Computers in Biology and Medicine 57 (2015) 54–65

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Pre-operative prediction of surgical morbidity in children: Comparison of five statistical models

Jennifer N. Cooper [a,*], Lai Wei [b], Soledad A. Fernandez [b], Peter C. Minneci [a,c], Katherine J. Deans [a,c]

[a] Center for Surgical Outcomes Research, The Research Institute at Nationwide Childrens Hospital, 700 Childrens Dr., Columbus, OH 43205, USA
[b] Center for Biostatistics, The Ohio State University, 2012 Kenny Road, Columbus, OH 43221, USA
[c] Department of Surgery, Nationwide Children's Hospital, 700 Childrens Dr., Columbus, OH 43205, USA

ARTICLE INFO

Article history:
Received 29 July 2014
Accepted 17 November 2014

Keywords:
Data mining
Machine learning
Prediction
Boosting
Random forests
Support vector machines
Logistic regression
Surgical morbidity
Pediatrics

ABSTRACT

Background: The accurate prediction of surgical risk is important to patients and physicians. Logistic regression (LR) models are typically used to estimate these risks. However, in the fields of data mining and machine-learning, many alternative classification and prediction algorithms have been developed. This study aimed to compare the performance of LR to several data mining algorithms for predicting 30-day surgical morbidity in children.
Methods: We used the 2012 National Surgical Quality Improvement Program-Pediatric dataset to compare the performance of (1) a LR model that assumed linearity and additivity (simple LR model) (2) a LR model incorporating restricted cubic splines and interactions (flexible LR model) (3) a support vector machine, (4) a random forest and (5) boosted classification trees for predicting surgical morbidity.
Results: The ensemble-based methods showed significantly higher accuracy, sensitivity, specificity, PPV, and NPV than the simple LR model. However, none of the models performed better than the flexible LR model in terms of the aforementioned measures or in model calibration or discrimination.
Conclusion: Support vector machines, random forests, and boosted classification trees do not show better performance than LR for predicting pediatric surgical morbidity. After further validation, the flexible LR model derived in this study could be used to assist with clinical decision-making based on patient-specific surgical risks.

1. Introduction

Data mining algorithms, sometimes called machine learning or statistical learning algorithms, have been increasingly used in biomedical research in recent years. Data mining is broadly defined as the process of selecting, exploring, and modeling large amounts of data to discover unknown and useful patterns or relationships [1,2]. Data mining algorithms arose from the fields of statistics and computer science, and are widely used in marketing, banking, engineering, and bioinformatics. Their application to clinical research, however, has been limited.

In clinical research, logistic regression models are by far the most commonly used algorithm for predicting the probability of occurrence of an event. While these models can provide unbiased estimates of the associations between predictors and the outcome, they have some limitations. First, they assume a particular parametric form of the relationships between the predictors and the outcome; namely, the assumption is made that the logit of the outcome is equal to a linear combination of the independent variables [3]. These models also assume additivity of the predictors' effects on the outcome. These assumptions are usually incorrect, though the extent to which they are incorrect varies. Furthermore, in small datasets, these assumptions may be necessary to avoid overfitting. In larger datasets, these assumptions can be circumvented by using transformations or splines to model continuous predictors and by including interactions between variables. These techniques can improve model fit, but they are infrequently used, partly because they tend to reduce model interpretability [4]. Another limitation of regression models is that they do not

* Correspondence to: Center for Surgical Outcomes Research and Center for Innovation in Pediatric Practice, The Research Institute at Nationwide Children's Hospital 700 Childrens Drive, JWest 4915 Columbus, OH 43205.
Tel.: +1 614 355 4526; fax: +1 614 722 6980.
E-mail addresses: jennifer.cooper@nationwidechildrens.org (J.N. Cooper), lai.wei@osumc.edu (L. Wei), soledad.fernandez@osumc.edu (S.A. Fernandez), peter.minneci@nationwidechildrens.org (P.C. Minneci), katherine.deans@nationwidechildrens.org (K.J. Deans).

http://dx.doi.org/10.1016/j.compbiomed.2014.11.009
0010-4825/

always provide optimal predictive accuracy. In clinical research, these models are typically built to describe the nature of the relationship between specific covariates and the outcome [2]. While estimating such relationships is clearly important in biomedical research, accurate prediction is also very important. In fact, in certain situations in which the primary aim is to achieve optimal predictive accuracy, a reduction in clinical interpretability may be acceptable.

One area of biomedical research in which data mining may be particularly useful is in outcome prediction using large clinical databases, such as the American College of Surgeons' National Surgical Quality Improvement Program (ACS NSQIP) database [2]. Of the few studies investigating the performance of data mining algorithms for predicting surgical morbidity or mortality, most have been small (several hundred or several thousand patients) [5–10], though a few larger studies have been reported [11–17]. These studies have been inconsistent in their findings, in that some have shown data mining algorithms to perform better than traditional logistic regression in terms of overall accuracy [13,14,16,18], discrimination [13,14,16], or calibration [11], whereas some have reported similar performance according to these measures [11,18–20]. Data from the ACS NSQIP has been used to create risk calculators to predict post-operative outcomes for adult surgery patients overall [21] and for patients undergoing specific procedures [22–25]. Several of these calculators are freely available online, and their use by both physicians and patients has the potential to improve shared decision making and informed consent [21–25]. All of these calculators are based on logistic regression models that are reported to have good discrimination and calibration. However, none of the studies in which these prediction models were derived reported investigating whether other statistical algorithms might perform as well as or better than logistic regression, and none included pediatric patients. The objective of this study was to compare the performance of five different statistical algorithms for predicting surgical morbidity in pediatric surgical patients. The algorithms evaluated were chosen because of their infrequent use in the clinical research literature and their straightforward implementation in freely available software and included (1) a logistic regression model that assumed linearity and additivity (simple logistic regression model) (2) a logistic regression model incorporating restricted cubic splines and interactions (flexible logistic regression model) (3) a support vector machine, (4) a random forest and (5) boosted classification trees.

2. Methods

This study used the 2012 NSQIP Pediatric (NSQIP-Peds) Participant Use Data File, which contains patient-level data on 51,008 pediatric surgery cases submitted in 2012 by 50 US and Canadian children's hospitals. The NSQIP-Peds program is a multi-specialty program with cases sampled from pediatric general/thoracic surgery, pediatric otolaryngology, pediatric orthopedic surgery, pediatric urology, pediatric neurosurgery, and pediatric plastic surgery. Launched in October 2008 with 4 sites, NSQIP-Peds has since expanded, with 50 institutions participating in 2012. The program provides peer-reviewed, risk-adjusted 30-day postoperative outcomes to participating institutions, for the purposes of benchmarking and quality improvement [26–28]. Included cases are selected based on Current Procedural Terminology codes using NSQIP 8-day cycle-based systematic sampling of 35 procedures per cycle. One hundred and twenty-nine variables are collected from the medical records and the patients and their families, including information on demographics, surgical profile, preoperative and intraoperative variables, and postoperative occurrences [26–28]. The conduct of this study was approved by Nationwide Children's Hospital Institutional Research Board with a waiver of informed consent.

In this study, we considered the question of which model most accurately predicts the occurrence of surgical morbidity within 30 days of surgery. Neonates were excluded, because of the known differences in risk of surgical morbidity between neonates and non-neonates and because of the relatively small number of neonates ($N=2919$) and larger amount of missing data in neonates compared to non-neonates ($N=48089$) in the 2012 NSQIP-Peds sample. The 49 preoperative variables in pediatric patients considered for inclusion in each model are shown in Table 1. This list consists of all preoperative patient characteristics available in the database, though some rare characteristics were eliminated or grouped with other similar characteristics. Of note, procedures that occurred concurrently with the principal operative procedure were not considered as predictors because whether these additional procedures would be performed was not necessarily known preoperatively. In addition, 60 individual procedures (designated by CPT codes) that were performed in the total cohort at least 200 times were also included as indicator variables in the models, resulting in a total of 109 predictor variables. A frequency of 200 times was chosen to maximize the external validity of the models by enabling the risk of surgical morbidity associated with each procedure to be estimated accurately in the training dataset. Many of the procedures performed less frequently had no associated cases of surgical morbidity in the sample, whereas all procedures performed 200 or more times were associated with at least one case of surgical morbidity. No observations were removed from the analyses due to the use of this criterion as each type of procedure included as a predictor was treated as an individual binary variable. The outcome variable was the occurrence of intra-operative or post-operative morbidity within 30 days of the surgery, which was defined as any of the following events: SSI (superficial, deep, or organ/space without open wound), wound disruption, pneumonia without preoperative pneumonia, unplanned intubation, pulmonary embolism, renal insufficiency or failure without preoperative renal failure or dialysis, urinary tract infection, central line associated bloodstream infection, coma >24 h without preoperative coma, seizure, nerve injury, any cerebral intra-ventricular hemorrhage, CVA/stroke or intracranial hemorrhage, cardiac arrest requiring CPR, venous thrombosis requiring therapy, bleeding/transfusion, graft/prosthesis/flap failure, or the development or worsening of sepsis [26,29]. Patients who died within 30 days of their surgery (0.1%) were included in all analyses because the outcome under examination, surgical morbidity, could occur either intraoperatively or postoperatively.

2.1. Statistical analysis

In order to avoid overfitting, which occurs when a model has excellent fit to the data used in model fitting but poor fit to external data [4,30], the 2012 NSQIP-Peds PUF dataset was split into training and validation datasets. Seventy percent of the observations were chosen randomly for use as the training dataset, and the other 30% were used as the test (validation) dataset. Each algorithm incorporated all 109 pre-operative variables of interest. The 5 statistical algorithms compared were: (1) a logistic regression model that assumed a linear relationship between each covariate and the log-odds of morbidity, with no interaction terms (simple logistic regression model), (2) a logistic regression model fit with the relationship between continuous variables and the log-odds of morbidity expressed using restricted cubic splines with decile knots and with interactions between any two predictors included if statistically significant at $p<0.01$ in stepwise selection when added to the model containing all main effects (flexible logistic regression model), (3) a support vector machine