Seoul Journal of Business
Volume 25, Number 2 (December 2019)
DOI: https://doi.org/10.35152/snusjb.2019.25.2.001

Is Expert Input Valuable? The Case of Predicting Surgery Duration*

ROUBA IBRAHIM**
University College London
London, U.K.

SONG-HEE KIM***
University of Southern California
Los Angeles, U.S.A.

ABSTRACT

Most data-driven decision support tools do not include input from people. We study whether and how to incorporate physician input into such tools, in an empirical setting of predicting the surgery duration. Using data from a hospital, we evaluate and compare the performances of three families of models: models with physician forecasts, purely data-based models, and models that combine physician forecasts and data. We find that combined models perform the best, which suggests that physician forecasts have valuable information above and beyond what is captured by data. We also find that applying simple corrections to physician forecasts performs comparably well.

Keywords: healthcare operations, operating room, predicting surgery duration, expert input, discretion

---

* The authors are grateful to Felipe Osorno, Joan Brown, and Manas Bhatnagar for providing access to the data sets used in this paper and for the helpful discussions.

** Associate Professor, School of Management, University College London. One Canada Square, London E14 5AB, U.K. Email: rouba.ibrahim@ucl.ac.uk.

*** Assistant Professor, Department of Data Sciences and Operations, Marshall School of Business, University of Southern California. Bridge Hall 307A, Los Angeles, CA 90089, U.S.A. Email: songheek@marshall.usc.edu. Phone: +1-213-821-4189.

INTRODUCTION

Using data-driven decision support tools to make operational decisions is becoming increasingly viable for hospitals due to the growing availability of electronic medical record data and hospital operational data (Macario 2010). Most data-driven tools do not ask people (e.g., clinicians) for their input (Berner 2009). People may often be inconsistent and biased, and hence by not incorporating people's input, one can ensure that data-driven tools make consistent and objective decisions. However, people may also observe important information that may remain unobserved in the collected data (Eijkemans et al. 2010). Furthermore, while the idea of data-driven practices is gaining support in healthcare, studies have shown that getting people to actually practice it is difficult (Bates et al. 2003), because of people's lack of confidence in the tools used, and the perception that these tools would reduce the decision-makers' autonomy (Cabana et al. 1999). People may become less resistant to adopting data-driven tools if their input can be effectively incorporated into those tools (Zhou et al. 2016).

In this study, we explore whether and how to incorporate the discretion/expertise/intuition of physicians (hereafter referred to as 'physician input') into data-driven decision support tools for improved operational decision-making in hospitals. We empirically investigate this question by using the surgery duration prediction data. Specifically, we address two research questions. First, does physician input offer predictive power beyond that of purely data-based models in predicting the surgery duration? If so, how can we make good use of it? Second, how much does the answer to the first question depend on the surgery characteristics (e.g., surgical specialty, surgeon, and task)? If the degree of heterogeneity is high, what can we do to improve the overall predictive accuracy?

Our choice of surgery duration prediction as the empirical setting is motivated by three reasons. First, the surgery duration prediction is a real and important problem. Accurately predicting the surgery case duration is directly linked to the efficient use of operating rooms that serve as the hospitals' most profitable as well as most expensive facilities (Cardoen, Demeulemeester, and Belien 2010). In fact, this research was motivated by a discussion with a hospital manager who wanted to replace the surgeons' predictions of surgery duration by those of purely data-based models, in the hope of

improving the hospital's use of operating rooms. When allocating operating room times to surgery cases, hospitals need to know how long the surgeries will take. Although the surgery duration prediction problem has been extensively studied, it still remains a challenge (Macario 2010; Zhou et al. 2016). Inaccurate predictions have a trickle-down effect on daily hospital operations, including wasting patients' and staffs' time and expensive operating room time.

Second, the surgery duration prediction problem offers a clean and objective way of measuring the performance of different approaches, including that of physician input. Unlike in many healthcare problems, where the outcome (e.g., quality of care) can be measured in multiple, and often, subjective ways (McGlynn 1997), the surgery duration prediction problem has a clean outcome measure—the actual surgery duration— to which the predicted values from the various methods can be compared.

Third, "there is at present no conclusive view on whether it is necessary to include the surgeons' subjective knowledge" to predict the surgery duration (Larsson 2013). Studies have shown that expert knowledge may be useful when the problem has structure, the performance can be evaluated via high-quality rapid feedback, and the expert has experienced many repetitions (Kahneman and Klein 2009). The surgeons' prediction of surgery duration meets all of these three conditions. As such, there is potential that the insights gleaned from this research can be used to transform the way in which hospitals predict surgery duration which, in turn, can save expensive operating room times and reduce delays in patient care.

Models of Surgery Duration

In this paper, we consider three families of models for surgery duration: (i) models with physician input, which rely on the physician input, either directly, after a correction, or by using a physician coefficients model; (ii) purely data-based models, which ignore the physician input altogether and rely on historical data only; and (iii) combined models, which incorporate both the physician input and purely data-based models. By considering a wide range of potential models and testing them with data, we provide a comprehensive and rigorous treatment of the surgery

duration prediction problem, which we believe is lacking in the literature.

Models with Physician Input. It is natural to begin our investigation by examining the predictive accuracy of physician input. Although physician input can be inconsistent and biased, it may also have an advantage over purely data-based models, as people are more flexible than purely data-based models in adapting to changing conditions and abnormal cases, and they are also able to evaluate variables that are difficult to measure objectively. In addition, as mentioned above, effectively incorporating physician input may help lower the barriers of physician adherence to data-driven decision support tools.

We consider three different physician input models: (a) a benchmark model where the surgeon's prediction of surgery case duration is used directly as the prediction for surgery duration; (b) a corrected model, which systematically removes the bias in the surgeon's prediction (Theil 1966); and (c) a physician coefficients model, which mimics the management coefficient model (Bowman 1963), where the relationship between the surgeon's prediction and the observed predictor variables are modeled via linear regression. Note that this family of models requires the elicitation of physician input (i.e., the surgeon's prediction of surgery duration) in order to be implemented in practice.

Purely Data-based Models. Our second family of models are purely data-based models, which do not need physician input and utilize historical data only. Purely data-based models have distinct advantages compared to models with physician input: They operate on observed information in a consistent and mechanical manner, and they optimally weigh the evidence. However, one loses the potential benefit of the information provided in physician input. In addition, purely data-based models can perform poorly when there is limited historical data, a known problem in predicting the surgery duration (Macario 2006), which we have also encountered in this paper. We consider two models: (a) a historical average model, which simply averages past surgery durations to predict the future surgery duration; and (b) a regression model, which models the relationship between the surgery duration and observed predictor variables via linear regression.

Combined Models. Our third family of models combines the physician input and purely data-based models. Studies have shown that combining forecasts in an effective manner generally leads to superior predictions compared to any one of the individual inputs (Timmermann 2006). Combined models leverage both the discretion/expertise/intuition in physician input and the consistency and unbiasedness of purely data-based models. In this paper, we first compute statistics that can help determine whether combined models can be useful. We then consider three models: (a) a simple regression combination model, which optimally weighs physician input and the output from the purely data-based model above; (b) a full regression combination model, which models the relationship between surgery duration and the observed predictor variables, including the surgeon's prediction, via linear regression; and (c) a heuristic model, which weighs the physician input and the output from the purely data-based model equally (Blattberg and Hoch 1990).

A Tailored Approach. In this study, we evaluate and compare the performance of various models for different groupings of the data. We do so because we observe considerable heterogeneity in the value of physician input when different groupings are used. The groupings that we consider are (a) surgical specialty, (b) surgeon, (c) procedure type, and (d) surgeon-procedure pair. We demonstrate that the predictive accuracy of the various models that we consider varies substantially depending on the grouping considered. By evaluating the performance for different groupings of the data, we take a tailored approach to the problem rather than a one-size-fits-all approach, and show that our tailored approach can lead to significant performance improvements for some of the models that we consider.

Main Contributions

We summarize our main contributions as follows:

- We propose a host of models to predict surgery duration and empirically compare their performances.
- The heterogeneity in the performance of surgeon's prediction across surgeons and across procedures has been pointed out

in the literature (Eijkemans et al. 2010). However, to the best of our knowledge, we are the first to empirically demonstrate the improvement in predictive accuracy resulting from different groupings of the data.

- Building on a theoretical framework developed in the expert judgment literature (Mincer and Zarnowitz 1969; Blattberg and Hoch 1990), we demonstrate how one can quantify the value of the surgeon's prediction. We then present several ways to leverage the surgeon's prediction. This stands in contrast to earlier papers where typically only a regression model combining the surgeon estimate and other predictor variables (similar to one of our combined models) was used.
- We find that physician input offers predictive power beyond that of purely data-based models in our empirical setting. The best performing model in our setting is the full regression combination model, under the condition that a single regression model is fitted to all surgeries. However, we find that the corrected physician input model performs comparably well when a correction model is fitted to each surgeon- procedure pair. The increase in the mean squared error, in this case, is only 5%. On the other hand, if the physician input is not used, the mean squared error of the best performing model increases by 17%.

LITERATURE REVIEW

We first discuss research that seeks to understand the effect of allowing expert input. Subsequently, we discuss papers that consider combining expert input and analytics-based models as well as relevant papers in operating room management and surgical scheduling.

Understanding the Benefits/Costs of Allowing Expert Input

There is a long line of research that examines the value of expert input in the judgment and decision-making literature. Some studies suggest that expert judgment has little predictive power beyond that of purely data-based models (Dawes, Faust, and Meehl 1989), whereas other studies suggest expert judgment can outperform purely data-based models (Bunn and Wright 1991). There are also studies that show systematized expert input—constructed by regressing expert input on observed covariates, also known as

the management coefficients model or the judgment bootstrapping model—can outperform expert input (Bowman 1963).

In the Operations Management literature, understanding the benefits/costs of allowing expert input in operational decision-making has been attracting interest in various application areas such as horizontal multimarket coordination (Anand and Mendelson 1997), ordering behavior in retail stores (Van Donselaar et al. 2010), and price setting (Phillips, Simsek, and van Ryzin 2015). In healthcare settings, Kim et al. (2015) examine physicians' hospital unit admission decisions and show that allowing physician input in their data-driven decision making can help improve the system's performance. On the other hand, Ibanez et al. (2018) show that when radiologists are allowed to deviate from their prescribed sequence, they deviate in a way that does not necessarily improve system performance. We contribute to this line of research by studying the effect of physician input on the use of operating rooms, and by showing how the physician input can be best leveraged to improve system performance.

Combining Expert Input with Analytics-based Models

Combining multiple forecasts to improve forecasting accuracy has been a popular topic in the statistics and management literature (Timmermann 2006). In general, the existing literature advocates combining forecasts if 1) the information sets for each forecast are not known, 2) the non-overlapping parts of information sets are important (e.g., low correlation in special cases), and 3) the forecasts are based on different loss functions (Timmermann 2006).

This study is most related to a specific category of forecast combinations where the experts' forecasts are combined with the forecasts from purely data-based models. Several papers have examined combining expert input with analytics-based models for the surgery duration prediction problem. Most of these papers combine the surgeons' prediction and analytics-based models by including the surgeons' prediction as a feature in their analytics-based models for predicting the surgery duration (Wright et al. 1996; Eijkemans et al. 2010; Stepaniak et al. 2009). They report that the surgeons' prediction is an important predictor of surgery duration. Zhou et al. (2016) is in the same spirit as this paper: It explores whether and how to involve surgeons in the prediction exercise.

In particular, the authors propose a method to detect whether a surgeon's prediction will overestimate or underestimate the surgery duration, which helps in deciding whether the surgeon's prediction should be used or not. We contribute to this stream of literature by proposing and empirically comparing the performances of a host of models that can be used to combine surgeon prediction with data.

Operating Room Management and Surgical Scheduling

Hospitals employ various methods to predict the surgery duration. Many hospitals rely on surgeons to provide forecasts (Macario 2010; Zhou et al. 2016), while others use the moving average of 5 to 10 previous cases of a similar nature (Ozen et al. 2016) or regression-based models based on the patient's and procedure's characteristics (Eijkemans et al. 2010). Some hospitals combine different methods, but in an unsystematic fashion in which the different methods are not optimally weighed in (Hosseini et al. 2015).

Studies have examined the accuracy of the surgeon's prediction. Laskin, Abubaker, and Strauss (2013) report that overestimating the surgery duration is more common than underestimating it. Travis et al. (2014) find heterogeneity in the degree of bias in the prediction and show that the sign and magnitude of the bias depends on the surgeon and the procedure. Larsson (2013) reports that while historical averages are more accurate than the surgeon's prediction in general, surgeons are better at identifying long cases.

Predicting surgery duration using purely data-based models has also been extensively studied (Strum et al. 2000; Eijkemans et al. 2010). When purely data-based models are used to predict the surgery duration, a major cause of inaccuracy has been found to be the lack of historical data. Macario (2006) reports that 50% of the surgery cases that need surgery duration prediction have less than five previous cases of the same procedure type and the same surgeon during the preceding year. Zhou and Dexter (1998) report that only 32% of their cases had two or more previous occurrences of the same procedure with the same surgeon. In this paper, we focus on surgery cases that have at least 20 previous cases of the same procedure and the same surgeon in the training sample to evaluate our proposed models. We show that even after restricting our sample in this way, the lack of historical data can be a major cause of inaccuracy for some of the models that we propose; in such

case, we show that models with physician input can be used.

EMPIRICAL SETTING

To address our research questions, we use the operating room scheduling and usage data from an academic hospital in a large metropolitan area of the United States. In what follows, we describe the relevant operating process at the hospital, the data, and the variables of interest.

Surgery Scheduling Process

When a physician deems that a surgery is needed, the physician works with a surgical scheduler within the clinical department to schedule a surgery. The departmental surgical scheduler submits an electronic surgery booking slip, in which the details of the surgery, including the surgeon's prediction of the surgery duration, the proposed date and time of the surgery, procedure name(s), patient information, and required pre-operative procedures, have to be entered. The hospital's surgical schedulers then schedule the surgery.

Data

We merge the electronic booking slips data, the patient information data, and the surgery information data of all the patients who had surgery from January 1, 2014 to December 31, 2016 at the study hospital to generate the dataset for this study. During our three-year study period, 24,037 surgery cases were performed in the 24 main operating rooms at the hospital. We removed all cases performed by cardiothoracic surgeons (2,492 cases) because 99.9% of their cases did not have the electronic booking slip data. We removed 5,733 additional cases with missing electronic booking slip data; the primary reason for missing electronic booking slip data was the urgent nature of the cases. We removed 9 surgeries with missing surgeon or procedure and 162 surgeries that lasted less than 15 minutes or longer than 720 minutes.

In this study, we consider various models for predicting the surgery duration, and we empirically compare their performances.

To provide an unbiased evaluation of model fits, we split our remaining data of the 15,641 surgery cases into two sets: the training set including 10,470 cases performed in 2014 and 2015, and the test set including 5,171 cases performed in 2016. In the following sections, we consider estimating separate models for different groups of the surgery cases. The smallest-sized groups are formed when we group surgery cases at the surgeon-procedure pair level. To ensure that we have enough samples to fit the parameters in the test set and that we have enough samples to have confidence in the performance evaluation in the training set, we remove surgeries that belong to the surgeon-procedure pairs with fewer than 20 (10) surgeries in the training (test) set. The resulting data consist of 6,705 surgery cases, 4,341 cases in the training set and 2,364 cases in the test set.

Variables

Surgery duration. We define surgery duration as the time from incision to closure. The average surgery duration was 217.3 minutes (SD 114.1) in the training set and 216.3 minutes (SD 112.0) in the test set (the summary statistics are also provided in table 1).

We next examine the average and the variability of the surgery duration across different groups: (a) by surgeon specialty, (b) by surgeon, (c) by procedure, and (d) by surgeon-procedure pair. Figure 1 shows the average surgery duration with its 95% confidence interval for each group in each of the four groupings; it shows that the average and the variability of surgery duration vary widely across different groups in each of the groupings.

We note that within each surgeon-procedure pair, the normal distribution provides a good fit for the surgery duration distribution. The Shapiro-Francia test, a statistical test for normality, fails to reject the null hypothesis that the surgery duration is normally distributed for 50 pairs (out of 81 pairs) at the 95% confidence level. Hence, we do not apply any data transformation to the surgery duration when we model it using linear regression models.

Surgeon's prediction of surgery duration. The average surgeon's prediction of surgery duration was 212.9 minutes (SD 100.8) in the training set and 215.3 minutes (SD 92.3) in the test set. If we compare only the averages and standard deviations, the surgeons'

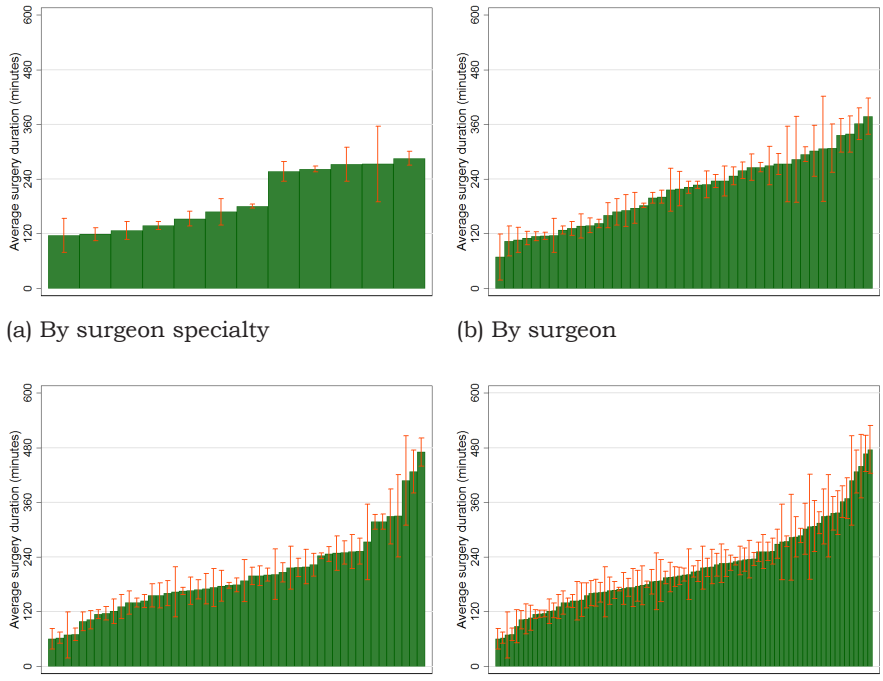
Table 1. Summary Statistics.

Variable	Training set (n=4,341)	Test set (n=2,364)
Surgery duration (minutes)	217.3 (114.1)	216.3 (112.0)
Surgeon estimate (minutes)	212.9 (100.8)	215.3 (92.3)
# Unique specialty	12	12
# Unique surgeon	42	42
# Unique procedure	49	49
# Unique surgeon-procedure pair	81	81
Age	61.5 (13.5)	61.7 (13.5)
Female	42%	42%
Race	7%	7%
Asian	5%	5%
Black	72%	72%
White	16%	16%
Other	3%	2%
ASA Level	44%	42%
0-1	51%	52%
2	3%	3%
3	95%	95%
4-6	7%	7%
Indicator for major anesthesia	8%	7%
Indicator for more than one surgeon		
Indicator for more than one procedure		

Notes: We report averages (standard deviation in parentheses) for continuous variables and percentages for binary or categorical variables. ASA Level is a six-level assessment of the fitness of the patient before surgery measured by the American Society of Anesthesiologists physical status classification system (American Society of Anesthesiologists 2014).

predictions do not seem to differ much from the actual surgery durations.

However, figure 2(a), a scatter plot of surgeon's prediction versus surgery duration for the test set along with a 45-degree line, tells a different story. We observe the presence of significant under-prediction (above the 45-degree line) as well as over-prediction (below the 45-degree line). About 29% (18%) of the surgeries ran over the surgeon's predicted time by more than 30 (60) minutes, and about 32% (19%) of the surgeries took shorter than the surgeon's predicted time by more than 30 (60) minutes. Figure 2(b) shows that the degrees of under-prediction and over-prediction vary widely across surgeon-procedure pairs, implying that there may be value in



(a) By surgeon specialty (b) By surgeon

(c) By procedure (d) By surgeon-procedure pair

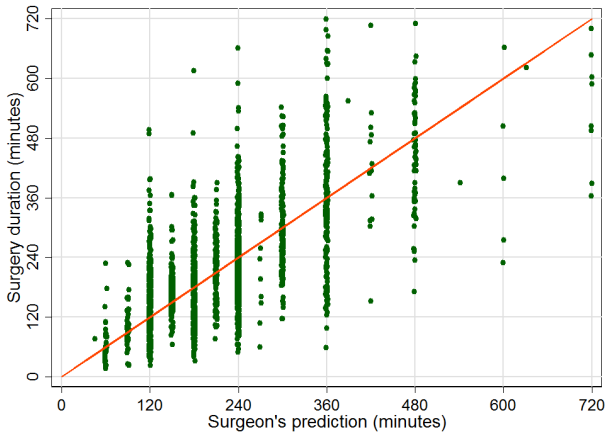
Figure 1. Average surgery duration and its 95% confidence interval by different groupings.

evaluating the surgeons' predictions by different groups.

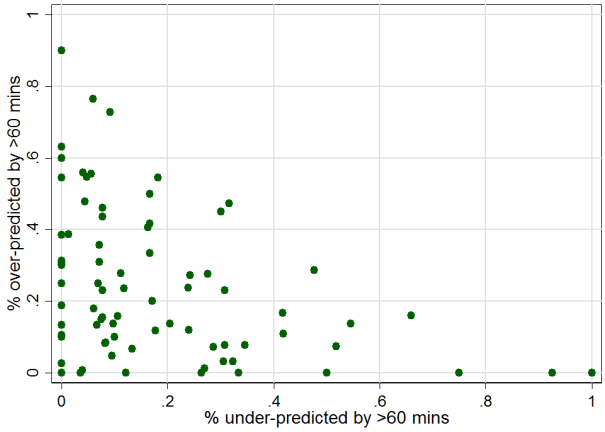
Observed predictors of surgery duration. Studies have identified various factors that affect surgery duration (e.g., see Strum et al. (2000) and Hosseini et al. (2015)). We follow the literature and include all the variables we have in our data as predictors in our surgery duration prediction models. Table 1 provides their summary statistics.

QUANTIFYING THE VALUE OF PHYSICIAN INPUT

In this section, we address our first research question. Namely, we investigate whether there is valuable information left in physician input above and beyond what can be captured by a purely data-



(a) Scatter plot of surgeon's prediction versus surgery duration.



(b) Scatter plot of % under-prediction versus % over-prediction. Each dot is a surgeon- procedure pair.

Figure 2. Comparing the surgeon's prediction and surgery duration.

based model.

Isolating the Physician's Residual Expertise

We first isolate the physician's residual expertise (see Blattberg and Hoch (1990) and references therein for related works). Let Y be the surgery duration, P be the surgeon's prediction of Y, and X be a vector of the observed predictor variables available to both the

surgeon and purely data-based models of surgery duration. Y given X can be estimated using the ordinary least squares (OLS) model:

$$Y = \beta_1 + \beta_2 X + \epsilon. \quad (1)$$

The residual, ϵ, captures the part of the surgery duration that is unexplained by the purely data-based model given in (1). Note that $\hat{M} = \hat{\beta}_1 + \hat{\beta}_2 X$ represents the information extractable from the observed predictor variables. Then, the physician's residual expertise, above and beyond what is captured by the purely data-based model in (1), can be isolated by regressing the surgeon's prediction (P) onto the purely data-based model's prediction, $\hat{M}$, as follows:

$$P = \gamma_1 + \gamma_2 \hat{M} + U. \quad (2)$$

We define U as the physician's residual expertise (Blattberg and Hoch 1990). That is, U contains the unique part of the physician's input which is composed of both valid intuition and random error. The valid intuition could result from the physician's ability to pick up omitted variables or nonlinearities and interactions that are not included in the purely data-based model. We also introduce the following additional equations:

$$Y = \theta_1 + \theta_2 P + \nu. \quad (3)$$

$$\hat{M} = \tau_1 + \tau_2 P + \omega. \quad (4)$$

Similar to ϵ capturing the part of the surgery duration that is unexplained by the purely data-based model in (1), ν captures the part of the surgery duration that is unexplained by the surgeon prediction. Correspondingly, ω contains the part of the prediction of the purely data-based model that is unexplained by the surgeon's prediction.

Three Statistics

Next, we compute three statistics to understand the value of the physician's residual expertise (e.g., see Mincer and Zarnowitz (1969) and Blattberg and Hoch (1990)). The following statistics will show: 1) whether we can use the physician's residual expertise, U, to improve

the surgery duration predictions; and 2) the relative predictive powers of the surgeon's prediction and purely data-based model, compared to each other.

- $r_{Y,U}$ is the correlation coefficient between the surgery duration, Y, and the physician's residual expertise, U. That is, it is the semipartial correlation between the surgery duration, Y, and the surgeon's prediction, P, after partialling the purely data-based model $\hat{M}$ out of P. Blattberg and Hoch (1990) call this statistic *the validity of expert intuition*. Whenever $r_{Y,U} \neq 0$, combining the surgeon's prediction with the purely data-based model output will be more accurate than either of the single inputs. (We note that this may not hold true when evaluating performance in the test set.)
- $r^2_{\epsilon,U}$ is the square of the correlation coefficient between the residual of the purely data-based model in (1), ϵ, and the physician's residual expertise, U. That is, it is the percent of surgery duration variance unexplained by the purely data-based model that can be explained by the surgeon's prediction. Having $r^2_{\epsilon,U} > 0$ means that the surgeon's prediction, P, contains predictive power based not only on the observed factors, but also on surgeon expertise.
- $r^2_{\nu,\omega}$ is the square of the correlation coefficient between the part of the surgery duration that is unexplained by the surgeon's prediction (ν in (3)) and the part of the prediction of the purely data-based model that is unexplained by the physician input (ω in (4)). That is, it is the percent of surgery duration variance unexplained by the surgeon's prediction and that can be explained by the purely data-based model. Having $r^2_{\nu,\omega} > 0$ means that the observed predictors contain a predictive power that was not used in the surgeon's prediction.

Results

For the vector of observed predictor variables, X, we include all of the observed predictor variables described in the Empirical Setting section and 81 dummy variables for each surgeon-procedure pair. As described in the Empirical Setting section, we use the training set to fit the parameters of the models in (1)-(4). We then use the estimated parameters to compute $r_{Y,U}$, $r^2_{\epsilon,U}$, and $r^2_{\nu,\omega}$ in the test set.

The values of the three statistics for the 2,364 surgery cases

Table 2. Quantifying the value of the surgeons' prediction by different groupings.

Measure	(1) One group	(2) By specialty	(3) By surgeon	(4) By procedure	(5) By surgeon-procedure
No. of groups	1	12	42	49	81
$r_{Y,U}$	.21	.31 (.25, -.21, .73)	.34 (.30, -.44, .92)	.29 (.29, -.38, .92)	.31 (.31, -.60, .92)
$r^2_{\epsilon,U}$	.12	.20 (.13, .04, .40)	.25 (.19, .00, .85)	.23 (.20, .00, .85)	.24 (.20, .00, .85)
$r^2_{v,\omega}$	.26	.11 (.15, .00, .49)	.09 (.12, .00, .59)	.13 (.16, .00, .63)	.11 (.14, .00, .63)

Notes: Averages (standard deviation, min, max) are reported.

in the test set are reported in column (1) of table 2. The validity of physician intuition $r_{Y,U}$ is .21, which shows that there is a substantial degree of physician intuition. As discussed above, because $r_{Y,U} \neq 0$, combining the surgeon's prediction with the purely data-based model output will be more accurate than either of the individual inputs. We find that $r^2_{\epsilon,U}$ is .12, meaning that surgeons' predictions explain 12% of the variance in the surgery duration not captured by the purely data-based model. Lastly, we find that $r^2_{v,\omega}$ is .26, i.e., the purely data-based model explains 26% of the variance in the surgery duration not captured by the surgeons' predictions.

The analyses in in the Empirical Setting section show that the average of surgery duration, the coefficient of variation of surgery duration, and the performance of the surgeon's prediction vary widely across different groups (e.g., groups by specialty, by surgeon, by procedure, and by surgeon-procedure). Motivated by this, we compute the three statistics for each group determined by specialty, by surgeon, by procedure, and by surgeon-procedure. Columns (2)-(5) of table 2 show the average, standard deviation, minimum value, and the maximum value of these statistics, by each grouping. They show substantial differences across the different groups, which suggests that the relative performance of the surgeon's prediction, the purely data-based model, and the combination of the two, will vary when we consider different groupings.

MODELS FOR PREDICTING SURGERY DURATION

We introduce models for predicting the surgery duration. The first type of models requires physician input, i.e., surgeons' predictions. The second type of models do not need physician input and utilizes historical surgery duration data only. The third type of models combines physician input models with purely data-based models. As before, we let Y be the surgery duration, P be the surgeon's prediction of Y, and X be a vector of the observed predictor variables available to both the surgeon and purely data-based models of surgery duration.

Models with Physician Input

In what follows, we propose three models that require the elicitation of physician input.

Physician input model. This model uses the surgeon's prediction, as is, for surgery duration prediction. That is, we let $M1 = P$.

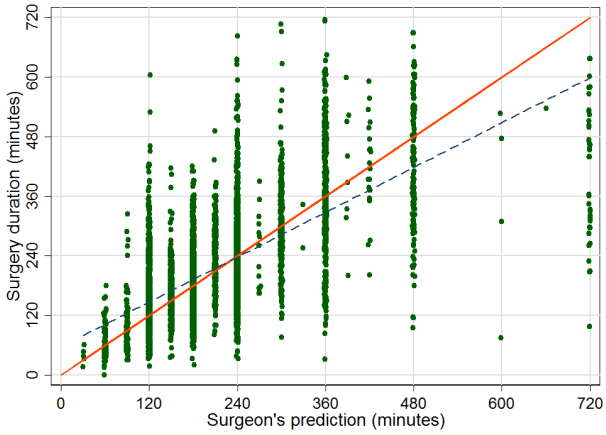
Corrected Physician Input Model. This model applies Theil's *optimal linear correction* (Theil 1966) to the surgeon's prediction, *P*. Consider the OLS model of the surgery duration *Y* on *P*:

$$Y = \alpha_1 + \beta_1 P + \epsilon_1. \tag{5}$$

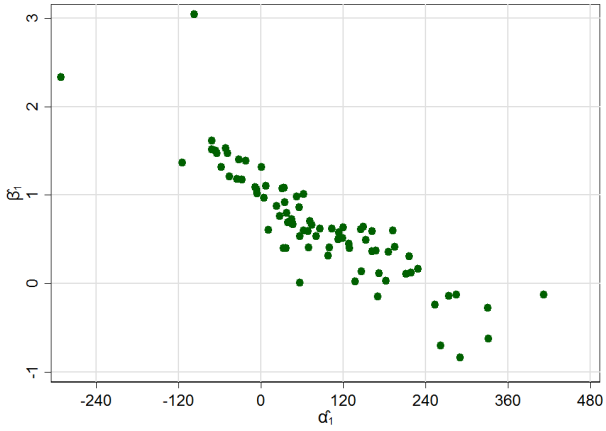
P is an unbiased prediction of *Y* if $\alpha_1 = 0$and is an efficient prediction of *Y* if $\beta_1 = 1$ (Mincer and Zarnowitz 1969). If $\alpha_1 \neq 0$ or $\beta_1 \neq 1$, then we can correct the surgeon's prediction using the estimated parameters: $M2 = (\hat{\alpha}_1 + \hat{\beta}_1 P$.

For example, fitting (5) to the training data, we obtain $\hat{\alpha}_1 = 57.00$ and $\hat{\beta}_1 = 0.75$; see figure 3(a). Having $\hat{\alpha}_1 > 0$ suggests that the surgeons are repeatedly underestimating *Y*. By adding $\hat{\alpha}_1$, the historically observed average error, we eliminate the bias. Having $\hat{\beta}_1 < 1$ suggests that surgeons are overestimating high values of *Y*, and underestimating low values of *Y*. By multiplying *P* by $\hat{\beta}_1$, we correct for this inefficiency.

The analyses of the previous sections suggest that the values of the correcting parameters $\hat{\alpha}_1$ and $\hat{\beta}_1$ could vary across different groups. As such, we fit (5) to each specialty, surgeon, procedure,



(a) Scatter plot of the surgeon's prediction versus surgery duration for the entire training set. The red line is the line of perfect predictions, and the blue dashed line is the regression line for equation (5) for the entire training set.



(b) Scatter plot of the estimated linear correction parameters of equation (5) for each surgeon- procedure pair. Each dot is a surgeon-procedure pair.

Figure 3. Corrected physician input model.

and surgeon-procedure groups: see figure 3(b). We observe that, as expected, the values of the correcting parameters vary widely.

Physician Coefficients Model. This model mimics what is known as the management coefficients model or the judgment bootstrapping model (e.g., see Bowman (1963), Camerer (1981), and Dawes, Faust,

and Meehl (1989)). A prediction model is constructed by regressing the surgeon's prediction onto the observed predictor variables as follows:

$$P = \alpha_2 + \beta_2 X + \epsilon_2. \tag{6}$$

Then the fitted values from the regression, $M3 = \hat{\alpha}_2 + \hat{\beta}_2 X$, are used to predict the surgery durations. This model systematizes surgeon judgment, and in so doing, it discards any intuition that the surgeon may have that is not consistent with the model. Hence, the physician coefficients model will outperform the surgeon's prediction when the residuals of the model in (6) consist mainly of random variance in the surgeon's prediction (Bowman 1963; Camerer 1981).

Purely Data-based Models

In what follows, we propose two models that do not require the elicitation of physician input.

Historical Average Model. This model simply uses the historical average of past surgery durations to predict the current surgery's duration. Given a group with N samples in the training set, the surgery duration prediction is given by $M4 = (\sum_i Y_i)/N$.

Regression Model. This model fits the OLS model of the surgery duration Y given the observed predictor variables X:

$$Y = \alpha_3 + \beta_3 X + \epsilon_3. \tag{7}$$

Note that this is the same model as (1). The fitted values from the regression, $M5 = \hat{\alpha}_3 + \hat{\beta}_3 X$, are used to predict the surgery duration.

Combined Models

Suppose now that we have access to both the surgeon's prediction and the data-based models. We propose four models which combine the two types of predictions.

Simple Regression Combination Model. This model assigns weights to $M1$ and $M5$ using the OLS model:

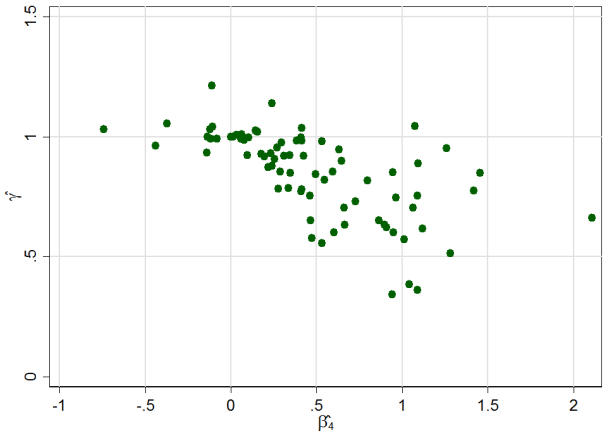


Figure 4. Simple Regression Combination Model. Scatter plot of the estimated parameters of equation (8) for each surgeon-procedure pair. Each dot is a surgeon-procedure pair.

$$Y = \alpha_4 + \beta_4 M1 + \gamma M5 + \epsilon_4. \tag{8}$$

Note that instead of just assigning weights to each prediction value, this model adds a constant term and does not constrain the weights to add to unity. Studies have shown that (8) tends to be the best linear combination method for combining multiple predictions in terms of minimizing the mean squared error of the prediction (Granger and Ramanathan 1984). The resulting predicted value is $M6 = \hat{\alpha}_4 + \hat{\beta}_4 M1 + \hat{\gamma} M5$.

Fitting (8) to the entire training data, we obtain $\hat{\alpha}_4 = -9.68$ ($p < 0.001$), $\hat{\beta}_4 = 0.32$ ($p < 0.001$), and $\hat{\gamma} = 0.73$ ($p < 0.001$). If M1 encompasses all of the features of M5, then we should have $\hat{\beta}_4 = 1$ and $\hat{\alpha}_4 = \hat{\gamma} = 0$ (Chong and Hendry 1986). Similarly, if $M5$ encompasses all of the features of M1, then we should have $\hat{\gamma} = 1$ and $\hat{\alpha}_4 = \hat{\beta}_4 = 0$. However, both M1 and M5 are assigned non-zero weights.

In fact, this is expected from the analysis in the previous section. We have $r^2_{\epsilon,U} = 0.12$, and $r^2_{v,\omega} = 0.26$ (see column (1) of table 2), which suggests that there is uncorrelated effective information in each of M1 and M5.

As before, we fit (8) to each specialty, surgeon, procedure, and surgeon-procedure group. We again observe that the estimated

weights differ significantly, as illustrated by a scatter plot of the estimated parameters $\hat{\beta}_4$ and $\hat{\gamma}$ for each surgeon-procedure pair in figure 4. There are surgeon-procedure pairs with $\hat{\beta}_4$ close to 0 and $\hat{\gamma}$ close to 1. These are the surgeon-procedure pairs with $r^2_{\epsilon,U}$ close to 0 (see column (5) of table 2). That is, because the percent of variance unexplained by the purely data-based model that can be explained by the surgeon's prediction is close to 0% for these pairs, M1 is assigned a very small weight.

Full Regression Combination Model. This model is similar to the purely data-based model in (7), with the only difference being the addition of the surgeon's prediction, M1, as a predictor variable:

$$Y = \alpha_5 + \beta_5 M1 + \theta X + \epsilon_5. \tag{9}$$

The resulting predicted value is $M7 = \hat{\alpha}_5 + \hat{\beta}_5 M1 + \hat{\theta} X$.

In the Simple Regression Combination Model, the weights given to each predictor variable in the Regression Model in (7), $\hat{\beta}_3$, are maintained. The weights are just scaled by $\hat{\gamma}$ when they are combined with the Physician Input Model in (8). In contrast, new weights, $\hat{\beta}_5$, are assigned to each predictor variable in the Full Regression Combination Model.

50% Physician + 50% Model. This model assigns equal weights to both the physician input and the purely data-based regression model. Assigning equal weights when combining forecasts is an attractive heuristic in practice because this method is intuitive and simple, and does not require estimating the optimal weights. Equal weights have also been shown to perform pretty well in practice (Blattberg and Hoch 1990). We define the following models: $M8 = 0.5M1 + 0.5M5$ where we combine the physician input model with the regression model, and $M9 = 0.5M2 + 0.5M5$ where we combine the corrected physician input model with the regression model.

PERFORMANCE OF THE MODELS

We evaluate and compare the performance of the models proposed in the previous section.

Surgery Duration Prediction Performance Measures

Given an actual surgery duration Y and its corresponding point prediction $\hat{Y}$, we quantify the predictive accuracy of a given model using six performance measures (Ibrahim and L'Ecuyer 2013): (1) correlation coefficient (Corr) between Y and $\hat{Y}$, (2) mean squared error (MSE), (3) root mean squared error (RMSE), (4) mean absolute percentage error (MAPE), (5) the percentage of surgeries for whom Y is within an 1-hour interval of the prediction $\hat{Y}$ (Cover-1hr), and (6) the percentage of surgeries for whom Y is within a 2-hour interval of the prediction $\hat{Y}$ (Cover-2hr). We note that Cover-1hr is the key performance measure used at our study hospital to evaluate their effectiveness in operating room time allocation.

Performance of Models with Physician Input

We first examine the performance of the physician input model M1. Column (1) of table 3 shows the accuracy of M1 when all the surgery cases are considered as one group. In columns (2)-(5) of

Table 3. Performance of the surgeon's prediction by different groupings.

	(1) One group	(2) By specialty	(3) By surgeon	(4) By procedure	(5) By surgeon-procedure
No. of groups	1	12	42	49	81
Corr	0.70	.61 (.10, .46, .73)	.58 (.20, .06, .92)	.39 (.27, -.49, .92)	.41 (.28, -.49, .92)
MSE	6550	7351 (4077, 3528, 15682)	7377 (6283, 438, 24214)	7300 (7016, 1213, 31177)	7660 (7645, 438, 31177)
RMSE	81	83 (22, 59, 125)	79 (34, 21, 156)	79 (34, 35, 177)	79 (37, 21, 177)
MAPE	33	47 (18, 20, 82)	39 (21, 11, 98)	44 (25, 17, 116)	39 (23, 11, 116)
Cover-1hr	0.39	.30 (.11, .11, .52)	.34 (.19, .00, .92)	.33 (.17, .00, .74)	.34 (.20, .00, .92)
Cover-2hr	0.64	.56 (.15, .33, .78)	.60 (.22, .03, 1.00)	.59 (.21, .10, 1.00)	.59 (.24, .00, 1.00)

Notes: Averages (standard deviation, min, max) are reported.

table 3, we compute the accuracy for each group and report their average, standard deviation, minimum value, and maximum value. As can be expected from the analyses in previous sections, we observe that the performance of the physician input model varies widely across the different groups. For instance, MSE(M1) is 438 for one surgeon-procedure pair and 31177 for a different surgeon-procedure pair.

Next, we consider the performance of M2. As discussed before, if we fit (5) to the entire training data, we obtain $\hat{\alpha}_1$ = 57.00 and $\hat{\beta}_1$ = 0.75. We can apply these correcting parameters to the entire test data. The resulting accuracy of M2 is reported in the first row under 'Corrected physician input model' in table 4. We observe that MSE(M2) (and consequently RMSE(M2)) decreases slightly compared to the physician input model (from 6550 to 6454). However, the performance declines slightly if we measure the accuracy using MAPE, Cover-1hr, or Cover-2hr.

Rather than using the same correcting parameters for all surgeries, one can use different correcting parameters for each group. The resulting accuracy of M2 is reported in the second row under 'Corrected physician input model' in table 4. Using different correcting parameters for each specialty improves the performance as compared to using the same correcting parameters (e.g., MSE(M2) decreases from 6454 to 6002). As we consider more granular groups, the performance continues to improve. If we separately estimate (5) for each of the 81 surgeon-procedure pairs and apply the resulting 81 sets of correcting parameters to the test set, the resulting MSE (*M*2) is 4823, which is a 25% decrease compared to using the same correcting parameters (from 6454 to 4823).

Lastly, we consider the performance of M3. If we fit (6) to the entire training set, M3 performs worse than M1 and M2 (see the first row under 'Physician coefficients model'). This result suggests that in our empirical setting, systematizing surgeon judgment—and hence discarding the discretion/expertise/intuition that is inconsistent with the physician coefficients model—leads to worse performance.

Note also that the performance of M3 deteriorates as we fit (6) to more granular groupings. There are at least 11 coefficients that need to be estimated in each estimation of (6). Because the group sizes in both the training set and the test set decrease as we consider more granular groupings—as described in the Empirical Setting section, the smallest group size in the training set is 20 and that in the test

Table 4. Accuracy of Surgery Duration Predictions.

	Corr	MSE	RMSE	MAPE	Cover-1hr	Cover-2hr
Physician input model (M1)	0.70	6550	81	33	0.39	0.64
	Corrected physician input model (M2)					
One model	0.70	6454	80	36	0.38	0.63
Model by specialty	0.72	6002	77	32	0.40	0.66
Model by surgeon	0.78	5020	71	28	0.46	0.72
Model by procedure	0.76	5344	73	30	0.41	0.69
Model by surgeon-procedure pair	0.78	4823	69	27	0.46	0.73
Physician coefficient model (M3)						
One model	0.64	7476	86	35	0.37	0.62
Model by specialty	0.64	7514	87	35	0.37	0.63
Model by surgeon	0.63	7838	89	36	0.38	0.62
Model by procedure	0.61	8287	91	37	0.37	0.62
Model by surgeon-procedure pair	0.61	8264	91	37	0.37	0.61
Historical average model (M4)						
One model	.	12539	112	55	0.23	0.46
Model by specialty	0.46	9917	100	45	0.31	0.54
Model by surgeon	0.58	8259	91	37	0.39	0.65
Model by procedure	0.69	6681	82	34	0.39	0.64
Model by surgeon-procedure pair	0.73	5849	76	30	0.44	0.70
Regression model (M5)						
One model	0.76	5364	73	29	0.44	0.70
Model by specialty	0.75	5571	75	29	0.44	0.70
Model by surgeon	0.72	6052	78	30	0.43	0.68
Model by procedure	0.69	6703	82	30	0.44	0.69
Model by surgeon-procedure pair	0.68	6946	83	31	0.44	0.68
Combined model: regression simple (M6)						
One model	0.79	4747	69	27	0.46	0.71
Model by specialty	0.79	4819	69	27	0.46	0.71
Model by surgeon	0.77	5052	71	28	0.45	0.72
Model by procedure	0.74	5708	76	28	0.46	0.71
Model by surgeon-procedure pair	0.73	5931	77	28	0.47	0.70

Table 4. (continued)

	Corr	MSE	RMSE	MAPE	Cover-1hr	Cover-2hr
Combined model: regression all (M7)						
One model	0.80	4577	68	27	0.45	0.73
Model by specialty	0.79	4711	69	27	0.45	0.72
Model by surgeon	0.76	5422	74	28	0.45	0.71
Model by procedure	0.73	6040	78	28	0.46	0.71
Model by surgeon-procedure pair	0.71	6517	81	29	0.45	0.70
Combined model: 50-1 (M8)						
One model	0.78	4870	70	28	0.45	0.71
Model by specialty	0.78	4892	70	28	0.45	0.71
Model by surgeon	0.78	5010	71	29	0.45	0.71
Model by procedure	0.77	5176	72	29	0.45	0.71
Model by surgeon-procedure pair	0.76	5221	72	29	0.45	0.70
Combined model: 50-2 (M9)						
One model	0.79	5080	71	30	0.43	0.70
Model by specialty	0.78	5011	71	29	0.45	0.71
Model by surgeon	0.78	4880	70	28	0.46	0.72
Model by procedure	0.77	5180	72	28	0.45	0.72
Model by surgeon-procedure pair	0.77	5115	72	27	0.47	0.72

set is 10—, overfitting and inaccurate performance evaluations are more likely to occur when we consider granular groupings, resulting in poor overall performance.

Performance of Purely Data-based Models

When we consider the surgeries as one group, M4 performs poorly (reported in the first row under 'Historical average model' in table 4). This poor performance is expected because using the same prediction for all surgeries, 217.3 minutes (see table 1), ignores the heterogeneity in the surgery duration across the different groups. The accuracy of M4 improves as we consider granular groupings. If we use the historical average at the surgeon-procedure level, the MSE decreases by 11% compared to using the physician input

model (from 6550 to 5849).

Next, we consider the performance of M5. When we fit (7) to the entire training set, the resulting accuracy (reported in the first row under 'Regression model') shows that the MSE decreases by 18% compared to the physician input model (from 6550 to 5364). As was the case for the performance of *M*3, the performance of M5 declines when we separately estimate (7) for granular groupings. This is again due to the overfitting and inaccurate performance evaluation caused by the small group sizes in granular groupings.

When model performance is evaluated at the surgeon-procedure level, the regression model outperforms the physician input model, i.e., MSE(M1) > MSE(M5), for 38 pairs out of the 81 surgeon-procedure pairs. As discussed before, one can expect the purely data-based model to outperform the physician input model as $r^2_{Y,U}$ and $r^2_{\epsilon,U}$ decrease and $r^2_{v,\omega}$ increases. For the 38 pairs above, compared to the remaining 43 pairs, the mean $r^2_{Y,U}$ in the training data is significantly lower (.06 vs .14; p = 0.002 in the t-test for the equality of means), and the mean $r^2_{\epsilon,U}$ in the training data is significantly lower (.09 vs .20; p = 0.003). The mean $r^2_{v,\omega}$ in the training data is lower but not statistically different (.24 vs .27; p = 0.420). However, we find that $r^2_{Y,U}$, $r^2_{\epsilon,U}$, and $r^2_{v,\omega}$ are not the only factors that determine whether the purely data-based models outperform the physician input model in the surgery duration problem. Because the performance of the purely data-based model depends on its power to correctly estimate its parameters, the purely data-based model would not perform well if the size of the training data is too small. In fact, for the 38 pairs in which the regression model outperforms the physician input model, the mean number of observations used to fit the purely data-based model in (7) is significantly higher (65.0 vs 43.5; p = 0.028) compared to the remaining 43 pairs in which the physician input model outperforms the regression model.

As discussed in the Literature Review section, the lack of historical data is a common problem in predicting surgery case duration using data-based models. Although one might speculate that this problem can be resolved by combining similar procedure codes, such approach is not practical because the surgery duration is likely to differ by a large amount even for a small change in the procedure code (Macario 2006). In such cases, using the physician input model or the corrected physician input model is a potential solution.

Performance of Combined Models

We first consider the performance of M6. When we fit (8) to the entire training set, the resulting accuracy (reported in the first row under 'Combined model: regression simple' in table 4) shows that the MSE decreases by 28% as compared to the physician input model (from 6550 to 4747).

The wide variation in the performance of M1 evaluated at the surgeon-procedure level (see table 3) suggests that the optimal weights for M1 and M5 in the combined model should be different for different surgeon-procedure pairs. Hence, one might expect to see the performance of M6 to improve as more granular groups are considered. However, table 4 shows that the performance of M6 declines when more granular groupings are considered. This is because the accuracy of M5, one of the two inputs for M6, decreases as more granular groupings are considered due to the small-sample problem.

The performance of M7 is similar to that of M6, and it is the best performing model among the combined models. When we fit (9) to the entire training set, the resulting accuracy (reported in the first row under 'Combined model: regression all' in table 4) shows that the MSE decreases by 30% as compared to the physician input model (from 6550 to 4577). As was the case for the performances of M3 and M5, the performance of M7 declines as more granular groupings are considered. This is again due to the overfitting and inaccurate performance evaluation caused by small group sizes in both the training and test sets.

Lastly, we note that the 50% Physician + 50% Models, M8 and M9 perform very well compared to the other combined models. This is in line with what the previous studies have found in other application areas (Blattberg and Hoch 1990), and is an interesting result especially given that such models do not require additional weight estimations.

Choosing the Best Model

In this subsection, we summarize the performance comparison of our different models. Our benchmark model is the model where the surgeon's prediction of the surgery case duration is used directly as the prediction for the surgery duration. The correlation value of this

model with the actual surgery duration in our test set is 0.70, MSE is 6550, RMSE is 81, MAPE is 33, Cover-1hr is 0.39, and Cover-2hr is 0.64.

Among the models with physician input, the corrected physician input model performs the best, under the condition that the correction model is fitted separately for each surgeon-procedure pair. Its performance measure values are correlation 0.78, MSE 4823, RMSE 69, MAPE 27, Cover-1hr 0.46, and Cover-2hr 0.73. Note that compared with the benchmark model, the MSE decreases by 26%. We emphasize that *fitting a separate model for each surgeon-procedure pair is critical*; this results from the fact that the value and accuracy of the surgeon's prediction vary widely across the different surgeon-procedure pairs, as observed in the previous sections. If a single correction model is used for all surgeons instead, then the resulting MSE will be 6454, which is not much lower than the MSE of the physician input model (equal to 6550).

Among the purely data-based models, the regression model performs the best, under the condition that a single regression model is fitted for all surgeries. Its performance measure values are correlation 0.76, MSE 5364, RMSE 73, MAPE 29, Cover-1hr 0.44, and Cover-2hr 0.70. Compared to the benchmark model, the MSE decreases by 18%. As opposed to the corrected physician input model, the regression model performs the best when a single model is fitted to all surgeries. This is due to the small sample sizes when groupings are used, which leads to overfitting.

Among the combined models, the full regression combination model performs the best, under the condition that a single regression model is fitted for all surgeries. Its performance measure values are correlation 0.80, MSE 4577, RMSE 68, MAPE 27, Cover-1hr 0.45, and Cover-2hr 0.73. Compared to the benchmark model, the MSE decreases by 30%. Similar to the regression model, it is important to fit a single model to all surgeries because of the small sample size problem when groupings are used.

Overall, the best performing model for our study hospital is the full regression combination model, under the condition that a single model is fitted for all surgeries. Note that this approach requires eliciting the surgeon's prediction for every surgery, as well as the historical data of the characteristics of past surgery cases, to fit the regression model. Also, the hospital will need to communicate to surgeons how their predictions will be combined with a purely data-

based model to produce the final predictions.

If the hospital does not have access to the historical data of the characteristics of the past surgery cases, or if surgeons prefer to rely solely on their own input, then the best model to use is the corrected physician input model, where the correction model is fitted separately to each surgeon-procedure pair. Compared with the full regression combination model, the MSE will increase by only 5%, suggesting that this model is a great alternative. Also, surgeons may be less resistant to using the corrected physician input model, compared with the full regression combination model, because the prediction depends solely on an individualized correction of their own input.

In the case where the hospital finds eliciting the surgeon's prediction for each surgery inconvenient, but has access to historical data of the past surgery cases, it can use the regression model, under the condition that a single regression model is fitted to all surgeries. Compared to the full regression combination model, the MSE will increase by 17%, which can be considered as the cost of forgoing the value of physician input.

CONCLUSIONS

The objective of this paper was to quantify the value that people can bring to operations. In particular, we focused on the context of predicting surgery durations in hospitals, and studied whether or not the physician's input should be considered in that prediction exercise. To do so, we considered a wide array of models, either including or excluding the surgeon's prediction of surgery duration, and compared these models in terms of predictive accuracy.

While it is clear that expert individuals, e.g., physicians in our context, may have key intuition or prior experience that should prove to be useful in operational decision-making, quantifying the value of that discretion/expertise/intuition remains, to a large extent, an open problem. In this paper, we took a step towards quantifying that value and, in so doing, derived some key insights. Importantly, we demonstrated that when studying the impact of people on operations, it is essential to account for the fact that people are, themselves, heterogeneous, e.g., some surgeons are clearly more accurate than others when predicting surgery

durations. We demonstrated how ignoring that heterogeneity leads to suboptimal operational decisions. We did this by comparing the predictive accuracies of tailored models (fitted to alternative groupings in the data) with aggregate, one-size-fits-all-type models that ignore such heterogeneity, and are fitted to the entire data set instead. Moreover, we proposed several easily implementable ways of accounting for that heterogeneity, e.g., we proposed correcting each surgeon's prediction differently, depending on the identity of the surgeon.

In studying the value of the physician's input, we can provide an answer to the question of whether or not to discard that input, as was initially proposed in our study hospital. The answer to that question, based on our analysis, is an emphatic no. Indeed, there is value in the physician's input that should not be disregarded, and doing so would lead to inferior operational decision-making in the hospital, as can be seen through our numerical study. Of course, it remains to test whether the conclusions of this paper would continue to hold in other healthcare settings, and with alternative data sets. In addition, future work may complement our results by exploring how using other machine learning methods can help further improve the accuracy of the different types of models we considered in this paper.

REFERENCES

American Society of Anesthesiologists (2014), "ASA Physical Status Classification System," https://www.asahq.org/standards-and-guidelines/asa-physical-status-classification-system.

Anand K. S. and H. Mendelson (1997), "Information and Organization for Horizontal Multimarket Coordination," *Management Science,* 43(12), 1609–1627.

Bates D. W., G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, and B. Middleton (2003), "Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality." *Journal of the American Medical Informatics Association,* 10(6), 523–530.

Berner E. S. (2009), "Clinical Decision Support Systems: State of the Art," AHRQ Publication No. 09-0069-EF. Rockville, Maryland: Agency for Healthcare Research and Quality.

Blattberg R. C. and S. J. Hoch (1990), "Database Models and Managerial

Intuition: 50% Model + 50% Manager," *Management Science,* 36(8), 887–899.

Bowman E. H. (1963), "Consistency and Optimality in Managerial Decision Making," *Management Science*, 9(2), 310–321.

Bunn D. and G. Wright (1991), "Interaction of Judgemental and Statistical Forecasting Methods: Issues & Analysis," *Management Science,* 37(5), 501–518.

Cabana M. D., C. S. Rand, N. R. Powe, A. W. Wu, M. H. Wilson, P. C. Abboud, and H. R. Rubin (1999), "Why Don't Physicians Follow Clinical Practice Guidelines?: A Framework for Improvement," *Journal of the American Medical Association,* 282(15), 1458–1465.

Cardoen B., E. Demeulemeester, and J. Belien (2010), "Operating Room Planning and Scheduling: A Literature Review," *European Journal of Operational Research*, 201(3), 921–932.

Chong Y. Y. and D. F. Hendry (1986), "Econometric Evaluation of Linear Macro-economic Models," *The Review of Economic Studies* 53(4), 671–690.

Dawes R. M., D. Faust, and P. E. Meehl (1989), "Clinical versus Actuarial Judgment," *Science* 243(4899), 1668–1674.

Eijkemans M. J., M. van Houdenhoven, T. Nguyen, E. Boersma, E. W. Steyerberg, G. Kazemier (2010), "Predicting the Unpredictable: A New Prediction Model for Operating Room Times Using Individual Characteristics and the Surgeon's Estimate," *The Journal of the American Society of Anesthesiologists*, 112(1), 41–49.

Granger C. W. and R. Ramanathan (1984), "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3(2), 197–204.

Hosseini N., M. Y. Sir, C. Jankowski, and K. S. Pasupathy (2015), "Surgical Duration Estimation Via Data Mining and Predictive Modeling: A Case Study," *AMIA Annual Symposium Proceedings,* Volume 2015, 640.

Ibanez M. R., J. R. Clark, R. S. Huckman, and B. R. Staats (2018), "Discretionary Task Ordering: Queue Management in Radiological Services," *Management Science*, 64(9), 4389–4407.

Ibrahim R. and P. L'Ecuyer (2013), "Forecasting Call Center Arrivals: Fixed-effects, Mixed-effects, and Bivariate Models," *Manufacturing & Service Operations Management*, 15(1), 72–85.

Kahneman D. and G. Klein (2009), "Conditions for Intuitive Expertise: A Failure to Disagree," *American Psychologist,* 64(6), 515.

Kim S. H., C. W. Chan, M. Olivares, and G. Escobar (2015), "ICU Admission Control: An Empirical Study of Capacity Allocation and Its Implication for Patient Outcomes," *Management Science*, 61(1), 19–38.

Larsson A. (2013), "The Accuracy of Surgery Time Estimations," *Production Planning & Control*, 24(10-11), 891–902.

Laskin D. M., A. O. Abubaker, and R. A. Strauss (2013), "Accuracy of

Predicting the Duration of a Surgical Operation," *Journal of Oral and Maxillofacial Surgery,* 71(2), 446–447.

Macario A. (2006), "Are Your Hospital Operating Rooms Efficient? A Scoring System with Eight Performance Indicators," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 105(2), 237–240.

_____ (2010), "Is it Possible to Predict How Long a Surgery Will Last?," *Medscape Anesthesiology*, 108(3), 681–685.

McGlynn E. A. (1997), "Six Challenges in Measuring the Quality of Health Care," *Health Affairs,* 16(3), 7–21.

Mincer, J. A., and V. Zarnowitz (1969), "The Evaluation of Economic Forecasts," *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, 3-46 (NBER).

Ozen A., Y. Marmor, T. Rohleder, H. Balasubramanian, J. Huddleston, and P. Huddleston (2016), "Optimization and Simulation of Orthopedic Spine Surgery Cases at Mayo Clinic," *Manufacturing & Service Operations Management,* 18(1), 157–175.

Phillips R., A. S. Simsek, and G. van Ryzin (2015), "The Effectiveness of Field Price Discretion: Empirical Evidence from Auto Lending," *Management Science,* 61(8), 1741–1759.

Stepaniak P. S., C. Heij, G. H. Mannaerts, M. de Quelerij, and G. de Vries (2009), "Modeling Procedure and Surgical Times for Current Procedural Terminology-anesthesia-surgeon Combinations and Evaluation in Terms of Case-duration Prediction and Operating Room Efficiency: A Multicenter Study," *Anesthesia & Analgesia*, 109(4), 1232–1245.

Strum D. P., A. R. Sampson, J. H. May, and L. G. Vargas (2000), "Surgeon and Type of Anesthesia Predict Variability in Surgical Procedure Times," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 92(5), 1454–1466.

Theil H. (1966), Applied Economic Forecasting (Rand-McNally & Co).

Timmermann A. (2006), "Forecast Combinations," *Handbook of Economic Forecasting* 1, 135–196.

Travis E., S. Woodhouse, R. Tan, S. Patel, J. Donovan, and K. Brogan (2014), "Operating Theatre Time, Where Does It All Go? A Prospective Observational Study," *BMJ,* 349, g7182.

Van Donselaar K. H., V. Gaur, T. Van Woensel, R. A. Broekmeulen, and J. C. Fransoo (2010), "Ordering Behavior in Retail Stores and Implications for Automated Replenishment," *Management Science,* 56(5), 766–784.

Zhou J. and F. Dexter (1998), "Method to Assist in the Scheduling of Add-on Surgical Cases-Upper Prediction Bounds for Surgical Case Durations Based on the Log-normal Distribution," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 89(5), 1228–1232.

Zhou Z., D. Miller, N. Master, D. Scheinker, N. Bambos, and P. Glynn (2016), "Detecting Inaccurate Predictions of Pediatric Surgical Durations," *2016*

IEEE International Conference on Data Science and Advanced Analytics (DSAA), 452–457.

Received September 5, 2019
Accepted September 17, 2019