A Decision Support System on Surgical Treatments for Rotator Cuff Tears

Journal:	*IIE Transactions on Healthcare Systems Engineering*
Manuscript ID:	UHSE-2014-0027
Manuscript Type:	Healthcare Informatics
Date Submitted by the Author:	02-Jul-2014
Complete List of Authors:	Guo, Weihong; University of Michigan, Industrial & Operations Engineering Jin, Jionghua (Judy); University of Michigan, Industrial and Operations Engineering Paynabar, Kamran; Georgia Institute of Technology, Miller, Bruce; University of Michigan Health System, Carpenter, James; University of Michigan Health System,
Keywords:	decision support, probabilistic prediction, data imputation, feature selection, logistic regression

A Decision Support System on Surgical Treatments for Rotator Cuff Tears

Weihong Guo[1*], Jionghua (Judy) Jin[1], Kamran Paynabar[2], Bruce Miller[3], and James Carpenter[3]

[1] *Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, USA*

[2] *School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA*

[3] *Department of Orthopaedic Surgery, University of Michigan Health System, University of Michigan, Ann Arbor, MI 48109-2117, USA*

[*] Corresponding author: E-mail address: graceguo@umich.edu, Tel.: (+1) 734 730 5978.

A Decision Support System on Surgical Treatments for Rotator Cuff Tears

ABSTRACT

Treatment of patients with rotator cuff tears usually starts with physical therapy, but some patients still need surgery eventually. Ineffective physical therapy increases the time and cost of treatment and pain for patients. The quality of treatment can be improved if patients who will not respond to physical therapy are identified at an early stage. However, there is little research available to systematically help physicians make a timely decision on whether a surgical treatment is eventually needed or not. In this research, we developed a decision support system that can predict the probability of eventually needing a surgical treatment by effectively analyzing the available patients' information at an early stage. Missing value imputation, variable selection, and regression methods are integrated in developing such a decision support system. The probability given by our model will either confirm physician's expert decision, or remind physician if there is any information ignored. This research has the potential to improve patient safety, reduce cost of unnecessary treatment, and help physicians prevent treatment errors.

Keywords: decision support, probabilistic prediction, data imputation, feature selection, logistic regression, rotator cuff tears

1 Introduction

A rotator cuff tear is a common cause of shoulder pain and disability among adults. According to America Academy of Orthopaedic Surgeons, in 2008, close to 2 million people in the United States went to their doctors because of a rotator cuff problem (AAOS, 2011). Although rotator cuff tears are not life-threatening, a torn rotator cuff weakens the shoulder and makes many daily activities such as hair-combing or getting dressed becomes difficult or painful.

The rotator cuff is a network of four muscles that come together as tendons to form a covering around the head of the humerus (upper arm bone). The rotator cuff attaches the humerus to the shoulder blade and helps to lift and rotate the arm. The tendons of rotator cuff can tear much like a piece of leather. The tendon may be only slightly damaged or irritated, or may has a complete tear, which means that the tendon has torn away from the bone. A substantial percentage of rotator cuff tears will get larger over time if not repaired.

There are two types of treatment for a rotator cuff tear: nonsurgical and surgical. Nonsurgical treatment includes various options such as rest, activity modification, and physical therapy. These treatments are often used first before considering surgery (Seida *et al.*, 2010). Surgery is recommended if shoulder pain does not improve with nonsurgical treatments for a certain period of time. Surgery is also often considered when there is a traumatic tear, a large tear, a long lasting period (6 to 12 months) of severe symptoms, or significant weakness or loss of function in the shoulder. According to unpublished data from the University of Michigan MedSport Clinic, about 75% of patients start with nonsurgical treatments, but 45% of these patients eventually need to go through surgery after taking nonsurgical treatments; thus about 75% × 45% = 33.75% of the patients receive delayed surgery decisions. Figure 1 illustrates these treatment options. In this study, the nonsurgical treatments on these 45% patients are wasted, which results in unnecessary therapy cost, aggravated patients' pain due to delayed surgical treatments, and increased patients' risk.

The best treatment option is different for every person. In planning treatment, in addition to considering the type and size of tear, physicians also need to consider each individual patient's activity level, general health, and many other factors. Patient medical records and initial examination results are available when the patient comes to the clinic. However, there is little research available to fully utilize this information, systematically analyze the massive data, and help physicians make a timely treatment decision. This motivates the research to develop a decision support system that can predict the probability of eventually needing a surgical treatment by effectively analyzing the available individual patient's information, including medical records and initial examination results, at an early stage. Physicians may combine the predicted probability along with their expertise judgment before making a treatment decision. If the predicted probability is around 50%, it indicates that the prediction is less certain and physicians should mostly rely on their expertise. If a very large or very small probability is given, it indicates a higher level of certainty in prediction and should draw more attention from physicians. If the predicted probability shows a consistent treatment decision with the physicians', they will be more confident with their decisions; otherwise, the prediction may suggest important information that might have been missed, or whether the model needs to be updated. This decision support system would help physicians avoid some overlooked mistakes, and meanwhile continuously improve model performance during the usage of the system.

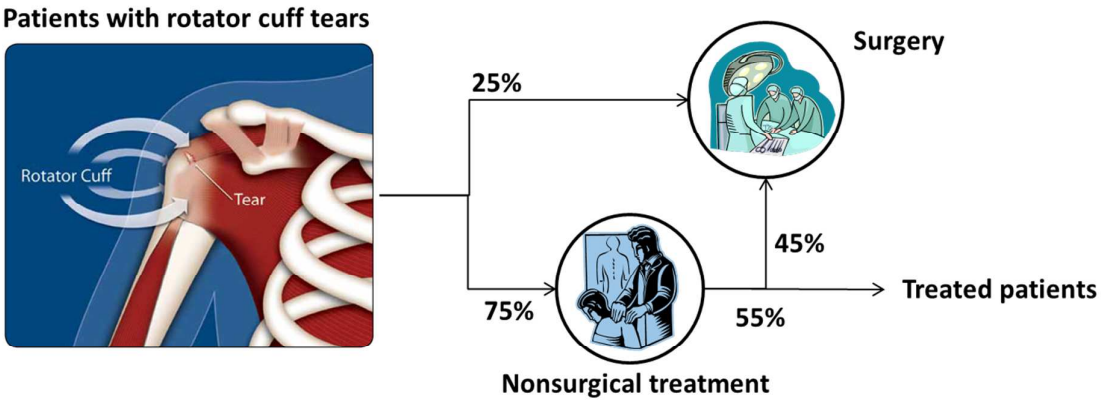

Figure 1. Treatment options for rotator cuff tears (Source: the University of Michigan MedSport Clinic)

Developing a decision support system for rotator cuff tears treatment based on patient information is very challenging. The patients' information provided by electronic medical records is high-dimensional and heterogeneous data. Moreover, there is a large amount of mixed-type missing data in medical records. Although managing and mining such a dataset is very difficult (Hu *et al.*, 2011; Johnstone and Titterington, 2009), effective analysis is needed for assessing the likelihood of patient eventually needing a surgical treatment.

The objective of this paper is to develop a decision support system that utilizes patients' medical records and initial examination data to predict at an early stage the probability of eventually needing surgery for rotator cuff tears patients. Specifically, the development of such a system needs to integrate missing data imputation, variable selection, and classification/regression methods. The developed decision support system will be used to help physicians make decisions on whether surgical treatments for rotator cuff tears will be eventually needed or not.

The remainder of this section reviews existing methods on healthcare-related classification/prediction problems and methods on dealing with missing data. Section 2 provides a methodology overview of the proposed decision support system followed by a description of the data used in this study in Section 3. Sections 4 and 5 then introduce detailed methodologies on missing data imputation and variable selection, respectively. Section 6 further demonstrates how the proposed decision support system works with a case study. Our conclusion is drawn in Section 7.

1.1 Literature review of the related work

In patient-centric healthcare delivery, data mining has been increasingly recognized as a potential value-added technique through new knowledge discoveries for improving decision-making in hospital operations, disease diagnosis, and medical treatments (Denton *et al.*, 2011). For example, Lavieri *et al.* (2012) proposed an innovative approach to help clinicians decide when to start radiation therapy in prostate cancer patients. In their method, the decision is based on predictions of the time

when the patient's prostate specific antigen (PSA) level reaches its lowest point. Claudio *et al.* (2014) proposed a dynamic multi-attribute utility theory–based decision support system for patient prioritization in the emergency department. Van der Pol and Cairns (2003) demonstrated how routinely collected data on breast screening could be used to predict demand from 65-67 year olds in Scotland as a result of introducing an additional breast screening round. They used logistic regression to model the attendance response and to predict demand. Alaeddini *et al.* (2011) developed a hybrid probabilistic model based on logistic regression and empirical Bayesian inference to predict the probability of no-shows in real time using both general patient social and demographic information and individual clinical appointments attendance records. Chen *et al.* (2011) explored the risk factors of preterm birth using data mining with the neural network and decision tree methods. However, these data mining techniques have not been widely applied to decision-making about surgical treatments for rotator cuff tears patients. With the advances in electronic medical record (EMR) data and standardized patient data collection procedure, it is highly demanding to develop a general framework that integrates various data mining techniques so that it can help doctors systematically analyze EMR data for making an efficient decision about surgical treatments.

Patients' examination data are collected from various surveys designed for rotator cuff tear patients (Michener *et al.*, 2002; Kirkley *et al.*, 2003; Raman and Macdermid, 2012). Such a data set includes mixed-type variables when both categorical variables (e.g. gender) and continuous variables (e.g. body mass index) present simultaneously. Due to the nature of hospital surveys, there is a large amount of mixed-type missing data in patients' examination data. Deleting the records with missing values would result in a huge waste of useful data and losing the majority of the collected data at the model training stage. Therefore, how to effectively impute missing values is considered as a critical first step in analyzing the EMR data.

Imputation techniques have been extensively studied over the last few decades, and a number of approaches have been proposed. For example, Ayuyev *et al.* (2009) proposed a dynamic clustering

imputation (DCI) algorithm for dealing with a large number of missing values in mixed-type data. Stekhoven and Bühlmann (2012) developed a non-parametric iterative imputation method, missForest, based on a random forest to cope with different types of variables simultaneously. Cadwell *et al.* (2007) described a Bayesian multiple imputation method which allows estimation and assessment of changes in heart disease mortality rates for men and women with diabetes using data from North Dakota between 1992 and 2003. Tarsitano and Falcone (2011) developed a single imputation method based on the popular nearest neighbor hot deck imputation, where "nearest" is defined in terms of a global distance obtained as a combination of the distance matrices computed for various types of variables involving numeric, ordinal, binary, and categorical variables. An overview of those imputation methods can be found at Kalton and Kasprzyk (1982) and Little and Rubin (2002). Some of these methods are already available in standard statistical software (or can easily be implemented) although there is little consensus as to the most appropriate technique to use for a particular situation.

The nearest neighbor hot deck imputation has been used for a number of years and enjoys high prestige for its theoretical and computational characteristics. Farhangfar *et al.* (2004) conducted experimental analysis of imputation methods and compared the performance between unsupervised imputation algorithms (mean imputation and hot deck imputation) and supervised imputation algorithms (maximum likelihood methods). They found that the unsupervised imputation methods are more stable with respect to increasing amount of missing information, and that their performance may be better for databases with large amounts of missing attributes. Their results also indicated that unsupervised imputation methods do not depend on the size of the input data, both in terms of the number of the attributes and the number of samples. Furthermore, Troyanskaya *et al.* (2001) compared a singular value decomposition (SVD) based method and a *k*-nearest neighbors (KNN) based method for missing value estimation for DNA microarrays. Their results indicated that although both KNN and SVD methods are robust to the increasing fraction of missing entries, KNN-based imputation shows less deterioration in performance with the increasing percent of missing data. They also found that the KNN-based imputation is less sensitive to the exact parameters used (number of

nearest neighbors), and it provides for a robust and sensitive approach to estimating missing data in biological fields. However, most existing studies related to the nearest neighbor hot deck imputation method are limited to the imputation of a single type of missing values. Although Tarsitano and Falcone (2011) recently applied nearest neighbor hot deck imputation to mixed-type of missing values, their work cannot be directly applied to the EMR data in our problem due to its failure in considering the correlation between different variables and the lack of methods to effectively combine the distance from various data types. Therefore, this paper aims to develop a systematic method to effectively impute missing values of mixed types for EMR data.

2 Proposed methodology framework

Figure 2 illustrates the framework of the proposed decision support system in this research. At the model training stage, historical patient data is analyzed through three steps, i.e., Step 1: missing data imputation, Step 2: variable selection, and Step 3: training logistic regression model. At Step 1, a new *k*-nearest neighbor hot deck imputation method is developed that can handle the mixed-type of missing values in EMR data. The proposed method is capable of dealing with categorical variables and continuous variables simultaneously, and determining the *k*-nearest neighbors based on imputation performance. At Step 2, an optimal subset of variables is selected so that data dimension and noise effects can be reduced, and that model interpretability and predictive ability can be improved. The LASSO (least absolute shrinkage and selection operator) method (Tibshirani, 1996) is adopted in this study for variable selection. At Step 3, a classification/regression model is trained through historical data based on the selected variables. Specifically in this study, a logistic regression (LR) model is developed to predict the probability that patient eventually needing a surgical treatment. The predicted probabilities are further used to classify the model decisions into certain decisions (either surgical or nonsurgical treatment) and uncertain decisions (no decisions suggested by the model). After the model training stage, a set of methods for imputing missing values, selecting

important variables, and predicting the probability of patient eventually needing a surgical treatment are established.

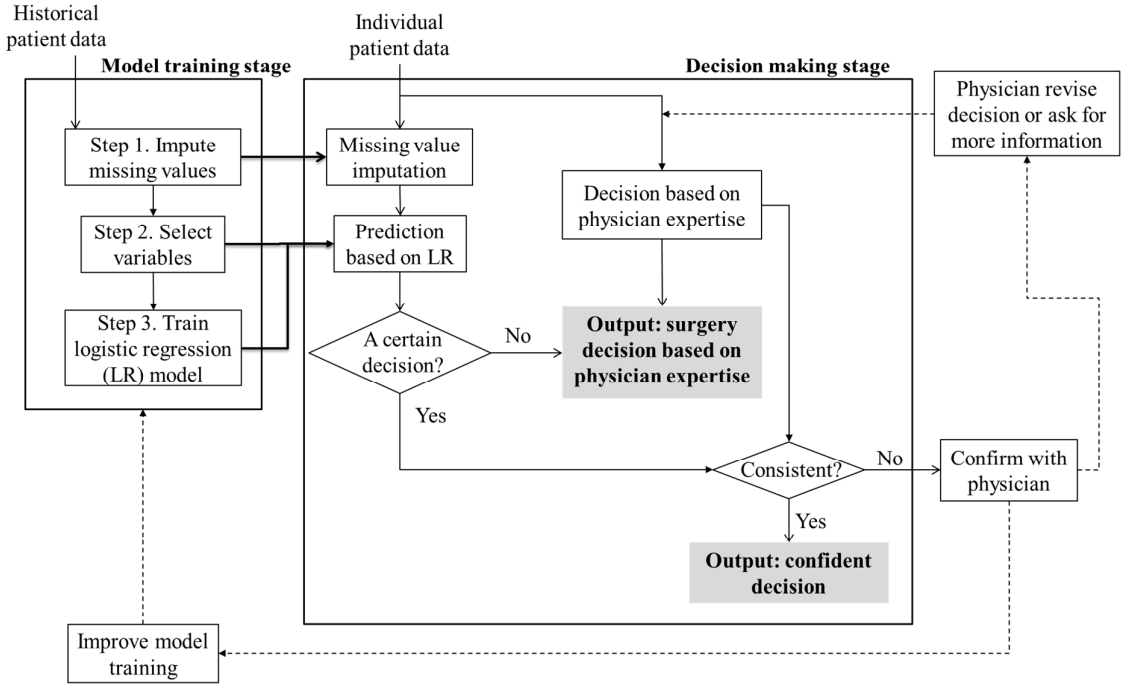

Figure 2. Framework of the proposed decision support system

At the physician's decision-making stage, each incoming individual patient record is firstly put through the missing value imputation module and then the regression model using the selected variables. The developed logistic regression model outputs the probability that the patient will eventually need surgery based on individual patient examination data, while physician also makes a decision from his/her expertise. If the predicted probability is around 50%, it indicates that the prediction is less certain and physicians should mostly rely on their expertise. If a very large or small probability is given, it indicates a higher level of certainty in prediction and should draw much attention from physicians. If the predicted probability shows consistent treatment decision with physician expertise judgment, physicians will be more confident with their decisions; otherwise, the prediction may remind physicians to check if they missed any information, or, the model needs to be

updated. Such decision outputs also provide physicians with explicit and reliable interpretation of the model outputs.

In Section 6, we demonstrate the effectiveness of the proposed decision support system with a case study. By developing a probability-based decision rule for the LR model, decision performance for surgical treatments will be evaluated.

3 Data description

Under the consent of the University of Michigan MedSport Clinic, we obtained a dataset recording rotator cuff tear patients from June 2009 to October 2012. The dataset contains 244 patient records including patient anthropometries, symptom information, scores from various surveys, and patients' surgical treatment decision. After removing irrelevant and redundant features in pre-screening (e.g., survey date is irrelevant to classification model, height and weight are redundant given BMI, etc.), we obtained a total of 22 features for our study, as described in Table 1. Among the $n = 22$ variables, there are $n_1 = 9$ categorical variables and $n_2 = 13$ continuous variables. Among 244 patient records, there are $N = 47$ records that do not contain missing values and $N_n = 197$ records that contain at least one missing value. Note that missing values only occur in 14 variables since there are no missing values in features f_1, f_2, and $f_5 \sim f_{10}$ required by the data collection procedure.

Table 1. Feature description

Feature Index	Feature Description		Variable Type	Number of missing values (in 244 records)
f_1	Surgery side		Categorical	0
f_2	Gender			0
f_3	Dominant side			53
f_4	Duration of symptoms (unit: month)		Continuous	53
f_5	Adhesive capsulitis	Diagnosis codes	Categorical	0
f_6	Impingement			
f_7	Osteoarthritis – glenohumeral joint			
f_8	Instability			

f_9	Osteoarthritis – acromioclavicular joint			
f_{10}	Partial thickness tear			
f_{11}	Body mass index (BMI)		Continuous	70
f_{12}	Survery12 Physical score	Scores from The Veterans Rand 12 Item Health Survey	Continuous	118
f_{13}	Survey12 Mental score			
f_{14}	ASES Total Score Opposite Side	Scores from ASES (American Shoulder and Elbow Surgeons) Shoulder Form	Continuous	135
f_{15}	ASES Total Score Affected Side		Continuous	55
f_{16}	ASES Total Score		Continuous	70
f_{17}	WORC Physical Score	Scores calculated based on the Western Ontario Rotator Cuff Index	Continuous	33
f_{18}	WORC Sport Rec Score		Continuous	33
f_{19}	WORC Work Score		Continuous	32
f_{20}	WORC Lifestyle Score		Continuous	32
f_{21}	WORC Emotions Score		Continuous	32
f_{22}	Shoulder Activity Score	Score from Shoulder Activity Level Form	Continuous	138

4 Missing value imputation

We propose a new method of imputation and reconstruction of missing values in more than one field of mixed data types on the basis of the popular nearest neighbor hot deck imputation (NNHDI). In NNHDI, "nearest" is defined based on a new concept of the "global distance", which is combined by the distance matrices computed for the mixture of two types of variables, continuous or categorical variables. As opposed to the concept of "global distance", the distance calculated from a single type of variable is denoted as the "partial distance". In this section, we first address the problem of defining partial distance for each type of variables, and then study how to determine a proper weight for combining the two partial distance matrices to obtain the global distance.

4.1 *k*-Nearest neighbor hot deck imputation (*k*-NNHDI)

The nearest neighbor hot deck imputation (NNHDI) method looks for the nearest subset of records that are most similar to the record with the missing values, where "nearness" is specified according to the minimal distances between the missing sample and the selected subset of samples without missing values.

Let $\boldsymbol{S}_{N+N_m}$ be a dataset consisting of $N + N_m$ sample records having n variables. Let $\boldsymbol{X}$ consists of all the complete records (N) in $\boldsymbol{S}_{N+N_m}$, while $\boldsymbol{M}$ consists of those records (N_m) that contain missing values:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_j^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix}_{N\times n}, \boldsymbol{x}_j^T = [x_{j,1}, \dots, x_{j,r}, \dots, x_{j,n}]; \boldsymbol{M} = \begin{bmatrix} \boldsymbol{m}_1^T \\ \vdots \\ \boldsymbol{m}_i^T \\ \vdots \\ \boldsymbol{m}_{N_m}^T \end{bmatrix}_{N_m\times n}, \boldsymbol{m}_i^T = [m_{i,1}, \dots, m_{i,r}, \dots, m_{i,n}].$$

Each record in $\boldsymbol{M}$ contains at least one missing value among n variables. The records in $\boldsymbol{M}$ are organized in the ascending order based on the number of missing values. For each record in $\boldsymbol{M}$, the NNHDI selects a neighborhood subset of k closest records, called donors, where k is a pre-fixed size of the neighborhoods, and these donors then provide a basis to determine the imputed values. If $k = 1$, all missing information is imputed from the single donor. In most cases, however, it is difficult to find one single donor that precisely matches the recipient record. Hence, we used the k-NNHDI method with $k > 1$, and its procedure is described in the pseudo-code in Table 2.

Table 2. *k*-NNHDI procedure

$i \leftarrow 1$

While $i \leq N_m$

1. Take record $\boldsymbol{m}_i$ in $\boldsymbol{M}$;
2. Calculate the distance between $\boldsymbol{m}_i$ and each record $\boldsymbol{x}_1, \dots, \boldsymbol{x}_j, \dots \boldsymbol{x}_N$ in $\boldsymbol{X}$, denote as $d_{ij}, j = 1, \dots, N$, respectively;
3. Among $d_{ij}, j = 1, \dots, N$, pick $j_1, \dots, j_k$ so that $d_{ij_1} \leq \cdots \leq d_{ij_k} \leq$ all other distances;
4. Impute the missing value(s) in $\boldsymbol{m}_i$ based on $x_{j_1}, \dots, x_{j_k}$:

$$m_{i,r} = \begin{cases} \text{mode of } x_{j_1,r}, \dots, x_{j_k,r}, \text{ if variable } r \text{ is a categorical variable} \\ \text{mean of } x_{j_1,r}, \dots, x_{j_k,r}, \text{ if variable } r \text{ is a continuous variable} \end{cases} \quad (1)$$

5. Update $\boldsymbol{X}$ by adding the completed $\boldsymbol{m}_i$ to the complete data set;
6. Update $\boldsymbol{M}$ by deleting record $\boldsymbol{m}_i$;
7. $i \leftarrow i + 1$; go to the next record in $\boldsymbol{M}$

End

Output: a data set $\boldsymbol{S}$ with no missing values

4.2 Distance measurement for mixed-type data

One key issue in using k-NNHDI is to define an adequate distance measure for the comparison of two records that share some, but not necessarily all, variables. To deal with the presence of mixed-type variables and different measurement scales, a general global distance is defined in this paper, which aims to effectively combine the partial distance of categorical and continuous variables together. Let C_1 be the class of categorical variables, i.e., $r \in C_1$ indicates variable r is a categorical variable; let C_2 be the class of continuous variables, i.e., $r \in C_2$ indicates variable r is a continuous variable. In this paper, we define the distance functions which have a range of [0, 1], irrespective of the number of variables. In this way, the defined distances are unaffected by the number of non-missing variables in each record.

The distance for categorical variables between a missing value record $\boldsymbol{m}_i$ and the complete record $\boldsymbol{x}_j$ is defined as

$$d_{1,ij} = \sqrt{\frac{\sum_{r=1,r\in C_1}^{n} h_{i,r} \cdot \delta_{i,j,r}}{n_{1,i}}} \tag{2}$$

where $h_{i,r} = \begin{cases} 1, \text{if variable } r \text{ is not missing in } \boldsymbol{m}_i \\ 0, \text{if variable } r \text{ is missing in } \boldsymbol{m}_i \end{cases}$, $\delta_{i,j,r} = \begin{cases} 1, \text{if } m_{i,r} \neq x_{j,r} \\ 0, \text{if } m_{i,r} = x_{j,r} \end{cases}$, $n_{1,i}$ is the number of non-missing categorical variables in $\boldsymbol{m}_i$.

If the continuous variables are assumed to be completely independent of each other, the distance for continuous variables between a missing value record $\boldsymbol{m}_i$ and the complete record $\boldsymbol{x}_j$ can be simply defined using the Euclidean distance:

$$d_{2,ij} = \sqrt{\frac{\sum_{r=1,r\in C_2}^{n} h_{i,r} \cdot \left(\frac{m_{i,r} - x_{j,r}}{R_r}\right)^2}{n_{2,i}}} \quad (3)$$

where $h_{i,r} = \begin{cases} 1, \text{if variable } r \text{ is not missing in } \boldsymbol{m}_i \\ 0, \text{if variable } r \text{ is missing in } \boldsymbol{m}_i \end{cases}$, $n_{2,i}$ is the number of non-missing continuous variables in $\boldsymbol{m}_i$; R_r is the observed range of variable r over all records:

$$R_r = \max_{\substack{j=1,\dots,N; \\ i=1,\dots,N_m}} \left(x_{j,r}, m_{i,r}\right) - \min_{\substack{j=1,\dots,N; \\ i=1,\dots,N_m}} \left(x_{j,r}, m_{i,r}\right), \forall r \in C_2 \quad (4)$$

In some cases, when the correlation between multiple continuous variables is not ignored, Mahalanobis distance can be generally used in Eq. (3) to preserve the existence of covariance matrix among variables. To deal with the correlation between continuous variables, we need to perform de-correlation before deriving a distance function in the format of Eq. (3). Denote $\widetilde{\boldsymbol{G}}\boldsymbol{m}_i$ to be an $n_{2,i} \times 1$ vector with only the non-missing continuous elements in $\boldsymbol{m}_i$; denote $\widetilde{\boldsymbol{G}}\boldsymbol{x}_j$ to be an $n_{2,i} \times 1$ vector with corresponding variables in $\boldsymbol{x}_j$. Matrix $\widetilde{\boldsymbol{G}} \in \boldsymbol{R}^{n_{2,i} \times n}$ is obtained after deleting the all-zero rows in matrix $\boldsymbol{G}_{n\times n} = diag_{r=1,\dots,n}(g_{r,i})$, where $g_{r,i} = 1$ if variable $r \in C_2$ is not missing in $\boldsymbol{m}_i$ and $g_{r,i} = 0$ otherwise. Let $\boldsymbol{\Sigma}$ be the sample covariance matrix calculated from $\boldsymbol{X}$ and also determined by $\widetilde{\boldsymbol{G}}$. The Mahalanobis distance between $\widetilde{\boldsymbol{G}}\boldsymbol{m}_i$ and $\widetilde{\boldsymbol{G}}\boldsymbol{x}_j$ is equivalent to the Euclidean distance between $\boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{m}_i$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{x}_j$, i.e., $\left(\widetilde{\boldsymbol{G}}\boldsymbol{m}_i - \widetilde{\boldsymbol{G}}\boldsymbol{x}_j\right)^T \boldsymbol{\Sigma}^{-1} \left(\widetilde{\boldsymbol{G}}\boldsymbol{m}_i - \widetilde{\boldsymbol{G}}\boldsymbol{x}_j\right) = \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{m}_i - \boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{x}_j\right)^T \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{m}_i - \boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{x}_j\right) = \left\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{G}}\boldsymbol{m}_i - \boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{G}}\boldsymbol{x}_j\right\|_2^2$, where $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{UD})^T$ if the eigenvalue decomposition of $\boldsymbol{\Sigma}^{-1}$ is given as $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U\Lambda U}^T$ and $diag(\boldsymbol{D}) = \sqrt{diag(\boldsymbol{\Lambda})}$.

Define $\widetilde{\boldsymbol{m}}_i = \boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{m}_i$, $\widetilde{\boldsymbol{x}}_j = \boldsymbol{\Sigma}^{-\frac{1}{2}}\widetilde{\boldsymbol{G}}\boldsymbol{x}_j$, both are $n_{2,i} \times 1$ vectors. The distance for continuous variables between $\boldsymbol{m}_i$ and $\boldsymbol{x}_j$ is

$$d_{2,ij} = \sqrt{\frac{\sum_{r=1,r\in C_2}^{n} h_{i,r} \cdot \left(\frac{\tilde{m}_{i,r} - \tilde{x}_{j,r}}{\tilde{R}_r}\right)^2}{n_{2,i}}} = \sqrt{\frac{\sum_{r=1,r\in C_2,h_{i,r}=1}^{n} \left(\frac{\tilde{m}_{i,r} - \tilde{x}_{j,r}}{\tilde{R}_r}\right)^2}{n_{2,i}}} \quad (5)$$

where $\tilde{m}_{i,r}$ is variable r in $\tilde{\boldsymbol{m}}_i$, $\tilde{x}_{j,r}$ is variable r in $\tilde{\boldsymbol{x}}_j$. In Eq. (5), $h_{i,r}$ can be omitted since $\tilde{\boldsymbol{m}}_i$ and $\tilde{\boldsymbol{x}}_j$ already do not contain missing variables. Similar to Eq. (4), $\tilde{R}_r$ is the observed range of variable r over all records on the rescaled data:

$$\tilde{R}_r = \max_{\substack{j=1,\dots,N;\\ i=1,\dots,N_m}} \left(\tilde{x}_{j,r}, \tilde{m}_{i,r}\right) - \min_{\substack{j=1,\dots,N;\\ i=1,\dots,N_m}} \left(\tilde{x}_{j,r}, \tilde{m}_{i,r}\right), \forall r \in C_2, h_{i,r} = 1 \quad (6)$$

In summary, Eq. (2) provides the partial distance $d_{1,ij}$ for categorical variables between a receptor record $\boldsymbol{m}_i$ and a donor record $\boldsymbol{x}_j$, and Eq. (3) and Eq. (5) provide the partial distance $d_{2,ij}$ for continuous variables. Specifically, Eq. (3) is preferred if the correlation between continuous variables is ignorable, while Eq. (5) is used when considering the variables' correlation. The global distance between record $\boldsymbol{m}_i$ and record $\boldsymbol{x}_j$, which considers both the categorical and continuous missing variables, is defined as:

$$d_{ij} = w \cdot d_{1,ij} + (1 - w) \cdot d_{2,ij} \quad (7)$$

where $w(0 \le w \le 1)$ is the weight between categorical variables' distance and continuous variables' distance. d_{ij}, $d_{1,ij}$, and $d_{2,ij}$ are all in range [0, 1], which will be obtained in the following subsection.

4.3 Determination of the weight coefficient

To use the combined distance presented in Eq. (7), a proper weight w needs to be determined. Intuitively from Eq. (7), a larger value of w is preferred if categorical variables are considered to be more important than continuous variables, and vice versa. We propose to use cross-validation approach to systematically determine the optimal weight, w^*, based on the entire dataset $\boldsymbol{X}$. Since $\boldsymbol{X}$ does not contain missing values, we randomly create missing values in some entries in $\boldsymbol{X}$, then

compare the imputed values from *k*-NNHDI to the true values. Cross-validation is used to estimate the differences between imputed values from true values, and the optimal weight is found as the weight that minimizes the imputation error. In this paper, a 10-fold cross-validation is performed in determining w^*. The detailed procedures are described as follows.

To create the simulated missing values for the cross-validation analysis, for each record $\boldsymbol{x}_i$ in the validation set, we artificially assign some missing values based on the binomial sampling outcome with probability q_r, i.e., let variable r in $\boldsymbol{x}_i$ to be missing if the random number generated from the binomial distribution with probability q_r is equal to 1, i.e., assign $x_{i,r}$ to be missing if $binom(q_r) = 1$. q_r is the percentage of missing values for variable r, which is calculated based on the entire dataset $\boldsymbol{S}_{N_m+N}$:

$$q_r = 1 - \frac{1}{N + N_m} \sum_{i=1}^{N_m} h_{i,r} \tag{8}$$

where $h_{i,r} = \begin{cases} 1, \text{if variable } r \text{ is not missing in } \boldsymbol{m}_i \\ 0, \text{if variable } r \text{ is missing in } \boldsymbol{m}_i \end{cases}$. Recall N is the number of records with complete values and N_m is the number of records with missing values.

For a given w, the missing values in the validation set are imputed using the *k*-NNHDI method presented in Eqs. (2), (5), and (7). Since 10-fold cross-validation is used, the training set which consists of 9 subsets of $\boldsymbol{X}$ serves as the donor set. The performance of imputation is estimated by imputation error, E, defined as follows:

$$E_{C_1} = \frac{\#\text{ of wrong categories}}{\#\text{ of categorical variables to impute}}$$

$$E_{C_2} = \frac{\sqrt{\sum_{\substack{\text{continuous} \\ \text{variables imputed}}} \left(\frac{\text{imputed value} - \text{true value}}{\text{mean value}}\right)^2}}{\#\text{ of continuous variables to impute}} \tag{9}$$

The imputation error on categorical variables, E_{C_1}, is quantified in terms of the percentage of wrong imputations, while the imputation error on continuous variables, E_{C_2}, is estimated in terms of the normalized Euclidean norm.

In order to assign the missing values over all variables, the above 10-fold cross-validation is repeated for N_s trials. An optimal weight w^* is found when both E_{C_1} and E_{C_2} are relatively small. In this paper, we establish a frontier based on the average value of E_{C_1} and the 95th percentile of E_{C_2} values over N_s trials. The average value of E_{C_1}, denoted as $\bar{E}_{C_1}$, gives the average imputation error for categorical variables, while the 95th percentile of E_{C_2} values, denoted as $E_{C_2,95}$ gives a threshold that the imputation error for continuous variables is guaranteed to be less than $E_{C_2,95}$ with a probability of 95%. w^* can be then determined based on the frontier of $\bar{E}_{C_1}$ and $E_{C_2,95}$. The entire procedure of finding the optimal w^* is described in the pseudo-code in Table 3.

Table 3. Simulation experiments and 10-fold cross validation for determining w^*

For $w = 0{:}\,0.05{:}\,1$
  For simulation run = $1{:}\,N_s$
    Randomly partition $\boldsymbol{X}$ into 10 subsamples;
    For $i = 1{:}10$
      1. Use subset i as the validation set, and the rest 9 subsets as the training set;
      2. For each record $\boldsymbol{x}_i$ in the validation set, assign $x_{i,r}$ to be missing if $binom(q_r) = 1$;
      3. Impute the missing values in the validation set with *k*-NNHDI;
      4. Calculate E_{C_1} and E_{C_2}.
    End
  End
End
Determine w^* based on the frontier of $\bar{E}_{C_1}$ and $E_{C_2,95}$.

We then use $w = w^*$ with Eq. (7) to estimate the global distance and impute the missing values in $\boldsymbol{M}$, until all missing values in dataset $\boldsymbol{S}$ are imputed. An example of determining optimal w^* will be shown in Section 6.1 with the case study.

5 The LASSO method for variable selection

This section discusses how to select an optimal subset of variables for designing a classification model to predict treatment decision. Variable selection provides the benefits of reducing data dimension, improving model interpretability, neglecting insignificant effects thus reducing noise effects, increasing model predictive capability, etc. The LASSO (least absolute shrinkage and selection operator) is a regression method proposed by Tibshirani (1996). Similar to Ordinary Least Squares regression, LASSO minimizes the residual sum of squares (RSS) but poses a constraint to the sum of the absolute values of the coefficients being less than a constant (Tibshirani, 1996). The LASSO penalties provide a natural criterion to enforce sparsity and simplicity in the solution and hence is effectively used for variable selection.

Let $\boldsymbol{s}_i$ be a medical record sample in dataset $\boldsymbol{S}$, y_i be the treatment decision for subject i, then the model is trained based on the following optimization formulation:

$$LASSO = \min \frac{1}{2} \sum_{i=1}^{N+N_m} \left(y_i - \boldsymbol{s_i^T \beta}\right)^2 + \lambda \sum_{r=1}^{n} |\beta_r| \tag{10}$$

where λ is a parameter determining the shrinkage level. In order to avoid untrustworthy approximations in determining λ, we select λ via cross-validation. However, due to possible inhomogeneity among subsamples in cross-validation, the selected λ and variable subset may be different among different trials of cross-validation. To overcome this randomness, we propose to perform multiple runs of cross-validation, say, C runs, and evaluate the percentage of runs in which each variable is selected. Suppose variable r is selected in $(\eta_r \times 100)\%$ of the cross-validation runs. The selected variable subset can be determined in two ways: (1) set a threshold η_0, and select the

variables with $\eta_r \geq \eta_0$, or (2) restrict the number of selected variables to be n_0 and select the top n_0 variables with the higher η_r values. In this paper, the case study utilizes rule (1) to select variables. It should be noted that the feature selection results need to be verified with physicians in order to make sure that the features are clinically meaningful.

6 Case study

In this section, we will apply the methods developed in previous sections to the dataset described in Section 3 and show the results in each of the following steps: (1) imputing the missing values, (2) selecting variables for classification, and (3) evaluating the prediction performance using LR and a probability-based decision rule.

6.1 Missing value imputation

Before imputing the missing values using *k*-NNHDI, we first determine the weight w based on dataset $\boldsymbol{X}$, which consists of 47 records without any missing values. For a given w, we performed $N_s = 10{,}000$ simulation runs with 10-fold cross-validation on $\boldsymbol{X}$ and record the imputation performance, E_{C_1} and E_{C_2}. *k*-NNHDI with $k = 5$ was implemented in this procedure. We searched w values between 0 and 1 in 0.05 increment. Figure 3 shows $\bar{E}_{C_1}$ and $E_{C_2,95}$ values under different weights. As mentioned in Section 4.3, $\bar{E}_{C_1}$ gives the average imputation error for categorical variables while $E_{C_2,95}$ gives a threshold that the imputation error for continuous variables is guaranteed to be less than $E_{C_2,95}$ with a probability of 95%. It can be seen from Figure 3 that $w \geq 0.20$ is generally preferred to achieve a small E_{C_1}, and $w \leq 0.80$ is generally preferred to achieve a small E_{C_2}. This is consistent with our intuition that a larger value of w is preferred if categorical variables are considered to be more important than continuous variables, and vice versa.

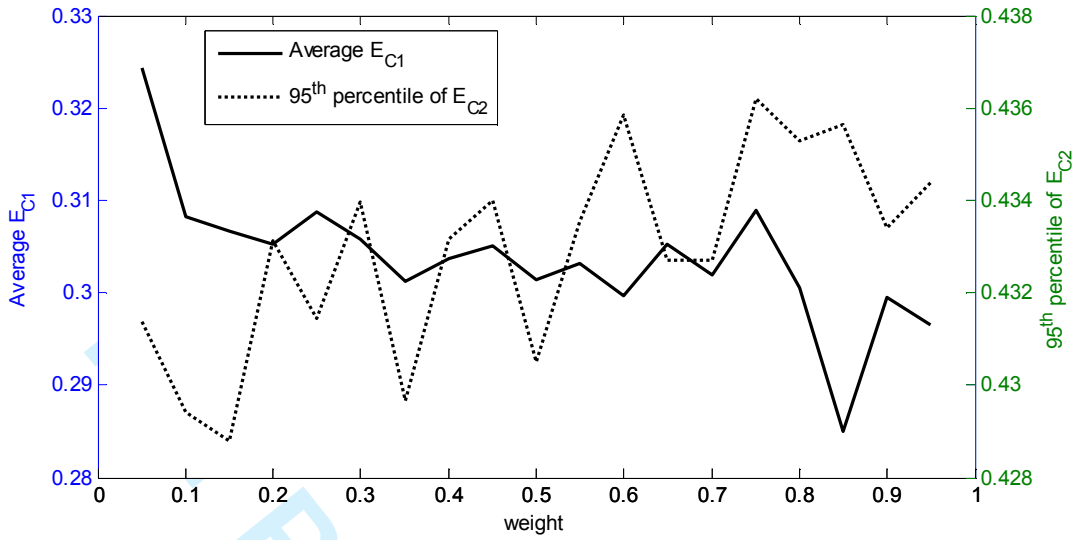

Figure 3. Imputation performance under different weights

Since the optimal weight w^* is desired when both E_{C_1} and E_{C_2} are relatively small, we establish a frontier of $\bar{E}_{C_1}$ and $E_{C_2,95}$ in Figure 4. The frontier of optimal weights is obtained by linking weights 0.15, 0.35, and 0.85. In this paper, we consider a balanced performance for both categorical and continuous variables, thus $w^* = 0.35$ is selected. We then use $w^* = 0.35$ with $k = 5$ for *k*-NNHDI to impute the missing values in $\boldsymbol{M}$, which contained 197 records with missing values.

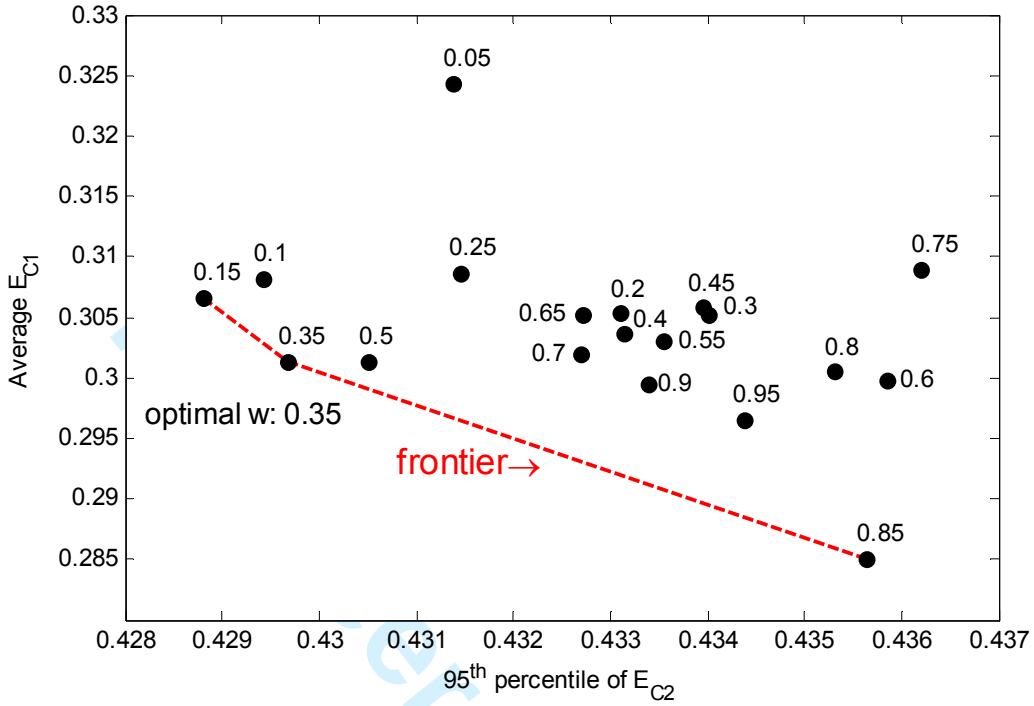

Figure 4. Imputation performance and the frontier

6.2 Variable selection

We performed LASSO method for variable selection with the inclusion of all 22 variables ($n_1 = 9, n_2 = 13$). A total of 100 runs of cross-validation are performed for variable selection. Figure 5 shows the percentage of runs in which each variable is selected. Setting the threshold at $\eta_0 = 50\%$, we obtained the selected variables as $f_{16}, f_{17}, f_{18}, f_{19}$, and f_4 (see Section 3 and Table 1 for the description of features). Specifically, the selected subset of variables includes:

- 4 features with 100% selection of cross-validation runs: 'ASES Total Score', 'WORC Physical Score', 'WORC Sport/Rec Score', 'WORC Work Score';
- 1 feature with 54% selection: 'Duration';

The selected continuous variables convey information gathered from the WORC form and the ASES form, whereas 'Duration' indicates the duration of the symptoms. After verifying with physicians, we will use these selected variables as input variables for classification models.

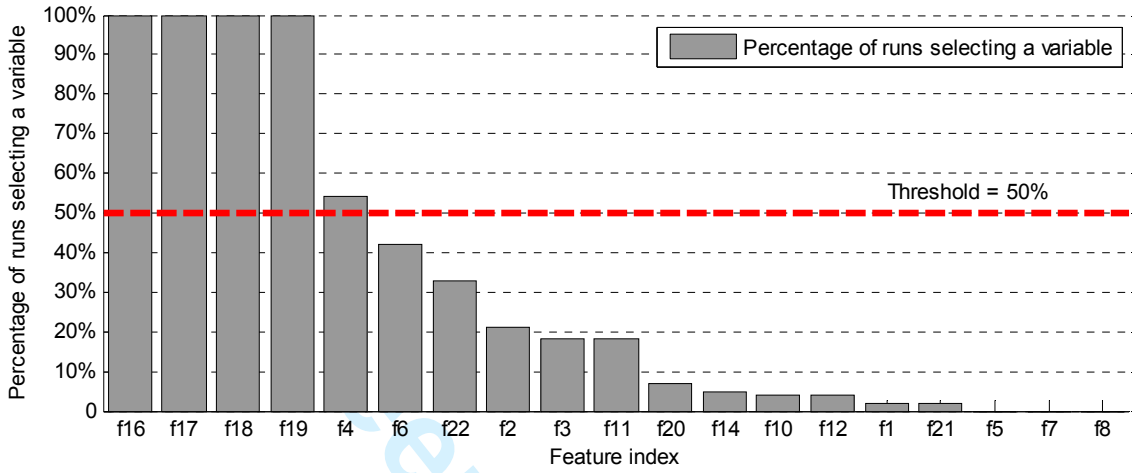

Figure 5. Variable selection results

6.3 Logistic regression and probability-based decision making

A logistic regression model is developed with the five predictors selected in Section 6.2. LR provides interpretable result in terms of the probability that a patient eventually needs surgery. The exclusion of noninformative variables in LR also ensures the model structure to be simple with a better interpretability. Based on all 244 patient records and the selected variables $f_{16}, f_{17}, f_{18}, f_{19}$, and f_4, the following LR model is developed:

$$ln\frac{\hat{p}_i}{1-\hat{p}_i} = -0.1963 - 0.3890s_{i,16} + 0.3502s_{i,19} - 0.1808s_{i,4} + 0.1590s_{i,18} + 0.1297s_{i,17} \quad (11)$$

where $\hat{p}_i$ is the predicted probability for subject i, $\boldsymbol{s}_i$ is the medical record for subject i, and $s_{i,16}, s_{i,19}, s_{i,4}, s_{i,18}$, and $s_{i,17}$ are the normalized values of $\boldsymbol{s}_i$ at variables $f_{16}, f_{19}, f_4, f_{18}$, and f_{17}.

The decision rule in conventional LR is based on comparing the predicted probability with 0.5. In our case, however, it is important to avoid jumping to a surgery decision if the predicted probability is only slightly above 0.5. Therefore, we propose a probability-based decision rule for our decision support system.

Under the probability-based decision rule, model decisions are made by comparing the predicted probability with an upper decision limit (UDL) and a lower decision limit (LDL). If the predicted probability given by LR, $\hat{p}$, is above UDL or below LDL, a more certain decision is suggested by the model, whereas if $\hat{p}$ is between LDL and UDL, a less certain decision is given by the model. In this way, uncertain predictions can be alarmed separately. As mentioned in Section 2 and Figure 2, if the decision support system gives an uncertain decision, physicians are suggested to make the surgery decision mostly based on their expertise; whereas if the decision support system gives a certain decision, physicians are suggested to consider model decision along with their expertise decision. Specifically, the probability-based decision rule is defined as follows: for subject i with predicted probability $\hat{p}_i$,

- Recommend surgery to subject i if $\hat{p}_i \geq UDL$;
- Recommend nonsurgical treatments to subject i if $\hat{p}_i \leq LDL$;
- Give an uncertain model decision to subject i and let physicians take the dominant role if $LDL < \hat{p}_i < UDL$.

Since we should avoid recommending a surgical treatment to patients who do not really need it, the decision limits need to be adaptively adjusted in order to keep the prediction error of having false positive low. In our data, among the total of 244 patient records, 113 had surgeries while 131 patients did not. Among the 131 patients who did not have surgeries, most of their predicted probabilities were less than 0.7. Therefore, setting $UDL = 0.7$ or higher is promising to have a small false positive rate. On the other hand, among the 113 patients who had surgeries, many of their

predicted probabilities were higher than 0.5. Therefore, setting $LDL = 0.5$ or lower is promising to have a relatively small false negative rate.

In Table 4 and Table 5, we present the average decision making performance using LR and the probability-based decision rule with $UDL = 0.7, LDL = 0.5$, and $UDL = 0.75, LDL = 0.5$, respectively. The results in Table 4 and Table 5 are the average results over 10 simulation trials. Under each trial, 80% of the data are randomly selected as the training dataset whereas 20% of the data are treated as the testing dataset. The results are reported in terms of patient percentage. Take Table 4 for example. In the training dataset, about 37.6% of the incoming patients are correctly identified that they do not need surgery, while 12.5% are correctly identified that they need surgery. The false positive rate in training dataset is 2.3%, indicating that 2.3% of the incoming patients are suggested to have a surgical treatment when they actually do not need surgery. The false negative rate in training dataset is 18.2%, indicating that 18.2% of the incoming patients are suggested to go through physical therapy but they will eventually need surgery.

Table 4. Decision making performance with UDL = 0.7, LDL = 0.5

Training		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	37.6%	2.3% (False positive rate)	13.4%
	Surgery	18.2% (False negative rate)	12.5%	16.0%
Testing		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	39.2%	2.8% (False positive rate)	13.1%
	Surgery	18.6% (False negative rate)	11.8%	14.5%

Table 5. Decision making performance with UDL = 0.75, LDL = 0.5

Training	Predicted as		
	Nonsurgical treatment	Surgery	Uncertain Decision

Actual	Nonsurgical treatment	37.6%	0.7% (False positive rate)	15.0%
	Surgery	18.2% (False negative rate)	7.4%	21.1%
Testing		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	39.2%	1.1% (False positive rate)	14.7%
	Surgery	18.6% (False negative rate)	7.2%	19.2%

It can be seen from Table 4 that decision limits $UDL = 0.7$ and $LDL = 0.5$ are able to keep the false positive rate under 3% and the false negative rate under 19%. We can further reduce the false positive rate to around 1% when UDL is adjusted to be 0.75, as shown in Table 5, but forcing more patients to have uncertain decisions. These probability-based decision limits can be flexibly adjusted to achieve satisfying prediction and decision making performance.

We further compare the decision making performance given by our decision support system and that in the MedSport Clinic. According to Figure 1, about 75% of patients start with nonsurgical treatments, but 45% of these patients eventually need to go through surgery after taking nonsurgical treatments. This gives the true decision rate on surgery to be 25% and the true decision rate on nonsurgical treatments to be $75\% \times (1 - 45\%) = 41.25\%$. The Clinic's false positive rate is simply assumed to be zero since there is no data to evaluate this rate. The false negative rate is $75\% \times 45\% = 33.75\%$ and this portion indicates the time and cost of treatment and pain for patients due to ineffective physical therapy before surgery. Comparing these data with Table 4 when $UDL = 0.7$ and $LDL = 0.5$, our decision support system manages to keep the false positive rate below 3%, which is an inevitable possibility from statistical methods. The false negative rate given by our decision support system is around 18~19%, which shows a significant reduction from the 33.75% given by the clinic.

Using the classification results and the probability-based decision rule has the advantage of providing the predicted probability of eventually needing surgery. Such a probability shows either a confident decision or a recommended decision from the model and thus provides the flexibility for decision makers to combine their own experiences. It should be noted that the results from our decision support system aim to provide support for physicians when making a treatment decision for rotator cuff tear patients instead of making decisions without physicians. The results from this case study further proves that our decision support system has the potential to improve patient safety, reduce cost of unnecessary treatment, and help physicians prevent treatment errors.

7 Conclusion

This research provides a prediction model to assist physicians in making an effective surgery decision at an early stage on whether or not a surgical treatment is eventually needed for rotator cuff tear patient. We developed a decision support system based on the integration of missing data imputation, variable selection, and classification/regression methods to predict the probability of requiring a surgical treatment. This research focuses on developing a new imputation method that deals with large amount of mixed-type missing data in patients' records by defining a global distance and selecting an optimal weight to combine partial distances. We demonstrated how the proposed decision support system works with a case study. A logistic regression model was developed based on the selected variables and a reliable probability-based decision rule was recommended. The proposed decision support system has the potential to be applied to other healthcare applications and also help to improve patient-centric healthcare delivery quality and patients safety.

Acknowledgements

This research is supported by the Laboratory for Optimization and Computation in Orthopaedic Surgery (LOCOS) at the Department of Orthopaedic, University of Michigan Medical

School. The data in this study is provided by the University of Michigan MedSport Clinic. The authors would like to thank Dr. Richard Hughes and Mr. Jaimee Gauthier for their help with background understanding and data collection.

References

American Academy of Orthopaedic Surgeons (2011) Rotator Cuff Tears, Retrieved online: http://orthoinfo.aaos.org/topic.cfm?topic=A00064.

Seida, J., Schouten, J. and Mousavi, S. (2010) Comparative Effectiveness of Nonoperative and Operative Treatments for Rotator Cuff Tears, in Comparative Effectiveness Reviews, No. 22., Rockville (MD): Agency for Healthcare Research and Quality (US).

Hu, H., Wang, H. and Zheng, B. (2011) Challenges in Managing and Mining Large, Heterogeneous Data, in Database Systems for Advanced Applications, Springer Berlin Heidelberg, pp. 462-462.

Johnstone, I.M. and Titterington, D.M. (2009) Statistical challenges of high-dimensional data. Phil Trans R Soc A,367, 4237-4253.

Denton, B.T., Alagoz, O., Holder, A. and Lee, E.K. (2011) Medical decision making: open research challenges. IIE Transactions on Healthcare Systems Engineering,1, 161-167.

Lavieri, M.S., Puterman, M.L., Tyldesley, S. and Morris, W.J. (2012) When to treat prostate cancer patients based on their PSA dynamics. IIE Transactions on Healthcare Systems Engineering,2, 62-77.

Claudio, D., Kremer, G.E.O., Bravo-Llerena, W. and Freivalds, A. (2014) A dynamic multi-attribute utility theory–based decision support system for patient prioritization in the emergency department. IIE Transactions on Healthcare Systems Engineering,4, 1-15.

Van der Pol, M. and Cairns, J. (2003) Predicting Attendance for Breast Screening Using Routinely Collected Data. Health Care Management Science,6, 229-236.

Alaeddini, A., Yang, K., Reddy, C. and Yu, S. (2011) A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Management Science,14, 146-157.

Chen, H.-Y., Chuang, C.-H., Yang, Y.-J. and Wu, T.-P. (2011) Exploring the risk factors of preterm birth using data mining. Expert Systems with Applications,38, 5384-5387.

Michener, L.A., McClure, P.W. and Sennett, B.J. (2002) American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: Reliability, validity, and responsiveness. Journal of Shoulder and Elbow Surgery,11, 587-594.

Kirkley, A.M.D.M.F., Alvarez, C.M.D.F. and Griffin, S.C.S.S. (2003) The Development and Evaluation of a Disease-specific Quality-of-Life Questionnaire for Disorders of the Rotator Cuff: The Western Ontario Rotator Cuff Index. Clinical Journal of Sport Medicine,13, 84-92.

Raman, J. and Macdermid, J.C. (2012) Western Ontario Rotator Cuff Index. Journal of Physiotherapy,58, 201.

Ayuyev, V.V., Jupin, J., Harris, P.W. and Obradovic, Z. (2009) Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data, in Proceedings of the Proceedings of the 11th International Conference on Data Warehousing and Knowledge Discovery, Springer-Verlag, City, pp. 366-377.

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

Cadwell, B., Boyle, J., Tierney, E. and Thompson, T. (2007) A Bayesian approach to assess heart disease mortality among persons with diabetes in the presence of missing data. Health Care Management Science,10, 231-238.

Tarsitano, A. and Falcone, M. (2011) Missing-Values Adjustment for Mixed-Type Data. Journal of Probability and Statistics,2011, 20 pages.

Kalton, G. and Kasprzyk, D. (1982) Imputing for missing survey responses. Proceedings of the Selection on Survey Research Methods, 22-31.

Little, R. and Rubin, D. (2002) Statistical Analysis with Missing Data, Second Edition, Wiley-Interscience.

Farhangfar, A., Kurgan, L.A. and Pedrycz, W. (2004) Experimental analysis of methods for imputation of missing values in databases. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series,5421.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics,17, 520-525.

Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological),58, 267-288.

List of figures

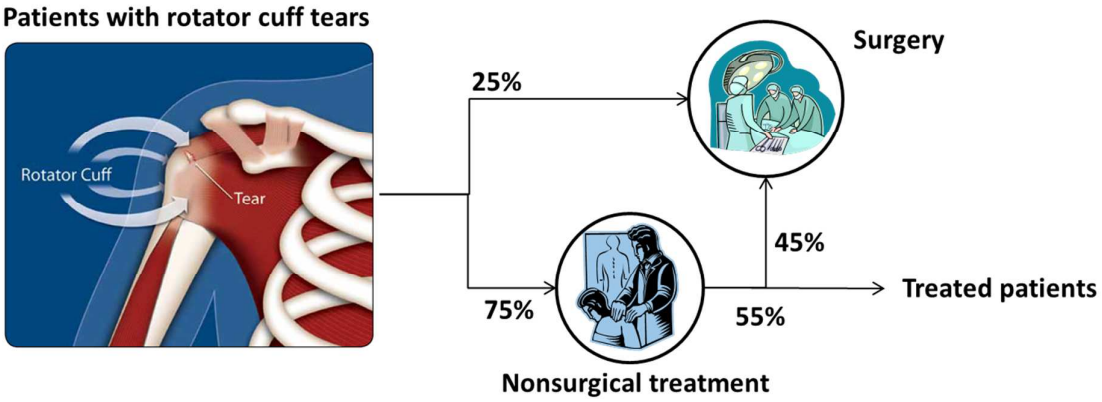

Figure 1. Treatment options for rotator cuff tears (Source: the University of Michigan MedSport Clinic)

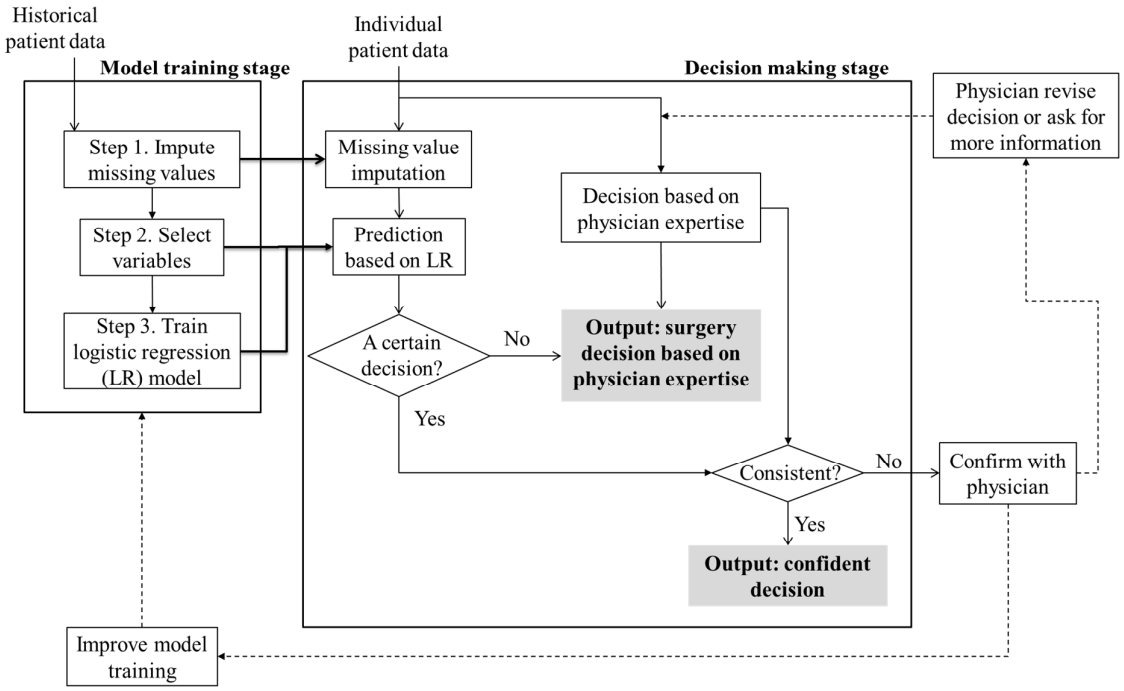

Figure 2. Framework of the proposed decision support system

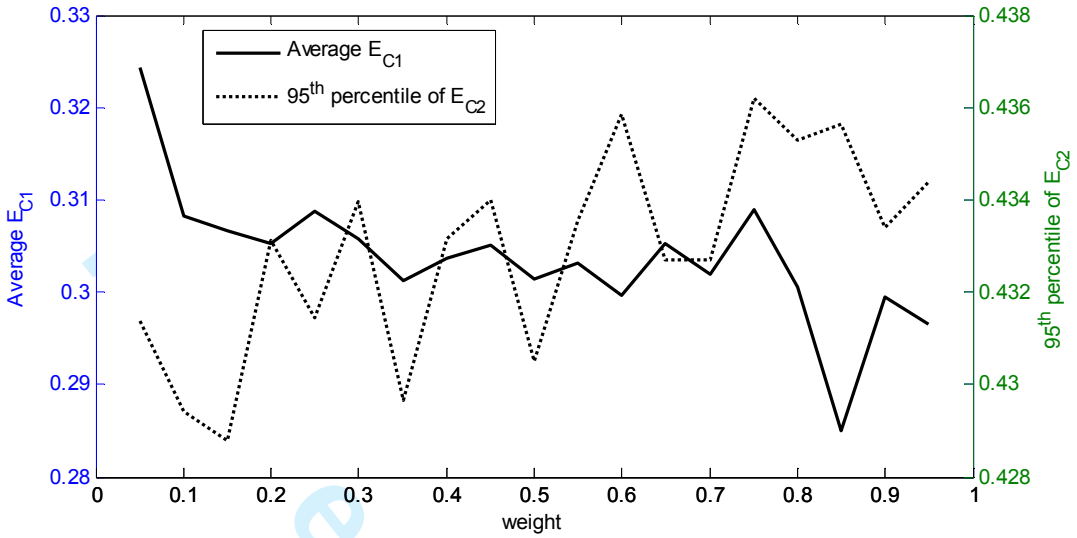

Figure 3. Imputation performance under different weights

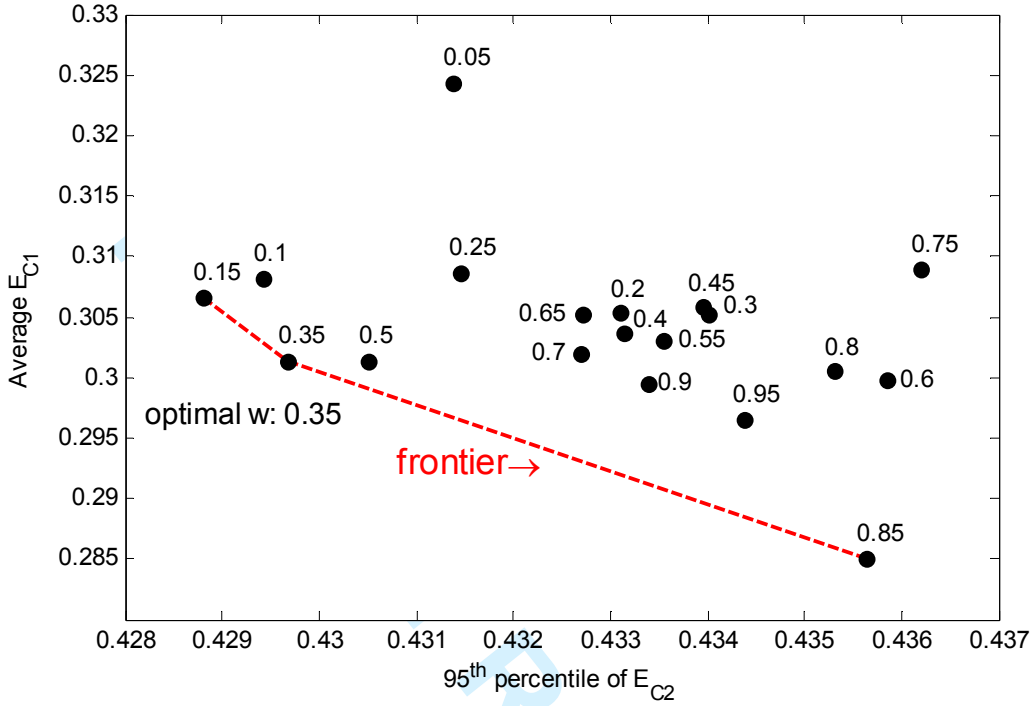

Figure 4. Imputation performance and the frontier

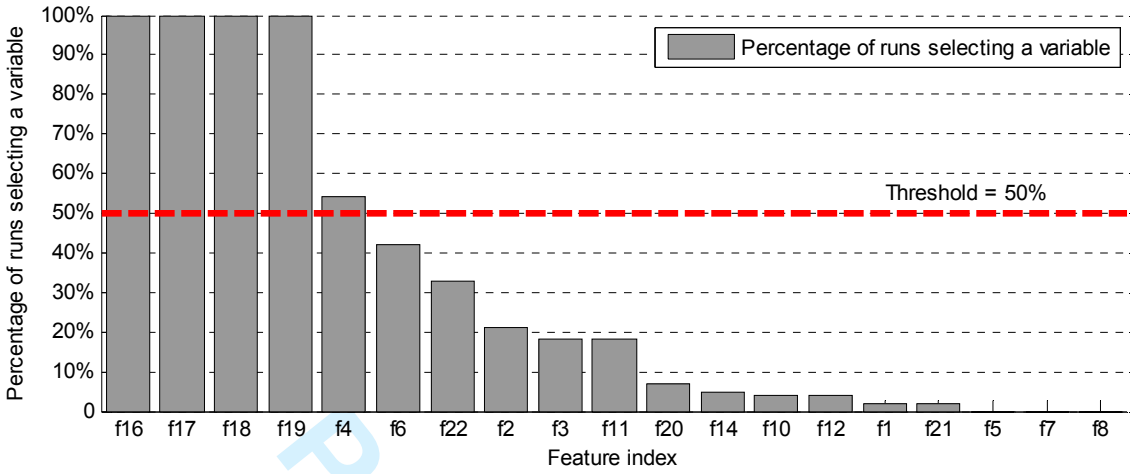

Figure 5. Variable selection results

List of tables

Table 1. Feature description

Feature Index	Feature Description		Variable Type	Number of missing values (in 244 records)
f_1	Surgery side		Categorical	0
f_2	Gender			0
f_3	Dominant side			53
f_4	Duration of symptoms (unit: month)		Continuous	53
f_5	Adhesive capsulitis	Diagnosis codes	Categorical	0
f_6	Impingement			
f_7	Osteoarthritis – glenohumeral joint			
f_8	Instability			
f_9	Osteoarthritis – acromioclavicular joint			
f_{10}	Partial thickness tear			
f_{11}	Body mass index (BMI)		Continuous	70
f_{12}	Survery12 Physical score	Scores from The Veterans Rand 12 Item Health Survey	Continuous	118
f_{13}	Survey12 Mental score			
f_{14}	ASES Total Score Opposite Side	Scores from ASES (American Shoulder and Elbow Surgeons) Shoulder Form	Continuous	135
f_{15}	ASES Total Score Affected Side		Continuous	55
f_{16}	ASES Total Score		Continuous	70
f_{17}	WORC Physical Score	Scores calculated based on the Western Ontario Rotator Cuff Index	Continuous	33
f_{18}	WORC Sport Rec Score		Continuous	33
f_{19}	WORC Work Score		Continuous	32
f_{20}	WORC Lifestyle Score		Continuous	32
f_{21}	WORC Emotions Score		Continuous	32
f_{22}	Shoulder Activity Score	Score from Shoulder Activity Level Form	Continuous	138

Table 2. *k*-NNHDI procedure

$i \leftarrow 1$

While $i \leq N_m$

1. Take record $\boldsymbol{m}_i$ in $\boldsymbol{M}$;
2. Calculate the distance between $\boldsymbol{m}_i$ and each record $\boldsymbol{x}_1, \dots, \boldsymbol{x}_j, \dots \boldsymbol{x}_N$ in $\boldsymbol{X}$, denote as $d_{ij}, j = 1, \dots, N$, respectively;
3. Among $d_{ij}, j = 1, \dots, N$, pick $j_1, \dots, j_k$ so that $d_{ij_1} \leq \cdots \leq d_{ij_k} \leq$ all other distances;
4. Impute the missing value(s) in $\boldsymbol{m}_i$ based on $x_{j_1}, \dots, x_{j_k}$:

$$m_{i,r} = \begin{cases} \text{mode of } x_{j_1,r}, \dots, x_{j_k,r}, \text{ if variable } r \text{ is a categorical variable} \\ \text{mean of } x_{j_1,r}, \dots, x_{j_k,r}, \text{ if variable } r \text{ is a continuous variable} \end{cases} \quad (1)$$

5. Update $\boldsymbol{X}$ by adding the completed $\boldsymbol{m}_i$ to the complete data set;
6. Update $\boldsymbol{M}$ by deleting record $\boldsymbol{m}_i$;
7. $i \leftarrow i + 1$; go to the next record in $\boldsymbol{M}$

End

Output: a data set $\boldsymbol{S}$ with no missing values

Table 3. Simulation experiments and 10-fold cross validation for determining w^*

For $w = 0:0.05:1$

For simulation run = $1:N_s$

Randomly partition $\boldsymbol{X}$ into 10 subsamples;

For $i = 1:10$

1. Use subset i as the validation set, and the rest 9 subsets as the training set;
2. For each record $\boldsymbol{x}_i$ in the validation set, assign $x_{i,r}$ to be missing if $binom(q_r) = 1$;
3. Impute the missing values in the validation set with *k*-NNHDI;
4. Calculate E_{C_1} and E_{C_2}.

End

End

End

Determine w^* based on the frontier of $\bar{E}_{C_1}$ and $E_{C_2,95}$.

Table 4. Decision making performance with UDL = 0.7, LDL = 0.5

Training		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	37.6%	2.3% (False positive rate)	13.4%
	Surgery	18.2% (False negative rate)	12.5%	16.0%
Testing		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	39.2%	2.8% (False positive rate)	13.1%
	Surgery	18.6% (False negative rate)	11.8%	14.5%

Table 5. Decision making performance with UDL = 0.75, LDL = 0.5

Training		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	37.6%	0.7% (False positive rate)	15.0%
	Surgery	18.2% (False negative rate)	7.4%	21.1%
Testing		Predicted as		
		Nonsurgical treatment	Surgery	Uncertain Decision
Actual	Nonsurgical treatment	39.2%	1.1% (False positive rate)	14.7%
	Surgery	18.6% (False negative rate)	7.2%	19.2%