Predicting Procedure Duration to Improve Scheduling of Elective Surgery

Zahra ShahabiKargar[1,2], Sankalp Khanna[2,1], Norm Good[2], Abdul Sattar[1],
James Lind[3], and John O'Dwyer[2]

[1] Institute for Integrated and Intelligent Systems,
Griffith University, Brisbane, Australia
Zahra.Shahabikargar@griffithuni.edu.au,
A.Sattar@griffith.edu.au
[2] The Australian e-Health Research Centre, CSIRO, Brisbane, Australia
{Sankalp.Khanna,Norm.Good,John.ODwyer}@csiro.au
[3] Gold Coast Hospital, Gold Coast, Australia
James.Lind@health.qld.gov.au

Abstract. The accuracy of surgery schedules depends on precise estimation of surgery duration. Current approaches employed by hospitals include historical averages and surgical team estimates which are not accurate enough. The inherent complexity of surgery duration estimation contributes significantly to increased procedure cancellations and reduced utilisation of already encumbered resources. In this study we employ administrative and perioperative data from a large metropolitan hospital to investigate the performance of different machine learning approaches for improving procedure duration estimation. The predictive modelling approaches applied include linear regression (LR), multivariate adaptive regression splines (MARS), and random forests (RF). Cross validation results reveal that the random forest model outperforms other methods, reducing mean absolute percentage error by 28% when compared to current hospital estimation approaches.

Keywords: Duration of procedure, Operating Room (OR), Random Forest, Linear Regression, Multivariate Adaptive Regression Splines (MARS).

1 Introduction

Operating Rooms (ORs) are of pivotal importance to hospitals as they are the main revenue and cost centre of the hospital [1, 2]. However, 10% to 40% of scheduled elective procedures often get cancelled before surgery [3, 4], with the primary reason for day of surgery cancellations being lack of theatre time due to over-run of other surgeries. Robust schedules require surgery duration estimations that are unbiased, highly accurate, and should minimise cases with large absolute errors [5]. Accurate surgery duration estimation is essential for efficient use of ORs in hospitals as optimal planning can be achieved only when reliable predictions are available [6].

D.-N. Pham and S.-B. Park (Eds.): PRICAI 2014, LNAI 8862, pp. 998–1009, 2014.

More accurate procedure time estimations can improve surgery scheduling by providing a better arrangement of cases throughout the ORs, leading to more efficient use of resources and reduced costs. Also, it may allow more surgeries to be done which can lead to increased revenue. Therefore, regardless of the method used to construct surgery schedules, having an accurate estimation of case duration is a prerequisite for matching demand to capacity and will help to reduce both underutilisation and overrunning of the planned schedules.

Despite its obvious importance, predicting surgery duration is not an easy task due to the variability of situations and several significant factors. The procedure code that represent the core actions during the surgery is the most significant factor for predicting surgery duration [7]. Surgeries with straightforward diagnosis and standardised procedures are more predictable than complex surgeries. Currently, many hospitals use the historical average time for the same procedure codes for planning surgeries [8, 9]. However, these estimates are not accurate enough and result in suboptimal use of surgical facilities. Other sources of variability including patient characteristics (e.g. age, gender, diagnosis, etc.), type of surgery, and individual surgeons and anaesthetists can affect the duration of surgery [7, 8] and need to be considered when building a predictive model [10, 11].

During the last two decades a wide range of statistical and machine learning techniques have been used for predicting surgery duration including Linear Regression (LR) [8, 12], ANOVA [11], Bayesian approaches [13, 14], Neural networks [10, 15], and Random forests [16]. However, while these current research efforts outperform current hospital estimation methods, the prediction error of the proposed models is still quite high and the majority of these models are either specialty specific or based on limited datasets which make them hard to use in practical situations.

In this study we used four years of elective surgery data including a wide range of predictors (e.g. patient, operation, and surgery team characteristics) across all specialties in a major metropolitan Australian hospital. These predictors were selected after an exhaustive review of available information sources and discussions with hospital clinicians and administrators. Our main focus is to build a model that can be applied as a standalone tool to provide a more accurate estimation to the administrator at the time of booking to help them improve their theatre utilisation. Such a tool can also be integrated within an intelligent hospital scheduling system to improve the allocation of theatre time and other resources.

The rest of this paper is organised as follows. Section 2 provides details about data collection and preparation, and briefly discusses the predictive modelling approaches employed. Section 3 presents our findings and discusses how the employed approaches compare to each other and to the baseline hospital estimates. Section 4 discusses how our work relates to the current state of the art approaches in surgery duration estimation. In section 5 we draw conclusions from our findings and indicate directions for future work.

2 Materials and Methods

2.1 Subject

The study employed 4 years (1/07/2008 to 31/06/2012) of administrative and perioperative data from a major metropolitan teaching hospital in Queensland, Australia with approximately 500 beds and catering to over 60,000 elective and emergency patients annually. Administrative data was obtained from the Hospital Based Corporate Information System (HBCIS) and perioperative information was obtained from the Operating Room Management Information System (ORMIS). Ethics approval for the study was obtained from the Gold Coast Hospital and Health Service Human Research Ethics Committee. Emergency surgical cases were not considered since the goal was estimating procedure time for planned, i.e. elective, surgeries.

2.2 Data Preparation

Administrative and perioperative elective surgery data was transformed to an individual procedure level. The data represented a wide range of details including patient characteristics, operation characteristics, and surgery team characteristics across 12 specialties. Potential predictors were chosen after an exhaustive review of available information sources and discussions with clinical experts and hospital administrators. Some of the input variables needed to be transformed before being employed for predictive modelling. Some additional features were also extracted from data. For instance, patient diagnosis (such as heart disease, AIDS, or cancer) was used to calculate the Charlson Comorbidity Index (CCI) [17] as an indicator of severity and complication of patient condition, and team size was calculated by summing up the number of people involved in the procedure. A list of resulting predictors that were then employed for predictive modelling is presented in Table 1. The initial dataset included 66233 individual procedures across 12 specialties. Those procedures that were performed less than 100 times during the period of this study, cases with no match between administrative and perioperative databases, and procedures that were not assigned to a surgical specialty were excluded. This left 38520 cases representing 104 different procedures. The output variable to be predicted was the total procedure time.

2.3 Modelling

Considering that the target output variable of our model (procedure time) is continuous, regression techniques were applied in this study. Linear Regression (LR), MARS (Multivariate Adaptive Regression Splines), and Random Forest (RF) algorithms were used for building the prediction models. Linear Regression was chosen as it is a common approach employed in prediction and can provide a good understanding of the relationships between variables. It is also a reasonable benchmark for evaluating other models. The MARS algorithm was chosen as it is able to search a large number of variables, their interactions, and all possible non-linear responses in a very efficient way and was well suited to the complexity of the problem at hand. Random forest was chosen because of its proven capability of being able to handle a large number of predictors efficiently and perform consistently well across a gamut of machine learning tasks. In this section, we briefly describe the techniques and how they were employed.

Table 1. Description of the predictors used for procedure time estimation

Predictor	Description (Type)	Distinct/ Mean values***
Patient characteristics		
Category	Urgency category of patient (Nominal)	6
Age	Age of patient (Numeric)	52
Gender	Patient gender (Nominal)	3
Type of admission	Patient Type of admission (Nominal)	4
Classification	Patient payment class (Nominal)	3
CCI	Charlson Comorbidity Index (Nominal)	21
Referral centre	Centre patient referred to (Nominal)	22
Operation characteristics		
Procedure indicator	Planned procedure (Nominal)	200
Unit	Hospital unit (Nominal)	65
Specialty	Hospital specialty (Nominal)	11
Theatre	Operating room number (Nominal)	22
Order	Operation order in session (Nominal)	25
Ward	Hospital ward (Nominal)	51
Sub specialty	Sub specialty code (Nominal)	69
Procedure	Procedure code (Nominal)	104
Primary	Is it the primary procedure? (Binary)	2
Procedure class*	Nominal	26
Session	Morning/ Afternoon session (Nominal)	2
Session type	Type of the session (Nominal)	5
Surgery team characteristics		
Consultant	Doctor who visited the patient (Nominal)	178
Con. category	Professional category of consultant (Nominal)	9
Surgeon	Surgeon in charge with operation (Nominal)	241
Surgeon category	Professional category of surgeon (Nominal)	11
Surgeon-Consultant	Is surgeon the same as consultant? (Binary)	2
Surgeons #	Number of surgeon involved (Nominal)	4
Anaesthetists #	Number of anaesthetists involved (Nominal)	4
Team size**	Total number of people involved (Nominal)	11

*This predictor has been extracted based on the historical average time of procedures before the period of this study. Those procedures that have similar average time have been grouped together.
** This number includes surgeons, anaesthetists, technicians, and nurses involved in the procedure.
*** For nominal and binary predictors, this represents the number of distinct levels/categories. For numeric predictors, this represents the mean value of the sample.

Linear Regression

Regression analysis is a common approach employed in prediction [18-21]. Generally, linear regression models the relationship between one or more input vectors

$X^T = (X_1, X_2, \ldots, X_p)$ and a dependent output variable Y using a linear equation. A regression model can be formularised as [22]:

$$f(X) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{1}$$

where β_i are the regression coefficients. Each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector of feature measurements for the i_{th} case. Here we have a set of training data (x_1, y_1). . . (x_n, y_n) where each x_i is a vector of predictors as listed in Table 1 and y_i is corresponding observed procedure time from which to estimate the parameters β.

The response variable was log-transformed for the regression analysis as it has been shown that the distribution of procedure duration best fits a lognormal distribution [6, 23]. The Statistics Toolbox in MATLAB R2014a was used for the modelling.

MARS (Multivariate Adaptive Regression Splines)

MARS is a non-parametric regression technique in which non-linear relationships between a response variable and the set of predictors are described by a series of piece-wise linear segments of differing slope.

$$f(x) = \sum_{i=1}^{n} c_i B_i(x) \tag{2}$$

The model is a weighted sum of basis functions $B_i(x)$. Basis functions are used to fit linear segments and added to the model in pairs using knots. Knots can be selected in a backwards/forward stepwise procedure to identify terms to be retained in the final model by evaluating a lack of fit function (e.g. maximising reduction in sum-of-squares residual error) [24]. MARS can search a large number of variables, their interactions, and all possible non-linear responses in a very efficient way. We implemented the ARESLab toolbox in MATLAB [25] to fit a MARS model to our data.

Random Forest

The Random Forest (RF) algorithm was developed by Leo Breiman [26] and is an ensemble learning method for classification and regression. The forest is built by combining numerous decision or regression trees in a random process in order to take advantage of the predictive power of each tree and boost the prediction performance and robustness of the algorithm. Each tree is grown on a bootstrap sample of the training cases and each node of the tree splits based on a random subset of the input variables. Random forest can handle a large number of predictors and its prediction performance compares well to other machine learning algorithms such as support vector machines (SVMs) [27], and artificial neural networks (ANN) [28]. Another advantage of random forests is a built-in estimate of accuracy, where for every tree, out of the bag (OOB) samples are used to get the estimated response of the corresponding tree for those cases. In regression the forest prediction is an unweighted average over all trees. In this study we used the Random Forest package in MATLAB R2014a [29].

2.4 Model Evaluation

We compared the performance of our models against each other and an existing current hospital estimation method. Ten-fold cross validation was employed and various error statistics, including Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), were measured.

RMSE quantifies the difference between the predicted values from a model ($f(x_i)$) and the actual values of the estimated variable (y_i).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2} \tag{3}$$

MAPE is another statistic which expresses accuracy as a percentage and can be used to measure of accuracy of a model.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - f(x_i)|}{y_i} \tag{4}$$

The R-squared value (R^2), also known as the coefficient of determination, was employed as a measure of goodness of fit. R^2 measures how well the fitted model can explain the variation of the data by measuring the correlation between the response value and the predicted values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

3 Results and Discussions

Figure 1 presents the distribution of our response variable, procedure time, across individual specialties including the mean and standard deviation (SD) of procedure times for each specialty. Analysis revealed considerable variation between specialties with the distribution being positively skewed with a long right tail in most specialties. In the data collected, 73% of procedure time estimates were based on the historic average time taken by the procedure in the past, 19% were based on estimates provided by surgeons, and 8% employed a default time in the absence of historical record and surgeon estimates.

Table 2. Performance of current method and prediction models

	RMSE		MAPE		R^2	
	Value	%Baseline*	Value	%Baseline*	Value	%Baseline*
Hospital	27.88	------	0.95	------	0.48	------
LR	28.12	-0.9	1.20	-26.3	0.46	-4.2
MARS	25.83	7.4	0.90	5.3	0.55	14.6
RF	22.78	18.3	0.68	28.4	0.65	35.4

*The %Baseline column shows the improvement in the performance metric as compared to the baseline performance, i.e. the hospital estimate of procedure time.

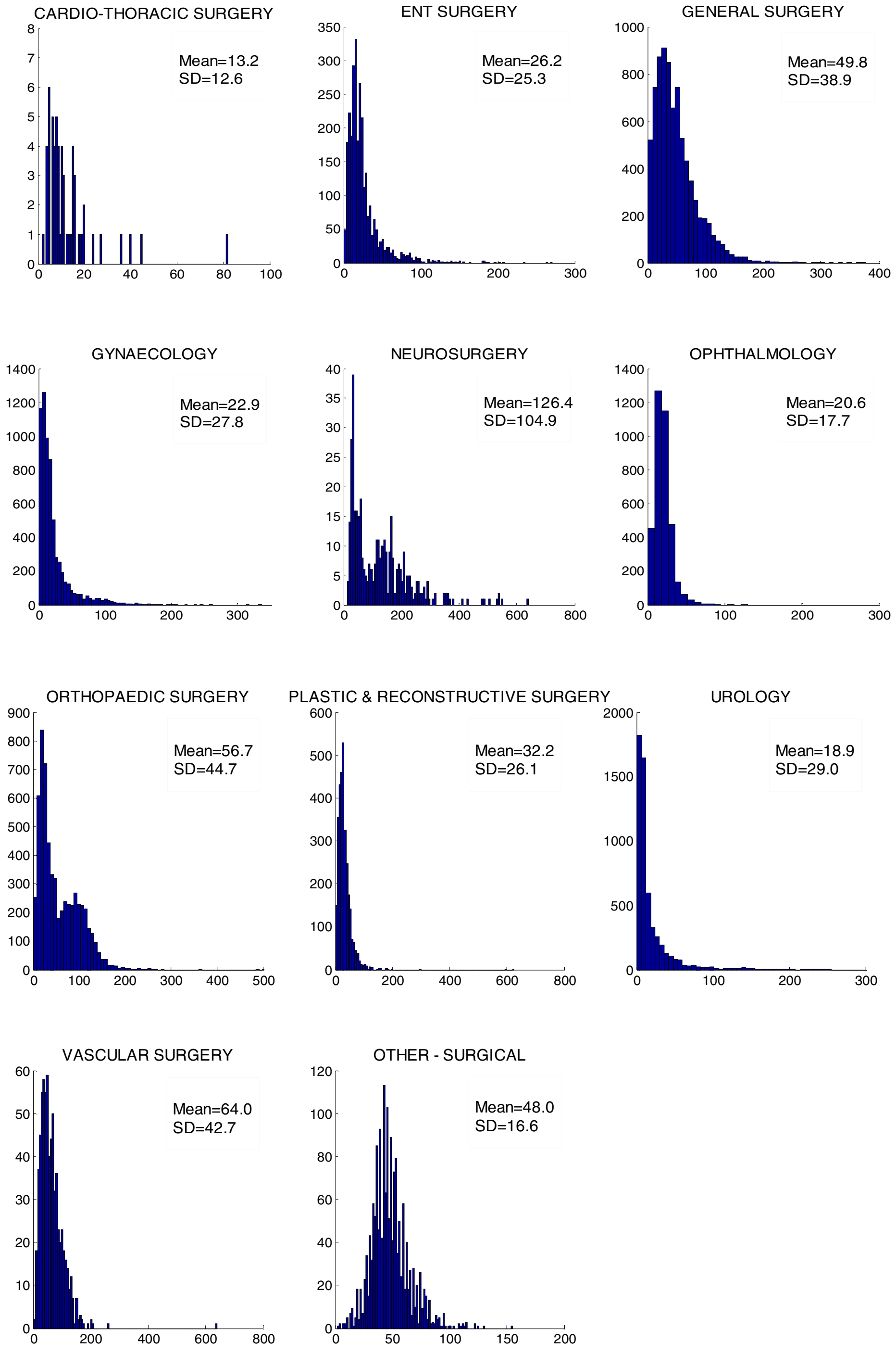

Fig. 1. Distribution of actual procedure times (min) in different specialties

Table 2 presents the performance of the predictive modelling approaches employed. The performance of the linear regression model was poor when compared to the baseline, i.e. the hospital estimate of procedure time. We speculate that the reason for this could be the fact that surgeons estimate the time based on their experience and implicitly consider the interactions between variables whereas these interactions were not taken into account by our linear regression model.

The performance of the MARS model, on the other hand, was much better, reducing RMSE and MAPE by 7.4% and 5.3% respectively when compared to the baseline, and provided a better fit, improving the R^2 value by 14.6% when compared to the baseline. This is likely because MARS is capable of fitting complex, nonlinear relationships between the response variable, procedure time in our case, and its predictors and can search a large number of variables and their interactions in a very efficient way.

The random forest model outperformed both linear regression and MARS models, delivering an improvement of 18.3% to the RMSE and 28.4% to the MAPE values when compared to the baseline. The R2 value increased by 35.4% when compared to the baseline, providing a significantly better fit than other approaches. Various configurations of the random forest were tried and the chosen configuration for our final random forest model was: number of trees=500, number of variables sampled at each split=9 and random sampling with replacement was used. We believe the superior performance of the random forest may come from its ability to boost the prediction performance by combining numerous regression trees in a random process. However, the higher accuracy of the random forest model comes at the expense of interpretability. Black box models such as random forest can't quantify the impact of each predictor to the predictions of the complex model. Therefore, the relationships between predictors and output variable can't be explained easily in the random forest model.

Table 3. Comparison of under/overestimated time of our best model with the current hospital method

	Current Method	Random Forest	Improvement
Underestimated time (min)	67534	47675	29.4%
Overestimated time (min)	61651	50819	17.6%

Further analysis of the random method algorithm investigated the overestimated and underestimated predicted procedure times. Table 3 presents these compared to the baseline while Figure 2 presents the residuals plot of the random forest model and the baseline method. It was observed that the random forest model significantly improved underestimations and overestimations by 29.4% and 17.6% respectively. Current hospital estimates tend to underestimate procedure times more frequently which often result in surgery cancellations in hospitals. The random forest model not only reduced both underestimated and overestimated times but also distributed estimation error more evenly around zero compared to the current hospital method.

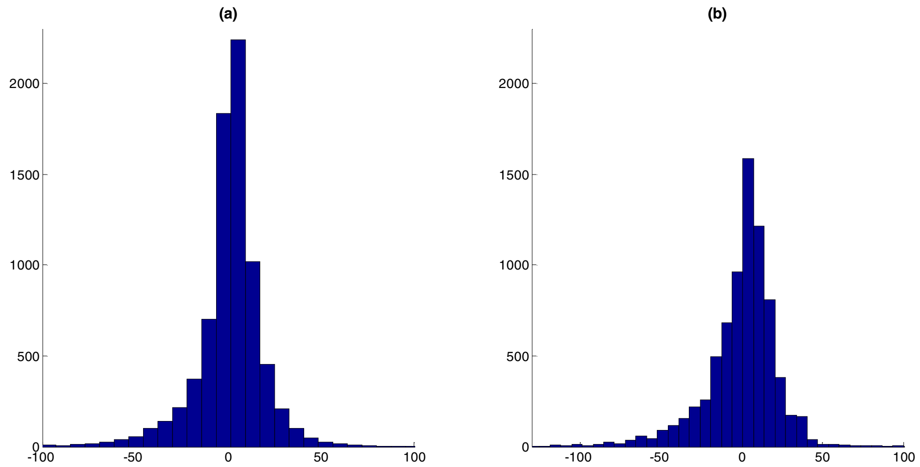

Fig. 2. (a) RF model residuals, (b) Current hospital method residuals

4 Related Work

Estimating surgery time has been the subject of many studies over the last decades and several statistical and machine learning techniques have been applied for predicting surgery duration. Although it has been shown by Strum et al. [23] that a log-normal distribution is well fitted to the total operation times because of its right skewness, the location parameter of log normal distributions vary from one procedure to another and it's not easy to estimate those parameters accurately. Wright et al. [8] developed a regression based model that combined a surgeon's estimate, scheduling software estimation, and several other predictors, and their results showed a 20% reduction in mean absolute error of their model predictions relative to surgeon estimates. However, since their study was performed within a tertiary referral centre the scope and sample size was constrained. Stepaniak et al.[11] built an Analysis of Variance (ANOVA) model of surgery duration, specifically investigating the effect of surgeon factors on surgery duration. They found team composition, experience, and time of the day to be the most significant factors effecting surgery duration and showed 15% improvement in mean absolute error by incorporating surgeon factors. However, the study ignored factors relating to patients. Eijkemans et al. [6] identified a wide range of potential factors related to the surgical team, surgeon, and patient and developed a linear mixed model on the logarithm of total OR time using the surgeon's estimate and other factors. Their results show a significant reduction in average overestimation and underestimation (respectively 12% and 25% improvement). Although they used a large sample of elective operations data over 12 years, they didn't consider the effect of surgery type on their results since all the selected operations for this study were from the department of general surgery. Gomes et al. [16] predicted surgery duration using three machine learning algorithms and compared them with the estimates made by surgeons and median duration of same-type surgeries in a general surgery department. Their results showed 36% improvement in surgery duration estimations compared to the surgeon's estimation. However, the scope of their study was restricted to operations in one department.

In this study we addressed some of the abovementioned shortcomings of current state of the art of surgery duration estimation. We collected four years of administrative and perioperative elective surgery data across all specialties of our case study hospital and the dataset represented a wide range of details including patient characteristics, operation characteristics, and surgery team characteristics. We employed machine learning algorithms that can handle complex datasets with a large number of predictors to build a more general model whose performance is comparable to or better than the current state of the art models. It is envisaged that the model will be delivered to hospital and can be used to improve current hospital practice.

5 Conclusion

This paper focuses on predicting procedure time for elective surgeries. We applied three different machine learning techniques to a large number of predictors including patient, operation and surgery team characteristics to build a procedure duration prediction model. The random forest model was found to outperform other models and delivered 28% improvement when compared to the current hospital method. In future work we will explore how predictions from our proposed model can be used in elective surgery scheduling and how much improvement can be delivered to theatre utilisation through these more accurate time estimations.

References

1. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: A literature review. European Journal of Operational Research 201(3), 921–932 (2010)
2. Macario, A., Vitez, T.S., Dunn, B., McDonald, T.: Where are the costs in perioperative care?: Analysis of hospital costs and charges for inpatient surgical care. Anesthesiology 83(6), 1138–1144 (1995)
3. Pandit, J.J., Carey, A.: Estimating the duration of common elective operations: Implications for operating list management. Anaesthesia 61(8), 768–776 (2006)
4. Schofield, W.N., Rubin, G.L., Piza, M., Lai, Y.Y., Sindhusake, D., Fearnside, M.R., Klineberg, P.L.: Cancellation of operations on the day of intended surgery at a major Australian referral hospital. Med. J. Aust. 182(12), 612–615 (2005)
5. Kayis, E., Wang, H., Patel, M., Gonzalez, T., Jain, S., Ramamurthi, R., Santos, C., Singhal, S., Suermondt, J., Sylvester, K.: Improving Prediction of Surgery Duration using Operational and Temporal Factors. In: AMIA Annu. Symp. Proc., pp. 456–462 (2012)
6. Eijkemans, M.J.C., Van Houdenhoven, M., Nguyen, T., Boersma, E., Steyerberg, E.W., Kazemier, G.: Predicting the unpredictable: A new prediction model for operating room times using individual characteristics and the surgeon's estimate. Anesthesiology 112(1), 41–49 (2010)

7. Dexter, F., Dexter, E.U., Masursky, D., Nussmeier, N.A.: Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. Anesthesia and Analgesia 106(4), 1232–1241 (2008)
8. Wright, I.H., Kooperberg, C., Bonar, B.A., Bashein, G.: Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon-provided estimates. Anesthesiology 85(6), 1235–1245 (1996)
9. Zhou, J., Dexter, F., Macario, A., Lubarsky, D.A.: Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. Journal of Clinical Anesthesia 11(7), 601–605 (1999)
10. Combes, C., Meskens, N., Rivat, C., Vandamme, J.P.: Using a KDD process to forecast the duration of surgery. International Journal of Production Economics 112(1), 279–293 (2008)
11. Stepaniak, P.S., Heij, C., De Vries, G.: Modeling and prediction of surgical procedure times. Statistica Neerlandica 64(1), 1–18 (2010)
12. Li, Y., Zhang, S., Baugh, R.F., Huang, J.Z.: Predicting surgical case durations using ill-conditioned CPT code matrix. IIE Transactions (Institute of Industrial Engineers) 42(2), 121–135 (2010)
13. Dexter, F., Ledolter, J.: Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. Anesthesiology 103(6), 1259–1267 (2005)
14. Dexter, F., Ledolter, J., Tiwari, V., Epstein, R.H.: Value of a scheduled duration quantified in terms of equivalent numbers of historical cases. Anesthesia & Analgesia 117(1), 205–210 (2013)
15. Devi, S.P., Rao, K.S., Sangeetha, S.S.: Prediction of surgery times and scheduling of operation theaters in optholmology department. Journal of Medical Systems 36(2), 415–430 (2012)
16. Gomes, C., Almada-Lobo, B., Borges, J., Soares, C.: Integrating data mining and optimization techniques on surgery scheduling. In: Zhou, S., Zhang, S., Karypis, G. (eds.) ADMA 2012. LNCS, vol. 7713, pp. 589–602. Springer, Heidelberg (2012)
17. Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R.: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J. Chronic Dis. 40(5), 373–383 (1987)
18. Palmer, P.B., O'Connell, D.G.: Regression Analysis For Prediction: Understanding the process. Cardiopulmonary Physical Therapy Journal 20(3), 23 (2009)
19. Heil, D.P., Freedson, P.S., Ahlquist, L.E., Price, J., Rippe, J.M.: Nonexercise regression models to estimate peak oxygen consumption, pp. 599–606. Williams & Wilkins, Baltimore (1995)
20. Dossey, J., Blum, W., Niss, M.: Using Mathematical Competencies to Predict Item Difficulty in PISA: A MEG Study. In: Research on PISA, pp. 23–37. Springer (2013)
21. Hedley, C.B., Yule, I.J.: A method for spatial prediction of daily soil water status for precise irrigation scheduling. Agricultural Water Management 96(12), 1737–1745 (2009)
22. Hastie, T., Tibshirani, R., Friedman, J.H.: The elements of statistical learning: Data mining, inference, and prediction. Springer, New York (2001)
23. Strum, D.P., May, J.H., Vargas, L.G.: Modeling the uncertainty of surgical procedure times: Comparison of log- normal and normal models. Anesthesiology 92(4), 1160–1167 (2000)

24. Friedman, J.H.: Multivariate Adaptive Regression Splines. Annals of Statistics 19(1), 1–141 (1991)
25. Jekabsons, G.: ARESLab: Adaptive Regression Splines toolbox for Matlab/Octave (2011), `http://www.cs.rtu.lv/jekabsons/`
26. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
27. Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
28. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation, DTIC Document (1985)
29. Liaw, A.: Breiman and Cutler's random forests for classification and regression (2012), `http://stat-www.berkeley.edu/users/breiman/RandomForests`