Integrating and Connecting Care
A. Ryan et al. (Eds.)

doi:10.3233/978-1-61499-783-2-133

Improved Prediction of Procedure Duration for Elective Surgery

Zahra SHAHABIKARGAR[a,b], Sankalp KHANNA[a,b], Adbul SATTAR[b], James LIND[c]
[a] *The CSIRO Australian e-Health Research Centre, Brisbane, Australia*
[b] *Institute for Integrated and Intelligent Systems, Griffith University, Australia*
[c] *Gold Coast University Hospital, Queensland Health, Australia*

Abstract. Accurate surgery duration estimation is essential for efficient use of hospital operating theatres and the scheduling of elective patients. This study focuses on analysing the performance of previously developed surgery duration prediction algorithms at a specialty level to gain further insight on their performance. We also evaluate algorithm performance after applying filtering to exclude unreliable data from modelling, and develop and validate new ensemble approaches for prediction. These are shown to significantly improve the prediction accuracy of the algorithms. Employing filtered data delivers a reduction in overall prediction error of 44% (Mean Absolute Percentage Error from 0.68 to 0.38) employing the Random Forests algorithm, while using the newly developed ensemble approach delivers a Mean Absolute Percentage Error of 0.31, a reduction of 55% when compared to the original error, and a reduction of 18% when compared to the Random Forests performance on filtered data.

Keywords. Ensemble methods, surgery duration prediction

Introduction

Improving the accuracy of surgery duration prediction is a necessary step in scheduling elective surgery patients at hospital since the accuracy of surgery schedules depends on precise estimation of surgery duration [1]. Studies have shown that the primary reason for day of surgery cancellations is lack of theatre time due to overrun of other surgeries which results in a large number of scheduled elective procedures being cancelled before surgery [2]. Scheduling "too long" or "too short" durations for surgeries leads to undesirable consequences such as idle time, overtime, or rescheduling of surgeries. Improving the accuracy of estimated procedure time would improve surgery scheduling by providing better arrangement of cases throughout the operating rooms, leading to more efficient use of resources and reduced costs and allowing more surgeries to be done which would increase revenue.

Previous studies implement a wide range of statistical and machine learning techniques for predicting surgery duration [3-6]. However, while these 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 previous work [7], we applied machine learning techniques to perioperative and administrative data from a large tertiary Australian public hospital to improve estimation of procedure duration for Elective Surgery scheduling. The developed prediction models

outperformed existing state of the art models and delivered 28% improvement when compared to the current hospital estimation method. The study however also identified that the accuracy of recorded timestamps was low for surgery cases where more than one procedure was performed. This work extends our previous study by exploring procedure duration data at a specialty level to identify how individual algorithms perform for various specialties. We also develop models that exclude the surgery cases meeting the low fidelity criteria identified above (i.e. multiple procedures per surgery). The predictive accuracy of these models is evaluated at a hospital and individual specialty level, and compared to previously developed models, and current hospital estimation methods.

1. Methods

Data for this study was sourced from two major hospital information systems, the inpatient administrative Hospital Based Corporate Information System (HBCIS), and the perioperative Operating Room Management Information System (ORMIS). The dataset represented a wide range of details about patients, bookings, operations, specialties, and surgery teams, for 60362 individual procedures performed between 01/07/2008 to 30/06/2012 (4 years) at the Gold Coast Hospital, a large metropolitan public hospital in south-east Queensland. This represented 104 different type of procedures across 11 surgical specialties. Ethics approval for the study was obtained from the Gold Coast Hospital and Health Service Human Research Ethics Committee.

For the first stage of this study, we evaluated the predictive accuracy of algorithms at an individual specialty level. We removed emergency surgical cases since our goal is to estimate procedure time for planned, i.e. elective surgeries. Also, surgical records with missing values, inconsistent values and duplicate data were removed. In addition, those procedures that were performed less than a hundred times during the period of this study, cases with no match between databases, and procedures that were not assigned to a surgical specialty were excluded. Potential predictors were chosen after an exhaustive review of literature and available data sources, and discussions with clinical experts and hospital administrators [7]. Potential predictors chosen can be categorised in three groups: patient characteristics, operation characteristics, and surgery team characteristics. Patient characteristics included patient age (years), gender, urgency category, type of admission, patient payment class, referral centre, and Charlson Comorbidity Index (CCI). All the predictors that related to hospital and operation including hospital unit, specialty, ward, theatre, and session are categorised as operation characteristics. Finally, all predictors associated with people who were involved in the surgery, such as number of surgeons, anaesthetists, their professional category and specialty are in surgery team category. In keeping with previous work, we developed Generalised Linear Model (GLM), Multivariate Adaptive Regression Splines (MARS) and Random Forests algorithms using the Statistics toolbox, ARESLab toolbox, and Jaiantilal Random Forests package in MATLAB respectively.

One of the findings in our previous work [7], and in our first stage models in this study, was that for the operations in the dataset that had more than one procedure performed (about 20% of all procedures), there was an overlap between procedure start and finish time. Clinical consultation revealed the lack of precise timestamping was attributed to operational complexity. To analyse model performance where precise timestamping was available, the second stage of this study employed filtering of patient records to exclude operations where more than one procedure had been performed. In

addition to employing this dataset to build models using algorithms employed in the first stage of the study, we extended our analysis to include and evaluate ensemble methods, applying three popular ensemble algorithms for regression models, namely M5 method, LS Boost, and Bagging algorithms. To build the M5 model we used M5PrimeLab in Matlab. The M5PrimeLab is a Matlab/Octave toolbox for building regression trees and model trees using M5 method and the built trees can also be linearised into decision rules either directly or using the M5 Rules method. We implemented LSBoost and Bagging models in Matlab Statistics and Machine Learning Toolbox which provides functions and apps to describe, analyse, and model data using statistics and machine learning.

Ten-fold cross validation was employed to evaluate the performance of our predictive models. To assess the prediction performance of these models, we used Mean Absolute Percentage Error (MAPE) as the statistic of choice. At the first stage, a detailed analysis of the surgery duration prediction was done in which the error analysis was broken down by specialties to explore whether for some specialties the surgery duration is more predictable than others. The other hypothesis tested was whether some prediction methods perform better than others for particular specialties. At the second stage, the performance measurements were used for comparison of different predictive models at the overall and specialty level on initial data and after filtering out operations with more than one procedure.

2. Results

Table 1 presents the prediction accuracy of our initial prediction models, which looked at all (i.e. unfiltered) surgery episodes broken down by specialties. It was observed that, for some surgical specialties e.g. Gynaecology and Cardio-Thoracic surgery, the performance is below the previously reported overall performance of these models [7], with MAPE of GLM model for Cardio-Thoracic surgery being almost double the overall MAPE. For some others, especially Neurosurgery and Vascular surgery, the prediction accuracy was significantly improved compared to the overall performance across all three prediction models evaluated.

Table 1. Performance of Prediction Models on Initial Data (unfiltered) – Overall, and by Specialty.

SPECIALTY	MAPE across individual algorithms		
	GLM	MARS	RF
OVERALL *	1.20	0.90	0.68
CARDIO-THORACIC	2.02	1.18	0.83
ENT	1.40	0.50	0.56
GENERAL	0.88	0.96	0.65
GYNAECOLOGY	2.45	1.31	0.93
NEUROSURGERY	0.37	0.50	0.43
OPHTHALMOLOGY	1.06	0.74	0.34
ORTHOPAEDIC	0.65	0.56	0.57
PLASTIC	0.57	0.86	0.73
UROLOGY	1.90	1.11	0.75
VASCULAR	0.53	0.39	0.39
OTHER SURGICAL	0.32	0.34	0.31

* Results from previous study [7]

Analysing the effect of multiple procedures in one surgery and overlap between procedures start and finish time revealed that inaccurate time records significantly affected the performance of prediction models. Table 2 presents the performance of Random Forests model when applied to initial data, before removing multiple procedures, and filtered data which includes surgeries with only one procedure. As shown in the table, the analysis was broken down to specialty level where we can see and compare the performance of prediction model for individual specialties as well as the overall performance across all specialties. The prediction accuracy of Random Forests model significantly increased with filtered data, reducing MAPE from 0.68 to 0.38 for overall episodes. For some specialties like Gynaecology the prediction error reduced even more sharply, from 0.93 to 0.33.

Table 2. Random Forests Model Performance on Initial and Filtered Data – Overall, and by Specialty.

SPECIALTY	MAPE-Initial Data	MAPE-Filtered Data	% Improvement
OVERALL	0.68	0.38	44%
CARDIO-THORACIC	0.83	0.63	24%
ENT	0.56	0.38	32%
GENERAL	0.65	0.31	52%
GYNAECOLOGY	0.93	0.33	65%
NEUROSURGERY	0.43	0.26	40%
OPHTHALMOLOGY	0.34	0.30	12%
ORTHOPAEDIC	0.57	0.31	46%
PLASTIC	0.73	0.37	49%
UROLOGY	0.75	0.47	37%
VASCULAR	0.39	0.28	28%
OTHER – SURGICAL	0.31	0.23	26%

Table 3. Ensemble Algorithm Performance on Filtered Data – Overall, and by Specialty.

SPECIALTY	MAPE - M5	MAPE - LSBoost	MAPE - Bagging
OVERALL	0.38	0.31	0.31
CARDIO-THORACIC	0.56	0.36	0.47
ENT	0.35	0.27	0.27
GENERAL	0.35	0.25	0.27
GYNAECOLOGY	0.41	0.26	0.26
NEUROSURGERY	0.43	0.28	0.22
OPHTHALMOLOGY	0.34	0.28	0.28
ORTHOPAEDIC	0.39	0.27	0.28
PLASTIC	0.39	0.36	0.32
UROLOGY	0.41	0.38	0.34
VASCULAR	0.39	0.26	0.25
OTHER – SURGICAL	0.26	0.25	0.23

Applying ensemble methods for building surgery duration prediction models also revealed significant improvements on the prediction accuracy. Table 3 shows the comparison of three ensemble algorithms, namely M5 rules, LSBoost, and Bagging Tree, for surgery duration prediction on overall and specialty level filtered data. As shown in the table, ensemble methods performed better, when compared to the previous models, for all episodes together, and for almost all specialties, with Bagging and LSBoost model reducing overall MAPE from 0.38 to 0.31 when compared to the performance of Random Forests on this data.

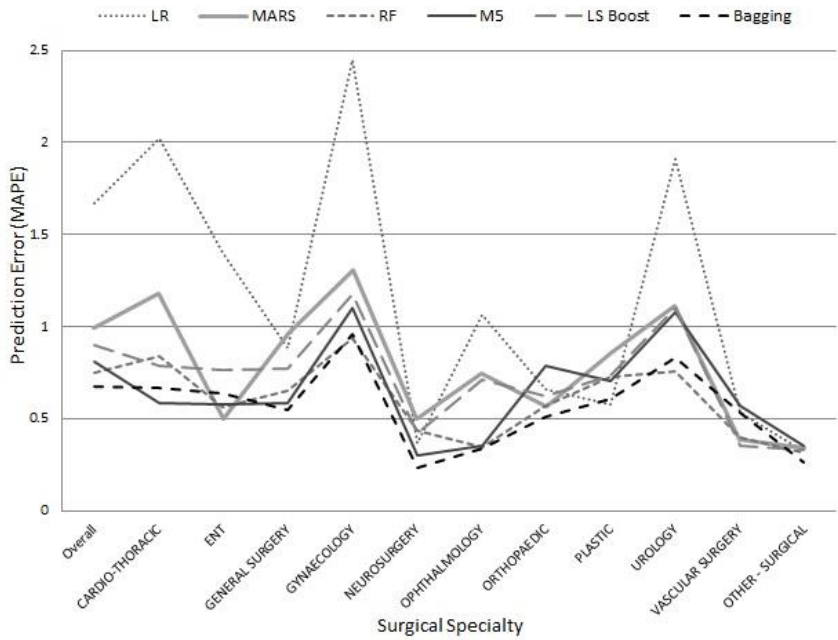

Figure 1. Performance Comparison of Different Prediction Models on initial Data by Specialty.

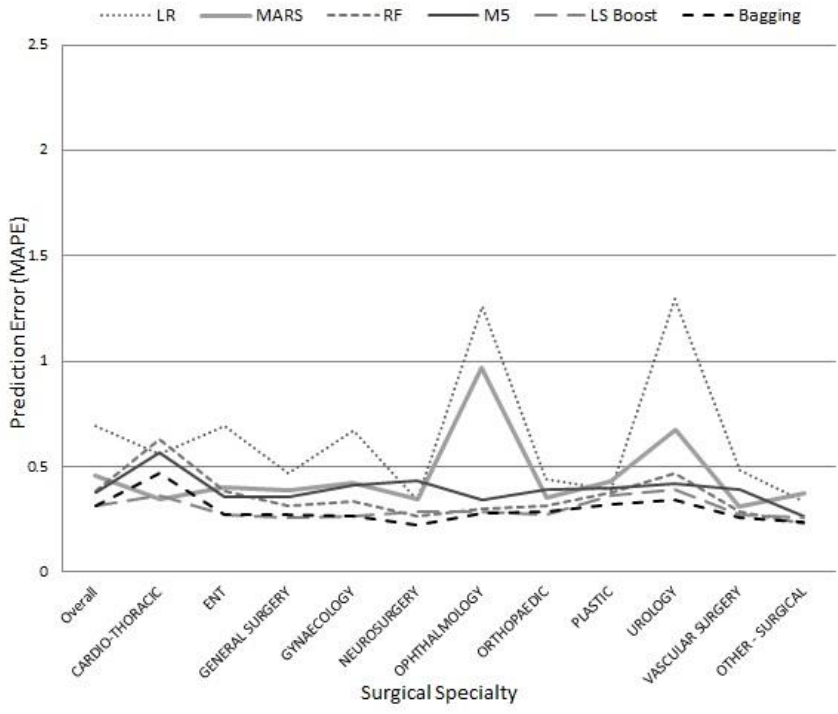

Figure 2. Performance Comparison of Different Prediction Models on Filtered Data by Specialty.

Figures 1 and 2 illustrate the comparison between the performance of all prediction models on overall and specialty level episodes, for the initial data and filtered data respectively. As shown in Figure 1 for initial data, the Bagging model has the best performance overall, and also for the majority of specialties. However, as seen with Figure 2, LSBoost and Bagging ensemble methods perform better for filtered data, returning a very good performance, overall and for individual specialties.

3. Discussion

In extending previously published work, we have investigated the performance of developed machine learning algorithms at a specialty level to evaluate whether these are better suited to certain specialties. The performance of prediction models varies significantly across different specialties. While the prediction error is higher than the overall performance of the model (from MAPE of 0.7 to 0.9 for Random Forests models) for some specialties, we identified others like Ophthalmology, Neurosurgery and Vascular that benefit from the application of these algorithms, given that the prediction error is significantly lower when compared to the overall performance.

We have also investigated the effect of excluding surgery episodes with multiple procedures, which make up 1 in 5 surgical episodes and are known to have poor quality timestamps, and the effect of ensemble classifiers. When applied to filtered data, the Random Forests algorithm delivers an overall performance improvement of 44% (12%-65% across specialties), reducing MAPE from 0.68 (when tested on unfiltered data) to 0.38. Among the newly developed ensemble approaches, Bagging and LSBoost further reduce the surgery duration prediction error significantly and deliver an overall MAPE of 0.31, an improvement of 18% over using Random Forests.

The current study is limited in that it explores surgical from a single hospital and the findings may not necessarily translate across other hospitals of varying size, serving varying populations. Proposed extensions of this work include exploring the performance of these algorithms across multiple hospitals, and developing algorithms that can work better for surgeries with multiple procedures.

References

[1] 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, 2010. **112**(1): p. 41-49.

[2] Pandit, J.J. and A. Carey, *Estimating the duration of common elective operations: Implications for operating list management.* Anaesthesia, 2006. **61**(8): p. 768-776.

[3] 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, 2012. **36**(2): p. 415-430.

[4] Dexter, F., et al., *Estimating surgical case durations and making comparisons among facilities: Identifying facilities with lower anesthesia professional fees.* Anesthesia and Analgesia, 2013. **116**(5): p. 1103-1115.

[5] 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), 2010. **42**(2): p. 121-135.

[6] Stepaniak, P.S., Heij, C., De Vries, G., *Modeling and prediction of surgical procedure times.* Statistica Neerlandica, 2010. **64**(1): p. 1-18.

[7] ShahabiKargar, Z., Khanna, S., Good, N., Sattar, A., Lind, J., O'Dwyer, J., *Predicting Procedure Duration to Improve Scheduling of Elective Surgery.* Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2014), 2014. p. 998-1009. Springer.