer prevention: a systematic review. J. Adv. Nurs. 54 (1), 94–110.

Schoonhoven, L., Grobbee, D.E., Donders, A.R., Algra, A., Grypdonck, M.H., Bousema, M.T., Buskens, E., 2006. Prediction of pressure ulcer development in hospitalized patients: a tool for risk assessment. Qual. Saf. Heathc. 15, 65–70.

Serpa, L.F., Santos, V.C., 2007. Assessment of the nutritional risk for pressure ulcer development through Braden scale. J. Wound. Ostomy Continence Nurs. 34 (3S), 4S–5S.

Serra, R., Caroleo, S., Buffone, G., Lugarà, M., Molinari, V., Tropea, F., deFranciscis, S., 2012. Low serum albumin level as an independent risk factor for the onset of pressure ulcers in intensive care unit patients. Int. Wound J. 1–5.

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am. J. Epidemiol. 179 (6), 764–774.

Stekhoven, D.J., Bühlmann, P., 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28 (1), 112–118.

Su, X.G., Azuero, A., Cho, J., Kvale, E., McNees, M.P., 2011. An introduction to tree-structured modeling with application to quality of life (QOL) data. Nurs. Res. 60 (4), 247–255.

Therneau, T.M., Atkinson, E.J., 2012. Package 'rpart'. *Recursive Partitioning*. CRAN-R, In: http://cran.r-project.org/web/packages/rpart/index.html.

Vangilder, C., Amlung, S., Harrison, P., Meyer, S., 2009. Results of the 2008–2009 international pressure ulver prevalence TM survey and a 3-year acute care, unit specific analysis. Ostomy Wound Manage. 55 (11), 39–47.

Vincent, D., Hastings-Tolsma, M., Effken, J., 2010. Data visualization and large nursing datasets. Online J. Nurs. Informatics 14 (2), 1–13.

Whittington, K.T., Briones, R., 2004. National prevalence and incidence study: 6-year sequential acute care data. Adv. Skin Wound Care 17 (9), 490–494.