REGULAR PAPER

Improving predictions of pediatric surgical durations with supervised learning

Neal Master[1] · Zhengyuan Zhou[1] · Daniel Miller[1] · David Scheinker[2] · Nicholas Bambos[1,3] · Peter Glynn[3]

Received: 30 December 2016 / Accepted: 3 May 2017 / Published online: 15 May 2017

Abstract Effective management of operating room resources relies on accurate predictions of surgical case durations. This prediction problem is known to be particularly difficult in pediatric hospitals due to the extreme variation in pediatric patient populations. We pursue two supervised learning approaches: (1) We directly predict the surgical case durations using features derived from electronic medical records and from hospital operational information. For this regression problem, we propose a novel metric for measuring accuracy of predictions which captures key issues relevant to hospital operations. We evaluate several prediction models; some are automated (they do not require input from surgeons) while others are semi-automated (they do require input from surgeons). We see that many of our automated methods generally outperform currently used algorithms and our semi-automated methods can outperform surgeons by a substantial margin. (2) We consider a classification problem in which each prediction provided by a surgeon is predicted to be correct, an overestimate, or an underestimate. This classification mechanism builds on the metric mentioned above and could potentially be useful for detecting human errors. Both supervised learning approaches give insights into the feature engineering process while creating the basis for decision support tools.

Keywords Supervised learning · Predictive modeling · Pediatric surgery · Surgical scheduling

1 Introduction

Operating rooms are a critical hospital resource that must be managed effectively in order to deliver high-quality care at a reasonable cost. Because of the expensive equipment and highly trained staff, operating rooms are very expensive with the average cost of operating room time in the US being roughly \$4000 per hour [18,26]. In addition, mismanaged operating rooms lead to cancelled surgeries with each cancellation decreasing revenue by roughly \$1500 per hour [11,12,19]. The management process is complicated and must account for heterogeneity of patient needs, uncertainty in patient recovery times, and uncertainty in surgical durations. In this paper, we aim to design statistical models that can be used to improve the accuracy of predictions of surgical durations. The motivating idea is that better predictions enable better operational decisions.

Specifically, we consider the problem of predicting pediatric surgical durations. Currently, many pediatric hospitals rely on surgeons to provide predictions and this alone

✉ Neal Master
nmaster@stanford.edu

Zhengyuan Zhou
zyzhou@stanford.edu

Daniel Miller
danielrm@stanford.edu

David Scheinker
dscheink@stanford.edu

Nicholas Bambos
bambos@stanford.edu

Peter Glynn
glynn@stanford.edu

[1] Department of Electrical Engineering, Stanford University, 350 Serra Mall, Stanford, CA 94305, USA

[2] Lucile Packard Children's Hospital Stanford, Palo Alto, CA, USA

[3] Department of Management Sciences and Engineering, Stanford University, Stanford, CA, USA

increases costs. Not only is a surgeon's individual time expensive, a surgeon may depend on the help of their staff to make the predictions. Each of these contributions may seem insignificant on its own, but these man-hours add up to increased costs. Consequently, although our primary goal is to increase prediction accuracy, automating the prediction process even without increasing accuracy can help reduce costs and improve efficiency.

A major reason that pediatric hospitals rely on surgeons' medical expertise is that accurately predicting pediatric surgical case lengths is a very difficult problem. It is considered to be a more difficult problem than predicting adult surgical durations because compared to patient populations at adult hospitals, patient populations at pediatric hospitals are characterized by extreme variation in patient age, size, and developmental level. This has been discussed in the academic medical literature for specific procedures [27] and is also supported by anecdotal evidence at Lucile Packard Children's Hospital (LPCH) Stanford. Although we use data from LPCH, our goal is to design models that apply to all pediatric hospitals. In particular, none of the features used by our models are specific to LPCH. Moreover, even though we consider a multitude of different procedure types, we only use features that are relevant to all kinds of surgical procedures. In this sense, we aim to provide a "one-size-fits-most" solution that is broadly applicable to pediatric hospitals regardless of size or case load profile.

Given this broad motivation, there are several papers on the topic of predicting surgical durations. However, the majority are focused on adult patient populations (e.g. [29, 31, 34]) with pediatric populations only being of very recent interest, e.g. [4, 16]. In addition, many of these studies rely on simple methods like linear regression [16, 29, 31, 34]; regression trees are the most "modern" technique considered in the literature [4]. For adult surgeries, researchers typically see "modest improvements in accuracy" [34] over human experts (e.g. surgeons and nurses). In contrast, for pediatric surgeries, the difficulty of the problem leads to negative conclusions with [4] reporting that "none of the variables previously associated with case time in adults were generally correlated with case time in our pediatric population." These papers demonstrate that compared to predicting adult surgical durations, predicting pediatric surgical durations is still a difficult open problem.

In light of the difficulty of the problem, rather than attempting to build a predictive model that fully replaces predictions provided by medical staff, there is also value in building a model that can provide guidance to human experts. Decision support tools can take many forms [20] and so we focus on the following specific problem: Given an expert prediction, can we reliably identify overestimates and underestimates? This classification problem is a "meta-prediction" problem in the sense that it makes a prediction about a prediction. Since the goal is to detect inaccurate predictions, we will refer to this as the detection problem and the original regression problem as the prediction problem. Though clearly related to the prediction problem, the detection problem is distinct and to our knowledge not previously studied in the academic literature.[1]

The remainder of our paper is organized as follows. In Sect. 2, we discuss how surgical duration predictions are made and used at LPCH. We also discuss current research on predictive models for pediatric surgical durations. In Sect. 3, we motivate and define a performance metric for prediction accuracy that models operational concerns in a hospital setting. This metric motivates a nonlinear data transformation for addressing the prediction problem and also allows us to present a concrete formulation of the detection problem. In Sect. 4, we present some empirical results on the prediction problem: We define some benchmark prediction methods, propose our own prediction models, and present a comparison. In Sect. 5, we present an empirical characterization of some models that address the detection problem. We discuss future avenues of work in Sect. 6 and conclude in Sect. 7.

2 The importance of accurate predictions and the state of the art

In this section, we provide background into the problems at hand. First we explain how predictions influence the scheduling of surgical environments at hospitals like LPCH. We then explain the methods that are currently used to make these predictions. We discuss the academic literature on the topic.

2.1 The state of practice

Scheduling a surgical procedure can be a very complicated process for the patient as well as for the hospital. Depending on the type of procedure, the scheduling process may begin several weeks before the procedure actually takes place. The patient and primary physician will coordinate with the appropriate surgical service to meet the clinical needs of the patient. Because patient preferences and hospital policies play a big role in determining these coarse-grained scheduling decisions, we will not describe them in detail. The most important feature of the coarse-grained scheduling is that operating rooms are shared across different surgical services with block scheduling. This means that a particular surgical service will have an operating room for a large block of time; at LPCH, a single block constitutes the entire day. Different block scheduling strategies can be used, e.g. [35], but regardless of how blocks are allocated across the week or month, block

[1] This paper is an extended version of results presented in a conference publication [36] and an unpublished technical report [21].

scheduling causes many similar procedures to be scheduled back-to-back.

The fine-grained scheduling decisions are made in the days immediately preceding a surgery. The two-step process we describe is somewhat specific to LPCH, but similar systems exist at other pediatric surgical hospitals. The first step is the prediction: Surgeons (with varying levels of assistance from their administrative staff) will need to predict the duration of each surgical procedure. The second step is the scheduling: A group of nurses and physicians use the predictions to schedule the operating rooms (ORs) and the Ambulatory Procedure Unit (APU).[2]

Although the scheduling is done manually rather than by an optimization algorithm, the nurses and physicians who make the scheduling decisions have several objectives in mind. One is to have all procedures completed by the end of the day; it is possible to run overtime but this is inconvenient and incurs additional staffing costs. Another objective is to have patients released to the recovery units at a regular rate so that recovery beds are available and the nursing staff is able to provide care.

The predictions impact all of these objectives and more. Consider a scenario in which some surgeries finish earlier than predicted and other surgeries finish later than predicted. Suddenly there is an unexpected spike in demand for recovery beds that the nursing staff is unable to accommodate. Patients will need to wait (in the ORs) for beds, and this delays other surgeries. If these delays are acute, surgeries will need to be rescheduled. This operational inefficiency reduces quality of care for patients and increases costs for the hospital. Patients whose surgeries have been cancelled may opt to have their procedures done at other hospitals. Thus, inaccurate predictions not only increase costs but also reduce revenue.

The prediction process is essential to delivering high-quality care while maintaining efficiency, but the current prediction methods are somewhat primitive. There are two competing methods available to most pediatric surgical hospitals. The first is historical averaging which predicts that the duration of a scheduled surgery is the average (arithmetic mean) of the past durations recorded of when that particular surgeon performed that particular procedure. If the particular surgeon has performed the procedure fewer than a certain number of times, say five times, then historical averaging will predict with the average across all surgeons' past performances. This method does not take into account the substantial variation in the patients and can be quite inaccurate. The second method is to rely on expert opinions. Surgeons (potentially with the assistance of their staff) can provide an estimate of how much time they need to complete a given procedure. This is the system currently used at LPCH. Although surgeons have extensive experience, their predictions are not necessarily accurate. One reason is that physicians are not using quantitative models so their predictions are merely "guesstimates." Another reason is that physicians may have financial incentives that cause them to systematically misestimate the surgical durations. For example, surgeons do not necessarily bear the overtime costs of surgeries running past the end of the day. As a result, in an effort to maximize the number of surgeries scheduled in a block (and hence maximize their personal compensation), surgeons may underestimate the amount of time required to complete a procedure.

Given the inadequacies of both historical averaging and expert predictions, it is not clear which method is superior. Not only does this comparison depend on the types of procedures and the population of patients, it also depends significantly on the surgical teams. At LPCH, expert prediction is currently used but this might not be the best choice for other hospitals. As we develop and evaluate our predictive models, we will need to compare our performance to both of these existing benchmark practices.

2.2 Literature review

Much of the applied statistics and academic medical literature on surgical procedure durations focuses on modeling problems rather than on prediction problems. For example, in [30] it was shown that lognormal distributions model the uncertainty of surgical procedure times better than normal distributions. A consequent line of research explores different methods for fitting lognormal distributions, e.g. [22,28]. Although this work does not directly address the prediction problem and is not focused on pediatrics, it does point out that surgical times tend to follow heavy-tailed distributions. This insight is valuable when designing predictive models and is discussed more in Sect. 3.

The literature on predictive modeling for pediatric surgeries is somewhat sparse. In the management science literature, LPCH has been previously used as a case study [16]. This previous work at LPCH focused on using features related to staff experience. The resulting linear model decreased the mean absolute error by 1.98 ± 0.28 min; an amount with limited practical significance. In the pediatric medical literature, the most relevant paper is based on data from Boston Children's Hospital [4]. This work identifies improving predictions of pediatric surgical durations[3] as a

[2] At LPCH, there are currently seven ORs and two APUs.

[3] We note that in [4], surgical durations are measured from "wheels in" to "wheels out," i.e. patient entry to the OR to patient exit from the OR. In our work, we focus on predicting the time from when the surgeon enters the OR to when the surgeon exits the OR. Although slightly different, we note that the dominant source of variability is the surgery. At LPCH, surgeons are asked to predict the amount of time they spend in the OR and this allows us to directly compare our methods to expert predictions.

key avenue of research. However, the paper is pessimistic: The authors find that “for most procedure types, no useful predictive factors were identified and, most notably, surgeon identity was unimportant.” They use surgeon identity, intensive care unit bed request, ASA status (explained in Sect. 4), patient age, and patient weight as features and conclude that “until better predictors can be identified, scheduling inaccuracy will persist.”

The negative results in [4] demonstrate that building predictive models for pediatric surgical durations is very difficult but we must raise some concerns with their statistical approach. Our primary criticisms are that the authors rely on a single learning algorithm and that they impose restrictions on this algorithm in a way that inhibits it’s performance. Specifically, the authors rely on the CART algorithm [7]. However, the authors opt to learn a separate tree for each procedure. This essentially forces the tree to split until each node has only one procedure type and then CART is used to learn the remainder of the tree. The motivating idea is that the procedure name is a very important feature but this model restriction unnecessarily fragments the data. A related issue is that for many procedures the authors had only 30 observations, creating training sets of 20 observations and testing sets of 10 observations. Given these small sample sizes, the authors restrict the learned trees to have depths of at most three. Although this model restriction may be appropriate for procedures with small sample sizes, for some procedures the authors had hundreds of observations and with larger sample sizes, deeper trees can be learned.

We also note that CART is inherently unstable, i.e. the learned tree is very sensitive to the training data [5]. As a result, although it is easy to draw conclusions from the topology of a learned decision tree, it is difficult to have confidence in these conclusions. This is exacerbated by small sample sizes. Consequently, the conclusions presented in [4] should be viewed with skepticism.

Despite our concerns with the statistical methodology of [4], we think that a more fundamental issue is the modeling methodology. The authors use root-mean-square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) as performance metrics. Although these metrics are statistically meaningful, they are not necessarily operationally meaningful. Given that scheduling decisions are made by human experts, we feel that performance metrics should be more easily interpretable by physicians and nurses. We address this concern in the following section by proposing a performance metric that is more tightly related hospital operations. Moreover, we see that this performance metric naturally leads to the aforementioned detection problem, a problem formulation that has been completely overlooked other academic researchers.

3 Operational and statistical performance metrics

Our study is motivated primarily by operational concerns, so we focus a performance metric that is operationally meaningful. Although RMSE and MAE each have a natural statistical meaning, neither has an obvious operational meaning. For example, it is unlikely that a physician would ask for the empirical root-mean-square prediction error of a proposed prediction model. On the other hand, consider another commonly used prediction metric, prediction R^2, which is defined as follows:

$$1 - \frac{\text{Prediction MSE}}{\text{Variance of the response}} \tag{1}$$

The variance of the response variable (i.e. Y) is essentially the prediction MSE from a model that predicted with the mean of Y. Hence, prediction R^2 is the reduction in MSE relative to a constant model. Prediction R^2 is a bit more interpretable than prediction RMSE: A value of 1.0 indicates perfect prediction while smaller values indicate lower predictive performance. However, the “gray area” of prediction R^2 values below 1.0 is still not easy to connect to operational impacts. For instance, suppose prediction errors are consistently equal to 5 min and the standard deviation of the surgical durations is also 5 min. An error of 5 min is unlikely to have major operational impacts, but the prediction R^2 could still be equal to zero. Moreover, it can be difficult to compare the performance across different procedure types because the variances for different procedures can differ substantially.

With this in mind, we consider a deceivingly simple question posed by a physician: How often is the predictive model correct? The nuance now is quantifying what it means for a prediction to be “correct.” Understanding the meaning of “correct” requires that we explore the operational impact of an “incorrect” prediction.

In the remainder of this section, we outline some hypothetical outcomes that illustrate how different kinds of prediction errors impact the hospital. We use these insights to propose a novel performance metric for the prediction problem. We also discuss how to leverage existing machine learning implementations to tailor our models to this performance metric. In particular, we will see that this performance metric motivates a particular nonlinear transformation of the data. We also use the insights from these hypotheticals to formulate the detection problem.

3.1 The operational impact of prediction errors

The predictions are used to set daily schedules; surgeries will be delayed or cancelled if the outcomes are sufficiently dissimilar from the predictions. Since similar kinds of procedures tend to be scheduled back-to-back within the same block, the threshold at which schedules "break down" depends on how long the procedures are predicted to take. We demonstrate this with the following hypotheticals. First suppose a particular procedure is predicted to take 150 min. If the procedure actually takes 165 min, then it is reasonable to say that the prediction was correct—the procedure only took 10% longer than was predicted and this will not significantly impact subsequent procedures which are scheduled for comparable amounts of time. Now suppose a different procedure in a different block is predicted to take 20 min. If the procedure actually takes 35 min, then it is not reasonable to claim that the prediction was correct. Short procedures are typically scheduled back-to-back and so if the procedure takes 75% longer than predicted, then this will undoubtedly cause operational problems. Note that in both cases the outcome was 15 min longer than the prediction but in one case the prediction was deemed correct and in the other case the prediction is incorrect. This demonstrates that the difference between the outcome and the prediction needs to be less than some percentage of the prediction in order for the prediction to be deemed accurate.

However, there are limits to this reasoning. Consider the same hypotheticals as above. If we have a procedure that is predicted to take 150 min, then having the outcome be within 10% (15 min) of this predicted time is reasonable. However, if we consider a procedure that is predicted to take 20 min, then requiring that the outcome be within 10% (now just 2 min) is no longer reasonable. Clearly, using a simple percentage is too restrictive for surgeries that are typically short. Similarly, using a simple percentage is too lax for surgeries that are typically quite long.

3.2 The prediction problem

To formalize the insights from these hypotheticals, we first consider the prediction problem. Consider the model

$$Y = f(X) + \epsilon \tag{2}$$

where X is a vector of features describing a particular surgical case, Y is the amount of time required for the surgeon to perform this procedure, $f(\cdot)$ is the target function, and ϵ is noise. We can use a learning procedure to predict Y with $\hat{Y} = \hat{f}(X)$ where $\hat{f}(\cdot)$ is an approximation to $f(\cdot)$. Given the discussion above, we propose the following metric for quantifying if this prediction is accurate. We say that the prediction is accurate (i.e. "correct") if

$$|Y - \hat{Y}| < \tau(\hat{Y}) \tag{3}$$

where

$$\tau(\hat{Y}) = \min\left\{\max\left\{p\hat{Y}, m\right\}, M\right\}, \tag{4}$$

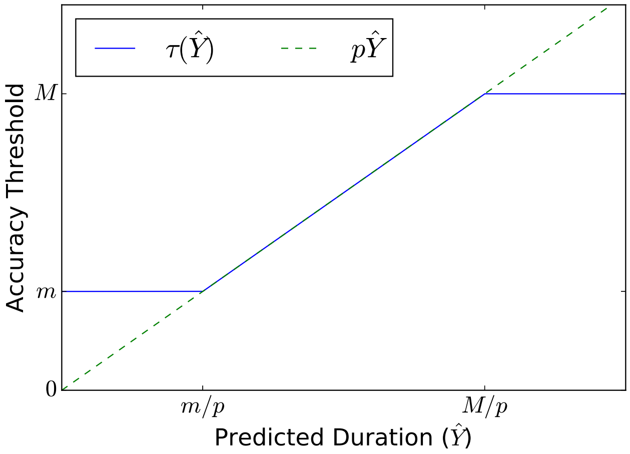

Fig. 1 A depiction of the $\tau(\hat{Y})$

$p \in (0, 1)$, and $M > m \geq 0$. We see that $\tau(\hat{Y})$ encapsulates the issues raised by our hypothetical examples from the previous section: $\tau(\hat{Y})$ is essentially a fraction p of $\hat{Y}$ that is restricted to being within $[m, M]$. Consider three examples with $p = 0.2, m = 15$ min, and $M = 60$ min. If $\hat{Y} = 50$ min, then we see that $p\hat{Y} = 10$ min while $\tau(\hat{Y}) = 15$ min; 20% of $\hat{Y}$ yields too stringent a threshold so the threshold become $m = 15$ min. If $\hat{Y} = 350$ min, then we see that $p\hat{Y} = 70$ min while $\tau(\hat{Y}) = 60$ min; 20% of $\hat{Y}$ yields too lax a threshold so the threshold become $M = 60$ min. Finally, if $\hat{Y} = 100$ min, then the threshold is simply $p\hat{Y} = 10$ min. This is depicted in Fig. 1.

One drawback of this metric is that it is binary and hence treats all incorrect predictions as equally incorrect. This is an artifact of our focus on quantifying accuracy rather than error. Given the critical nature of the patient scheduling, all inaccurate predictions should be avoided regardless of the magnitude of inaccuracy. A benefit of this approach is that we can easily measure accuracy on the unit interval. This is particularly helpful when presenting results to non-technical healthcare professionals.

Another drawback of this performance metric is the induced loss function:

$$\ell(Y, \hat{Y}) = \begin{cases} 1, & |Y - \hat{Y}| \geq \tau(\hat{Y}) \\ 0, & |Y - \hat{Y}| < \tau(\hat{Y}) \end{cases} \tag{5}$$

The discontinuous nature of the loss function leads to several problems. Not only is the loss function non-differentiable but on the regions where it is differentiable, the derivative is zero. This makes it difficult to apply gradient-based learning techniques. Furthermore, because the loss function takes discrete values, it is not suitable as an information criterion for building trees: when searching for the best split there will typically be many splits that maximize the information gain. Moreover, the discontinuity is sensitive to p, m, and M.

Although we developed $\tau(\hat{Y})$ [and hence $\ell(Y, \hat{Y})$] based on expert input from healthcare professionals, translating these qualitative insights into precise parameter values is fraught with difficulties. Methods like the Delphi technique could be used to translate expert input into parameter values, but such methods are not always reliable [24].

To alleviate these issues, we would like to "massage" this loss function into a form that is less sensitive to p, m, and M. We sketch the idea as follows. Suppose that m is sufficiently small and M is sufficiently large so that $\tau(\hat{Y}) = p\hat{Y}$. Then $\ell(Y, \hat{Y}) = 0$ when

$$|Y - \hat{Y}| < \tau(\hat{Y}) = p\hat{Y}. \tag{6}$$

Dividing by $\hat{Y}$ gives us that

$$|Y/\hat{Y} - 1| < p \tag{7}$$

which is equivalent to

$$1 - p < Y/\hat{Y} < 1 + p \tag{8}$$

and taking logarithm shows that this is equivalent to

$$\log(1 - p) < \log(Y) - \log(\hat{Y}) < \log(1 + p). \tag{9}$$

If we let $\epsilon(p) = \min\{-\log(1 - p), \log(1 + p)\}$, then this shows that

$$(\log(Y) - \log(\hat{Y}))^2 < \epsilon(p)^2 \implies \ell(Y, \hat{Y}) = 0. \tag{10}$$

So if we aim to minimize $(\log(Y) - \log(\hat{Y}))^2$, then we will likely also have $\ell(Y, \hat{Y})$ equal to zero.[4] Although taking the logarithm of $\hat{Y}$ is not quite the same as estimating $\log(Y)$, this sketch suggests that given our operational performance metric, it is reasonable to perform the prediction in log-space under mean-square loss. Specifically, if we let $Z = \log(Y)$, then we can use the model

$$Z = g(X) + \eta \tag{12}$$

where $g(\cdot)$ is the target function and η is the error. We can then learn $g(\cdot)$ to get $\hat{Z} = \hat{g}(X)$. We can then use $\exp(\hat{Z})$ as a prediction for Y.

We note that this sketch merely suggests that using a logarithmic transformation is appropriate, but it does not provide any guarantees regarding $\ell(\cdot, \cdot)$. In particular, we see that by doing the learning in log-space under mean-square loss, we are not taking into account the asymmetry of the original loss function. Our sketch suggests that when m is small and M is large this approach is reasonable but it by no means optimal. Though we sacrifice optimality (with respect to the loss criterion), there are several practical benefits to this transformation. As mentioned earlier, a key benefit is that our models are no longer sensitive to the parameters p, m, and M. These parameters are necessarily subjective but our model training procedures are not. Furthermore, by transforming the data and relying on mean-square loss, we can now apply existing implementations of a variety of machine learning algorithms. This is useful not only for research and prototyping but for eventual deployment as well.

The logarithmic transformation can also be motivated with more traditional applied statistics methodology. For example, consider the histograms in Fig. 2. The original quantity has a heavy right tail (i.e. a positive skew), but after the logarithmic transformation the histogram is fairly symmetric. This also agrees with the lognormal models discussed in Sect. 2.2. Although logarithmic transformations are common, they are not always acceptable; see [14] for some examples of when logarithmic transformations can actually introduce skew. However, as noted in [3], for many practical problems logarithmic transformations can be very useful.

With this discussion in mind, we will train our prediction models in log-space under mean-square loss to learn $\hat{g}(\cdot)$. When evaluating a prediction model on a test set $\{(X_i, Y_i)\}_{i=1}^N$, the average prediction error will be defined as

$$\frac{1}{N}\sum_{i=1}^{N} \ell\left(Y_i, e^{\hat{g}(X_i)}\right) \tag{13}$$

and the average prediction accuracy will be defined as

$$\frac{1}{N}\sum_{i=1}^{N} \left(1 - \ell\left(Y_i, e^{\hat{g}(X_i)}\right)\right). \tag{14}$$

It is important to note that p, m, and M reflect operational concerns and hence can vary from hospital to hospital. We are using these parameters to quantitatively answer the question posed above: How often is the predictive model correct? Since the term "correct" is necessarily subjective, there is no "optimal" choice for these parameters. Based on input from nurses and physicians at LPCH, we have chosen $p = 0.2$, $m = 15$ min, and $M = 60$ min. Most procedures are LPCH; the operational management team can typically adapt to a 20% error in predicted surgical duration. We have $m = 15$ min because at LPCH, delays that are less than

[4] Technical note: One might think that because the logarithm is continuous, we can find some $\delta(p) > 0$ such that

$$|Y - \hat{Y}| < \delta(p) \implies |\log(Y) - \log(\hat{Y})| < \epsilon(p) \tag{11}$$

which would suggest that minimizing $(Y - \hat{Y})^2$ will also likely give us $\ell(Y, \hat{Y}) = 0$. However, no such $\delta(p)$ exists that is independent of both Y and $\hat{Y}$. This is because $x \mapsto \log(x)$ is not uniformly continuous on $(0, \infty)$. See [25, Chapter 4] for more details regarding uniform continuity.

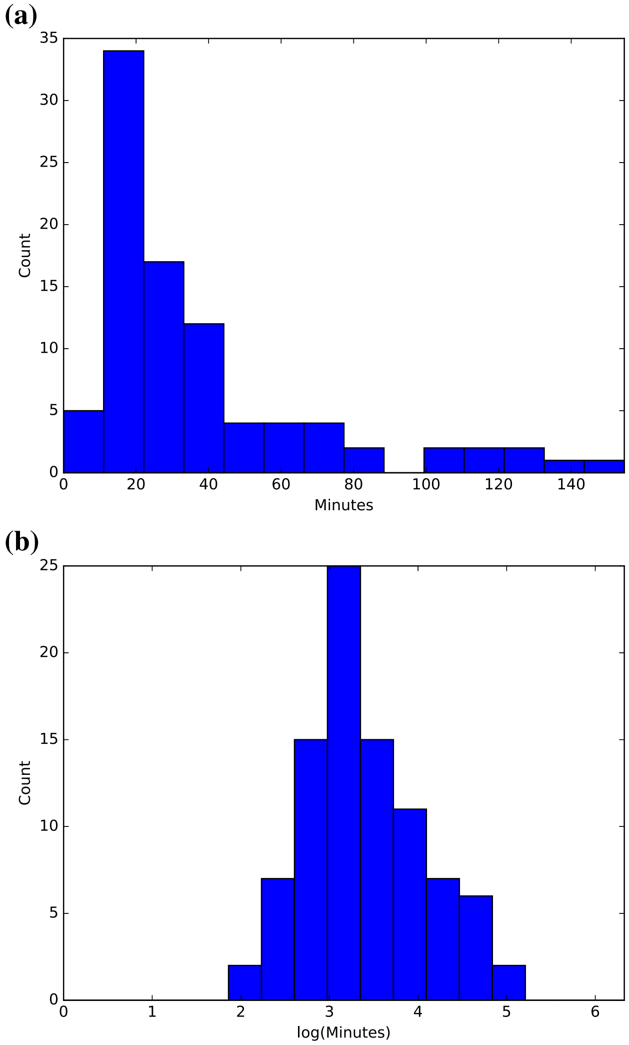

Fig. 2 A demonstration of the effect of a logarithmic transformation on surgical procedure durations. **a** A histogram of the durations (in minutes) of lumbar punctures (with intrathecal chemotherapy). Note the heavy tail and the positive skew. **b** A histogram of the natural logarithm of durations (in minutes) of lumbar punctures (with intrathecal chemotherapy). Note that the histogram is fairly symmetric

15 min are generally acceptable. We have $M = 60$ min because there are some procedures that take several hours (e.g. organ transplants) and in these cases a 20% prediction error could be well over an hour and would lead to delays. These parameter choices are subjective so we will later explore how sensitive our empirical results are to these parameters.

We will also empirically evaluate the impact of the logarithmic transformation by considering the effect of not applying the transformation. In this case, the average prediction accuracy is given by

$$\frac{1}{N}\sum_{i=1}^{N}\left(1-\ell\left(Y_i, \hat{f}(X_i)\right)\right). \tag{15}$$

3.3 The detection problem

The loss function proposed above also gives us a natural formulation for the detection problem. Suppose $\hat{Y}$ is a prediction provided by a surgeon and Y is the realized outcome. We can provide a label W that indicates overestimates, underestimates, and correct prediction as follows:

$$W = \begin{cases} \text{underestimate}, & \hat{Y} < Y - \tau(\hat{Y}) \\ \text{overestimate}, & \hat{Y} > Y + \tau(\hat{Y}) \\ \text{correct}, & \text{otherwise} \end{cases}$$

If we have a data set $\{(X_i, Y_i)\}_{i=1}^{N}$, we can label the observations so that we have $\{(X_i, W_i)\}_{i=1}^{N}$. We can use the labels for the purposes of building a classifier that will detect underestimates and overestimates.

There are a variety of ways of comparing different classifiers. The misclassification rate is an important metric but because of the different operational impacts of each class, we also consider the following metrics:

- For the “correct” class, we want to know the *true negative rate* (*TNR*) and the *negative predictive value* (*NPV*). TNR gives us an estimate of the probability that an inaccurate prediction is detected. NPV gives us an estimate of the probability that a prediction is inaccurate given that it is classified as such. We want both TNR and NPV to be high. These two values allow us to assess how well a classifier can inform decisions regarding with the expert prediction is correct versus incorrect.
- For the “underestimate” class and the “overestimate” class, we want to know the *true positive rate* (*TPR*) and the *positive predictive value* (*PPV*). TPR gives us an estimate of the probability that an underestimate/overestimate is detected and classified as an underestimate/overestimate. PPV gives an estimate of the probability that a prediction is an underestimate/overestimate given that it is classified as such. While the previous metrics focus on a classifiers performance with respect to correct predictions, TPR and PPV highlight performance with respect to underestimates and overestimates.

4 Predicting surgical durations

In this section, we focus on the prediction problem. We describe two benchmark methods that model the current state of practice (i.e. historical averaging by surgeon and expert prediction) and propose several tree-based prediction models. Some of these models provide an automated prediction: they use features that are available in electronic records and they do not require input from surgeons. Other models provide what we refer to as a “semi-automated” prediction: they

Table 1 Definition of the American Society of Anesthesiologists (ASA) physical status classification system as described in [1]

ASA classification	Definition	Examples
ASA I	A normal healthy patient	Healthy, non-smoking, no/minimal alcohol use
ASA II	A patient with mild systemic disease	Smoking, social alcohol drinker, obesity
ASA III	A patient with severe systemic disease	Active hepatitis, alcohol dependence/abuse, morbid obesity
ASA IV	A patient with severe systemic disease that is a constant threat to life	Ongoing cardiac ischemia or severe valve dysfunction, sepsis
ASA V	A moribund patient who is not expected to survive without the operation	A ruptured abdominal/thoracic aneurysm, massive trauma, intracranial bleed with mass effect
ASA VI	A declared brain-dead patient whose organs are being removed for donor purposes	

use features that are available in electronic records but they also take advantage of input from surgeons. We causally divide our data into a training set and a testing set to compare the performance of these models for several pediatric surgical procedure types. Averaged over all procedures, the automated ensemble methods outperform expert prediction. Given the fact that surgeons have extensive medical expertise and can potentially use a patient's entire medical record, it is remarkable that we can achieve this high level of performance without relying on complicated features. Augmenting these ensemble methods with input from surgeons further improves performance.

4.1 Benchmark prediction methods

As described in Sect. 2.1, there are currently two options for predicting surgical case durations:

1. *Historical averaging* If a surgeon is planning on performing a particular surgical procedure, we take the average (arithmetic mean) time the surgeon took to perform that particular procedure in the training set and use this value as the prediction. In the case that a surgeon has not performed this particular surgery at least five times, we use the average of all surgeons' past times as the prediction. We refer to this prediction method as **AVG**.
2. *Expert predictions* A surgeon (perhaps with the assistance of his staff) gives an expert prediction of how long a surgical procedure will last. This amount of time is recorded in the data set. Since these are the predictions that were used to make actual scheduling decisions, we refer to this prediction method as **SCH**.

4.2 Tree-based prediction methods

Given our critique in Sect. 2.2 of the regression tree method used in [4], we propose three tree-based automated prediction methods. Motivated by the discussion in Sect. 3, for each of these methods, we perform the prediction in log-space and transform the result back to a linear scale by exponentiating.

Each of the proposed models uses the following features:

- Gender of the patient (male vs. female)
- Weight of the patient (in kilograms)
- Age of the patient (in years)
- American Society of Anesthesiologists (ASA) physical status/score of the patient as described in Table 1[5]
- Primary surgeon identity
- Location (in an OR vs. in the APU)
- Patient class (in-patient vs. out-patient)
- Procedure name

We did not "mine" our data to choose these features; each of these features is motivated by our domain knowledge. For example, the first four features (gender, weight, age, and ASA score) provide a crude summary of the patient's clinical state. Although previous studies [4] reported that the surgeon identity was not useful, conventional wisdom suggests that surgeon identity is useful and so we opt to include it as a feature. The location and patient classification provide some basic information about the expected complexity of the procedure—procedures performed in the APU are typically shorter and simpler; out-patient procedures also tend to be less complex.

[5] In our data set, there were no patients with ASA VI.

The procedure name has obvious predictive power but is actually quite nuanced. The procedure display names that are currently used for operational purposes do not necessarily fully distinguish different procedures. For example, several cases in our data set were scheduled with the procedure name "Radiation Treatment." This name does not include the type of radiation treatment (i.e. internal vs. external) or the part of the body. Current Procedural Terminology (CPT) codes provide a detailed and standardized way of describing procedures. Although a set of potential CPT codes is known to the surgeon *ex ante*, the particular CPT code used is only recorded *ex poste*. Consequently, we rely on the procedure display name rather than CPT code as a feature.

Each of the proposed models is based on regression trees. The simplest model is a single decision tree regressor [7], denoted **DTR**. We also consider ensembles of trees. In particular, we use a random forest regressor [6], denoted **RFR**, and an ensemble of gradient boosted regression trees [15], denoted **GBR**. For each of these methods, we rely on the implementations provided by the `scikit-learn` package [23].[6] Note that while **DTR** may seem like the same method that was used (unsuccessfully) in [4], recall that we are fitting our trees in log-space and we are also measuring performance according to an alternative criterion.

DTR, **RFR**, and **GBR** each provide an automated prediction method: The aforementioned features are easily pulled from electronic medical records and can be plugged into the learned models. However, we can also use these methods in a semi-automated fashion. In addition to the aforementioned features, we can also use the prediction provided by the surgeon as a feature. The idea is that the surgeon can still provide expert input and the model can use the other features to adjust the expert prediction. Since the expert prediction is the output of **SCH**, we refer to **DTR-SCH**, **RFR-SCH**, and **GBR-SCH** as **DTR**, **RFR**, and **GBR** with the additional feature of the expert prediction. The potential benefit of this approach is improved prediction accuracy, but we immediately lose the benefits of automation. Another downside of this approach is that surgeon behavior may adapt and the model may need to be more frequently re-trained. This semi-automated approach is just one way of incorporating expert information into a statistical model; we discuss other approaches in Sect. 6.

4.3 Prediction results

Our data set includes all surgical procedures performed at LPCH from May 5, 2014, through January 11, 2016. This data set includes 4898 unique procedure names, but to avoid the small sample problems in previous work [4] we consider only the 10 most common. Although this limits the breadth of our study, it also focuses our study on procedure types with the most significant operational impact. We causally split the data into a training set and a testing set: observations that precede April 5, 2015, are used for training and the remaining observations are used for testing. This cutoff was chosen so that the training and testing sets are roughly equal in size while having a nontrivial number of observations of each procedure in both the training set and the testing set. Our previous work [21,36] used k-fold cross-validation to evaluate our different methods. This removed the possibility of concept drift and hence was an optimistic view. By causally splitting the data, we present a more realistic view of our proposed models.

Descriptive statistics (of the entire data set) are given in Table 2. Note that for each procedure (and also overall), the sample median is less than the sample mean. This is consistent with the discussion above regarding the heavy-tailed distributions that are typically used to model surgical durations. This further motivates the logarithmic transformation used in our predictive models.

We use the testing set to estimate the average prediction accuracy for each method and also provide a breakdown based on each procedure name; the results are shown in Table 3. We use the shorthand Acc(Method 1) to denote the estimated mean prediction accuracy of Method 1. When Acc(Method 2) > Acc(Method 1), we say that Method 2 outperforms Method 1. Overall, we see that

$$\begin{aligned}\text{Acc}\,(\mathbf{DTR}) &< \text{Acc}\,(\mathbf{AVG}) = \text{Acc}\,(\mathbf{DTR\text{-}SCH})\\ &< \text{Acc}\,(\mathbf{SCH}) < \text{Acc}\,(\mathbf{RFR})\\ &< \text{Acc}\,(\mathbf{RFR\text{-}SCH})\\ &< \text{Acc}\,(\mathbf{GBR}) = \text{Acc}\,(\mathbf{GBR\text{-}SCH}). \qquad (16)\end{aligned}$$

Although **DTR** does not outperform either benchmark, **RFR** and **GBR** both outperform the benchmarks. By including expert predictions as a feature to these methods, we increase prediction accuracy with the semi-automated prediction models **DTR-SCH** and **RFR-SCH** outperforming their automated counterparts. **GBR-SCH** and **GBR** have the same overall performance which suggests that the additional expert information is not necessary for achieving high prediction accuracy.

Given that **GBR** and **GBR-SCH** are the two best performing prediction methods overall, let us now consider how they compare to **AVG** and **SCH** on each individual proce-

[6] To use these implementations we apply a "one-hot" encoding to all categorical features. By embedding the categorical variables into Euclidean space, we implicitly restrict categorical splits but note that the CART [7] algorithm still does not impose a metric on the feature space. Other methods (e.g. kernel regression) require a metric which can be somewhat artificial for categorical features. We discuss this issue more in Sect. 7.
Each of these methods has a handful of parameters that could be tuned. Because of our relatively small sample size, we opt to be less aggressive with parameter tuning and simply use the default settings.

Table 2 Descriptive statistics for each procedure

	Sample size (training/testing)	Mean	Standard deviation	Median	IQR
Overall	1713 (820/893)	44.13	37.06	31.00	36.00
Lumbar puncture (LP) with intrathecal chemotherapy	90 (51/39)	38.50	31.16	26.50	24.75
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	281 (142/139)	47.83	34.80	33.00	49.00
Tonsillectomy and adenoidectomy	455 (189/266)	30.32	17.55	26.00	15.00
Dental rehabilitation	114 (72/42)	119.04	42.48	115.00	52.75
Laparoscopic appendectomy	146 (74/72)	54.52	15.82	52.00	17.00
Myringotomy with tubes	104 (94/10)	38.96	44.71	21.00	30.50
Adenoidectomy	96 (38/58)	25.78	21.55	20.00	18.50
Bilateral myringotomy with tubes	212 (51/161)	28.33	35.81	16.00	18.25
Colonoscopy with biopsy	112 (62/50)	68.42	27.27	65.00	32.25
Portacath removal	103 (47/56)	30.82	11.40	29.00	11.00

Table 3 Estimated average prediction accuracy (based on accuracy metric discussed in Sect. 3)

	AVG	SCH	DTR	RFR	GBR	DTR-SCH	RFR-SCH	GBR-SCH
Overall	0.61	0.64	0.58	0.66	0.73	0.61	0.69	0.73
Lumbar puncture (LP) with intrathecal chemotherapy	0.26	0.67	0.67	0.64	0.62	0.51	0.62	0.64
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	0.27	0.58	0.42	0.47	0.53	0.55	0.58	0.60
Tonsillectomy and adenoidectomy	0.86	0.71	0.67	0.81	0.86	0.66	0.80	0.83
Dental rehabilitation	0.38	0.36	0.36	0.36	0.33	0.33	0.29	0.40
Laparoscopic appendectomy	0.75	0.19	0.50	0.60	0.79	0.60	0.57	0.67
Myringotomy with tubes	0.80	0.90	0.30	0.80	0.80	0.60	0.90	0.90
Adenoidectomy	0.76	0.74	0.76	0.84	0.86	0.60	0.81	0.83
Bilateral myringotomy with tubes	0.48	0.81	0.58	0.65	0.78	0.71	0.79	0.86
Colonoscopy with biopsy	0.36	0.42	0.34	0.30	0.50	0.52	0.40	0.44
Portacath removal	0.93	0.79	0.79	0.80	0.93	0.59	0.70	0.77

dure. First note that for each for 9 of the 10 procedures, **GBR** outperforms **AVG**. For the one procedure that **AVG** outperforms **GBR**, the difference is small (0.05). Moreover, when **GBR** outperforms **AVG**, the difference is often substantial: For both LPs and EGDs the accuracy of **GBR** is roughly double the accuracy of **AVG**. Consequently, it seems that **GBR** could serve as a useful automated alternative to **AVG**. Similarly, we see that **GBR-SCH** outperforms **SCH** for 8 out of 10 procedures. For the two procedures that **SCH** outperforms **GBR-SCH**, the difference is again quite small ($\leq$0.03). As a result, it seems that **GBR-SCH** could plausibly be used to augment expert predictions.

We also note that while **SCH** outperforms **AVG** overall and one might think that an expert would always outperform a strategy as simple as historical averaging, Table 3 shows that this is not the case. Laparoscopic appendectomies provide an extreme example of this: **AVG** accurately predicts the durations more than 3 times as often as **SCH**. As a more modest example, we see that **AVG** also outperforms **SCH** for portacath removals. Of course there are procedures for which **SCH** drastically outperforms **AVG**: For EGDs, **SCH** is more than twice as accurate as **AVG**.

Now instead of considering prediction accuracy as defined in Sect. 3, consider the estimated prediction R^2 (in linear prediction space) and recall our reasoning for why R^2 was an imperfect performance metric. The results in Table 4 demonstrate why this is the case. Let R^2(Method 1) denote the estimated prediction R^2 for Method 1. Overall we see that

$$\begin{aligned} R^2(\textbf{DTR}) &< R^2(\textbf{AVG}) < R^2(\textbf{RFR}) \\ &< R^2(\textbf{DTR-SCH}) < R^2(\textbf{GBR}) \\ &< R^2(\textbf{SCH}) \\ &< R^2(\textbf{RFR-SCH}) < R^2(\textbf{GBR-SCH}). \end{aligned} \tag{17}$$

Table 4 Estimated prediction R^2

	AVG	**SCH**	**DTR**	**RFR**	**GBR**	**DTR-SCH**	**RFR-SCH**	**GBR-SCH**
Overall	0.34	0.46	0.28	0.38	0.44	0.42	0.57	0.61
Lumbar puncture (LP) with intrathecal chemotherapy	−0.05	0.33	−0.01	0.09	−0.09	0.03	0.25	0.34
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	−0.04	0.42	−0.62	−0.06	−0.03	0.18	0.38	0.43
Tonsillectomy and adenoidectomy	0.05	0.17	−0.30	−0.03	0.09	0.17	0.38	0.34
Dental rehabilitation	0.04	−0.35	−0.58	−0.41	−0.14	−0.39	−0.14	0.06
Laparoscopic appendectomy	−0.03	−7.20	−1.98	−0.50	0.00	−1.91	−1.39	−0.90
Myringotomy with tubes	−0.06	0.89	0.74	0.73	0.81	0.83	0.96	0.96
Adenoidectomy	−0.01	0.49	0.43	0.01	0.38	0.13	0.53	0.50
Bilateral myringotomy with tubes	−0.03	0.78	0.35	0.50	0.50	0.64	0.79	0.77
Colonoscopy with biopsy	−0.10	−0.50	−0.30	−0.59	−0.21	−0.53	−0.33	−0.17
Portacath removal	−0.30	−1.13	−1.79	−0.66	−0.06	−9.46	−4.08	−1.58

Qualitatively, this ordering has some similarities with the ordering for Acc(·): **DTR** and **AVG** perform poorly; the semi-automated methods outperform their automated counterparts; **GBR-SCH** is the best performing prediction methods. However, the story becomes less clear when we breakdown the performance by procedure. For example, although $R^2(\textbf{AVG}) > 0$, the prediction R^2 when conditioned on procedure is negative or close to zero for most procedures. Recall that prediction R^2 is defined as

$$1 - \frac{\text{Prediction MSE}}{\text{Variance of the response}}. \tag{18}$$

Because the variance conditioned on the procedure type can vary drastically, it is difficult to use R^2 to compare performance per procedure. More importantly, it is not clear what different values of R^2 indicate operationally.

Finally, we will consider the impact of the logarithmic transformation. For each of the six prediction methods, we use the suffix "**-LIN**" to denote the corresponding method *without* the logarithmic transformation. The results are given in Tables 5 and 6. First consider the performance using the performance metric from Sect. 3. If we compare the results in Table 5 to the results in Table 3, we see that while the logarithmic transformation generally increases performance across the board, it is most useful for the automated methods. As discussed earlier, the logarithmic transformation is useful not only for addressing the operational metric but also for handling the skew of the underlying distribution. Our results show that the expert information used by the semi-automated methods is equally useful for handling the tail behavior. Now consider R^2 as a performance metric. If we compare Table 6 to Table 4, we see that the logarithmic transformation degrades the performance measured by R^2. This is not surprising since R^2 corresponds to mean-squared loss.

4.4 Feature importance

Because we are using tree-based methods, we can also use the mean decrease in risk across splits as a heuristic for relative feature importance [7]. For each method, this heuristic provides a non-negative score for each feature with these scores summing to one (although there are some round-off errors). The results are shown in Table 7. Note that because the semi-automated methods have an additional feature, the relative importance scores of the automated methods should not be compared directly to the relative importance scores of the semi-automated methods.

First consider the automated methods. For **DTR**, **RFR**, and **GBR**, the procedure name, patient weight, and primary surgeon identity are the most important features. Procedure name and primary surgeon identity are basic pieces of information that have obvious predictive value; indeed, this is why historical averaging is currently so common. This contradicts the conclusion in [4] that surgeon identity is not a useful feature.

It may be surprising that patient weight is such an important feature, but we offer two explanations. We first note that age is typically used to indicate developmental status in children and weight correlates strongly with age; in our data the Pearson correlation coefficient between weight and age is 0.84. This also explains why age does not have as high an importance score. Secondly, we note that childhood obesity is an increasingly pervasive public-health issue [13] and obesity is known to lead to complications during surgery [9]. These ideas are supported by Fig. 3 which shows patient weight as a function of age.[7] Figure 3 shows a strong cor-

[7] It may seem that the oldest patients are too old for a pediatric hospital but it actually common for patients to continue seeing the same physicians into early adulthood.

Table 5 Estimated average prediction accuracy (based on accuracy metric discussed in Sect. 3) for each method *without* the logarithmic transformation

	DTR-LIN	RFR-LIN	GBR-LIN	DTR-SCH-LIN	RFR-SCH-LIN	GBR-SCH-LIN
Overall	0.56	0.63	0.64	0.62	0.71	0.72
Lumbar puncture (LP) with intrathecal chemotherapy	0.49	0.56	0.46	0.56	0.62	0.56
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	0.47	0.40	0.39	0.57	0.60	0.61
Tonsillectomy and adenoidectomy	0.62	0.75	0.85	0.69	0.81	0.83
Dental rehabilitation	0.40	0.33	0.40	0.36	0.48	0.52
Laparoscopic appendectomy	0.53	0.74	0.81	0.56	0.57	0.64
Myringotomy with tubes	0.80	0.80	0.80	0.70	0.80	0.90
Adenoidectomy	0.69	0.74	0.78	0.72	0.79	0.83
Bilateral myringotomy with tubes	0.58	0.64	0.46	0.73	0.80	0.78
Colonoscopy with biopsy	0.34	0.38	0.44	0.32	0.52	0.42
Portacath removal	0.64	0.73	0.88	0.61	0.71	0.75

Table 6 Estimated prediction R^2 for each method *without* the logarithmic transformation

	DTR-LIN	RFR-LIN	GBR-LIN	DTR-SCH-LIN	RFR-SCH-LIN	GBR-SCH-LIN
Overall	0.19	0.42	0.48	0.43	0.60	0.63
Lumbar puncture (LP) with intrathecal chemotherapy	−0.05	0.17	0.09	0.07	0.36	0.42
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	−0.53	−0.06	0.04	0.20	0.42	0.47
Tonsillectomy and adenoidectomy	−0.61	−0.06	0.13	0.24	0.36	0.43
Dental rehabilitation	−0.45	−0.11	0.03	−0.53	0.19	0.13
Laparoscopic appendectomy	−1.10	−0.31	0.07	−1.31	−1.53	−1.06
Myringotomy with tubes	0.87	0.97	0.96	0.89	0.93	0.97
Adenoidectomy	−0.60	−0.01	0.17	0.25	0.46	0.51
Bilateral myringotomy with tubes	0.34	0.53	0.50	0.65	0.75	0.75
Colonoscopy with biopsy	−0.51	−0.38	0.00	−0.59	−0.17	−0.07
Portacath removal	−3.75	−1.32	−0.75	−9.71	−4.57	−2.59

Table 7 Relative feature importance for the prediction models

	DTR	RFR	GBR	DTR-SCH	RFR-SCH	GBR-SCH
Gender	0.02	0.02	0.00	0.01	0.01	0.00
Weight	0.20	0.18	0.16	0.09	0.08	0.13
Age	0.08	0.08	0.07	0.05	0.04	0.09
ASA score	0.04	0.03	0.03	0.02	0.02	0.03
Primary surgeon	0.18	0.17	0.35	0.09	0.09	0.33
Location	0.02	0.02	0.04	0.02	0.01	0.01
Patient class	0.12	0.12	0.05	0.04	0.04	0.04
Procedure name	0.36	0.37	0.30	0.06	0.07	0.12
Expert prediction	N/A	N/A	N/A	0.63	0.63	0.29

relation between age and weight, but it also shows that the distribution of weights is positively skewed, particularly for teenage patients.

We can also gain insights about the features with low relative importance. Recall that location and patient class are included as features because they contain some information

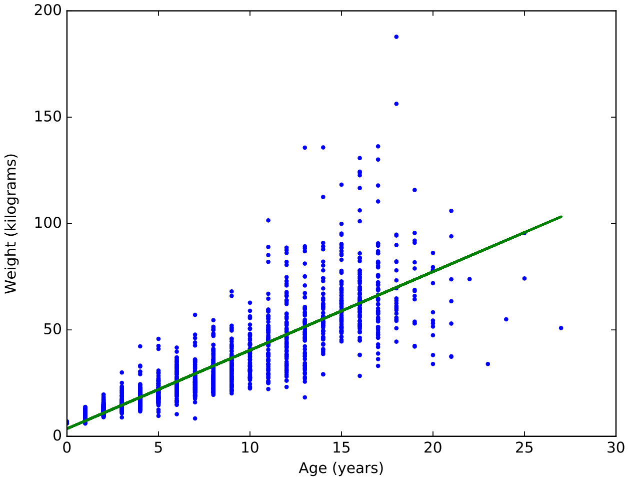

Fig. 3 Weight versus age. A scatter plot along with the ordinary least squares regression line

about the complexity of the operation. Although each of these features has fairly low importance, for **RFR** and **DTR** the combined importance of these features is comparable to the importance of patient weight. This suggests that location and patient class are fairly effective proxies for procedure complexity. We see that the patient ASA score has a low relative importance. We conjecture that this is because information encoded in the ASA score is better represented by other features. In particular, Table 1 shows that obesity is part of the ASA score, but this information is better represented by the patient weight. We also see that patient gender has low predictive power. Patient gender is typically not a useful predictor for surgical times; in fact, patient gender was not used in [4].

Now consider the semi-automated methods. We see that expert prediction is by far the most important feature to **DTR-SCH** and **RFR-SCH**. However, the next three most important features for **DTR-SCH** and **RFR-SCH** are procedure name, primary surgeon, and weight. **GBR-SCH** is qualitatively different in that expert prediction has roughly the same importance as primary surgeon identity. Procedure name and patient weight are the next most important features. We note that for all semi-automated methods gender and ASA score are not very important features.

4.5 Sensitivity to the performance metric

Finally, we make a brief comment regarding the performance metric. As noted in Sect. 3, the choice of p, m, and M is inherently subjective and the estimated prediction accuracy of each method is sensitive to these parameters. In Fig. 4, we plot the estimated prediction accuracy as p varies with $m = 15$ and $M = 60$ fixed. We only concern ourselves with $p \leq 0.5$ because prediction errors of more than 50% would never be considered acceptable in a patient scheduling scenario.

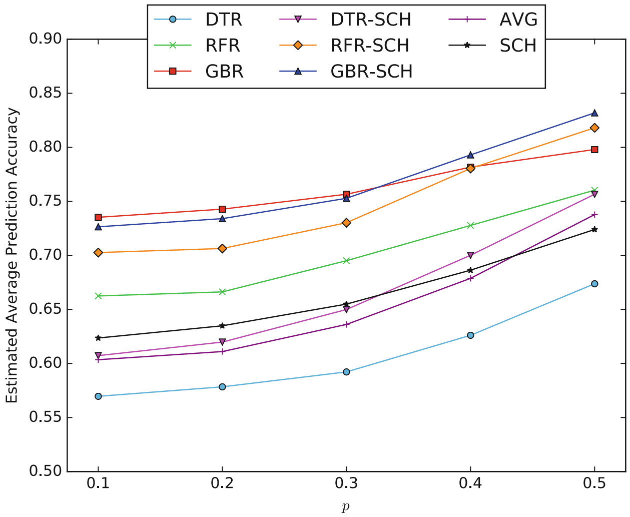

Fig. 4 Estimated overall prediction accuracy versus p. We vary p while holding $m = 15$ and $M = 60$

As p increases, the performance requirements become more lax and the estimated accuracies of all methods increase. We note the following trends that hold for all p:

- **DTR** is by far the least accurate method
- **AVG**, **DTR-SCH**, and **SCH** perform comparably and all outperform **DTR**
- **RFR** outperforms **AVG**, **DTR-SCH**, and **SCH**
- **GBR**, **RFR-SCH**, and **GBR-SCH** perform comparably and all outperform **RFR**

This suggests that although the estimated prediction accuracy depends on the choice of parameters, the general trends that we have noted should hold for a wide range of parameter choices.

5 Detecting inaccurate predictions

In this section, we propose several models for detecting inaccurate predictions including one that directly uses one of the regression models from the previous section. Each detection model uses the same features as the semi-automated prediction models. Observations are labeled as “correct,” “underestimates,” or “overestimates” according to the scheme discussed in Sect. 3.

Before we discuss our detection models, we discuss the labeling scheme. As before we use $p = 0.2$, $m = 15$ (min), and $M = 60$ (min). In Fig. 5, we see how the overall class breakdown is affected by the parameter p. We see that when $p = 0.2$, roughly 60% of the expert predictions are correct and as expected this proportion varies monotonically with p. We also see that when $p = 0.2$, roughly 10% of

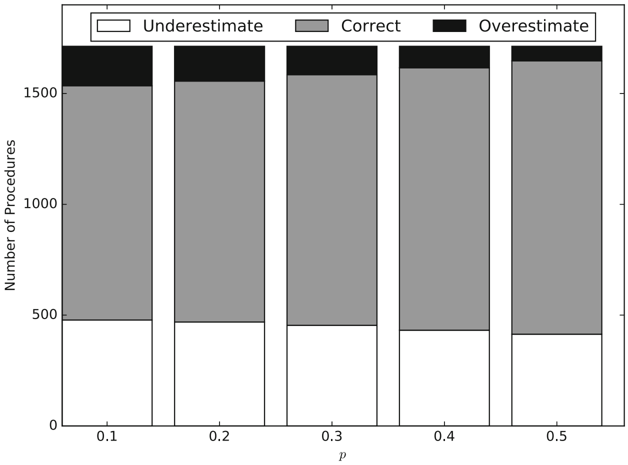

Fig. 5 Classification labels versus p. We vary p while holding $m = 15$ and $M = 60$

predictions are overestimates while roughly 30% are underestimates. This is consistent with our previous remark that surgeons actually have a financial incentive to underestimate surgical durations.

5.1 Detection methods

As discussed earlier, the detection problem is intimately related to the prediction problem and so it is natural to leverage the semi-automated prediction models. Suppose $\tilde{Y}$ is a prediction provided by one of the semi-automated prediction models and $\hat{Y}$ is the corresponding expert prediction. We can define the prediction $\hat{W}$ as follows:

$$\hat{W} = \begin{cases} \text{underestimate}, & \hat{Y} < \tilde{Y} - \tau(\hat{Y}) \\ \text{overestimate}, & \hat{Y} > \tilde{Y} + \tau(\hat{Y}) \\ \text{correct}, & \text{otherwise} \end{cases}$$

Since **GBR-SCH** was the best performing semi-automated prediction method, we will use **GBR-SCH** to provide $\tilde{Y}$. The resulting classification scheme will be referred to as **GBR-SCH-C**.

In addition to **GBR-SCH-C**, we can use the classification equivalents to the regression methods used above. Hence, we consider a decision tree classifier (**DTC**), a random forest classifier (**RFC**), and an ensemble of gradient boosted classification trees (**GBC**). We train these models with the labeled data using the same training set as the semi-automated prediction methods. As before, these tree-based prediction methods are well-suited to the categorical features. We will discuss other methods in Sect. 7.

5.2 Detection results

We now discuss the metrics outlined in Sect. 3: misclassification error, true negative rate (TNR) for the "correct" class, negative predictive value (NPV) for the "correct" class, true positive rate (TPR) for the "underestimate" class, TPR for the "overestimate" class, positive predictive value (PPV) for the "underestimate" class, and PPV for the "overestimate" class. We note that when the metrics are broken down by procedure, values are reported as Not A Number (NAN) whenever the denominator of the expression is zero.

The misclassification error for each detection method is shown in Table 8. Recall from Table 3 that the predictions are accurate 64% of the time, so we would like the overall misclassification error to be no greater than 0.36. **DTC** exactly achieves this baseline error while **RFC** and **GBC** are slightly better. **GBR-SCH-C** has the lowest misclassification error and is the only method with a misclassification error that is less than 0.30. Moreover, **GBR-SCH-C** is the only detection method that achieves a misclassification error that is at most 0.50 for each procedure. For each procedure, we see that either **GBR-SCH-C** outperforms the other methods or is comparable to the best method.

The TNR and NPV for the "correct" class are shown in Table 9. Recall that the TNR for the "correct" class is the empirical probability that an inaccurate prediction is detected while the NPV for the "correct" class is the empirical probability that a prediction is inaccurate given that it is classified as such. We want both of these values of be high, but given the class imbalance, this is difficult to do. Overall, we see that the detection methods with the lower misclassification error tend to have higher NPVs and lower TNRs. Hence, when the more accurate methods detect an inaccurate prediction, we can be confident that the prediction is indeed inaccurate but when a prediction is not classified as inaccurate, we should not be confident that it is necessarily accurate.

The TPRs for the "underestimate" and "overestimate" classes are shown in Table 10. Recall that TPR is the empirical probability that an underestimate/overestimate is detected and classified as an underestimate/overestimate. Overall, we see that all detection methods have fairly low TPR for both the "underestimate" as well as the "overestimate" class. However, we see that for laparoscopic appendectomies all methods have high TPR for the overestimate class—we can very reliably detect overestimates for this procedure.

The PPVs for the "underestimate" and "overestimate" classes are shown in Table 11. Recall that PPV is the empirical probability that a prediction is an underestimate/overestimate given that it is classified as such. Overall we see that **DTC** has the lowest PPVs; **RFC** outperforms **DTC**; **GBC** outperforms **RFC**; and **GBR-SCH-C** outperforms **GBC**. We also see that PPV for the "overestimate" class is generally higher than the PPV for the "underesti-

Table 8 Misclassification error

	DTC	RFC	GBC	GBR-SCH-C
Overall	0.36	0.33	0.30	0.28
Lumbar puncture (LP) with intrathecal chemotherapy	0.36	0.36	0.31	0.31
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	0.43	0.40	0.40	0.37
Tonsillectomy and adenoidectomy	0.36	0.30	0.28	0.29
Dental rehabilitation	0.62	0.48	0.43	0.45
Laparoscopic appendectomy	0.28	0.35	0.22	0.15
Myringotomy with tubes	0.20	0.20	0.00	0.10
Adenoidectomy	0.31	0.24	0.24	0.24
Bilateral myringotomy with tubes	0.27	0.29	0.20	0.19
Colonoscopy with biopsy	0.48	0.54	0.62	0.50
Portacath removal	0.32	0.23	0.21	0.21

Table 9 True negative rate (TNR) and the negative predictive value (NPV) for the “correct” class

	DTC	RFC	GBC	GBR-SCH-C
Overall	0.50, 0.57	0.41, 0.60	0.33, 0.74	0.29, 0.84
Lumbar puncture (LP) with intrathecal chemotherapy	0.23, 0.50	0.08, 0.33	0.23, 1.00	0.08, 1.00
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	0.55, 0.59	0.52, 0.64	0.33, 0.59	0.24, 0.70
Tonsillectomy and adenoidectomy	0.27, 0.38	0.14, 0.44	0.09, 0.70	0.04, 0.60
Dental rehabilitation	0.52, 0.61	0.59, 0.76	0.56, 0.83	0.56, 0.79
Laparoscopic appendectomy	0.79, 0.88	0.81, 0.81	0.78, 0.94	0.84, 0.96
Myringotomy with tubes	1.00, 0.33	0.00, 0.00	1.00, 1.00	0.00, NAN
Adenoidectomy	0.33, 0.42	0.20, 0.60	0.07, 1.00	0.07, 1.00
Bilateral myringotomy with tubes	0.26, 0.30	0.26, 0.28	0.03, 0.50	0.06, 0.67
Colonoscopy with biopsy	0.72, 0.64	0.52, 0.56	0.41, 0.48	0.24, 0.70
Portacath removal	0.67, 0.53	0.00, 0.00	0.08, 0.50	0.00, NAN

Table 10 True positive rate (TPR) for the “underestimate” and “overestimate” classes

	DTC	RFC	GBC	GBR-SCH-C
Overall	0.28, 0.41	0.24, 0.38	0.28, 0.30	0.31, 0.26
Lumbar puncture (LP) with intrathecal chemotherapy	0.50, 0.09	0.00, 0.09	0.00, 0.09	0.00, 0.09
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	0.10, 0.47	0.05, 0.50	0.10, 0.37	0.05, 0.32
Tonsillectomy and adenoidectomy	0.25, 0.21	0.25, 0.14	0.50, 0.05	0.25, 0.01
Dental rehabilitation	0.50, 0.00	0.55, 0.14	0.60, 0.00	0.60, 0.00
Laparoscopic appendectomy	0.32, 0.97	0.32, 0.97	0.37, 0.97	0.58, 0.97
Myringotomy with tubes	NAN, 1.00	NAN, 0.00	NAN, 1.00	NAN, 0.00
Adenoidectomy	0.00, 0.36	0.25, 0.18	0.25, 0.00	0.25, 0.00
Bilateral myringotomy with tubes	0.00, 0.22	0.00, 0.19	0.00, 0.00	0.25, 0.04
Colonoscopy with biopsy	0.50, 0.61	0.17, 0.57	0.00, 0.48	0.00, 0.30
Portacath removal	0.12, 0.00	0.00, 0.00	0.00, 0.25	0.00, 0.00

mate” class. In this sense, when a detection method reports that a prediction is an “overestimate,” it is more likely to be correct than when it reports that a prediction is an “underestimate.”

5.3 Feature importance

As with the prediction methods, we can discuss the relative feature importance for each detection model. The results are

Table 11 Positive predictive value (PPV) for the “underestimate” and “overestimate” classes

	DTC	RFC	GBC	GBR-SCH-C
Overall	0.37, 0.45	0.53, 0.50	0.57, 0.70	0.63, 0.90
Lumbar puncture (LP) with intrathecal chemotherapy	0.50, 0.25	NAN, 0.33	0.00, 0.50	NAN, 1.00
Esophagogastroduedenoscopy (EGD) diagnostic with biopsy	0.15, 0.44	0.33, 0.43	0.22, 0.61	0.20, 0.80
Tonsillectomy and adenoidectomy	0.14, 0.31	1.00, 0.42	1.00, 0.50	0.33, 0.50
Dental rehabilitation	0.53, 0.00	0.58, 0.50	0.67, NAN	0.63, NAN
Laparoscopic appendectomy	0.60, 0.90	0.46, 0.84	0.70, 1.00	0.85, 1.00
Myringotomy with tubes	NAN, 0.33	NAN, 0.00	NAN, 1.00	NAN, NAN
Adenoidectomy	NAN, 0.33	1.00, 0.55	1.00, NAN	1.00, NAN
Bilateral myringotomy with tubes	0.00, 0.24	NAN, 0.17	NAN, 0.00	1.00, 0.50
Colonoscopy with biopsy	0.31, 0.61	0.33, 0.54	0.00, 0.46	0.00, 0.78
Portacath removal	0.50, 0.00	NAN, 0.00	NAN, 0.50	NAN, NAN

Table 12 Relative feature importance for the detection models

	DTC	RFC	GBC	GBR-SCH-C
Gender	0.03	0.05	0.01	0.00
Weight	0.21	0.21	0.16	0.13
Age	0.12	0.14	0.08	0.09
ASA score	0.05	0.04	0.03	0.03
Primary surgeon	0.20	0.18	0.35	0.33
Location	0.02	0.03	0.02	0.01
Patient class	0.03	0.05	0.03	0.04
Procedure name	0.16	0.13	0.08	0.12
Expert prediction	0.19	0.18	0.23	0.29

given in Table 12; note that the relative feature importance for **GBR-SCH-C** is the same as that for **GBR-SCH**. In general, we see that the most important features for all of the methods are the expert prediction, the procedure name, the primary surgeon, the patient weight, and the patient age. Based on the relative feature importance of the prediction models, the intuition behind these features is clear.

Although all of the detection models heavily rely on these features, since **GBC** and **GBR-SCH-C** outperformed **RFC** and **DTC**, it is interesting to see the differences between the feature importances. In particular, we see that **GBC** and **GBR-SCH-C** rely more heavily on primary surgeon and the expert prediction than **DTC** and **RFC**. This suggests that better detection models rely more heavily on idiosyncrasies of particular surgeons and/or their teams. A downside of this is that the performance of these detection models may not be robust to personnel changes in the hospital. On the other hand, **DTC** and **RFC** rely more heavily on patient age and patient weight. Although these detection models do not perform quite as well as **GBC** and **GBR-SCH-C**, since they rely more heavily on patient characteristics, their performance may be more robust to changes in hospital staff.

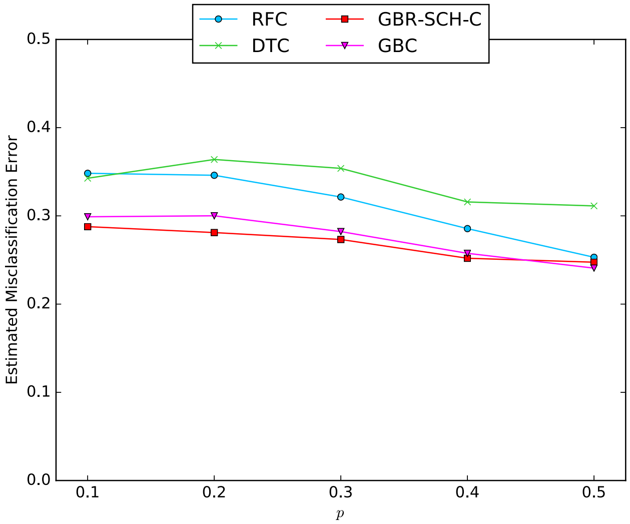

Fig. 6 Misclassification error versus p. We vary p while holding $m = 15$ and $M = 60$

5.4 Sensitivity to the performance metric

The labeling scheme outlined in Sect. 3 depends on the parameters p, m, and M. As a result, these parameters directly affect the training process for **DTC**, **RFC**, and **GBC**. In contrast, p, m, and M affected the evaluation of the prediction models but not the training. To demonstrate that our conclusions are robust to the parameters, we plot the misclassification error as a function of p for each detection method. The results are shown in Fig. 6. We see that for each p, **GBR-SCH-C** is typically slightly better than **GBC**; **GBC** outperforms **RFC**; and **RFC** generally outperforms **DTC**. For the smallest and largest p, this ranking is not preserved but we see that the best performing detection method is always either **GBC** or **GBR-SCH-C** while the worst is always **DTC** or **RFC**.

6 Directions of future work

Our current work suggests some new directions. Although we chose to focus on tree-based methods because of our categorical features, there are other nonparametric methods (e.g. nearest neighbor regression and kernel regression [2] and support vector machines [10]) that are also worth exploring. An obstacle for applying these methods is that they rely on the feature space being endowed with a metric. While we can heuristically embed categorical features into Euclidean space with a scheme like one-hot encoding, the resulting metric is not very meaningful. For example, our study considers both myringotomies and bilateral myringotomies. We would expect these procedures to be related and hence "closer" than unrelated procedures, but the Euclidean metric applied to a one-hot encoded feature space does capture this intuition. This suggests that metric learning [17] should be applied to make use of this intuition. Metric learning can often be used to determine a transformation of the feature space (e.g. [33]), and studying this transformation can potentially give us additional insights into the features space. This would allow us to quantitatively understand what it means for different procedures (or different surgeons) to be similar.

There are also many other methods of incorporating expert opinions into our models. We elected to use expert predictions as a feature for frequentist methods, but Bayesian methodologies could provide a broader framework for incorporating expert knowledge from surgeons and nurses. Bayesian methods can also help us deal with smaller samples sizes. This can help us broaden the applicability of our results to procedures that are less common.

In addition to considering different methods, it may also be useful to consider different features. Our current feature set is intentionally generic: The features that we consider can be used at any pediatric hospital and for any procedure, giving our models broad applicability. However, it may be worth sacrificing this broad applicability to use more specific features that yield improved prediction accuracy. For example, in teaching hospitals it is known that having a resident in the OR will lead to longer surgeries [8,32]. For specific procedures, it may be useful to have more detailed clinical information about the patient. Feature engineering can be an open-ended process, but more extensive feature engineering is likely to improve the predictive power of our models.

7 Conclusions

Motivated by operational problems in hospitals, we have studied the problem of building prediction models for pediatric surgical case durations. We have proposed a novel performance metric for prediction in this application. Not only does this performance metric capture issues relevant to hospital operations, it also motivates a nonlinear transformation of the data. We have also proposed a related classification problem that aims to detect inaccurate expert predictions. We demonstrate that contrary to the medical literature, our prediction models outperform currently used algorithms and are often on par with human experts. When we take advantage of expert opinions, our models can significantly outperform surgeons. We also present empirical evidence that our detection models could form the basis for decision support tools to assist experts when making predictions and scheduling surgeries. These positive results point to new directions of research that will ultimately enable automated and semi-automated prediction methods to be deployed in pediatric hospitals.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

1. American Society of Anesthesiologists—ASA Physical Status Classification System. https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system. Accessed 06 April 2016
2. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992)
3. Bland, J., Altman, D., Rohlf, F.: In defence of logarithmic transformations. Stat. Med. **32**(21), 3766 (2013)
4. Bravo, F., Levi, R., Ferrari, L.R., McManus, M.L.: The nature and sources of variability in pediatric surgical case duration. Pediatr. Anesth. **25**(10), 999–1006 (2015)
5. Breiman, L.: Heuristics of instability and stabilization in model selection. Ann. Stat. **24**(6), 2350–2383 (1996)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
8. Bridges, M., Diamond, D.L.: The financial impact of teaching surgical residents in the operating room. Am. J. Surg. **177**(1), 28–32 (1999)
9. Byrne, T.K.: Complications of surgery for obesity. Surg. Clin. N. Am. **81**(5), 1181–1193 (2001)
10. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
11. Dexter, F., Blake, J.T., Penning, D.H., Lubarsky, D.A.: Calculating a potential increase in hospital margin for elective surgery by changing operating room time allocations or increasing nursing staffing to permit completion of more cases: a case study. Anesth. Analg. **94**(1), 138–142 (2002)
12. Dexter, F., Marcon, E., Epstein, R.H., Ledolter, J.: Validation of statistical methods to compare cancellation rates on the day of surgery. Anesth. Analg. **101**(2), 465–473 (2005)
13. Ebbeling, C.B., Pawlak, D.B., Ludwig, D.S.: Childhood obesity: public-health crisis, common sense cure. Lancet **360**(9331), 473–482 (2002)
14. Feng, C., Wang, H., Lu, N., Tu, X.M.: Log transformation: application and interpretation in biomedical research. Stat. Med. **32**(2), 230–239 (2013)

15. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
16. Kayış, E., Khaniyev, T.T., Suermondt, J., Sylvester, K.: A robust estimation model for surgery durations with temporal, operational, and surgery team effects. Health Care Manag. Sci. **18**(3), 222–233 (2015)
17. Kulis, B.: Metric learning: a survey. Found. Trends Mach. Learn. **5**(4), 287–364 (2012)
18. Macario, A.: What does one minute of operating room time cost? J. Clin. Anesth. **22**(4), 233–236 (2010)
19. Macario, A., Dexter, F., Traub, R.D.: Hospital profitability per hour of operating room time can vary among surgeons. Anesth. Analg. **93**(3), 669–675 (2001)
20. Marakas, G.M.: Decision Support Systems in the 21st Century, vol. 134. Prentice Hall, Upper Saddle River (2003)
21. Master, N., Scheinker, D., Bambos, N.: Predicting pediatric surgical durations. arXiv preprint arXiv: 1605.04574 (2016)
22. May, J.H., Strum, D.P., Vargas, L.G.: Fitting the lognormal distribution to surgical procedure times. Decis. Sci. **31**(1), 129–148 (2000)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
24. Powell, C.: The delphi technique: myths and realities. J. Adv. Nurs. **41**(4), 376–382 (2003)
25. Rudin, W.: Principles of mathematical analysis Vol. 3. New York, McGraw-Hill (1964)
26. Shippert, F.R.D.: A study of time-dependent operating room fees and how to save $100000 by using time-saving products. Am. J. Cosmet. Surg. **22**(1), 25–34 (2005)
27. Smallman, B., Dexter, F.: Optimizing the arrival, waiting, and npo times of children on the day of pediatric endoscopy procedures. Anesth. Analg. **110**(3), 879–887 (2010)
28. Spangler, W.E., Strum, D.P., Vargas, L.G., May, J.H.: Estimating procedure times for surgeries by determining location parameters for the lognormal model. Health Care Manag. Sci. **7**(2), 97–104 (2004)
29. Stepaniak, P.S., Heij, C., De Vries, G.: Modeling and prediction of surgical procedure times. Stat. Neerl. **64**(1), 1–18 (2010)
30. Strum, D.P., May, J.H., Vargas, L.G.: Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. J. Am. Soc. Anesthesiol. **92**(4), 1160–1167 (2000)
31. Strum, D.P., Sampson, A.R., May, J.H., Vargas, L.G.: Surgeon and type of anesthesia predict variability in surgical procedure times. J. Am. Soc. Anesthesiol. **92**(5), 1454–1466 (2000)
32. Vinden, C., Malthaner, R., McGee, J., McClure, J., Winick-Ng, J., Liu, K., Nash, D., Welk, B., Dubois, L.: Teaching surgery takes time: the impact of surgical education on time in the operating room. Can. J. Surg. **59**(2), 87 (2016)
33. Weinberger, K.Q., Tesauro, G.: Metric learning for kernel regression. In: AISTATS, pp. 612–619 (2007)
34. Wright, I.H., Kooperberg, C., Bonar, B.A., Bashein, G.: Statistical modeling to predict elective surgery timecomparison with a computer scheduling system and surgeon-provided estimates. J. Am. Soc. Anesthesiol. **85**(6), 1235–1245 (1996)
35. Zenteno, A.C., Carnes, T., Levi, R., Daily, B.J., Price, D., Moss, S.C., Dunn, P.F.: Pooled open blocks shorten wait times for nonelective surgical cases. Ann. Surg. **262**(1), 60–67 (2015)
36. Zhou, Z., Miller, D., Master, N., Scheinker, D., Bambos, N., Glynn, P.: Detecting inaccurate predictions of pediatric surgical durations. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA) (2016)