Journal of Medical Systems (2020) 44: 42
https://doi.org/10.1007/s10916-019-1501-4

SYSTEMS-LEVEL QUALITY IMPROVEMENT

An Evaluation of the Hybrid Model for Predicting Surgery Duration

K. W. Soh[1] · C. Walker[1] · M. O'Sullivan[1] · J. Wallace[2]

Received: 2 July 2019 / Accepted: 14 November 2019 / Published online: 2 January 2020

Abstract

The degree of accuracy in surgery duration estimation directly impacts on the quality of planned surgical lists. Model selection for the prediction of surgery duration requires technical expertise and significant time and effort. The result is often a collection of viable models, the performance of which varies across different strata of the surgical population. This paper proposes a prediction framework to be used after a comprehensive model selection process has been completed for surgery duration prediction. The framework produces a partition of the surgical cases and a "hybrid model" that allocates different predictors from the collection of viable models to different parts of the surgical population. The intention is a flexible prediction process that can reassign models and adapt as surgical processes change. The framework is tested via a simulation study, and its utility is demonstrated by predicting surgery durations for Ear, Nose and Throat surgeries in a New Zealand hospital. The results indicate that the hybrid model is effective, performing better than standard model selection in two of the three simulation studies, and marginally worse when the selected model was the true underlying process.

Keywords Linear regression · Prediction · Hybrid model · Cross-validation · Simulation

Introduction

Accurate estimates of surgery duration are essential in the management of surgical theatres, which are some of the most expensive resources in a hospital. Streamlined surgical schedules are attainable when scheduling decisions are embedded within a surgery booking process [1]. Computer simulations performed by [2] demonstrate that scheduling decisions supported by improved predictions of surgery duration have the real-world potential to increase the productivity of surgical theatres. This in turn realises the benefits to stakeholders that include cost savings, increased patient throughput and reduced waiting times.

Numerous statistical and machine learning tools have been implemented in the literature to predict surgery durations. Examples of the approaches used include Bayesian methods [3], neural networks [2], random forests [4] and regression techniques [2, 4–6]. The scope of application for these tools differs considerably by the types of surgeries (elective versus emergency), the number of surgical specialities included and the number of hospitals involved. Briefly, [2] considered only three types of elective surgeries in the ophthalmology department of a hospital, namely, cataract surgery, corneal transplant surgery and oculoplastic surgery. Case studies by [4, 5, 7] examine all elective surgeries that took place at a single hospital, while [6] considered all surgeries, including emergencies, that occurred across six hospitals. The various techniques used for the prediction of surgery durations in these studies present competing approaches that each can perform well. However, the technical expertise required to monitor their performance as the underlying surgery processes evolve is a significant drawback. Although the initial selection process to determine and evaluate the numerous candidate prediction models is unavoidable, we propose an automatic prediction framework to better employ these candidate models as patient characteristics and surgery processes change. This framework enables different subsets of the

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

K. W. Soh
skia593@aucklanduni.ac.nz

[1] Department of Engineering Science, University of Auckland, Auckland, New Zealand

[2] North Shore Hospital, Auckland, New Zealand

surgeries needing to be scheduled to have their durations predicted by different models, making use of all available data for training of all candidate predictors.

This approach is motivated by consideration of the hospital's surgical process, where new decisions implemented at the policy level, changes in staffing and the procurement of new medical equipment are continual, and may affect the actual surgery durations. As the use of a single model for prediction may not be resilient to these changes, the combination of several models will extend the utility of the prediction model to a wider range of scenarios.

In the remainder of this paper, we aim to address the above shortcoming by proposing a prediction framework that constructs the hybrid model. This framework is elaborated upon in Section "Prediction framework and the hybrid model", and the resulting hybrid model allows the use of several prediction models simultaneously. The performance of the hybrid model is further examined with the aid of computer simulations in Section "Computer simulations for investigating hybrid model performance". Then in Section "Case study: An actual surgery dataset", the prediction framework is applied in a clinical setting to predict the duration of elective surgeries. In general, the hybrid model allows any type of prediction model to be used as candidate models. However, for the purpose of illustration, only linear regression models shall be used in the computer simulations and the case study. Finally, Section "Conclusion" concludes with our views on the utility of the hybrid model.

Prediction framework and the hybrid model

This section presents a prediction framework that utilises models from one or more prediction tools, including some of the approaches mentioned in the introduction section (Section "Introduction"). The framework leads to a hybrid model that can be used for prediction. In essence, given a set of viable models for predicting surgery duration, the framework partitions the set of historical surgeries and then allocates to each part one of the viable models trained on all the data, to be used for predicting the duration of future surgeries from that part. The framework is intended to be run repeatedly at the same regularity as surgical lists are planned. With each booking cycle, the most recently completed surgeries are added to the set of historical surgeries, the parameters of the candidate prediction models are updated, the partition of the historical surgeries constructed, and the viable models allocated to the parts of the data. This section describes the framework and the resulting hybrid model.

Consider a dataset which is split into two sets: (1) the training set which is to be used to develop the prediction framework, and (2) the test set which is used to evaluate the performance of the resulting hybrid model. Suppose that there are N observations in the training data and a predetermined set $F = \{f_1, f_2, \ldots, f_M\}$ that comprises M distinct prediction models, where $M \geq 2$. The models within F may use different explanatory variables. These models may be chosen based on expert opinion and/or related literature. Let $\mathbb{X}$ and $\mathbb{W}$, respectively, be the set of all categorical and continuous explanatory variables used by at least one prediction model in F. While $\mathbb{W}$ may be an empty set, assume that $\mathbb{X}$ contains V categorical explanatory variables, where $V > 0$. Without loss of generality, let $\mathbb{X} = \{X_1, X_2, \ldots, X_V\}$. Under this notation, each observation $n \in \{1, 2, \ldots, N\}$ comprises a V-tuple $(x_{1n}, x_{2n}, \ldots, x_{Vn}) \in X_1 \times X_2 \times \ldots \times X_V$, i.e., the V-dimensional space denoting the observed values of all the categorical variables.

The first step of the prediction framework involves partitioning the observations into P subsets by the values of all the categorical explanatory variables used in the set of prediction models F. Each subset that corresponds to a unique V-tuple is called a *bucket*, and the number of observations contained in the p^{th} bucket is K_p^F. Here, the superscript and subscript of K specify the model(s) involved in the partitioning of the training data and a particular subset of the partition respectively, so $\sum_{p=1}^{P} K_p^F = N$. In particular, if the indices of the observations in the p^{th} bucket are denoted by $i_{p,1}, i_{p,2}, \ldots, i_{p,K_p^F}$, then $\bigcup_{p=1}^{P} \{i_{p,1}, i_{p,2}, \ldots, i_{p,K_p^F}\} = \{1, 2, \ldots, N\}$. Note that the buckets $\{1, 2, \ldots, P\}$ can also be defined by subsets resulting from the partition of the observations by individual models. These subsets are referred to as the *model subsets*. For a given model $f_m \in F$, the observations are partitioned into model subsets by all the categorical variables that the model uses and the values of each observation for those variables. For example, if the model f_m uses only a categorical variable X_1, then the observations will be partitioned by the values taken by X_1 (within the observations) with one model subset for each value taken. In particular, if there are four observations with values for X_1 of $x_{11} = 1, x_{12} = 2, x_{13} = 5$ and $x_{14} = 2$, then these observations will be partitioned into 3 model subsets, namely, $\{1\}, \{2, 4\}$ and $\{3\}$ for the observed values 1, 2, and 5 (of X_1) respectively. Now, a similar notation can be used to describe the model subsets of f_m such that there are $K_q^{(m)}$ observations in each model subset $q = 1, 2, \ldots, Q^{(m)}$ and the observations' indices are $j_{q,1}^{(m)}, \ldots, j_{q,K_q^{(m)}}^{(m)}$. The superscript of j may be dropped if it is clear which model is being referred to. Finally, the buckets can be determined from the model subsets as follows. An observation $n \in \{1, 2, \dots, N\}$ is in the bucket $p \in \{1, 2, \dots, P\}$ if

$$n = i_{p,k} \text{ for some } k \in \{1, 2, \dots, K_p^F\}.$$

Denote this bucket by $p(n)$. For a given model f_m, an observation $n \in \{1, 2, \dots, N\}$ is in the model subset $q \in \{1, \dots, Q^{(m)}\}$ if

$$n = j_{q,k} \text{ for some } k \in \{1, 2, \dots, K_q^{(m)}\}.$$

Denote this model subset by $q^{(m)}(n)$. Then, the bucket containing observation n is the intersection of all the model subsets containing the same observation, i.e.,

$$p(n) = \bigcap_{m=1}^{M} q^{(m)}(n).$$

It follows that $K_{p(n)}^F \leq K_{q^{(m)}(n)}^{(m)}$ for each $n \in \{1, 2, \dots, N\}$. Observe that

$$q^{(m)}(i_{p,1}) = q^{(m)}(i_{p,2}) = \dots = q^{(m)}(i_{p,K_p^F}),$$

Hence, we can define the model subset $p^{(m)} = q^{(m)}(i_{p,1})$ for a given bucket $p \in \{1, 2, \dots, P\}$. It follows that $K_p^F \leq K_{p^{(m)}}^{(m)}$. For simplicity, $K_{p^{(m)}}^{(m)}$ may be written as $K_p^{(m)}$. This value will be used in the second step of the framework.

The second step of the framework allocates a prediction model to each bucket (of the training set). This allocation is determined by computing the leave-one-out cross-validation (LOOCV) residuals as follows. Consider the bucket p and the set of prediction models F. For the k^{th} observation in bucket p, or equivalently $i_{p,k}$, denote the value of the response variable for this observation as $y_{i_{p,k}}$. For the m^{th} model fitted on the training dataset with observation $i_{p,k}$ deleted, the predicted value of the response variable for observation $i_{p,k}$ is calculated. This value is denoted by $\hat{y}_{[i_{p,k}]}^{(m)}$. The sum of squared residuals for the m^{th} model across bucket p is then given by $r_{pm}^2 = \sum_{k=1}^{K_p^F} (y_{i_{p,k}} - \hat{y}_{[i_{p,k}]}^{(m)})^2$. These residuals are used to select the best prediction model for each bucket p after excluding the overfitted models.

In order to safeguard against overfitting, the number of observations related to each bucket should be sufficiently large relative to the number of explanatory variables included in the m^{th} model. A threshold can be employed to reduce the likelihood of overfitted models being assigned to a bucket. The use of such models for prediction tends to cause large out-of-sample prediction errors. For linear regression analyses, a measure that is used as a threshold is the minimum number of subjects per variable (SPV), which is defined as the ratio of the sample size to the number of explanatory variables used to build the linear regression model. Currently, there is no consensus in the literature on the recommended SPV. [8] proposes a minimum of 5 SPV for stable regression models, though 20 SPV is preferable. [9] suggests a minimum of 10 SPV for prediction models. [10] recommends a range of 15 to 20 SPV depending on the context for fitting a regression model. [11] suggest that “a minimum of only two SPV is required for adequate estimation of regression coefficients”, although their experiments are less extensive. Similar to the SPV criterion, define the number of subjects related to the bucket per variable (SBPV) as the ratio of the value $K_p^{(m)}$ to the number of explanatory variables used in the m^{th} prediction model. If the SBPV-based threshold is not met, then the model will not be chosen to represent observations in the p^{th} bucket regardless of the value r_{pm}^2. The threshold is implemented using the number of explanatory variables u_m in $\mathbb{X} \cup \mathbb{W}$ included in f_m, and the value s chosen for the threshold.

The collection and selection of prediction models may now be represented by the following integer programming formulation:

$$\min \sum_{m=1}^{M} \sum_{p=1}^{P} z_{pm} r_{pm}^2$$

$$\sum_{m=1}^{M} z_{pm} = 1 \qquad \forall p \qquad (1)$$

$$z_{pm} \left(K_p^{(m)} - s u_m \right) \geq 0 \qquad \forall p, m \qquad (2)$$

$$z_{pm} \in \{0, 1\} \qquad \forall p, m.$$

The binary variable z_{pm} indicates if the m^{th} model is selected by the p^{th} bucket. Note that constraint (1) restricts the number of models selected by each bucket to one, while constraint (2) imposes the threshold that will not allow the m^{th} model to be chosen by the p^{th} bucket if $K_p^{(m)} \leq s u_m$. We shall see that constraint (2) may not be required in the subsequent computer simulations. Note that the solution to the programming problem can be systematically obtained without requiring any optimisation technique to solve it because it decomposes into separate, simple formulations for each bucket. The model selection procedure is: for each bucket, identify all the model(s) that do not violate the threshold. After that, choose the model that corresponds to the smallest r_{pm}^2 for the bucket; set z_{pm} to 1 for that particular value of m, and 0 otherwise.

It should be noted that the objective function in the integer program can also be written as $\sum_p \sum_k \left(y_{i_{p,k}} - \sum_m z_{pm} \hat{y}_{[i_{p,k}]}^{(m)} \right)^2$, an expression that resembles the jackknife model average [12]. The jackknife model averaging is one of the many methods classified as "stacked generalisation" or stacking, an idea – popularised by [13] – to describe methods that combine different model estimates to improve predictions. For the hybrid model, it is emphasised that the corresponding selected models could be different for each of the buckets because there is a binary variable z_{pm} for each "(bucket, model)" combination. The hybrid model should not be classified as stacking because there is no interaction between the predicted variables from the models in F to make predictions for each bucket.

The last step of the prediction framework is to remove all unused prediction models f_m from F and any explanatory variables not used by at least one prediction models remaining in F. If at least one categorical explanatory variable is removed, then the steps of the prediction framework are repeated to update the buckets and the subsequent assignment of a prediction model to each bucket. This is an important step as categorical explanatory variable(s) that are not used for prediction should not affect the data segmentation into buckets. If no categorical explanatory variable is removed, then the set $\mathbb{X}$ is unchanged and the resulting collection of prediction models F is the hybrid model. In Fig. 1, an example using $F = \{f_1, f_2, f_3, f_4\}$ and $\mathbb{X} = \{X_1, X_2\}$ is demonstrated, where X_1 and X_2 are binary explanatory variables.

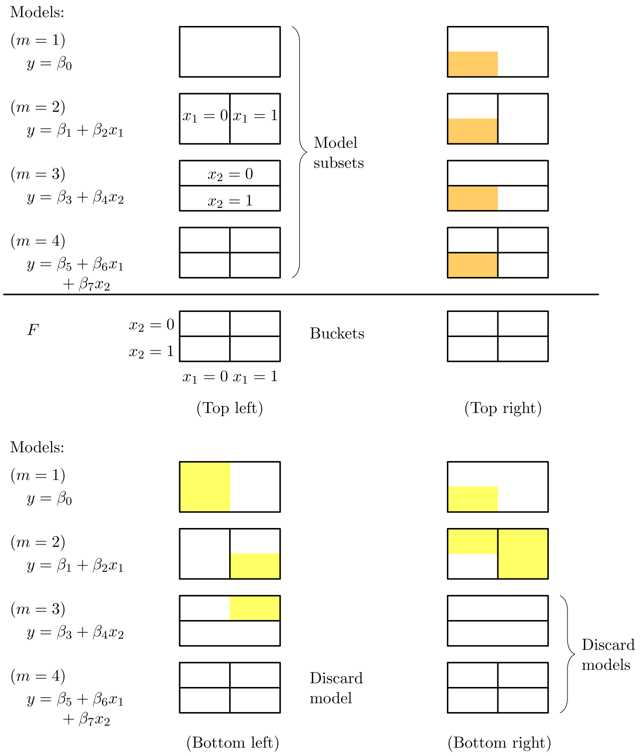

Fig. 1 An illustration of the prediction framework with the assumptions that the training dataset contains all combinations of values for the binary explanatory variables X_1 and X_2, and $K_p^{(m)}$ is sufficiently large for all p, m. In (top left), the model subsets and buckets are shown. In (top right), the highlighted regions show the possible models that could be selected by the bucket taking values $(x_1, x_2) = (0, 1)$. The remaining figures are two possible outcomes of the selection of prediction models by the buckets. The selected model for each bucket is highlighted. The hybrid model (bottom left) proceeds to the testing phase. However, the hybrid model (bottom right) requires the buckets to be refined as X_2 is not used

Figure 2 shows a flow diagram that summarises the steps taken to obtain a hybrid model from the prediction framework. Even though the hybrid framework can adapt to new data such as (1) a new surgeon performing a routine surgery; or (2) an experienced surgeon performing a new procedure, it is dependent on the choice of prediction model set F. If the new data cannot be effectively modelled with F, then new prediction models may need to be included, i.e., F may need to be extended, or existing prediction models may cause overfitting and need to be excluded, i.e., F may need to be contracted. A periodic review of the set of prediction models F is therefore recommended.

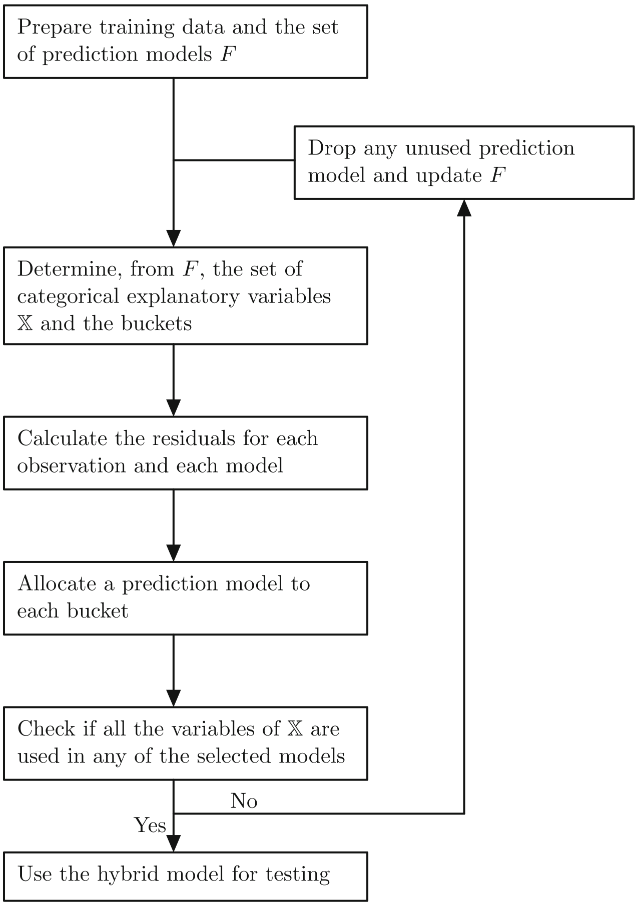
Prepare training data and the set of prediction models F
Drop any unused prediction model and update F
Determine, from F, the set of categorical explanatory variables $\mathbb{X}$ and the buckets
Calculate the residuals for each observation and each model
Allocate a prediction model to each bucket
Check if all the variables of $\mathbb{X}$ are used in any of the selected models
No
Yes
Use the hybrid model for testing

Fig. 2 A flow diagram of the prediction framework to create a hybrid model for prediction

In the subsequent sections, the prediction framework shall be applied on several synthetic datasets during computer simulations and on a surgery dataset in the case study. The finalised hybrid model will be used on the test set to compute the out-of-sample prediction errors.

Computer simulations for investigating hybrid model performance

The hybrid model enables the integration of several different prediction tools. For the purpose of illustration, only linear regression models are considered here. In what follows in Section "Linear regression modelling", we introduce linear regression analysis and explain our treatment with regards to the modelling assumptions and the computation of residuals for each bucket of the hybrid model. Then in Sections "Investigating the performance of the hybrid model" and "Investigating the SBPV-based threshold of the hybrid model", the utility of the hybrid model is illustrated with the aid of computer simulations and linear regression models.

Linear regression modelling

Regression analysis is a predictive modelling technique which estimates the relationship between two or more variables. This modelling technique requires the following key assumptions about the data:

1. the true relationship between the response variable and the explanatory variable(s) is linear in its parameters,
2. the residuals are normally distributed,
3. no two or more explanatory variables are highly linearly related (no multicollinearity),
4. there is no auto-correlation and
5. the residuals are evenly scattered across all values of the explanatory variables (homoscedasticity).

Since the intent for applying linear regression is for prediction, only the predictive ability of the regression models is of interest. The normality, the multicollinearity, the auto-correlation and the homoscedasticity assumptions will not be checked. In other words, some or all of these assumptions are allowed to be violated.

In Section "Prediction framework and the hybrid model", the use of the LOOCV criterion for computing the residuals r^2_{pm} is mentioned. For linear regression models, there is a computationally less intensive method to calculate the residuals. This method does not require models to be fitted repeatedly (with an observation deleted each time) to predict the response variable for the n^{th} observation, $\hat{y}^{(m)}_{[n]}$. Suppose that the linear regression model m that is fitted on the training set of N observations is expressed as $Y = X\beta + \epsilon$. Let $H = X(X^T X)^{-1}X^T$. This matrix H is commonly referred to as the "hat-matrix", with the diagonal elements denoted by $h_{11}, h_{22}, \ldots, h_{NN}$. The sum of squared residuals can thus be expressed as

$$r^2_{pm} = \sum_{n=1}^{N} \delta_n \left(\frac{y_n - \hat{y}^{(m)}_n}{1 - h_{nn}} \right)^2 ,$$

where $\hat{y}^{(m)}_n$ is the value obtained by using the linear regression model m to predict the response variable of the n^{th} observation, y_n, and

$$\delta_n = \begin{cases} 1 & \text{if observation } n \text{ is assigned to bucket } p, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the new expression for r^2_{pm} is derived from the predicted residual error sum of squares (PRESS) statistic, which is proposed in [14]. For linear regression models, a proof demonstrating the equivalence of the PRESS statistic and the LOOCV statistic is given in [15]. Therefore, the PRESS statistic shall be used instead of the LOOCV statistic to calculate the residuals in the second step of the prediction framework.

Investigating the performance of the hybrid model

In this subsection, computer simulations are used to study the performance of the hybrid model in a controlled setting with synthetic datasets. These synthetic datasets are constructed with simplicity kept in mind for the purpose of analysis, but resemble clinical datasets which are usually generated from several underlying complex processes.

Three synthetic datasets will be simulated with the use of surgical procedure, surgeon and (for the first two datasets) physical status rating of a patient to determine the simulated surgery durations. The construction of these synthetic datasets is motivated from the Ear, Nose and Throat (ENT) elective surgeries performed at two hospitals in New Zealand; the actual surgery dataset is investigated later in Section "Case study: An actual surgery dataset". As the true relationship describing a surgery dataset is often unknown or takes the form of a complicated expression, the proposed relationship for each of the synthetic datasets in the simulations reflects different levels of simplification. In the first and third synthetic datasets, a multiplier effect related to surgeons on the surgery duration is introduced. Such an effect may be observed among experienced surgeons, for example, who always have complicated cases assigned to them. These surgeons may require more time to complete each stage of a surgery, and therefore a multiplier effect may be more appropriate than an additive effect. The first dataset also includes effect of the physical status ratings on the surgery duration, but this is removed in the third dataset. The justification for removing this effect is based on the prediction method adopted by the two hospitals in New Zealand, which does not consider these ratings in their prediction of surgery durations. Finally, the second synthetic dataset is deliberately oversimplified such that all effects of the explanatory variables on the surgery duration are additive, and the data is best fitted by the hyperplane of the form $y = \beta_0 + \beta_1 p + \beta_2 s + \beta_3 r$. This simplification is made to observe how well the hybrid model works when predicting purely additive effects (versus a situation in which multiplicative effects exist).

The first dataset demonstrates the utility of the hybrid model and the remaining two datasets further examine the model performance under different scenarios as described in the previous paragraph. Following the procedure described at the beginning of Section "Prediction framework and the hybrid model", each dataset is partitioned into a training set and a test set. The measure of performance is the root-mean-squared error of prediction (RMSEP), which is computed using the out-of-sample prediction errors from the test set. In order to assess the utility of the hybrid model on each dataset, we will compare the RMSEP values computed on the test set for the following models: (1) a trained hybrid model and (2) a linear regression model that only contains the main effects and that has been selected using an exhaustive model selection procedure on the training set. Model (2) is denoted by the *baseline* model (referred to simply as *baseline* for brevity) which fits a single linear regression model to the dataset. The choice of a single linear regression model is appropriate for benchmarking the hybrid model, as it has been widely accepted for use in many real-world applications such as in the context of water resources management [16] and education [17].

It is hypothesised that the different levels of simplification will change the relative performance of the hybrid model with respect to the baseline. In each of the three synthetic datasets, 6 surgical procedures, 5 surgeons and 3 ratings that indicate the physical status of a patient are considered. A surgery may be labelled by a 3-tuple (p, s, r), where $p \in P = \{1, 2, \ldots, 6\}$ denotes the procedure, $s \in S = \{1, 2, \ldots, 5\}$ denotes the surgeon and $r \in R = \{1, 2, 3\}$ denotes the rating. Assume that each of the 90 possible combinations of a 3-tuple is equally likely to occur. The response variable is the duration of a surgery in minutes which is denoted by y. In what follows, the truncated normal distribution that is derived from the normal distribution with mean μ and standard deviation σ is used. For the synthetic datasets, this normal distribution is truncated by imposing the requirement that each realisation from the truncated normal distribution cannot deviate from its mean by more than 2.5 times its standard deviation. The resulting truncated normal distribution shall be denoted by $\mathrm{TN}(\mu, \sigma^2)$. The values of the parameters μ and σ are also chosen such that the duration of a surgery, y, is greater than 0 minutes, i.e., 2.5 times the standard deviation is less than the mean. Finally, the resulting synthetic durations of surgeries are rounded to the nearest minute.

To briefly summarise,

1. the durations in the first dataset collectively depend on the procedure, surgeon (effected by a multiplier) and patient;
2. the durations in the second dataset depend on the procedure, surgeon (additive effect) and patient; and
3. the durations in the third dataset collectively depend on the procedure and surgeon (effected by a multiplier) only.

Description of the first synthetic dataset: The truncated normal distributions of the first three procedures $p \in \{1, 2, 3\}$ have means set as $45p$, i.e., 45, 90 and 135 mins respectively, and standard deviations set as $6p-1$, i.e., 5, 11 and 17 mins respectively. Both the surgeon and the rating do not affect the duration of surgery. For the remaining three procedures $p \in \{4, 5, 6\}$, their distributions have means set as $15p - 35$, i.e., 25, 40, 55 mins respectively, and standard deviations set as $2p - 6$, i.e., 2, 4 and 6 mins respectively. The surgeon performing the surgery has a multiplicative effect on the duration of procedure, with the multipliers set as $1 + 0.05(s - 3)$, i.e., 0.9, 0.95, 1, 1.05 and 1.1 for the respective surgeons. The physical status rating of patient has an additive effect on the resulting duration of surgery. The means of the 3 ratings are set as $10(r - 1)$, i.e., 0, 10 and 20 mins respectively, and the corresponding standard deviations are set as 2^r, i.e., 2, 4 and 8 mins respectively. Hence, the simulated duration of surgeries for the first synthetic dataset may be mathematically expressed as

$$y \sim \begin{cases} \mathrm{TN}(45p, (6p-1)^2) & \forall p \in \{1, 2, 3\}, \forall s \in S, r \in R, \\ [1 + 0.05(s-3)] \times \mathrm{TN}(15p - 35, (2p-6)^2) & \\ \qquad\qquad + \mathrm{TN}(10(r-1), (2^r)^2) & \forall p \in \{4, 5, 6\}, \forall s \in S, r \in R. \end{cases}$$

Description of the second synthetic dataset: The duration of each surgery is the sum of the individual contributions of the procedure, the surgeon and the rating. The simulated duration of surgeries for the second synthetic dataset are

$$y \sim \text{TN}\,(30p - 10, (4p-2)^2) + \text{TN}\,(s-3, 2^{-2}) \\ + \text{TN}\,(10(r-1), 2^{2r}) \forall (p,s,r) \in P \times S \times R.$$

Description of the third synthetic dataset: The third synthetic dataset is almost identical to the first synthetic dataset except that the ratings do not alter the duration of surgeries. In other words, the expression for the simulated duration of surgeries is:

$$y \sim \begin{cases} \text{TN}\,(45p, (6p-1)^2) & \forall p \in \{1,2,3\}, \forall s \in S, r \in R, \\ [1 + 0.05(s-3)] \times \text{TN}\,(15p - 35, (2p-6)^2) & \forall p \in \{4,5,6\}, \forall s \in S, r \in R. \end{cases}$$

The software R [18] is used to perform the computer simulations and the subsequent calculations. Each run of a simulation involves the generation of 10,000 surgeries from one of the synthetic datasets. Each 3-tuple (p, s, r) is equally likely to be chosen when a surgery is generated for the run. The first 5,000 surgeries form the training set, and all the remaining surgeries are allocated to the test set. The training set determines the baseline and the hybrid model. For the hybrid model, only linear regression models are considered with only the main effects and no interaction terms. The value for the SBPV-based threshold, which was earlier defined within the second step of the prediction framework in Section "Prediction framework and the hybrid model", is set to 0. In fact, any reasonably small value will not enforce the threshold because of the large sample size. On average, each bucket in the training set comprises 56 surgeries. After that, a pair of values of the RMSEP is calculated using the test set, one from the baseline and the other from the hybrid model. The simulation is repeated 30 times to obtain 30 pairs of values. Finally, a paired-sample t-test is performed with the alternative hypothesis that there is a difference in the means of the values of the RMSEP from each model.

In each of the three synthetic datasets, the results indicate that the hybrid model uses the fitted linear regression model $y = \beta_0 + \beta_1 p + \beta_2 s + \beta_3 r$ for several, but not all, buckets. As all the explanatory variables are required in the hybrid model, it is not necessary to revise the model. All the paired-sample t-tests reject the null hypothesis in favour of the alternative hypothesis with p-values smaller than 10^{-5}. Reporting the 95% confidence intervals, the hybrid model performs better than the baseline by 0.57 ± 0.02 and 0.045 ± 0.006 minutes in the first and third synthetic datasets respectively. In the second synthetic dataset, the baseline performs better than the hybrid model by 0.04 ± 0.01 minutes. The results demonstrate that neither the hybrid model nor the baseline is superior. The performances of these models depend on the construction of the synthetic datasets, that is, the form of the formula that states the relationship between the explanatory variables and the response variable. Indeed, a learning algorithm will have datasets whose performances are underachieving or excellent [19].

While it is acknowledged that the difference in the predicted values of the baseline and hybrid models is too small to be of any practical significance, such as in the planning of surgical lists, the reader is reminded that the RMSEP is still ultimately a test statistic. To achieve a small improvement of 0.1 minutes in the value of RMSEP will require a similar improvement for all the 5,000 predictions in the test set. This is a feat as compared to improving the prediction of a single surgery duration by 0.1 minutes. In addition, the value of the true RMSEP should also be considered. This value can be computed using the population means of the relevant truncated normal distributions as follows.

$$\text{Value of the true RMSEP} = \sqrt{\frac{\sum_{i=1}^{I} (y_i - \mu_i)^2}{I}}. \tag{3}$$

For the i^{th} surgery in the above expression, μ_i is the population mean of the statistical distribution that generates y_i, the simulated actual duration, and the number of surgeries in the test set is denoted by I. Note that the value of the true RMSEP may be interpreted as the absolute minimum that is attained by the "perfect" prediction tool. Figure 3 compares the values of the true RMSEP from the computer simulations of each synthetic dataset with the values of the RMSEP that indicate the performance of the baseline and the hybrid model. The graphs clearly demonstrate that it is very difficult to observe large improvements in the accuracy of predictions when the errors are inherent from the statistical distributions.

Earlier, it is observed from the first and third synthetic datasets that the hybrid model performs better than the baseline, but the same cannot be said for the second synthetic dataset. This observation may be attributed to

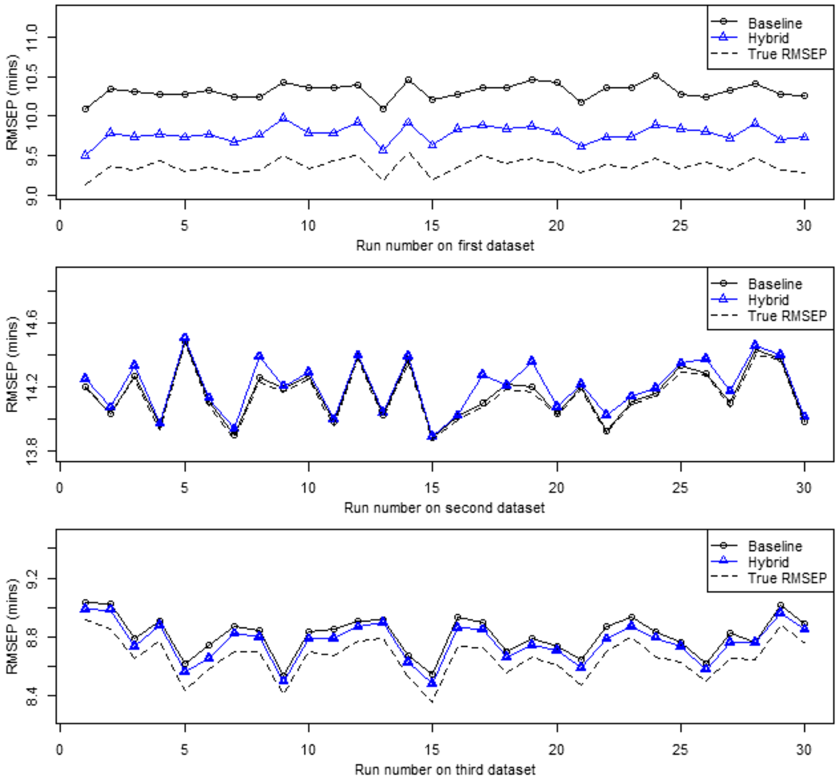

Fig. 3 Values of the RMSEP in minutes from each simulated run and each prediction tool (the baseline and the hybrid model). A smaller value of the RMSEP indicates a better performance of the prediction tool. For each dataset, the dotted line in the respective charts shows the value of its true RMSEP

the presence of the multiplier effect by surgeons on the surgery duration which is absent in the second synthetic dataset. This agrees with the hypothesis that different levels of simplifications will affect the relative performance of the hybrid model with respect to the baseline.

Furthermore, for the second synthetic dataset, the baseline correctly identifies the simulated process $y = \beta_0 + \beta_1 p + \beta_2 s + \beta_3 r$ through its exhaustive model selection procedure. However, Table 1 shows that the hybrid model has buckets assigned to other fitted linear regression models, such as, $y = \beta_4 + \beta_5 p + \beta_6 s$ and $y = \beta_7 + \beta_8 p$. The results from the second dataset show that the out-of-sample prediction errors are higher for surgeries that are assigned to the incorrect fitted linear regression models. A further analysis of each bucket does not reveal any unusual finding. In fact, the sampling distribution of the surgeries for each bucket in the training set may significantly differ from that in the test set, and the two sampling distributions often do not resemble the population distribution. To illustrate this, the reader is referred to Fig. 4 which compares, for a chosen bucket of the first synthetic dataset, the violin plots of the training set and the test set. When the size of dataset is increased by 5 times, the results using particular runs indicate that the proportion of surgeries that choose the correct linear regression model increases from 45.6% to 58.2%. A larger sample size is expected to further reduce the error rate. Therefore, the failure for a significant number of buckets to be assigned to the linear regression model that represents the true process may be attributed to the inherent variability in the surgery duration. Consequently, for the second synthetic dataset, the hybrid model always performs no better than the baseline as long as the baseline identifies the correct simulated process.

Finally, note that the hybrid model should not be used to identify whether an explanatory variable is part of the causal relationship that describes the response variable. From Table 1, the results of the third dataset indicate that a significant number of buckets are allocated to one of the fitted linear regression models that include the patient's physical status rating as an explanatory variable, even though the simulated process does not depend on this explanatory variable.

In the context of predicting surgery duration, it is unlikely for a dataset to be generated by processes whose effects on the surgery duration are purely additive. Due to the complex interactions of events that occur before and during the surgery, the causal relationship that determines the duration of surgeries is usually nontrivial. The hybrid model has the potential to improve the accuracy of predictions especially when the processes involved are complex and the prediction

Table 1 Summary of out-of-sample prediction errors by prediction models used in the hybrid model

Explanatory variable(s) in fitted model	First dataset			Second dataset			Third dataset		
	Number of surgeries assigned	Squared error (mins^2)	Root-mean-square error (mins)	Number of surgeries assigned	Squared error (mins^2)	Root-mean-square error (mins)	Number of surgeries assigned	Squared error (mins^2)	Root-mean-square error (mins)
-	173	14619.6	9.19	0			66	3296.9	7.07
p	1333	118836.8	9.44	357	109122.8	17.48	1711	165324.7	9.83
s	53	2392.9	6.72	0			0		
r	55	4982.7	9.52	0			64	2943.3	6.78
p, s	1118	130084.3	10.79	646	154839.8	15.48	1692	130009.8	8.77
p, r	980	86038.5	9.37	1717	374309.5	14.76	386	38921.5	10.04
s, r	149	16434.6	10.50	0			0		
p, s, r	1139	100093.6	9.37	2280	413068.7	13.46	1081	26011.2	4.91

The results are computed from the fifth run of each dataset. The variables p, s and r denote the surgical procedure, surgeon and patient's physical status rating respectively

tools are carefully selected for the hybrid model. Further studies and testing using the hybrid model on real world datasets are therefore recommended.

Investigating the SBPV-based threshold of the hybrid model

In this subsection, the first and third synthetic datasets introduced in Section "Investigating the performance of the hybrid model" are used to investigate the SBPV-based threshold. Recall that the third dataset is identical to the first dataset with the only exception that the simulated durations in the former do not depend on the physical status rating of patients. A factorial study is implemented by varying both the sample sizes and the values for the SBPV-based threshold. Conclusions are drawn from the observed impact of these changes on the values of the RMSEP by the hybrid model.

The software R [18] is used to perform the computer simulations and the subsequent calculations. In each run of a simulation, a synthetic dataset with a specified sample size between 500 to 3,000 is generated. The dataset is evenly split to obtain the training set and the test set. For each of the values 0, 5, 10, 15 and 20 for the SBPV-based threshold, the same training set is used to train the hybrid model. Similarly, the same test set is used to compute the values of the RMSEP from the hybrid model. By choosing the reference to be the value of the RMSEP when the value of the SBPV-based threshold is set to 0, the difference in the value of the RMSEP is computed for each non-zero value of the SBPV-based threshold. This procedure is repeated 30 times and the 95% confidence intervals for the differences in the values of RMSEP are reported. The results of the computer simulations are given in Tables 2 and 3.

For the first and third synthetic datasets, the results indicate that each of the three explanatory variables, P, S and R, is used by at least one bucket in the hybrid model. From Tables 2 and 3, there is a decline in the performances of the hybrid model when larger values of the SBPV-based threshold take effect. High prediction errors are expected when the sample size is small, since most of the linear regression models in the collection would be discarded. However, this adverse impact is magnified particularly for the first synthetic dataset. This observation can be explained by studying the effect of the SBPV-based threshold on the collection of linear regression models for the hybrid model. Setting a SBPV-based threshold leads to the following linear regression model being discarded from the collection: $y \approx \beta_0 + \beta_1 p + \beta_2 s + \beta_3 r$. This linear regression model has the most explanatory variables, so by the definition of the SBPV-based threshold, it will be first discarded by the hybrid model. The first synthetic dataset, whose surgery durations depend on all the three explanatory variables, will

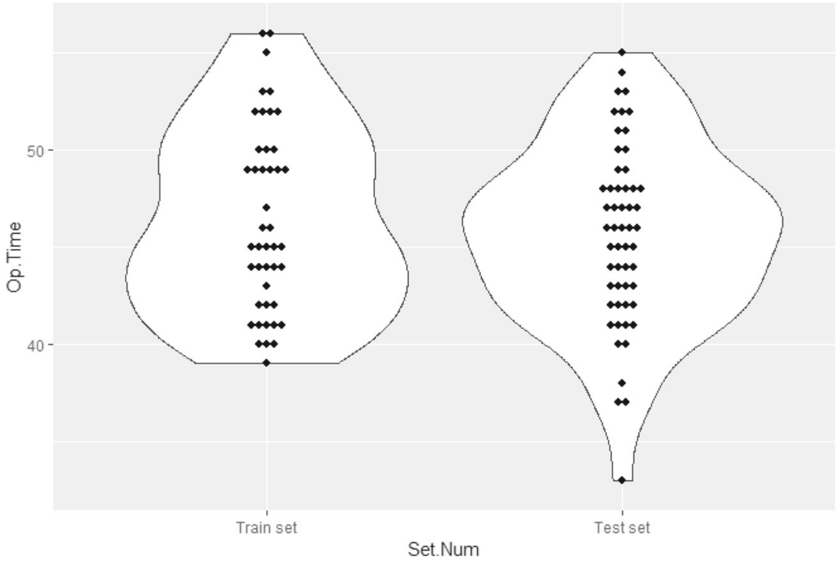

Fig. 4 A comparison of the training set durations and the test set durations for a particular bucket that is sampled from the first synthetic set

Table 2 The 95% confidence intervals from 30 generations of the first dataset

Sample size	Value of the SBPV-based threshold				
	0	5	10	15	20
500	0	−0.13 ± 0.07	−0.89 ± 0.12	−0.96 ± 0.12	−0.96 ± 0.12
1000	0	−0.17 ± 0.13	−0.21 ± 0.13	−0.64 ± 0.14	−0.93 ± 0.16
1500	0	−0.11 ± 0.02	−0.12 ± 0.02	−0.14 ± 0.04	−0.45 ± 0.08
2000	0	−0.09 ± 0.01	−0.11 ± 0.02	−0.11 ± 0.02	−0.14 ± 0.03
2500	0	−0.07 ± 0.01	−0.11 ± 0.02	−0.11 ± 0.01	−0.12 ± 0.02
3000	0	−0.03 ± 0.01	−0.11 ± 0.01	−0.11 ± 0.01	−0.11 ± 0.01

The intervals report the differences in the performance of the hybrid model when compared to the chosen reference, that is, the same sample size is used and the value of the SBPV-based threshold is 0. When the difference is positive (resp. negative), the resulting performance is better (resp. worse) than the reference

Table 3 The 95% confidence intervals from 30 generations of the third dataset

Sample size	Value of the SBPV-based threshold				
	0	5	10	15	20
500	0	−0.04 ± 0.06	−0.06 ± 0.05	−0.06 ± 0.05	−0.06 ± 0.05
1000	0	−0.02 ± 0.02	−0.10 ± 0.06	−0.12 ± 0.06	−0.13 ± 0.06
1500	0	0	−0.01 ± 0.01	−0.09 ± 0.02	−0.13 ± 0.04
2000	0	0	0	−0.03 ± 0.01	−0.09 ± 0.01
2500	0	0	0	−0.01 ± 0.01	−0.05 ± 0.01
3000	0	0	0	0	−0.01 ± 0.01

The intervals report the differences in the performance of the hybrid model when compared to the chosen reference, that is, the same sample size is used and the value of the SBPV-based threshold is 0. When the difference is positive (resp. negative), the resulting performance is better (resp. worse) than the reference bear a closer resemblance to the discarded linear regression model than the third synthetic dataset. Therefore, higher prediction errors from computer simulations of the first synthetic dataset are observed.

To conclude from Sections "Investigating the performance of the hybrid model" and "Investigating the SBPV-based threshold of the hybrid model", the simulation results suggest that the hybrid model may perform well on a real world dataset where the underlying processes are far more complex than that modelled in a simulation. The performance of the hybrid model depends on the sample size and the choice of prediction tools used in the hybrid model. However, the SBPV-based threshold may not be necessary; imposing such a threshold may increase the out-of-sample prediction errors and correspondingly decrease the performance of the hybrid model.

Case study: An actual surgery dataset

We have access to 2 years of Ear, Nose and Throat (ENT) elective surgery data comprising 3692 observations performed in New Zealand at either North Shore Hospital (within the main building or the Elective Surgery Centre) or Waitakere Hospital. Both hospitals, which are within the Waitemata District Health Board, are embarking on a surgery redesign project to scrutinise the booking processes for elective surgeries. This project also involves a simplification of the surgery classifications and a review of the current method for predicting the duration of upcoming surgeries. An electronic booking system is under development and will eventually be adopted. The predictions of surgery duration are currently based on the averages of recent historical surgeries of the same type performed by the same surgeon (henceforth referred to as the Method of Taking Averages - MTA). The hybrid model can be integrated within the booking system to predict the duration of surgeries.

The objectives of this case study are to investigate the hybrid model when applied to the surgery dataset and compare the hybrid model's performance with that of the baseline (prediction method) and the MTA, in Sections "Comparing the baseline with the hybrid models" and "Comparing the method of taking averages with the adjusted hybrid model" respectively. The specific details regarding the models will be given in their respective subsections. All data analysis and statistical calculations are performed in R [18].

Comparing the baseline with the hybrid models

A surgery comprises a procedure or a combination of procedures to be performed by a surgeon on a patient. Each surgery is described using the following categorical explanatory variables: (i) the procedure(s) to be performed, (ii) the primary surgeon and (iii) the five-category physical status classification system adopted by the American Society of Anesthesiologists (ASA). These explanatory variables are abbreviated "Proc", "PS" and "ASA" respectively. All the three explanatory variables are used to form the set $\mathbb{X}$ (with $V = 3$) and its corresponding buckets.

From the dataset, the 1872 surgeries occurring during the first year are selected for the training set. The 302 distinct combinations of procedures, 27 primary surgeons and 5 distinct physical status ratings of the training set form 834 buckets. Ignoring all the buckets that comprise only a single surgery, a histogram of the standard deviation of surgeries in each bucket is shown in Fig. 5. The remaining 1820 surgeries in the second year of the dataset are assigned to the test set. This choice for splitting the dataset by year is based on the logical assumption that the profile of the surgery types performed are similar across each year. The test set should therefore satisfy the important criterion for the assessment of prediction errors that it is a comprehensive coverage of all the surgery types present in the training set.

As mentioned before, only linear regression models are used for the hybrid model. For the linear regression models fitted on the training set, which has a large number of combinations of procedures relative to its size of dataset, a sensible choice for the collection is to include all the 2^V possible linear regression models with only the main effects, and no interaction or higher order terms. Therefore, only the three main effects are considered so that the maximum number of fitted linear regression models is eight. All the

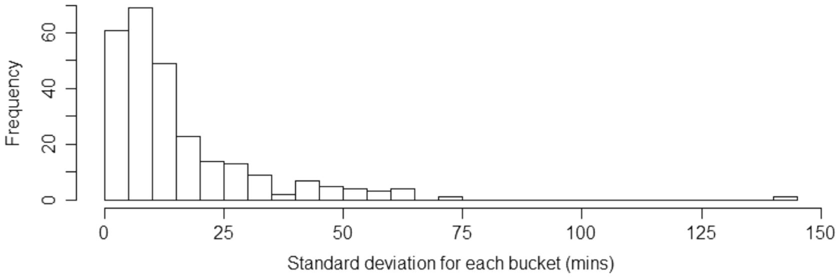

Fig. 5 A histogram of the standard deviations of surgeries in each bucket of the training set

eight linear regression models are initially included in the collection. Following the prediction framework described in Section "Prediction framework and the hybrid model", the result of the training phase indicates that all of the eight models are selected by at least one of the buckets. The collection is therefore retained and used for the (full) hybrid model. The reader should note that two different hybrid models will be compared, and this particular construction of the hybrid model shall subsequently be referred to as the *full hybrid* model.

The performance of the full hybrid model is evaluated by calculating the RMSEP for all surgeries in the test set. Since the magnitudes of the prediction errors are heavily dependent on the data, reporting only the magnitudes of error will give little information about the performance of a new learning algorithm. Instead, the relative performance of a new learning algorithm is compared with that of an algorithm that has been widely accepted for implementation. In particular, the prediction errors obtained from the full hybrid model are compared with that obtained from the multiple linear regression model with an exhaustive model selection, which is chosen as the baseline.

The results for the models are presented in the first two non-header rows of Table 4. For each model, the value of the RMSEP is computed when the value of the SBPV-based threshold is set to 0, 5, 10 and 20 respectively. Note that the earlier definition of a bucket can also be extended to the baseline, but the SBPV-based threshold should be omitted as it is different from the SPV threshold as mentioned in the introduction, so SBPV-based thresholds are not part of the accepted approach for the baseline.

For the full hybrid model, the results indicate that increasing the SBPV-based threshold leads to a systematic reduction in the number of linear regression models each bucket may choose. This consequently increases the value of the RMSEP, a result that is consistent with simulations performed in Section "Investigating the SBPV-based threshold of the hybrid model". Therefore from this point forth, the performance of the hybrid model shall be as assessed when the SBPV-based threshold is set to 0 (i.e., all models will be considered for each bucket). Note that while the SBPV-based threshold appears unnecessary for this particular dataset, this threshold may generally be required when a different dataset is considered.

A first glance at the first two non-header rows of Table 4 show that the full hybrid model seems to perform significantly better than the baseline. However, it is observed from the 2nd column that the baseline predicts 441 more surgery durations from the test set than the full hybrid model. This is a consequence from partitioning the training set into buckets, which in this case isolates an additional 441 surgeries from the test set that do not fit into any of these buckets. As there is a possibility that a smaller number of predicted surgery durations leads to a reduction in the value of RMSEP, the further step of identifying the 1059 surgeries from the test set that were predicted successfully by the full hybrid model is taken in order to make a fair comparison. The baseline is then used to predict these identified surgeries. The result obtained from the baseline after filtering the surgeries in the test set is presented in the second last row of Table 4. Although the baseline performance is improved, the results now show that the full hybrid model still remains the best performing model, albeit not by much.

It is speculated that the inclusion of a greater number of linear regression models in the hybrid model does not necessarily lead to more accurate predictions. In particular, if a careful selection of a smaller number of linear regression models is performed before applying the prediction framework for the hybrid model, then the resulting model may possibly lead to better predictive abilities. The reason for this improvement is that the selection of models (as described in Section "2") is less likely to be adversely affected because: 1) there are less misspecified models to choose from; and 2) the variability in the data will have less chance to select a misspecified model. Indeed, this observation holds true for the surgery dataset. To illustrate this, four of the linear regression models are removed from the full hybrid model. There is no particular criterion that determines which linear regression

Table 4 Values of the RMSEP of surgery duration in minutes by each model on the test set (or subset of). In the last 3 rows, the same subset of the test set is used

Model	Size of dataset for testing	Value for the SBPV-based threshold			
		0	5	10	20
Baseline	1500	24.466			
Full hybrid	1059	20.927	21.763	27.501	30.617
Baseline	1059	20.934			
Adjusted hybrid	1059	20.504	22.323	29.842	35.214

model should be removed, but the removal of models was determined by thorough experimentation with different combinations of models. However, $V = 3$ is required because the original partition of the observations should not be altered. The resulting hybrid model, which is denoted by the *adjusted hybrid* model, comprises the following linear regression models:

1. $y \approx \beta_0$,
2. $y \approx \beta_1 + \beta_2(\text{Proc})$,
3. $y \approx \beta_3 + \beta_4(\text{Proc}) + \beta_5(\text{PS})$ and
4. $y \approx \beta_6 + \beta_7(\text{Proc}) + \beta_8(\text{PS}) + \beta_9(\text{ASA})$.

After checking that these linear regression models are used by at least one of the buckets, the resulting adjusted hybrid model is used to calculate the values for the RMSEP. The results are shown in the last row of Table 4. Even though the difference in the magnitudes of the RMSEP between the two hybrid models, 0.423 minutes, is small, it is explained earlier in Section "Investigating the performance of the hybrid model" that from the perspective of model performance, this difference may still be significant relative to the value of the true RMSEP. However, the population parameters and the exact distributions that generate the surgery durations are unknown. It is therefore necessary to estimate the population parameters and rewrite expression (3) – from Section "Investigating the performance of the hybrid model" – in terms of these population parameters. First, it is better to estimate the population parameters from the entire dataset instead of just the test set, even though it is assumed that the profiles of the training set and the test set are similar. Then for each surgery type $t \in \{1, 2, \ldots, T\}$, suppose that Y_t is the random variable representing its duration. Assume also that all the random variables are mutually independent of each other. Furthermore, when referring to the surgery type of a particular surgery i in the test set or n in the clinical dataset, the surgery type t shall be referred to as $t(i)$ or $t(n)$ respectively. Finally, n_t is the number of type t surgeries in the clinical dataset. It follows that

$$\begin{aligned}
\text{estimate for the value of the true RMSEP} &= \sqrt{\frac{\sum_{i=1}^{I}(y_i - \mathbb{E}(Y_{t(i)}))^2}{I}} \\
&\approx \sqrt{\frac{\sum_{n=1}^{N}(y_n - \mathbb{E}(Y_{t(n)}))^2}{N}} \\
&= \sqrt{\frac{\sum_{t=1}^{T} n_t \mathbb{E}(Y_t - \mathbb{E}(Y_t))^2}{\sum_{t=1}^{T} n_t}} \\
&= \sqrt{\frac{\sum_{t=1}^{T} n_t \text{Var}(Y_t)}{\sum_{t=1}^{T} n_t}}.
\end{aligned}$$

The above approximation is applied to the clinical dataset. The result indicates that the estimate for the value of the true RMSEP is 18.73 minutes. This estimate may be viewed as the absolute minimum that can be achieved by a "perfect" prediction model. It supports the viewpoint that it is difficult to observe a significant decrease in the magnitude of the RMSEP when the full or the adjusted hybrid model is used instead of the baseline.

The prediction errors of the last two scenarios given in Table 4 are examined. Plots of the prediction errors are as shown in Fig. 6. From the plots, it is observed that there are observations where each of the prediction models severely overestimates or underestimates the actual surgery duration by at least 100 minutes (in fact, when all 1500 observations are included this deteriorates to an underestimate of 217.6 minutes). The large prediction errors may be partially attributed to the presence of large variability in durations (see Fig. 5) that are not accounted for in the surgery dataset, such as differences in human anatomy, unexpected theatre events and heavy bleeding during surgery. There is also a possibility that the extreme observations are the result of an error in data collection, of which it is difficult or impossible to ascertain. The modeller should therefore be aware that the study of the RMSEP alone may not provide sufficient information about the accuracy of predictions for all the buckets. The empirical probability density function gives more information about the distribution of errors and can also be useful when selecting a prediction model.

Comparing the method of taking averages with the adjusted hybrid model

We revisit the surgery redesign project as mentioned at the beginning of Section "Case study: An actual surgery dataset". Recall that the hospitals are currently using the Method of Taking Averages (MTA) to estimate the upcoming surgery duration. In particular, all identical surgeries that are historically performed by the same

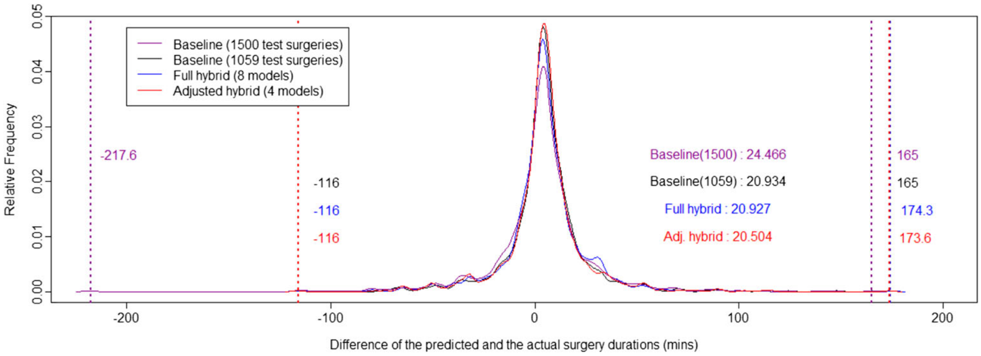

Fig. 6 An empirical probability density plot to compare the prediction errors by each model. The vertical dotted lines represent the extreme (minimum and maximum) values of the prediction errors in minutes

surgeon are used to predict the duration. It is emphasised that the physical status of patient, which is abbreviated "ASA", is not considered in their current method. This exclusion is considered in the analysis presented in the next paragraph. We evaluate the MTA on the same training and test set that is used earlier in our adjusted hybrid model and compare the results. The adjusted hybrid model is chosen because it is the best performing model as demonstrated in Section "Comparing the baseline with the hybrid models".

From the 1059 surgeries in the test set, the results indicate that the MTA gives 22.843 minutes as the value of the RMSEP. This is considerably larger than 20.504 minutes, the RMSEP value that is computed from the adjusted hybrid model. It remains to check that the increase in the value of RMSEP is not attributed to the exclusion of the explanatory variable "ASA" from the MTA. The same calculation is performed when "ASA" is factored into the MTA. The result is worse; the value of the RMSEP is 25.728 minutes. Henceforth, the explanatory variable "ASA" shall be excluded from the MTA.

Figure 7 also shows, for each bucket, the difference of the RMSEP values between the MTA and the adjusted hybrid model. A positive value for the difference indicates that the adjusted hybrid model performs better than the MTA for that particular bucket. It is observed that the adjusted hybrid model tends to perform better than the MTA when the number of surgeries in the bucket is large, i.e., for those buckets of low index/high frequency, to the left in Fig. 7. There are notable instances of singleton buckets where the MTA has done better than the hybrid model. In particular, the large negative spike for the singleton bucket with index labelled 178 corresponds to a rare surgeon-procedure pairing. The models within the set of viable regression models from which the prediction framework can choose do not perform well on this particular surgeon-procedure pairing, so the absolute mean of all surgeries is chosen as the best model. However, the MTA approach is equivalent to one of these regression models with the addition of an interaction term for the surgeon-procedure combinations. Without significant surgeon and procedure aggregation, there is insufficient data to support this model in general. However, given the model with this interaction term included to choose from, the prediction framework would produce the same estimate as the MTA. Of course, the

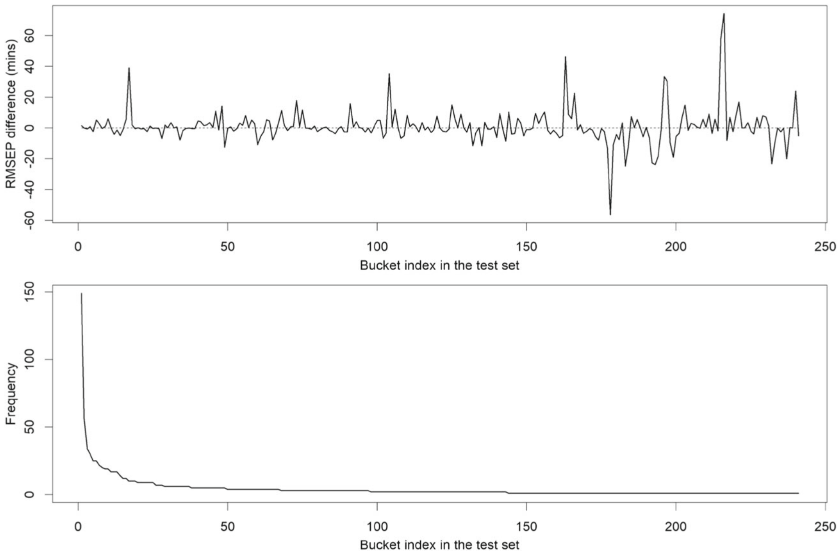

Fig. 7 The difference of RMSEP values and their corresponding number of surgeries for each bucket are presented in the top and bottom plots respectively. The differences are computed using **a** the prediction method that takes the averages of historical surgeries and **b** the adjusted hybrid model. Note that the buckets are arranged by the number of surgeries in decreasing order. A positive value for the difference indicates that the adjusted hybrid model performs better than the MTA for the particular bucket

use of this prediction model without sufficient data support results in the large positive spikes that are also evident in Figure 7, so a high degree of care and expertise is needed when selecting the prediction models. This highlights that although the prediction framework presented here has the potential to improve the prediction process post model selection, it is not a replacement for the model selection step. It is possible for the hospital we are working with that significant improvement in the prediction process is attainable by including a clustering analysis as part of the model selection process to aggregate procedure and surgeons. Such an analysis is part of the redesign of the hospital's electronic booking system for elective surgeries currently underway.

Conclusion

This paper describes a prediction framework that constructs a partition of the set of historical surgeries using a collection of viable prediction models, and allocates one of these models to each part of the historical surgical population. For ease of application, in this study the prediction models are restricted to linear regression models, and the resulting prediction approach is referred to as the "hybrid model". The performance of the hybrid model is evaluated, with the results indicating that in some situations, the hybrid model performs better than a standard modelling approach, i.e., the baseline model that is the multiple linear regression model with model selection. It also performs better than current practice (the MTA) when evaluated against a real hospital's data. For the synthetic datasets, created via computer simulations, and the case study, a threshold based on the number of subjects related to the bucket per variable (SBPV) is not required. In addition, the inclusion of a larger number of regression models does not lead to predictions with higher accuracy.

The main drawback of this approach is the inability to estimate surgery duration for new surgery types. In the case study presented here, cases where new procedures or surgeons were introduced into the test set resulted in no corresponding part being present in the training set, so the hybrid model could make no estimate of the duration. Of course, this is the case for standard model selection also, and in practice is overcome by introducing a duration default, such as a "surgeon estimate" or reverting to a hierarchical historical average depending on the level of historical data available to support the estimate. In practice, procedure aggregation would be performed via a clustering approach to reduce the number of possible procedures from 302 to a more manageable number. Likewise, surgeons may also be clustered based on experience or a similar indicator of expected duration (different surgeons may perform the same procedure using different equipment, which could affect the surgery duration). If the collection of viable prediction models is based on procedure and surgeon aggregations, then new procedures and surgeons can be introduced to the problem via inclusion in the appropriate bin. As time rolls on, new surgeries are included in the historical collection and the prediction framework can adapt, allocating different models to some parts or adjusting model coefficients as the proportions in casemixes vary. It is expected that expert model selection analysis would need to be carried out occasionally to ensure the covariates included in the candidate models continue to perform well.

Other approaches for estimating the duration of new surgery types within the prediction framework presented here are the subject of current research. Approaches to aggregate "buckets" analagous to aggregation of factor levels are under investigation. Another research avenue currently under exploration is an approach to model averaging. Relaxation of the z_{pm} variables in the integer programming formulation presented in Section "Prediction framework and the hybrid model" yields an extension from model selection to model averaging, where the model weights for the various viable prediction models can differ from part to part across the historical surgical partition. Preliminary results are promising, and are the subject of a forthcoming paper.

The conclusions of this paper are specific to the datasets presented with surgery duration prediction via linear regression models. Studies using different types of datasets and other learning algorithms need to be conducted to determine whether these results generalise.

Compliance with Ethical Standards

Conflict of interests All the authors declare that they have no conflicts of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Graue, R. M., Prediction and optimization techniques to streamline surgical scheduling, Master's thesis, Massachusetts Institute of Technology United States, 2013.
2. Devi, S. P., Rao, K. S., and Sangeetha, S. S., Prediction of surgery times and scheduling of operation theaters

in optholmology department. *J. Med. Sys.* 36(2):415–430. https://doi.org/10.1007/s10916-010-9486-z, 2012.

3. Dexter, F., Epstein, R. H., Bayman, E. O., and Ledolter, J., Estimating surgical case durations and making comparisons among facilities: identifying facilities with lower anesthesia professional fees. *Anesth. Analg.* 116(5):1103–1115. https://doi.org/10.1213/ANE.0b013e31828b3813, 2013.
4. ShahabiKargar, Z., Khanna, S., Good, N., Sattar, A., Lind, J., and O'Dwyer, J., Predicting procedure duration to improve scheduling of elective surgery. In: Pham, D.-N., and Park, S.-B. (Eds.) *PRICAI 2014: Trends in Artificial Intelligence. Cham: Springer, ISBN 978-3-319-13560-1, 998–1009*, 2014. https://doi.org/10.1007/978-3-319-13560-1_86.
5. Hosseini, N., Sir, M. Y., Jankowski, C. J., and Pasupathy, K., Surgical duration estimation via data mining and predictive modeling: a case study, AMIA Annual Symposium Proceedings, 640–648, 2015.
6. Edelman, E. R., van Kuijk, S. M. J., Hamaekers, A. E. W., de Korte, M. J. M., van Merode, G. G., and Buhre, W. F. F. A., Improving the prediction of total surgical procedure time using linear regression modeling. *Frontiers in Medicine* 4:85. https://doi.org/10.3389/fmed.2017.00085, 2017.
7. ShahabiKargar, Z., Khanna, S., Sattar, A., and Lind, J., Improved prediction of procedure duration for elective surgery. In: Ryan, A., Schaper, L. K., and Whetton, S. (Eds.) *Integrating and Connecting Care, vol. 239 of Studies in Health Technology and Informatics, IOS Press Ebooks, 133–138*, 2017. https://doi.org/10.3233/978-1-61499-783-2-133.
8. Green, S. B., How many subjects does it take to do a regression analysis. *Multivariate Behav. Res.* 26(3):499–510. https://doi.org/10.1207/s15327906mbr2603_7, 2010.
9. Harrell, F, *Regression modeling strategies*. 1 ed. New York: Springer, 2001. https://doi.org/10.1007/978-1-4757-3462-1.
10. Schmidt, F. L., The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educ. Psychol. Meas.* 31(3):699–714. https://doi.org/10.1177/001316447103100310, 1971.
11. Austin, P. C., and Steyerberg, E. W., The number of subjects per variable required in linear regression analyses. *J. Clin. Epidemiol.* 68(6):627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014, 2015.
12. Hansen, B. E., and Racine, J. S., Jackknife model averaging. *J. Econom.* 167(1):38–46. https://doi.org/10.1016/j.jeconom.2011.06.019, 2012.
13. Wolpert, D. H., Stacked generalization. *Neural Netw.* 5(2):241–259, 1992.
14. Allen, D. M., The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16:125–127, 1974.
15. Seber, G. A. F., and Lee, A. J., *Linear regression analysis*. 2 ed. New York: Wiley, 2003.
16. Khanmohammadi, N., Rezaie, H., Montaseri, M., and Behmanesh, J., The application of multiple linear regression method in reference evapotranspiration trend calculation. *Stochastic Environmental Research and Risk Assessment* 32(3):661–673. https://doi.org/10.1007/s00477-017-1378-z, 2018.
17. Theobald, R., and Freeman, S., Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE Life Sciences Education* 13(1):41–48. https://doi.org/10.1187/cbe-13-07-0136, 2014.
18. R Core Team, R., A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2018.
19. Alpaydin, E., *Introduction to machine learning*. 2 ed. Cambridge: The MIT Press, 2010. ISBN 9780262012430.