Predicting Surgery Duration: Physician Input, Statistical Models, and Combined Models

Rouba Ibrahim
UCL School of Management, University College London, rouba.ibrahim@ucl.ac.uk

Song-Hee Kim
Marshall School of Business, University of Southern California, songheek@marshall.usc.edu

This version: October 12, 2018

Problem definition: Most data-driven tools do not ask people for their input. We study whether and how to incorporate the discretion/expertise/intuition of physicians into data-driven decision support tools for improved operational decision-making in hospitals, in an empirical setting of predicting the surgery duration.

Academic/Practical Relevance: Understanding the benefits/costs of allowing expert input in decision-making has been attracting the interest of the OM community in various application areas. We contribute to this line of research by studying the effect of incorporating physician input on operating room use. For hospital managers, we show the potential value of incorporating physician input in data-driven decision support tools.

Methodology: We consider three families of models for predicting the surgery duration: models with physician input, statistical models, and models that combine the two. Using the operating room scheduling and usage data from an academic hospital collected over three years, we empirically evaluate and compare the performances of the different models.

Results: We find that physician input offers predictive power beyond that of statistical models in our empirical setting. The best performing model is the combined model which includes the physician input as a feature in a statistical model with other predictors. The corrected physician input model, which applies a simple correction to the physician input, performs comparably well (the mean squared error increases by 5%), when corrections are applied at the surgeon-procedure level. Without the physician input, the mean squared error of the best performing model increases by 17%. We also compare the performances of the operating-room schedules resulting from our models, and show that the results carry over.

Managerial Implications: Our findings suggest that hospital managers should consider eliciting physician input and incorporating it into data-driven decision support tools. We also show how physician input can be best leveraged in the context of predicting the surgery duration.

Key words: empirical research, healthcare operations, operating room reservation, surgery duration, expert input, discretion, decentralization

1. 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, Hong et al. 2015).

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 three research questions. First, does physician input offer predictive power beyond that of statistical 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? Third, what would be the operational implications (i.e., the impact on operating room use) of changing the surgery duration prediction methods based on the findings of this paper?

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 et al. 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 statistical 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, Erdogan et al. 2011, Zhou et al. 2016). Inaccurate predictions have a trickle-down effect on daily hospital operations, including wasting patients' and surgery 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.

1.1. 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) statistical models, which ignore the physician input altogether and rely on historical data only; and (iii) combined models, which incorporate both the physician input and statistical 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 statistical models, as people are more flexible than statistical 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 judgement bootstrapping model (Armstrong 2001), 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.

Statistical Models. Our second family of models are statistical models, which do not need physician input and utilize historical data only. Statistical 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, statistical 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 statistical 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 statistical models. In this paper, we first compute statistics that can help determine whether combined models can be useful and then quantify their benefit. We consider three models: (a) a simple regression combination model, which optimally weighs physician input and the output from the statistical 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 statistical 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 (see §4). The groupings that we consider are (a) surgical specialty, (b) surgeon, (c) procedure type, and (d) surgeon-procedure pair. In §6, we demonstrate that the predictive accuracy of the various models that we consider varies substantially depending on the grouping considered, and this leads to different operational performance benefits, as shown in §7. 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.

1.2. Main Contributions

The goal of this paper is to develop a data-driven and tailored framework for quantifying the value of physician input, and to demonstrate the operational performance improvements when insights from that framework are used. 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 (Wright et al. 1996, 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 statistical 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%.
- Finally, we demonstrate the implications of our models and the different groupings on operating room use, which has not been addressed in the extant literature.

1.3. Organization

The rest of this paper is organized as follows. In §2, we review the relevant literature. In §3, we describe the empirical setting of this paper. In §4, we quantify the value of physician input compared to statistical models. In §5, we describe our models for predicting surgery duration, and we quantify their predictive accuracy in §6. We quantify the impact on operational performance in §7, and we conclude this paper in §8.

2. 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. Lastly, we discuss relevant papers in operating room management and surgical scheduling.

2.1. 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 statistical models (e.g., see Dawes et al. (1989) and references therein), whereas other studies suggest expert judgment can outperform statistical models (e.g., see Bunn and Wright (1991) and references therein). 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 (e.g., see Bowman (1963), Camerer (1981) and references therein).

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. For example, it has been examined in the context of horizontal multimarket coordination (Anand and Mendelson 1997), ordering behavior in retail stores (Van Donselaar et al. 2010), capacity supply decisions in service operations (Campbell and Frei 2011), sales forecasting (Osadchiy et al. 2013), price setting (Phillips et al. 2015), and earnings forecasting (Ball and Ghysels 2018). 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) use the data from radiological diagnoses and 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.

2.2. 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 (e.g., see the survey papers Armstrong (2001) and 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 statistical models (see Blattberg and Hoch (1990) and references therein). This idea has been explored in various application settings, including supply chain planning (Gaur et al. 2007) and predicting the supreme court's decisions (Bommarito and Blackman 2014).

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, Joustra et al. 2013, Master et al. 2016). 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 statistical modelling.

2.3. Operating Room Management and Surgical Scheduling

There is a wealth of literature on operating room management and surgical scheduling (e.g., see Gupta (2007), May et al. (2011), Guerriero and Guido (2011) and references therein). Studies have used various methods to optimize surgery scheduling, including integer programming (e.g., Benchoff et al. (2017)), stochastic optimization (e.g., Denton et al. (2010)), and discrete-event simulation (e.g., Adan et al. (2009)).

In order to apply the techniques developed in the aforementioned studies, one needs the distributional information about the surgery duration. Currently, 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' advantages and disadvantages are not optimally weighed in (e.g., a regression-based model provides a forecast and a surgeon decides whether to accept it or not) (Hosseini et al. 2015). In this paper, we adopt a stochastic-optimization framework from Jafarnia-Jahromi and Jain (2017) to solve the scheduling problem of surgeries in the operating room (see §7).

Studies have examined the predictive accuracy of the surgeons prediction. For example, Laskin et al. (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 surgeon's prediction and show that the sign and magnitude of the bias depends on the type of 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. Wright et al. (1996) and Roque et al. (2015) describe that surgeons are generally more accurate than statistical models.

Predicting surgery duration using statistical models has also been extensively studied (Strum et al. 2000, Dexter and Ledolter 2005, Eijkemans et al. 2010), and factors such as age, gender, ASA (American Society of Anesthesiology classification of physical status) risk level, surgeon, surgical team, procedure type, and anesthesia type have been identified to be the important features in accurately predicting surgery case durations (Strum et al. 2000, Dexter et al. 2008). Some studies have used machine-learning methods to predict surgery duration and they show that their methods generally outperform surgeons' predictions (Kargar et al. 2013, Gomes et al. 2012).

When statistical 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. For surgeries with little or no historical data, studies suggest using the mean duration of cases of the same procedure that was performed by other surgeons (Macario and Dexter 1999), pooling historical data from different facilities (Dexter et al. 2002), and combining physician input and historical data in a Markov Chain Monte Carlo model (Luangkesorn and Eren-Doğu 2016). 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.

3. 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.

3.1. Surgery Scheduling Process

When a physician sees a patient and deems that surgery is needed, the physician works with the patient and a surgical scheduler within the clinical department (e.g., Department of Neurology) 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. When the hospital's surgical schedulers receive the electronic surgery booking slip, the hospital's surgical schedulers use the surgeon's prediction of the surgery duration to schedule the surgery, but they may add or subtract time based on need and experience. They then communicate the confirmed time and place to the departmental surgical scheduler and the surgery takes place on the scheduled surgery date.

3.2. 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. For each surgery, we have the surgeon's prediction for the surgery duration, the scheduled surgery start and end dates and times, and the actual surgery start and end dates and times. The actual surgery start and end dates and times enable us to evaluate the performance of the different predictions of the surgery duration compared to the actual surgery duration. We have various surgery-level characteristics including the procedure names (multiple names if more than one procedure is performed during the surgery) captured by the SurgiNet procedure codes (Cerner 2017), the type of anesthesia used for the surgery (major or not), the surgeon identifier for the primary surgeon, an indicator for whether the surgery was performed by more than one surgeon, and an assessment of the fitness of the patient before surgery measured by the American Society of Anesthesiologists (ASA) physical status classification system. The ASA classification system has six levels, where ASA 1 is a normal healthy patient, ASA 2 is a patient with a mild systemic disease, ASA 3 is a patient with a severe systemic disease, ASA 4 is a patient with a severe systemic disease that is a constant threat to life, ASA 5 is a moribund patient who is not expected to survive without the operation, and ASA 6 is a declared brain-dead patient whose organs are being removed for donor purposes; see American Society of Anesthesiologists (2014) for descriptions of example patients with different ASA levels. In addition, we have patient-level characteristics including age, gender, and race.

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, which includes the surgeon's prediction of the surgery duration. We removed 5,733 additional cases with missing electronic booking slip data. We note that one of the main causes of missing electronic booking slip data is the urgent nature of the cases. For example, about 62% of the 5,733 cases were created either on the day of the surgery or the day before the surgery, and 2,151 cases of the 5,733 cases were scheduled to take place outside of the 24 main operating rooms, in an "add-on" unit. Next, we removed nine surgeries with missing surgeon or procedure identifiers. We also removed 162 outliers, i.e., 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. As is the standard practice (James et al. 2013), we use the training set to fit the parameters of the various models that we consider. We use the fitted models to predict the duration of surgeries in the test set, and evaluate the predictive performances of our models.

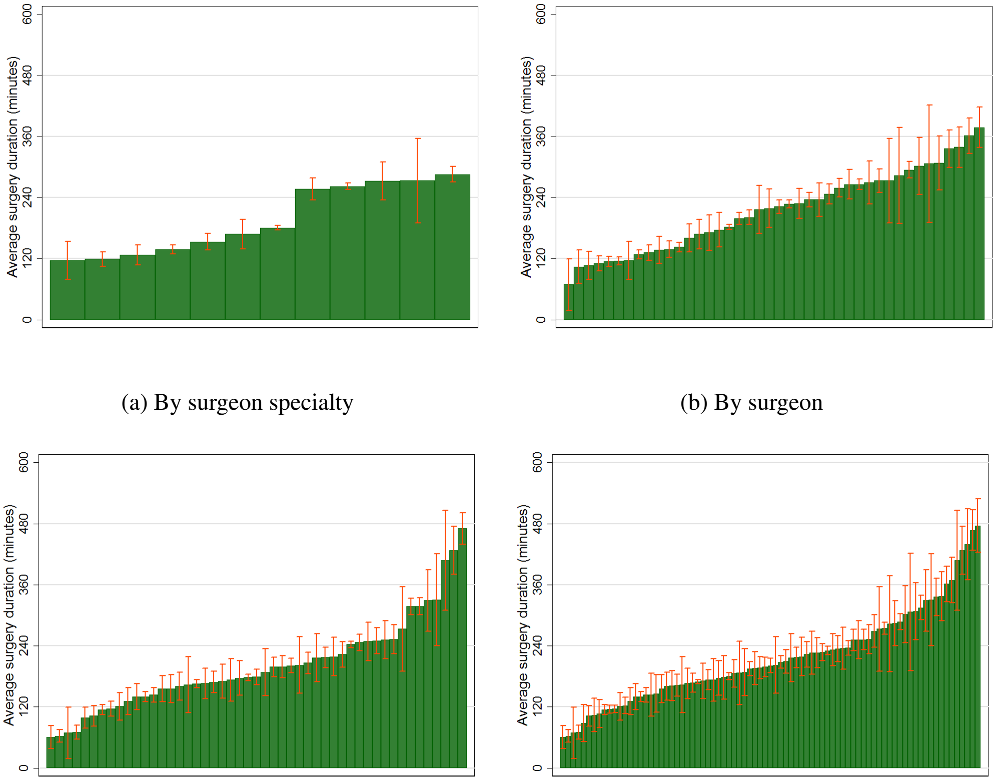
In the following sections, we consider estimating separate models for different groups of the surgery cases. The smallest-sized groups are formed when we group surgery cases at the surgeon-procedure pair level (i.e., surgeries with the same procedure name and done by the same surgeon). To ensure that 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 surgeries in the training set or fewer than 10 surgeries in the 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.

3.3. Variables

3.3.1. Surgery duration. We define surgery duration as the time since a patient enters the operating room to the time the patient leaves the operating room. The average surgery duration was 217.3 minutes (SD 114.1) or 3.6 hours in the training set and 216.3 minutes (SD 112.0) or 3.6 hours in the test set (the summary statistics are also provided in Table A1 in the online supplement).

In what follows, we examine the average and the coefficient of variation (CV) of the surgery duration across different groups. The first grouping we consider is by surgeon specialty. Surgeons belong to different specialties, and we have 12 unique specialties in the data: colorectal surgery, ENT (ear, nose and throat) surgery, general surgery, neurosurgery, obstetrics and gynecology surgery, orthopedic surgery, plastic surgery, spine surgery, thoracic surgery, transplant and hepatobiliary surgery, urologic surgery, and vascular surgery. In the test set, each specialty had an average of 197.0 surgeries (SD 241.2, MIN 12, MAX 775),

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



(a) By surgeon specialty

(b) By surgeon

(c) By procedure

(d) By surgeon-procedure pair

3.5 (2.7, 1, 9) different surgeons, 4.4 (3.2, 1, 11) different procedures, and 6.8 (6.0, 1, 18) different surgeon-procedure pairs. Figure 1(a) shows the average surgery duration with its 95% confidence interval for each specialty. The average surgery duration by specialty varies from 116.6 minutes to 285.8 minutes, and the CV of surgery duration by specialty varies from .31 to .73.

The second grouping we consider is by surgeon. We have 42 unique surgeons in the data. In the test set, each surgeon performed an average of 56.3 surgeries (SD 56.6, MIN 10, MAX 297) and 1.9 (1.0, 1, 4) different procedures. Figure 1(b) shows the average surgery duration with its 95% confidence interval for each surgeon. The average surgery duration by surgeon varies from 69.0 minutes to 378.0 minutes, and the CV varies from .12 to 1.21.

The third grouping we consider is by procedure. We have 49 unique procedures in the data. In the test set, each procedure was performed on an average of 48.2 times (SD 69.6, MIN 10, MAX 299) by 1.7 (1.1, 1, 6) different surgeons. Figure 1(c) shows the average surgery duration with its 95% confidence interval for each procedure. The average surgery duration by procedure varies from 60.4 minutes to 470.7 minutes, and the CV varies from .14 to 1.21.

The fourth grouping is by surgeon-procedure pair. We have 81 unique surgeon-procedure pairs in the data. In the test set, each surgeon-procedure pair occurred an average of 29.1 times (SD 26.9, MIN 10, MAX 155). Figure 1(d) shows the average surgery duration with its 95% confidence interval for each surgeon-procedure pair. The average surgery duration for each surgeon-procedure pair varies from 60.4 minutes to 476.3 minutes, and the CV varies from .13 to 1.21.

We note that within each surgeon-procedure pair, the normal distribution provides quite a good fit for the surgery duration distribution when examined graphically. 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 actual surgery duration when we model it using linear regression models. For robustness checks, we also try fitting the lognormal distribution for the surgery duration, and obtain consistent results (provided in the online supplement).

In sum, the analysis above shows that the average and CV of surgery duration vary widely across different groups.

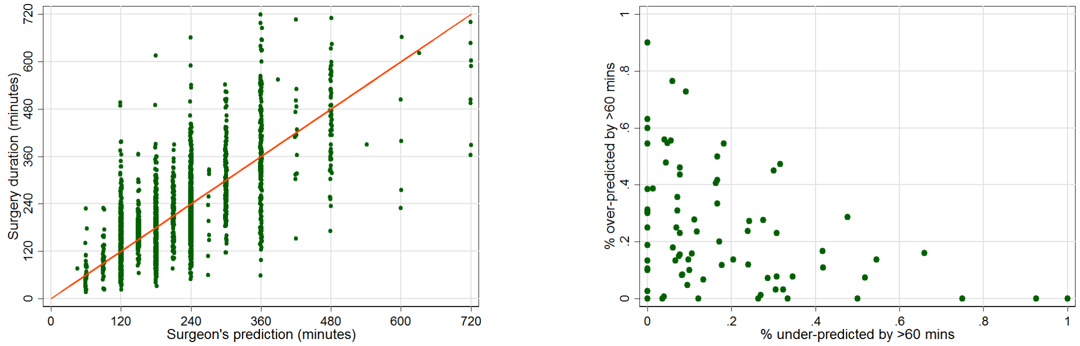
3.3.2. Surgeon's prediction of surgery duration. The average surgeon's prediction of surgery duration was 212.9 minutes (SD 100.8) or 3.5 hours in the training set and 215.3 minutes (SD 92.3) or 3.6 hours in the test set. If we compare only the averages and standard deviations, the surgeons' predictions 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. In §6, we further evaluate the performance of the surgeon's prediction compared to other prediction methods.

Figure 2(b) shows that the degrees of under-prediction and over-prediction vary widely across surgeon-procedure pairs. For example, there are three surgeons who never under-predict or over-predict by more than 60 minutes (the dots at (0,0)), a surgeon who never under-predicts by more than 60 minutes but over-predicts by more than 60 minutes 90% of the time (the dot at (0,.9)), a surgeon who always under-predicts by more than 60 minutes (the dot at (1,0)), and a surgeon who under-predicts by more than 60 minutes 32% of the time and over-predicts by more than 60 minutes 47% of the time (the dot at (.32,.47)). This scatter plot implies that there may be value in evaluating the surgeons' predictions by different groups (e.g., by surgeons and by surgeon-procedure pairs).

As mentioned in §3.1, the surgeon's prediction of surgery duration is not necessarily equal to the amount of time scheduled for each surgery, because the hospital's surgical schedulers may add or subtract time

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



(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.

based on need and experience. Among the 6,705 surgeries in our training and test sets, we observe that the scheduled duration is longer than the surgeon's prediction for 66% of the surgeries, equal to the surgeon's prediction for 31% of the surgeries, and shorter than the surgeon's prediction for 3% of the surgeries. Throughout the paper, we refer to the surgeon's prediction of surgery duration, and not the actual scheduled time, as the physician input.

3.3.3. Observed predictors of surgery duration. Studies have identified various factors that affect surgery duration (e.g., see Strum et al. (2000), Hosseini et al. (2015), Eijkemans et al. (2010), Kayis 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. The variables we have access to include age, gender, race, physical status classification measured by the ASA system, whether the patient received major anesthesia for the surgery, whether the surgery was performed by more than one surgeon, and whether more than one procedure was performed. Their summary statistics are provided in Table A1 in the online supplement.

4. 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 statistical model.

4.1. 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 statistical models of surgery duration. Y given X can be estimated using the ordinary least squares (OLS) model:

$$Y = \beta_1 + \beta_2 X + \varepsilon. \tag{1}$$

The residual, ε, captures the part of the surgery duration that is unexplained by the statistical 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 statistical model in (1), can be isolated by regressing the surgeon's prediction (P) onto the statistical model's prediction, $\hat{M}$, as follows:

$$P = \gamma_1 + \gamma_2 \hat{M} + U. \tag{2}$$

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 statistical model. We also introduce the following additional equations:

$$Y = \theta_1 + \theta_2 P + \nu, \tag{3}$$

$$\hat{M} = \tau_1 + \tau_2 P + \omega. \tag{4}$$

Similar to ε capturing the part of the surgery duration that is unexplained by the statistical 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 statistical model that is unexplained by the surgeon's prediction.

4.2. 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 statistical 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 statistical 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 statistical 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.)

- $r^2_{\varepsilon,U}$ is the square of the correlation coefficient between the residual of the statistical model in (1), ε, and the physician's residual expertise, U. That is, it is the percent of surgery duration variance unexplained by the statistical model that can be explained by the surgeon's prediction. Having $r^2_{\varepsilon,U} > 0$ 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 statistical 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 statistical 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.

In essence, how well the surgeon's prediction, the statistical model, or the combination of the two performs compared to each other will depend on the values of $r_{Y,U}$, $r^2_{\varepsilon,U}$, and $r^2_{\nu,\omega}$.

4.3. Results

For the vector of observed predictor variables, X, we include all of the observed predictor variables described in §3.3.3 and 81 dummy variables for each surgeon-procedure pair. As described in §3.2, we use the training set to fit the parameters of the models in (1)-(4). We then use the estimated parameters to compute $r_{Y,U}$, $r^2_{\varepsilon,U}$, and $r^2_{\nu,\omega}$ in the test set.

The values of the three statistics for the 2,364 surgery cases in the test set are reported in column (1) of Table 1. (As a robustness check, we report the corresponding results when $log(Y)$, instead of Y, is used in (1) and (3) in Table A2 in the online supplement. We find that the results are similar.) 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 statistical model output will be more accurate than either of the individual inputs. We have that $r^2_{\varepsilon,U} = .12$, meaning that surgeons' predictions explain 12% of the variance in the surgery duration not captured by the statistical model. We also have that $r^2_{\varepsilon,U} = .26$, i.e., the statistical model explains 26% of the variance in the surgery duration not captured by the surgeons' predictions.

Table 1 Quantifying the value of the surgeons' prediction by different groupings.

	(1)	(2)	(3)	(4)	(5)
Measure	One group	By specialty	By surgeon	By procedure	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)
$r^2_{\varepsilon,U}$	.12	.20 (.13, .04, .40)	.25 (.19, .00, .85)	.23 (.20, .00, .85)	.24 (.20, .00, .85)
$r^2_{\nu,\omega}$	.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.

The analyses in §3.3 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. That is, we estimate separate models for each group in the training set, and then use the estimated parameters to compute $r_{Y,U}$, $r^2_{\varepsilon,U}$, and $r^2_{\nu,\omega}$ for each group in the test set.

Columns (2)-(5) of Table 1 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. For example, the validity of physician intuition, $r_{Y,U}$, varies from -.21 to .73 when we consider each specialty separately, and from -.60 to .92 when we consider each surgeon-procedure pair separately. Figure 3 shows the distributions of $r_{Y,U}$ by each grouping (the second row of Table 1) in greater detail. As can be expected, the range of $r_{Y,U}$ becomes wider as the grouping becomes more granular, i.e., as we move from studying at the specialty level to the surgeon-procedure level.

Figure 3 Distribution of $r_{Y,U}$ by different groupings.

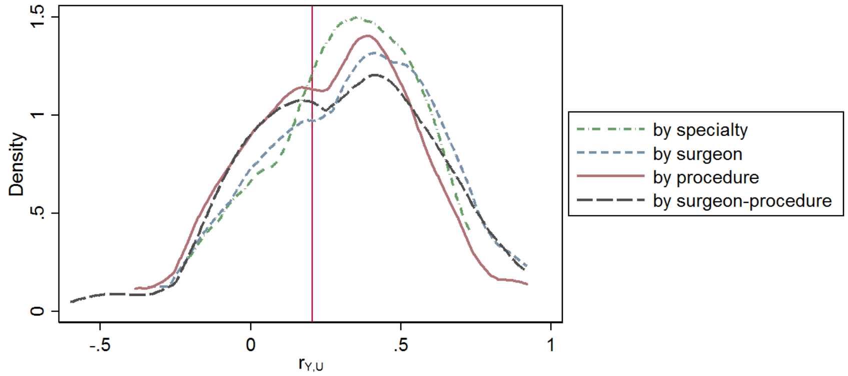


Figure A1 in the online supplement shows the distributions of $r^2_{\varepsilon,U}$ and $r^2_{\nu,\omega}$; they also vary widely across different groups. For example, for one of the 42 surgeons in the data, $r^2_{\varepsilon,U}$ is 0, meaning that this surgeon's predictions explained 0% of the variance in the surgery duration not captured by the statistical model. On the other hand, for another surgeon, $r^2_{\varepsilon,U}$ is .85, suggesting that this surgeon's predictions explained 85% of the variance in the surgery duration not captured by the statistical model.

The results in Table 1, Figure 3, and Figure A1 all suggest that the relative performance of the surgeon's prediction, the statistical model, and the combination of the two, will vary when we consider different groups. We continue exploring this fact in the following sections.

5. Models for Predicting Surgery Duration

In this section, we introduce models for predicting the surgery duration, summarized in Table 2. The first type of models discussed in §5.1 requires physician input, i.e., surgeons' predictions. The second type of models discussed in §5.2 do not need physician input and utilizes historical surgery duration data only. The third type of models discussed in §5.3 combines physician input models with statistical models.

As in §4, 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 statistical models of surgery duration.

Table 2 Summary of Models.

Models with Physician Input	
Physician input model	$M_{physician\_input} = P$
Corrected physician input model	$M_{corrected\_physician\_input} = \hat{\alpha_1} + \hat{\beta_1}P$
Physician coefficient model	$M_{physician\_coefficient} = \hat{\alpha_2} + \hat{\beta_2}X$
Statistical Models	
Historical average model	$M_{historical\_avg} = \frac{\sum_i Y_i}{N}$
Regression model	$M_{regression\_model} = \hat{\alpha_3} + \hat{\beta_3}X$
Combined Models	
Combined model: regression simple	$M_{combined\_reg\_simple} = \hat{\alpha_4} + \hat{\beta_4}M_{physician\_input} + \hat{\gamma}M_{regression\_model}$
Combined model: regression all	$M_{combined\_reg\_all} = \hat{\alpha_5} + \hat{\beta_5}M_{physician\_input} + \hat{\theta}X$
Combined model: 50-1	$M_{combined\_50\_1} = .5 \times M_{physician\_input} + .5 \times M_{regression\_model}$
Combined model: 50-2	$M_{combined\_50\_2} = .5 \times M_{corrected\_physician\_input} + .5 \times M_{regression\_model}$

Notes. Y is the surgery duration, P is the surgeon's prediction of Y, and X is a vector of the observed predictor variables available to both the surgeon and models of surgery duration.

5.1. Models with Physician Input

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

5.1.1. Physician input model. This model uses the surgeon's prediction, as is, for surgery duration prediction. That is, we let $M_{physician\_input} = P$.

5.1.2. 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:

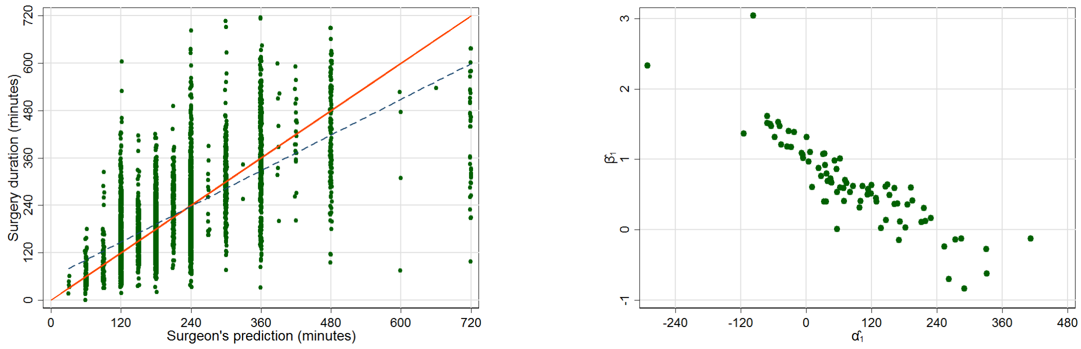
$$Y = \alpha_1 + \beta_1 P + \varepsilon_1. \tag{5}$$

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: $M_{corrected\_physician\_input} = \hat{\alpha_1} + \hat{\beta_1}P$.

For example, fitting (5) to the training data, we obtain $\hat{\alpha_1} = 57.00$ ($p < 0.001$) and $\hat{\beta_1} = 0.75$ ($p < 0.001$) with $R^2 = .44$; see Figure 4(a). Having $\hat{\alpha_1} > 0$ means 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, and surgeon-procedure groups. In §6, we compare their performances. We observe that, as expected, the values of the correcting parameters vary. For example, Figure 4(b) provides a scatter plot of the correcting parameters for each surgeon-procedure pair. We observe that there are surgeon-procedure pairs for which surgeons provide unbiased and efficient predictions ($\hat{\alpha_1}$ close to 0 and $\hat{\beta_1}$ close to 1). On the other hand, there are also surgeons whose predictions are too large by more than 300 minutes ($\hat{\alpha_1} > 300$) or too small by more than 120 minutes ($\hat{\alpha_1} < -120$). Also, $\hat{\beta_1}$ varies from -1 to 3.

Figure 4 **Corrected physician input model.**



(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 (5) for the entire training set.

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

5.1.3. 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), Dawes et al. (1989)). A prediction model is constructed by regressing the surgeon's prediction onto the observed predictor variables as follows:

$$P = \alpha_2 + \beta_2 X + \varepsilon_2. \tag{6}$$

Then the fitted values from the regression, $M_{physician\_coefficient} = \hat{\alpha_2} + \hat{\beta_2} X$, 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).

5.2. Statistical Models

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

5.2.1. 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 $M_{historical\_avg} = \frac{\sum_i Y_i}{N}$.

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

$$Y = \alpha_3 + \beta_3 X + \varepsilon_3. \tag{7}$$

Note that this is the same model as (1). The fitted values from the regression, $M_{regression\_model} = \hat{\alpha_3} + \hat{\beta_3} X$, are used to predict the surgery duration.

5.3. Combined Models

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

5.3.1. Simple Regression Combination Model. This model assigns weights to $M_{physician\_input}$ and $M_{regression\_model}$ using the OLS model:

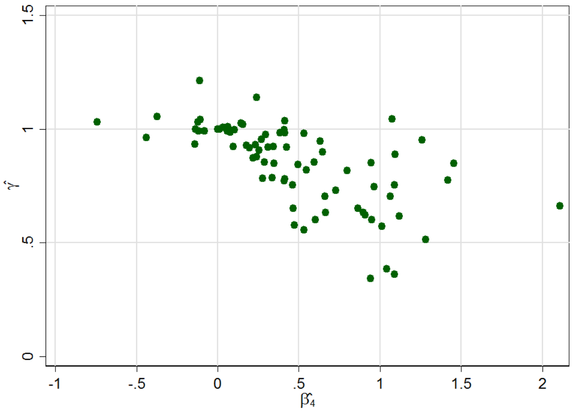
$$Y = \alpha_4 + \beta_4 M_{physician\_input} + \gamma M_{regression\_model} + \varepsilon_4. \tag{8}$$

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 $M_{combined\_reg\_simple} = \hat{\alpha_4} + \hat{\beta_4} M_{physician\_input} + \hat{\gamma} M_{regression\_model}$.

Fitting (8) to the entire training data, we obtain $\hat{\alpha_4} = -9.68$ ($p < 0.01$), $\hat{\beta_4} = 0.32$ ($p < 0.001$), and $\hat{\gamma} = 0.73$ ($p < 0.001$) with $R^2 = .59$. If $M_{physician\_input}$ encompasses all of the features of $M_{regression\_model}$, then we should have $\hat{\beta_4} = 1$ and $\hat{\alpha_4} = \hat{\gamma} = 0$ (Chong and Hendry 1986). Similarly, if $M_{regression\_model}$ encompasses all of the features of $M_{physician\_input}$, then we should have $\hat{\gamma} = 1$ and $\hat{\alpha_4} = \hat{\beta_4} = 0$. However, both $M_{physician\_input}$ and $M_{regression\_model}$ are assigned non-zero weights. In fact, this is expected from the analysis in §4. We have $r^2_{\varepsilon,U} = .12$ and $r^2_{\varepsilon,U} = .26$ (see column (1) of Table 1), which suggests that there is uncorrelated effective information in each of $M_{physician\_input}$ and $M_{regression\_model}$.

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 5. 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 $r^2_{\varepsilon,U}$ close to 0 (see column (5) of Table 1). That is, because the percent of variance unexplained by the statistical model that can be explained by the surgeon's prediction is close to 0% for these pairs, $M_{physician\_input}$ is assigned a very small weight.

Figure 5 **Simple Regression Combination Model. Scatter plot of the estimated parameters of (8) for each surgeon-procedure pair. Each dot is a surgeon-procedure pair.**



5.3.2. Full Regression Combination Model. This model is similar to the statistical model in (7), with the only difference being the addition of the surgeon's prediction, $M_{physician\_input}$, as a predictor variable:

$$Y = \alpha_5 + \beta_5 M_{physician\_input} + \theta X + \varepsilon_5. \tag{9}$$

The resulting predicted value is $M_{combined\_reg\_all} = \hat{\alpha_5} + \hat{\beta}_5 M_{physician\_input} + \hat{\theta} X$.

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.

5.3.3. 50% Physician + 50% Model. This model assigns equal weights to both the physician input and the regression model described in §5.2.2. 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: $M_{combined\_50\_1} = .5 \times M_{physician\_input} + .5 \times M_{regression\_model}$ where we combine the physician input model with the regression model, and $M_{combined\_50\_2} = .5 \times M_{corrected\_physician\_input} + .5 \times M_{regression\_model}$ where we combine the corrected physician input model with the regression model.

6. Performance of the Models

In this section, we evaluate and compare the performance of the models proposed in §5. In what follows, we define the performance measures that we use to evaluate the predictive accuracy of the alternative models.

6.1. 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, James et al. 2013). The first performance measure is the *correlation coefficient* (Corr) between Y and $\hat{Y}$, given by

$$Corr = \frac{\sum_{i=1}^{K}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{K}(Y_i - \bar{Y})^2 \sum_{i=1}^{K}(\hat{Y}_i - \bar{\hat{Y}})^2}}, \tag{10}$$

where $\bar{Y}$ is the average value of Y_i's, and K is the number of predictions that we are considering. The second performance measure is the *mean squared error* (MSE), defined by

$$MSE = \frac{1}{K}\sum_{i=1}^{K}(Y_i - \hat{Y}_i)^2. \tag{11}$$

As is the standard practice, the MSE will be our main performance measure when we compare the performances of our different models.

The third performance measure is the *root mean squared error* (RMSE), given by

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(Y_i - \hat{Y}_i)^2}. \tag{12}$$

In addition to the MSE and the RMSE, the fourth performance measure is a relative measure of accuracy, the *mean absolute percentage error* (MAPE), defined by

$$MAPE = 100 \cdot \frac{1}{K}\sum_{i=1}^{K}\frac{|Y_i - \hat{Y}_i|}{Y_i}. \tag{13}$$

The fifth and sixth performance measures evaluate whether Y is within a 1-hour or 2-hour interval of the prediction $\hat{Y}$. Specifically, we define the *1-hour cover* (Cover-1hr) as

$$Cover - 1hr = \frac{1}{K}\sum_{i=1}^{K} I(Y_i \in (\hat{Y}_i - 30, \hat{Y}_i + 30)), \tag{14}$$

where $I(\cdot)$ is the indicator function. Note that here we assume that the units for Y and $\hat{Y}$ are in minutes. Similarly, the *2-hour cover* (Cover-2hr) is defined as

$$Cover - 2hr = \frac{1}{K}\sum_{i=1}^{K} I(Y_i \in (\hat{Y}_i - 60, \hat{Y}_i + 60)). \tag{15}$$

6.2. Performance of Models with Physician Input

We first examine the performance of the physician input model $M_{physician\_input}$. Column (1) of Table 3 shows the accuracy of $M_{physician\_input}$ when all the surgery cases are considered as one group. In columns (2)-(5) of Table 3, we compute the accuracy for each group and report their average, standard deviation, minimum value, and maximum value. For example, there are 12 specialty groups and for each group, we compute $Corr(M_{physician\_input})$. The average $Corr(M_{physician\_input})$ of the 12 groups is .61 with standard deviation .10. 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, $MSE(M_{physician\_input})$ is 438 for one surgeon-procedure pair and 31177 for a different surgeon-procedure pair.

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

	(1)	(2)	(3)	(4)	(5)
	One group	By specialty	By surgeon	By procedure	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.

Next, we consider the performance of the corrected physician input model, $M_{corrected\_physician\_input}$. As discussed in §5.1.2, if we fit (5) to the entire training data, we obtain $\hat{\alpha_1} = 57.00$ ($p < 0.001$) and $\hat{\beta_1} = 0.75$ ($p < 0.001$). We can apply these correcting parameters to the entire test data. The resulting accuracy of $M_{corrected\_physician\_input}$ is reported in the first row under 'Corrected physician input model' in Table 4. (As a robustness check, we report the corresponding results when $log(Y)$, instead of Y, is used in (5), (7), (8), and (9) in Table A3 in the online supplement. We find that the results are similar and that in general using Y results in superior performance than using $log(Y)$.) We observe that $MSE(M_{corrected\_physician\_input})$ (and consequently $RMSE(M_{corrected\_physician\_input})$) 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. That is, one can separately estimate (5) for each of the 12 specialty groups in the training set, and apply the resulting 12 sets of correcting parameters to the corresponding specialty in the test set. The resulting accuracy of $M_{corrected\_physician\_input}$ 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., $MSE(M_{corrected\_physician\_input})$ decreases from 6454 to 6002). As we consider more granular groups (from by specialty to by surgeon-procedure pair), the performance continues to improve. If we separately estimate (5) for each of the 81

Table 4 Accuracy of Surgery Duration Predictions.

	Correlation	MSE	RMSE	MAPE	Cover-1hr	Cover-2hr
Physician input model	0.70	6550	81	33	0.39	0.64
Corrected physician input model						
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						
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						
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						
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						
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
Combined model: regression all						
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						
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						
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

surgeon-procedure pairs and apply the resulting 81 sets of correcting parameters to the test set, the resulting $MSE(M_{corrected\_physician\_input})$ is 4823, which is a 25% decrease compared to using the same correcting parameters (from 6454 to 4823).

Lastly, we consider the performance of the physician coefficients model $M_{physician\_coefficient}$. If we fit (6) to the entire training set, $M_{physician\_coefficient}$ performs worse than $M_{physician\_input}$ and $M_{corrected\_physician\_input}$ (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 $M_{physician\_coefficient}$ 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 and more granular groupings—as described in §3.2, the smallest group size in the training set is 20 and that in the test set is 10—, overfitting and inaccurate performance evaluations are more likely to occur when we consider granular groups, resulting in poor overall performance.

6.3. Performance of Statistical Models

When we consider the surgeries as one group, $M_{historical\_avg}$ exhibits a poor performance (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 A1 in the online supplement), ignores the heterogeneity in the surgery duration across the different groups (see §3.3). The accuracy of $M_{historical\_avg}$ 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 the regression model $M_{regression\_model}$. 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 $M_{physician\_coefficient}$, the performance of $M_{regression\_model}$ declines when we separately estimate (7) for smaller groups. This is again due to the overfitting and inaccurate performance evaluation caused by the small group sizes in granular groupings, in both the training and test sets.

When model performance is evaluated at the surgeon-procedure level, the regression model outperforms the physician input model, i.e., $MSE(M_{physician\_input}) > MSE(M_{regression\_model})$, for 38 pairs out of the 81 surgeon-procedure pairs. As discussed in §4, one can expect the statistical model to outperform the physician input model as $r^2_{Y,U}$ and $r^2_{\varepsilon,U}$ decrease and $r^2_{\nu,\omega}$ 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 $r^2_{\varepsilon,U}$ in the training data is significantly lower (.09 vs .20; $p = 0.003$). The mean $r^2_{\nu,\omega}$ in the training data is lower but not statistically different (.24 vs .27; $p = 0.420$). However, we find that $r^2_{Y,U}$, $r^2_{\varepsilon,U}$, and $r^2_{\nu,\omega}$ are not the only factors that determine whether the statistical models outperform the physician input model in the surgery duration problem. Because the performance of the statistical model depends on its power to correctly estimate its parameters, the statistical model would not perform well if the size of the training data is too small. In fact, for the 38 pairs compared to the remaining pairs, the mean number of observations used to fit the statistical model in (7) is significantly higher (65.0 vs 43.5; $p = 0.028$),

As discussed in §2.3, 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, that is not a practical solution as 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.

6.4. Performance of Combined Models

We first consider the performance of the simple regression combination model $M_{combined\_reg\_simple}$. 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 $M_{physician\_input}$ evaluated at the surgeon-procedure level (see Table 3) suggests that the optimal weights for $M_{physician\_input}$ and $M_{regression\_model}$ in the combined model should be different for different surgeon-procedure pairs. Hence, one might expect to see the performance of $M_{combined\_reg\_simple}$ to improve as more granular groups are considered. However, Table 4 shows that the performance of $M_{combined\_reg\_simple}$ declines when more granular groupings are considered. This is because the accuracy of $M_{regression\_model}$, one of the two inputs for $M_{combined\_reg\_simple}$, decreases as more granular groupings are considered due to the small-sample problem.

The performance of the full regression combination model $M_{combined\_reg\_all}$ is similar to that of $M_{combined\_reg\_simple}$, 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 $M_{physician\_coefficient}$ and $M_{regression\_model}$, the performance of $M_{combined\_reg\_all}$ 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, $M_{combined\_50\_1}$ and $M_{combined\_50\_2}$ 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.

6.5. 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 statistical 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 statistical 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.

The above results show that physician input is indeed valuable in our empirical setting. In the case where the hospital finds eliciting the surgeon's prediction for each surgery inconvenient, but has access to historical data about the characteristics 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.

7. Impact on Operating Room Usage

In this section, we investigate the operational implications of relying on the different models that were introduced in earlier sections of the paper. We focus primarily on the problem of scheduling surgeries, as we explain next.

7.1. Stochastic Appointment Scheduling Problem

We envision the distributional and point predictions, based on the different models, being used as inputs in the scheduling of surgeries at the hospital. The problem of scheduling can be easily stated: Assume that there is a finite number, k, of surgeries that are to be scheduled in a given operating room. Surgery durations are random, and are assumed to be normally distributed (the mean and variance of that normal distribution depend on the model at hand). If a surgery is completed before the scheduled starting time of the following surgery, then the operating room will remain idle in between. Conversely, if a surgery runs over the scheduled starting time of the following surgery, then the patient and the surgery staff of the next surgery will have to wait until the operating room is available.

Hospital schedulers encounter two problems. The first is a *sequencing* problem, which amounts to determining a cost-minimizing sequence of surgeries, whereby costs are assigned to both idle time and delay. The second is a *scheduling* problem, where, for a given sequence of surgeries, a cost-minimizing schedule of surgery starting times is determined. We learned that there is little room to alter the sequence of surgeries in any given operating room in our study hospital since these sequences are intimately tied to the seniority of the surgeons involved. Thus, we focus strictly on the scheduling problem in what follows.

7.2. Problem Statement

To formulate our scheduling problem, we focus on a single operating room, on a given day. We can do this because we are taking the sequence of surgeries in a day as given and, conditional on this information, the problem objectives across different operating rooms and days are separable. We let $\mathbf{S} = (S_1, \cdots, S_k)$ denote a vector of the random durations of the k surgeries that are to be performed. We let $\mathbf{t} = (t_2, \cdots, t_k)$

denote the scheduled starting times of the k surgeries. Without loss of generality, we assume that $t_1 \equiv 0$. We let E_i be the random variable denoting the end time of surgery i, $1 \leq i \leq k$. Moreover, we let $\mathbf{E} \equiv (E_1, E_2, \cdots, E_k)$ denote the corresponding vector of end times. As described in Jafarnia-Jahromi and Jain (2017), the sequence of random end times is given by the following two equations:

$$E_1 = S_1, \tag{16}$$

$$E_i = \max\{E_{i-1}, t_i\} + S_i, \quad i = 2, \cdots, k. \tag{17}$$

Equations (16) and (17) can be used to recursively determine the sequence of ending times which results from a given sequence of scheduled start times, along with realized surgery duration. The stochastic scheduling problem can then be formulated as

$$\inf_{\mathbf{t} \in \mathcal{T}} \mathbb{E}[(E_{i-1} - t_i)^2], \tag{18}$$

where $\mathcal{T} \equiv \{(t_2, \cdots, t_k) \in \mathbf{R}^{k-1} | t_1 \equiv 0 \leq t_2 \leq t_3 \leq \cdots \leq t_k.\}$. We note that the problem formulation in (18) coincides with problem formulation (6) in Jafarnia-Jahromi and Jain (2017). In that paper, the authors demonstrate that the solution to this optimization problem does indeed exist and is unique; see Theorem 1 in Jafarnia-Jahromi and Jain (2017). Importantly, the authors of that paper also demonstrate that a sample average approximation (SAA) to problem (18) is asymptotically accurate, in that SAA is a consistent estimator for the scheduling problem in (18); see Theorem 2 in Jafarnia-Jahromi and Jain (2017). In this paper, we rely on that theoretical evidence as a justification for the usage of a sample average approximation for the solution to problem (18). In particular, we let $(\mathbf{S}^j)_{j=1}^m$ denote an independent and identically distributed random sample of size m for surgery durations $\mathbf{S}$, and $(\mathbf{E}^j)_{j=1}^m$ denote the vectors of end times, corresponding to each of the m samples. We then numerically solve the following problem instead of (18):

$$\inf_{\mathbf{t} \in \mathcal{T}} \frac{1}{m} \sum_{j=1}^{m} (E_{i-1}^j - t_i)^2. \tag{19}$$

7.3. Scheduling Results

Description of the numerical experiments. To solve the numerical approximation in (19), we must decide on an appropriate sample size. In our simulations, we let $m = 10,000$ independent samples, because we observed in numerical tests that this number was large enough (the optimal solution is roughly constant beyond that number). In order to test the operational performance of the various models, we focus on a sample of the data for which we have predictions under each model, i.e., we consider the same training and test sets described in §3.2. From the data, we observe the actual duration of each surgery. We also observe the actual duration (time slot) assigned to that surgery in the corresponding operating room. This enables us to assess and check whether our proposed alternative schedules, based on the different models, would lead to any improvement as compared to the current scheduling process at the hospital. Additionally, we are able to quantify how the predictive accuracy (of future surgery duration) impacts the scheduling performance.

Upon inspecting the current schedule at the hospital, we observed that the current practice is to schedule relatively few surgeries per day (on average, roughly 1.54 surgeries per day per operating room). This resulted in roughly 293 minutes per day of operating room under-use, i.e., on average, each operating room is empty during its operating hours for roughly 293 minutes per day. Given this substantial under-use of resources, and to be able to conduct a detailed comparison between the alternative models, we opted to concatenate surgeries across different days into a "master" hypothetical schedule. In particular, we opted to concatenate groups of 5 consecutive surgeries in the data (that took place in the same operating room, potentially on consecutive days) as being part of the same hypothetical day. We believe that this is appropriate because doing so would enable us to eliminate, to some extent, the effect of the current mismanagement of the operating room schedule from the alternative schedules that we obtain based on each of our models. Among 2,364 surgeries in our test set, we removed the remaining surgeries in each operating room after creating groups of 5 consecutive surgeries. As a result, we evaluate the operational performance of our models based on 2,295 surgeries across 20 operating rooms.

For each of the models considered in §5, we use distributional predictions for the surgery durations, simulate m independent samples of corresponding surgeries (recall that the sequence of surgeries is assumed to be fixed), and numerically solve for the optimal sequence of scheduled starting times in (19). Lastly, we compare the performance of the different schedules using the real surgery durations observed in the data. In addition, we compare them to schedules constructed based on the actual scheduled durations. For simplicity, we hereafter denote such schedules by 'Current'. Note that when solving for the optimal scheduled starting times, we do not take into consideration the time needed for cleaning and disinfecting the operating rooms. We do this because including a constant time in between surgeries should not alter performance significantly.

We let SE_i denote the scheduled end time of surgery i. Note that $SE_i = t_{i+1}$ for all surgeries except for the last surgery of each day. We assume that $SE_i = t_i + M_i$ for the last surgery of each day, where M_i is the point prediction given by each of our models. We let $I_i \equiv \max\{0, SE_i - E_i\}$ denote the idle time corresponding to surgery i, $D_i \equiv \max\{0, E_i - SE_i\}$ denote the delay corresponding to surgery i, and $Error_i = I_i + D_i$, all measured in minutes. We define 'under-use per day per operating room' as the sum of I_i of all surgeries i scheduled for that day in the specific operating room. We define 'over-use per day per operating room' as the value of D_i for the last surgery of that day in the specific operating room. The average error, $\overline{Error}$, is defined as the average of the errors across all surgeries i. The average squared error, $\overline{Error^2}$, is defined as the average of the squared errors across all surgeries i. Lastly, we also compute the average proportions (per day) of surgeries i that started on-time, with a 15 minute grace period. That is, a surgery i starts on time if $D_{i-1} \leq 15$, for all surgeries except for the first surgery of each day (we assume that the first surgery of each day in each operating room always starts on time and, hence, we do not include it while computing the average proportions of surgeries that start on-time). Furthermore, we compute the average proportions (per day) of surgeries i that ended on-time, with a 15 minutes grace period before and after the scheduled end time i.e., $\max\{I_i, D_i\} \leq 15$.

Summary of results. In Table 5, we present the performance of each schedule, based on the numerical solution of problem (19). In Table 6, we present the performance of a heuristic schedule where the scheduled duration allocated to each surgery coincides with the point prediction of its duration i.e., this approach assumes that there is no variability in surgery duration. While admittedly suboptimal, we consider this heuristic alternative because it mimics the current practice at our study hospital. Moreover, since this method is simple, it can be easily implemented at other hospitals. Indeed, comparing Tables 5 and 6 illustrates the advantage of incorporating the variability of surgery duration into the schedule.

We can make several noteworthy observations based on Table 5. We notice that different models may be superior depending on the performance measure at hand. Nevertheless, we can note the superiority, in general, of both the combined model of (9), i.e., $M_{combined\_reg\_all}$, and the tailored corrected physician input model, $M^t_{correcetd\_physician\_input}$, where the subscript t denotes a tailored model at the surgeon-procedure level i.e., the correction model is fitted to each surgeon-procedure pair. For example, $M^t_{correcetd\_physician\_input}$ yields the smallest absolute error across all models (equal to 61 minutes, on average). On the other hand, $M_{combined\_reg\_all}$ yields the largest proportion of surgeries with on-time start, across all days and all operating rooms. The superiority of those two models, which both exploit the physician input in some way, indicates that overlooking physician input altogether, which was believed to be the best alternative at our study hospital, would lead to a deterioration in both predictive accuracy and in operational performance. Furthermore, we also note that the "one-size-fits-all" corrected physician input model, $M_{correcetd\_physician\_input}$, performs reasonably well, yet it is outperformed by $M^t_{correcetd\_physician\_input}$ for every performance measure considered. Thus, our results both confirm and quantify the importance of accounting for heterogeneity across people when considering people-oriented operations, as we do in this paper.

It is also interesting to see that we can significantly improve on the current schedule by relying on the alternative schedules that we propose. For example, by relying on model $M^t_{correcetd\_physician\_input}$, we are able to reduce the average error per surgery, per day, by 22 minutes. We would also be able to significantly reduce the amount of time that the operating room is underused per day by almost half (from 300 minutes to 164 minutes). We also note that relying strictly on the physician input (with no correction) does not yield superior performance: Model $M_{phsician\_input}$ is outperformed by both its tailored and "one-size-fits-all" corrected model counterparts, $M^{t}_{correcetd\_physician\_input}$ and $M_{correcetd\_physician\_input}$.

Lastly, in comparing the performance of the optimal schedules based on (19) in Table 5 to the heuristic schedules in Table 6, we note a significant deterioration in performance. We note that this provides additional support to the results in Jafarnia-Jahromi and Jain (2017). The absolute errors are roughly between 20% and 30% greater than relying on the optimal solutions of (19). However we continue to observe the superior performance of $M^{t}_{correcetd\_physician\_input}$ and $M_{combined\_reg\_all}$ compared to other models.

Table 5 Alternative performance measures for our various model-based schedules, based on the solution of problem (19).

	Current	$M_{\substack{physician\\input}}$	$M_{\substack{corrected\\physician\\input}}$	$M^{t}_{\substack{corrected\\physician\\input}}$	$M_{\substack{regression\\model}}$	$M^{t}_{\substack{regression\\model}}$	$M_{\substack{combined\\reg\\all}}$	$M^{t}_{\substack{combined\\reg\\all}}$
$\overline{Error}$	83	83	78	61	69	71	64	66
$\overline{Error^2}$	11345	14128	11508	7192	9060	10300	7612	9129
Ontime start	0.84	0.66	0.78	0.78	0.80	0.77	0.81	0.77
Ontime end	0.09	0.14	0.13	0.17	0.14	0.17	0.15	0.18
Total scheduled mins/day	1352	1148	1239	1201	1218	1211	1223	1213
Total underuse mins/day	300	149	213	164	188	184	184	179
Total overuse mins/day	29	82	56	44	51	54	43	49

Notes. We add superscript t to denote a tailored model, i.e., one which is fit at the (surgeon, procedure) level grouping.

Table 6 Alternative performance measures for our various model-based schedules, based on a scheduling heuristic.

	Current	$M_{\substack{physician\\input}}$	$M_{\substack{corrected\\physician\\input}}$	$M^{t}_{\substack{corrected\\physician\\input}}$	$M_{\substack{regression\\model}}$	$M^{t}_{\substack{regression\\model}}$	$M_{\substack{combined\\reg\\all}}$	$M^{t}_{\substack{combined\\reg\\all}}$
$\overline{Error}$	83	94	90	72	77	83	70	77
$\overline{Error^2}$	11345	18848	17701	11531	13756	16006	11111	13566
Ontime start	0.84	0.58	0.61	0.64	0.62	0.61	0.65	0.62
Ontime end	0.09	0.15	0.15	0.20	0.19	0.18	0.21	0.20
Total scheduled minutes per day	1352	1073	1093	1093	1084	1086	1099	1098
Total underuse minutes per day	300	101	109	88	88	97	90	98
Total overuse minutes per day	29	110	98	76	85	92	72	82

Notes. We add superscript t to denote a tailored model, i.e., one which is fit at the (surgeon, procedure) level grouping.

8. Conclusions

In this paper, our overriding goal 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 both predictive accuracy and scheduling performance.

Key takeaways. 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. In general, one essential message that can be gleaned from our analysis is that both statistical and human inputs should guide the final operational decisions in a hospital. Combining these inputs may be done in the context of simple models, as in the combined regression model that we considered in this paper.

Future research directions. In our analysis, we restricted attention to the scheduling problem faced in our study hospital. We did so because we learned that the sequence of surgeries in a given day is difficult to alter. In future work, it would be interesting to consider both the scheduling and sequencing problems concurrently, and to study the operational implications of using the physician input on both problems. Our study of operational implications was also strictly numerical. In the future, we hope to derive theoretical results on the scheduling performance of incorporating physician input. Finally, there remains to test whether the conclusions of this paper would continue to hold in other healthcare settings, and with alternative data sets.

Acknowledgments

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.

References

Adan I, Bekkers J, Dellaert N, Vissers J, Yu X (2009) Patient mix optimization and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science* 12(2):129–141.

American Society of Anesthesiologists (2014) ASA physical status classification system. URL https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system.

Anand KS, Mendelson H (1997) Information and organization for horizontal multimarket coordination. *Management Science* 43(12):1609–1627.

Armstrong JS (2001) *Principles of forecasting: A handbook for researchers and practitioners*, volume 30 (Springer Science & Business Media).

Ball RT, Ghysels E (2018) Automated earnings forecasts: Beat analysts or combine and conquer? *Management Science* 64(10):4936–4952.

Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B (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.

Benchoff B, Yano CA, Newman A (2017) Kaiser Permanente Oakland Medical Center optimizes operating room block schedule for new hospital. *Interfaces* 47(3):214–229.

Berner ES (2009) Clinical decision support systems: State of the art. AHRQ Publication No. 09-0069-EF. *Rockville, Maryland: Agency for Healthcare Research and Quality* .

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model+ 50% manager. *Management Science* 36(8):887–899.

Bommarito MJ, Blackman J (2014) Using data to predict supreme court's decisions. URL: http://msutoday.msu.edu/news/2014/using-data-to-predict-supreme-courts-decisions.

Bowman EH (1963) Consistency and optimality in managerial decision making. *Management Science* 9(2):310–321.

Bunn D, Wright G (1991) Interaction of judgemental and statistical forecasting methods: Issues & analysis. *Management science* 37(5):501–518.

Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PC, Rubin HR (1999) Why don't physicians follow clinical practice guidelines?: A framework for improvement. *Journal of the American Medical Association* 282(15):1458–1465.

Camerer C (1981) General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance* 27(3):411–422.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing & Service Operations Management* 13(1):2–19.

Cardoen B, Demeulemeester E, Beliën J (2010) Operating room planning and scheduling: A literature review. *European journal of operational research* 201(3):921–932.

Cerner (2017) Surginet solution. URL https://cerner.com/Solutions/Hospitals_and_Health_Systems/Perioperative/SurgiNet/?LangType=3081.

Chong YY, Hendry DF (1986) Econometric evaluation of linear macro-economic models. *The Review of Economic Studies* 53(4):671–690.

Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial judgment. *Science* 243(4899):1668–1674.

Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research* 58(4):802–816.

Dexter F, Dexter EU, Masursky D, Nussmeier NA (2008) Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesthesia & Analgesia* 106(4):1232–1241.

Dexter F, Ledolter J (2005) Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. *The Journal of the American Society of Anesthesiologists* 103(6):1259–1167.

Dexter F, Traub RD, Fleisher LA, Rock P (2002) What sample sizes are required for pooling surgical case durations among facilities to decrease the incidence of procedures with little historical data? *Anesthesiology: The Journal of the American Society of Anesthesiologists* 96(5):1230–1236.

Eijkemans MJ, Van Houdenhoven M, Nguyen T, Boersma E, Steyerberg EW, Kazemier G (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.

Erdogan SA, Denton BT, Cochran JJ, Cox LA, Keskinocak P, Kharoufeh JP, Smith JC (2011) Surgery planning and scheduling. *Wiley Encyclopedia of Operations Research and Management Science, Wiley Online Library* .

Gaur V, Kesavan S, Raman A, Fisher ML (2007) Estimating demand uncertainty using judgmental forecasts. *Manufacturing & Service Operations Management* 9(4):480–491.

Gomes C, Almada-Lobo B, Borges J, Soares C (2012) Integrating data mining and optimization techniques on surgery scheduling. *International Conference on Advanced Data Mining and Applications*, 589–602 (Springer).

Granger CW, Ramanathan R (1984) Improved methods of combining forecasts. *Journal of Forecasting* 3(2):197–204.

Guerriero F, Guido R (2011) Operational research in the management of the operating theatre: A survey. *Health Care Management Science* 14(1):89–114.

Gupta D (2007) Surgical suites' operations management. *Production and Operations Management* 16(6):689–700.

Hong CS, Hwang AS, Ferris TG (2015) Finding a match: How successful complex care programs identify patients. *California Healthcare Foundation* .

Hosseini N, Sir MY, Jankowski C, Pasupathy KS (2015) Surgical duration estimation via data mining and predictive modeling: a case study. *AMIA Annual Symposium Proceedings*, volume 2015, 640 (American Medical Informatics Association).

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Ibrahim R, L'Ecuyer P (2013) Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management* 15(1):72–85.

Jafarnia-Jahromi M, Jain R (2017) Non-indexability of the stochastic appointment scheduling problem. *arXiv preprint arXiv:1708.06398* .

James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*, volume 112 (Springer).

Joustra P, Meester R, van Ophem H (2013) Can statisticians beat surgeons at the planning of operations? *Empirical Economics* 44(3):1697–1718.

Kahneman D, Klein G (2009) Conditions for intuitive expertise: A failure to disagree. *American Psychologist* 64(6):515.

Kargar Z, Khanna S, Sattar A (2013) Using prediction to improve elective surgery scheduling. *The Australasian medical journal* 6(5):287.

Kayis E, Khaniyev TT, Suermondt J, Sylvester K (2015) A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Management Science* 18(3):222–233.

Kim SH, Chan CW, Olivares M, Escobar G (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 DM, Abubaker AO, Strauss RA (2013) Accuracy of predicting the duration of a surgical operation. *Journal of Oral and Maxillofacial Surgery* 71(2):446–447.

Luangkesorn KL, Eren-Doğu Z (2016) Markov chain monte carlo methods for estimating surgery duration. *Journal of Statistical Computation and Simulation* 86(2):262–278.

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.

Macario A (2010) Is it possible to predict how long a surgery will last? *Medscape Anesthesiology* 108(3):681–685.

Macario A, Dexter F (1999) Estimating the duration of a case when the surgeon has not recently scheduled the procedure at the surgical suite. *Anesthesia & Analgesia* 89(5):1241–1245.

Master N, Scheinker D, Bambos N (2016) Predicting pediatric surgical durations. *arXiv preprint arXiv:1605.04574* .

May JH, Spangler WE, Strum DP, Vargas LG (2011) The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management* 20(3):392–405.

McGlynn EA (1997) Six challenges in measuring the quality of health care. *Health affairs* 16(3):7–21.

Mincer JA, Zarnowitz V (1969) The evaluation of economic forecasts. *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, 3–46 (NBER).

Osadchiy N, Gaur V, Seshadri S (2013) Sales forecasting with financial indicators and experts' input. *Production and Operations Management* 22(5):1056–1076.

Ozen A, Marmor Y, Rohleder T, Balasubramanian H, Huddleston J, Huddleston P (2016) Optimization and simulation of orthopedic spine surgery cases at Mayo Clinic. *Manufacturing & Service Operations Management* 18(1):157–175.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* 61(8):1741–1759.

Roque DR, Robison K, Raker CA, Wharton GG, Frishman GN (2015) The accuracy of surgeons' provided estimates for the duration of hysterectomies: A pilot study. *Journal of Minimally Invasive Gynecology* 22(1):57–65.

Stepaniak PS, Heij C, Mannaerts GH, de Quelerij M, de Vries G (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 DP, Sampson AR, May JH, Vargas LG (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, Woodhouse S, Tan R, Patel S, Donovan J, Brogan K (2014) Operating theatre time, where does it all go? A prospective observational study. *Bmj* 349:g7182.

Van Donselaar KH, Gaur V, Van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Science* 56(5):766–784.

Wright IH, Kooperberg C, Bonar BA, Bashein G (1996) Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon-provided estimates. *The Journal of the American Society of Anesthesiologists* 85(6):1235–1245.

Zhou J, Dexter F (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, Miller D, Master N, Scheinker D, Bambos N, Glynn P (2016) Detecting inaccurate predictions of pediatric surgical durations. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 452–457 (IEEE).

Online Supplement

Table A1 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		
Asian	7%	7%
Black	5%	5%
White	72%	72%
Other	16%	16%
ASA Level		
0-1	3%	2%
2	44%	42%
3	51%	52%
4-6	3%	3%
Indicator for major anesthesia	95%	95%
Indicator for more than one surgeon	7%	7%
Indicator for more than one procedure	8%	7%

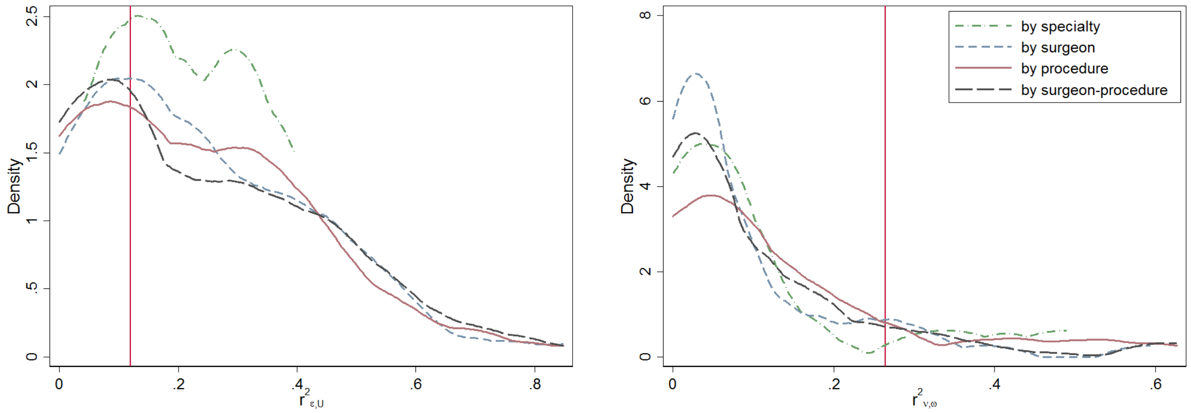
Notes. We report averages (standard deviation in parentheses) for continuous variables and percentages for binary or categorical variables.

Table A2 Quantifying the value of the surgeon's prediction by different groupings. Instead of Y, $log(Y)$ is used in (1) and (3).

	(1)	(2)	(3)	(4)	(5)
Measure	One group	By specialty	By surgeon	By procedure	By surgeon-procedure
No. of groups	1	12	42	49	81
$r_{Y,U}$	.17	.29 (.21, -.11, .67)	.32 (.30, -.44, .82)	.27 (.28, -.32, .87)	.29 (.31, -.60, .87)
$r^2_{\varepsilon,U}$	.09	.19 (.12, .05, .40)	.24 (.18, .00, .66)	.20 (.16, .00, .64)	.22 (.18, .00, .66)
$r^2_{\nu,\omega}$	.35	.12 (.14, .00, .43)	.09 (.11, .00, .52)	.13 (.17, .00, .66)	.10 (.14, .00, .66)

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

Figure A1 **Distribution of $r^2_{\varepsilon,U}$ and $r^2_{\nu,\omega}$ by different groupings.**



(a) Distribution of $r^2_{\varepsilon,U}$

(b) Distribution of $r^2_{\nu,\omega}$

Table A3 **Accuracy of Surgery Duration Predictions. Instead of Y, $log(Y)$ is used in (5), (7), (8), and (9).**

	Correlation	MSE	RMSE	MAPE	Cover-1hr	Cover-2hr
	Corrected physician input model					
One model	0.64	8007	89	33	0.36	0.64
Model by specialty	0.70	6380	80	30	0.42	0.68
Model by surgeon	0.75	5497	74	27	0.46	0.73
Model by procedure	0.74	5688	75	29	0.43	0.70
Model by surgeon-procedure pair	0.77	5088	71	27	0.46	0.73
	Regression model					
One model	0.74	5692	75	29	0.43	0.70
Model by specialty	0.73	5877	77	29	0.44	0.71
Model by surgeon	0.70	6430	80	30	0.44	0.69
Model by procedure	0.68	6971	83	30	0.44	0.70
Model by surgeon-procedure pair	0.67	7300	85	31	0.44	0.69
	Combined model: regression simple					
One model	0.72	6692	82	27	0.46	0.73
Model by specialty	0.77	5144	72	26	0.47	0.72
Model by surgeon	0.67	7789	88	28	0.46	0.72
Model by procedure	0.67	7706	88	27	0.47	0.71
Model by surgeon-procedure pair	0.58	11145	106	29	0.47	0.71
	Combined model: regression all					
One model	0.77	5030	71	26	0.46	0.73
Model by specialty	0.78	4924	70	26	0.46	0.73
Model by surgeon	0.65	8781	94	28	0.46	0.71
Model by procedure	0.70	6840	83	28	0.47	0.72
Model by surgeon-procedure pair	0.60	10214	101	30	0.46	0.70
	Combined model: 50-1					
One model	0.78	4935	70	28	0.45	0.72
Model by specialty	0.78	4948	70	28	0.46	0.71
Model by surgeon	0.77	5079	71	29	0.46	0.71
Model by procedure	0.76	5246	72	29	0.45	0.71
Model by surgeon-procedure pair	0.76	5303	73	29	0.45	0.71
	Combined model: 50-2					
One model	0.77	5314	73	28	0.45	0.72
Model by specialty	0.78	5065	71	27	0.47	0.72
Model by surgeon	0.78	4968	70	27	0.47	0.72
Model by procedure	0.76	5294	73	28	0.45	0.72
Model by surgeon-procedure pair	0.77	5194	72	27	0.47	0.73