The Impact of Surgeon Daily Workload and its Implications for Operating Room Scheduling

Yiwen Shen
Information Systems, Business Statistics and Operations Management,
Hong Kong University of Science and Technology, yiwenshen@ust.hk

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School, cwchan@columbia.edu

Fanyin Zheng
Analytics and Operations, Imperial College Business School, f.zheng@imperial.ac.uk

Michael Argenziano
Department of Surgery, New York Presbyterian/Columbia University Irving Medical Center, ma66@cumc.columbia.edu

Paul Kurlansky
Department of Surgery, New York Presbyterian/Columbia University Irving Medical Center, pk2245@cumc.columbia.edu

Abstract: In many service systems, an individual server's workload can have a substantial impact on service time and quality. Such effects are particularly important in healthcare systems which often operate under resource and time constraints. In much of the literature, this effect of workload has been primarily considered at the system and instantaneous level rather than the individual and cumulative level. In this study, we investigate this relationship in the context of cardiac surgery, i.e., how surgery duration and patient outcomes are affected by the individual surgeon's daily workload. Using a detailed data set of more than 5,600 cardiac operations in a large hospital, we quantify how individual surgeon daily workload (the number of operations performed by the focal surgeon) affects surgery duration and patient outcomes. To handle the endogeneity of surgeon daily workload, we construct instrumental variables using operational factors of the cardiac surgery department, including the regular surgery schedule of surgeons. We find that high daily workload for the focal surgeon is associated with longer surgery duration as well as post-surgery length-of-stay in the ICU and hospital. These results highlight the potential negative impact of high individual surgeon workload. We develop a surgical scheduling model that incorporates the estimated impact of surgeon daily workload. We solve the model by mixed-integer quadratic programming and show that our proposed schedule can substantially reduce total OR time and post-surgery length-of-stay. Our results suggest that hospitals should take into account the effects of individual surgeon daily workload when managing their ORs. Specifically, they can substantially improve patient flow and patient outcomes by smoothing individual surgeon's workload across days.

***Keywords**:* healthcare delivery, empirical operations management, behavioral operations, operating room scheduling, surgeon workload, quality of care

---

1. Introduction

Operating rooms are significant cost sinks as well as revenue generators of US hospitals (Rothstein and Raval 2018). They are extremely expensive medical resources, with estimated costs of up to 37 dollars per minute (Childers and Maggard-Gibbons 2018). At the same time, operating room generate nearly 50% of hospitals' total revenue (McDermott et al. 2017). Thus, improving the efficiency of how operating rooms are utilized is essential for hospitals to achieve their clinical and financial goals. Not surprisingly, various efforts have been devoted to operating room management from hospital administrators, policy makers, and healthcare professionals. Common recommendations include increasing capacity, better training of surgeons and staff, as well as managing schedules in order to facilitate more timely care for patients (see, e.g., May et al. 2011). However, most of the literature on surgical scheduling focus on the modeling and algorithmic side without identifying causal determinants of the duration and the outcomes of the operation. In this study, we focus on a novel factor that affects the operating room performance. Specifically, we investigate how individual surgeon's daily workload impacts the surgery duration and outcomes in the context of cardiac operations, and, in turn, how such impacts can be utilized to improve surgery scheduling.

The relationship between system workload and service performance has drawn increasing attention in the operations management community. Traditional operations management models generally assume service time is fixed and independent of system workload. However, a growing body of empirical research shows that the service time of human-serviced systems can be endogenously impacted by the overall system workload (e.g. Staats and Gino 2012 and Tan and Netessine 2014). Such effects are particularly important in healthcare, where resources are often constrained, and timely access to medical services is important. System workload has been shown to affect service time in different healthcare settings, such as intensive care units (ICUs) (Kc and Terwiesch 2012), patient transportation and cardiac surgery (Kc and Terwiesch 2009), emergency departments (Kc 2014, Batt and Terwiesch 2016), and paramedic teams (Bavafa and Jónasson 2024). Beyond service time, the effect of system workload on the quality of care, such as mortality and readmission, has also been investigated in both the operations management and medical communities (e.g., Kc and Terwiesch 2009, Kc and Terwiesch 2012, Needleman et al. 2011). In this paper, we investigate the impact of workload in the setting of operating rooms. Our empirical findings provide potential directions for how hospitals can improve surgical scheduling by incorporating such effects.

We consider a novel type of workload using a detailed data set of cardiac operations. In particular, we measure individual surgeon's daily workload by the total number of operations performed on a given day. In most of the existing literature in healthcare settings, the workload is measured at the system level in an instantaneous way. A common example of this type is the hospital unit's bed occupancy at the time of a patient's admission or discharge (e.g., Kc and Terwiesch 2012, Kuntz et al. 2015, and Berry Jaeker and Tucker 2017). In contrast, we focus on the *cumulative* workload at the *individual* surgeon level, which, to our best knowledge, is the first in the field of operations management. For cardiac operations, surgeons often have high ownership of their patients, thus the workload at the individual level would be a more relevant measure than that at the system (department) level. In addition, the effect of workload at the individual level may be subject to more behavioral and operational variations. On the other hand, cardiac operations usually take a long time to complete and are highly demanding for surgeons. Consequently, high cumulative workload in a day may have negative impacts on surgeon's performance. For these reasons, the effect of individual surgeon's cumulative workload may differ from that of the system-level, instantaneous workload. Understanding such effects can provide new operational levers for hospitals to improve their performance. For example, while the department-level workload is hard to change, the hospitals may have more flexibility in changing daily workload of individual surgeons in their scheduling.

In our study hospital, it is common for a cardiac surgeon to perform multiple operations a day: the median surgeon's daily workload is two operations, and the maximum is four operations. On average, each operation takes more than seven hours to complete. Although some parts of the operation can be done by other members of the medical team, performing multiple operations a day is a heavy physical and cognitive burden for the surgeon. With long working hours, surgeons can suffer from physical and mental fatigue, which may lead to worse medical outcomes (Janhofer et al. 2019). In addition, high surgeon workload may strain other ancillary resources, such as nurses and post-surgery recovery beds. In this study, our goal is to understand the effect of surgeon workload on operating time and surgery outcomes. Due to data limitations, we are not able to fully differentiate the driving factors leading to such effect, e.g., surgeon fatigue, operational constraints, or their combination. Based on our empirical findings, we then develop and solve an illustrative surgical scheduling model that incorporates the estimated effect of surgeon daily workload. It provides a potential direction for improving the scheduling procedure.

We examine the impact of surgeon daily workload using a data set of cardiac surgery from a large academic medical center. Our data comes from the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database for our partner hospital and contains detailed information of more than 5,600 cardiac operations that are performed over a horizon of four years. We measure the impact of surgeon daily workload – at the individual and cumulative level – on multiple outcomes. First, we examine how surgeon daily workload affects the surgery duration of each case, as measured by its Operating Room (OR) time or incision time. This sheds light on the relationship between individual server's workload and service time in the context of cardiac surgery. Next, we analyze the effects of surgeon daily workload on the patient's post-surgery length-of-stay (LOS) in the ICU and in the hospital. The post-surgery LOS is important for the hospital as it affects the demand for downstream resources (e.g., ICU and ward beds) and overall throughput efficiency. We also examine whether the treatment effect of surgeon workload is heterogeneous for different types of patients, e.g., elective patients and non-elective patients.

Our detailed data set allows us to control for a comprehensive set of demographic, risk, and operative factors that may also affect the surgical outcomes. However, we still face a major challenge in identifying the causal effect of surgeon daily workload. That is, the surgeon daily workload can be endogenous. The endogeneity arises when there are risk factors which are considered by the surgeons when they schedule their cases but are not fully observed in the data (e.g., patient's cognitive status). These unobserved factors will affect both surgeon daily workload and surgical outcomes, thus violating the exogeneity condition for identification. For example, a surgeon may schedule more cases by packing in low risk, "easy" cases. If some measures of low risk are unobserved in the data, it will generate a negative bias in the estimated effect of surgeon daily workload. Alternatively, a high risk case may be squeezed in so as to avoid delaying care for that patient, resulting in a positive bias in the estimated effect of surgeon daily workload.

We handle the endogeneity bias by utilizing an instrumental variable (IV) approach, which is essential for our identification. A valid IV in our study should influence the surgical outcomes only via the surgeon daily workload, conditional on the other covariates in the model. We construct two IVs by leveraging operational factors in cardiac surgery. The first IV is the number of cases performed by other cardiac surgeons on the same day. As many resources (staff, operating rooms, ICU beds, etc.) are shared by surgeons in the cardiac department, more operations by *other* surgeons tend to limit the daily workload of the focal surgeon. The second IV is the proportion of the focal surgeon's elective cases performed on the same weekday as the focal case over a long horizon. Surgeons tend to schedule their operations on certain weekdays in a week, especially for elective patients. The second IV thus captures the impact of such long-term working pattern on surgeon daily workload. We show that the two IVs significantly impact surgeon's daily workload, after controlling for a comprehensive set of demographic, risk, and operational factors. In addition, we provide evidence that the two IVs are unlikely to be correlated with patient's unobservable factors.

We find that higher daily workload for a surgeon is associated with longer surgery duration and post-surgery length-of-stay of their patients. Specifically, we estimate that adding one more case to a surgeon's daily workload increases the OR time by 27.3 minutes for each case performed by the surgeon in the day. This is a 6.5% relative increase. Similar impact is also observed for the incision time. These results highlight how workload affects service time when the workload level is already high, which is consistent with the second tipping point empirically observed in Berry Jaeker and Tucker (2017). Higher surgeon daily workload also leads to longer post-surgery LOS of patients in both the ICU and the hospital: when the surgeon does one more case in a day, the affected patients are expected to stay in the ICU (resp. hospital) for 1.1 (resp. 1.2) more days after their operations on average. These results highlight the negative impacts of high surgeon daily workload on surgery duration and patient outcomes. We further show that there is substantial heterogeneity in the effect of surgeon daily workload for elective and non-elective (urgent and emergent) patients. We find that the effects of surgeon daily workload on post-surgery LOS are significant only for non-elective patients and have larger magnitudes than those for the full sample. One possible explanation for such heterogeneity is that the non-elective patients are generally more severe, thus their surgical outcomes are more sensitive to surgeon workload.

Based on the empirical results, we develop an illustrative surgery scheduling model that incorporates the effect of surgeon daily workload. In most of the existing literature, the surgery duration and patient outcomes are assumed to be exogenous with deterministic or stochastic distributions. However, as shown by our study, they can endogenously depend on surgeon daily workload, which is determined by the surgical schedule itself. We thus propose a scheduling model that accounts for such effect. We consider the intervention of changing the surgery dates to mitigate the negative impact of high daily workload for a surgeon. This intervention does not require any expansion of the OR capacity, and, thus, is a cost neutral intervention. We formulate and solve the model as a mixed-integer quadratic programming problem. Using our estimated effects, we find the new schedules from our model can substantially reduce the total OR time and post-LOS, which are economically important for the hospital. This highlights the benefits of accounting for the impact of surgeon daily workload in surgery scheduling. Our work reveals the benefit of load balancing at the *individual* level in the context of cardiac surgery scheduling. The proposed scheduling model is intended to be an illustrative example rather than prescriptive, but in combination with our empirical findings, sheds light on the potential direction of improvements that could be achieved with better scheduling approaches.

The rest of the paper is organized as follows. The next section is a brief overview of related literature. Section 2 describes the data and clinical setting of our study. In Section 3, we develop the econometric framework and estimation methodology. Section 4 provides the main empirical findings. We discuss our surgery scheduling model in Section 5. Section 6 concludes the paper and discusses future directions. The Electronic Companion includes variable definitions, model formulation details, and supplementary tables.

1.1. Literature Review

Our study is related to four primary streams of literature: (1) effect of system workload on service rate and quality, (2) volume-outcome relationship, (3) impact of surgeon fatigue, and (4) operating room scheduling.

While traditional models usually assume a constant and exogenous service rate, there is rich literature, both analytical and empirical, which focuses on the relationship between system workload and service rate. The dynamic queueing literature has studied the optimal policies and system performance when service rate is adaptive to system state (e.g., Delasay et al. 2016). To examine how human workers actually behave under varying workload, various empirical research has been conducted using observational data in real-world settings, and the results are mixed. In some healthcare settings, it is found that the service rate increases when the servers face high workload (Kc and Terwiesch 2009, Kc and Terwiesch 2012). However, the opposite direction of the impact is also observed empirically. For example, Green and Nguyen (2001) show patient's LOS can increase when patient load becomes higher. The seemingly opposite effects can be partially reconciled by an inverted-U shape pattern. That is, the service time first increases and then decreases with the workload level. Empirical evidence for this inverted-U shape pattern is found using restaurant chain data in Tan and Netessine (2014), and in the healthcare setting in Batt and Terwiesch (2016) and Berry Jaeker and Tucker (2017). Different mechanisms have been proposed to explain the impact of workload, such as server speedup (Kc and Terwiesch 2009), task reduction (Oliva and Sterman 2001), multitasking (Freeman et al. 2017), and server fatigue (Kuntz et al. 2015).

There is also a rich literature studying the effect of workload on servers' behavior and quality. Freeman et al. (2017) find that gatekeeper-providers would alter their service configuration and referral decisions in response to their workload. In multiple healthcare settings, the quality of care is found to suffer under high workload, such as higher mortality and readmission rate (Kc and Terwiesch 2009, Kuntz et al. 2015, and Berry Jaeker and Tucker 2017), as well as longer LOS and higher likelihood of transfer-up (Kim et al. 2015). The positive linkage between hospital workload and mortality is also observed in the medical literature (e.g., Neuraz et al. 2015). Our study contributes to this line of literature by considering a novel type of workload in healthcare settings, i.e., number of operations performed in a day by the focal surgeon. We find a surgeon's high workload is associated with longer surgery duration and post-surgery length-of-stay of their patients.

In most of the existing healthcare literature, workload is measured at the system level, e.g. bed occupancy in different hospital units (e.g., Kc and Terwiesch 2012, Kuntz et al. 2015, and Berry Jaeker and Tucker 2017). In contrast, the impact of individual server's workload is relatively understudied. Different from these works, we consider the workload at the individual surgeon level, as it is more relevant in the cardiac surgery setting. On the other hand, the existing literature largely focuses on the *current* workload level using the instantaneous, or at least recent, system state, e.g., the unit's bed occupancy at the time of patient's admission or discharge (e.g., Kc and Terwiesch 2012, Kim et al. 2015, and Berry Jaeker and Tucker 2017). Instead, we study the effect of surgeon's *cumulative* workload using the total number of operations performed in a day. In a survey paper, Delasay et al. (2019) develop a general framework to describe the impact of workload on service times. Our workload measure resembles the *extended load* in their framework, which tracks the history of workload, but is measured at the individual level.

Two related papers in healthcare operations management also consider the workload at the individual level. Kc (2014) uses operational data at the individual level like in our study. However, Kc (2014) studies how multitasking (caring for multiple patients at the same time) of ED physicians affects service time and outcomes, i.e., the impact of processor-sharing. This is a very different workflow from our work on the impact of surgeon daily workload. Bavafa and Jónasson (2024) measure the fatigue level of paramedic crews using the cumulative number of prior jobs completed during a shift. They show that workers' fatigue increases the mean and uncertainty of the service time. However, their clinical setting and empirical approach are very different from ours (ambulance service versus cardiac operation). In addition to average service time, we also reveal the impact of surgeon daily workload on the surgical outcomes and develop a scheduling model that incorporates such effects.

Next, our work is related to the literature on volume-outcome relationship in healthcare management. In the medical community, there is vast evidence supporting a positive relationship between a surgeon's (or a hospital's) volume and surgical outcomes (e.g., Bashir et al. 2017). The volume in these studies usually refers to the number of operations performed by the surgeon in a relatively long period (e.g., the past one year). Research in different empirical settings has been conducted to investigate the drivers and mechanisms behind the relationship, e.g., learning and specialization. Relevant works in this area include Kc and Terwiesch (2011), Kc and Staats (2012), and Staats and Gino (2012) among others. Recent work by Pourghannad and Wang (2025) shows that the effects of surgical volume on surgery duration is heterogeneous across patients. Complementing this line of research, we investigate the impact of a surgeon's short-term volume, i.e., cases performed in a day, on surgery duration and surgical outcomes.

Our work also contributes to the literature on surgeon fatigue. Medical and psychological literature has shown that high cumulative workload can lead to server fatigue (Thomas et al. 2012, Hockey 2013), which is commonly measured by the total number of task completions In addition, the cumulative number of task (Bavafa and Jónasson 2024). As the work of a surgeon is highly demanding both physically and mentally, the potential negative impact of surgeon fatigue has long been a focus of the medical community (see a survey in Janhofer et al. 2019). However, empirical results on the relation between surgeon fatigue and worse patient outcomes are mixed in the medical literature (see, e.g., Thomas et al. 2012, Govindarajan et al. 2015). Our work sheds light on this important problem using a detailed data set of cardiac surgery and rigorous econometric analysis. Different from existing medical literature, which focuses primarily on correlational analysis, we use IVs to control for the endogeneity to generate causal estimates. As a limitation of our study, our IV-based method cannot identify the exact factors driving the estimated effects of surgeon daily workload (e.g., through surgeon fatigue, operational constraints, or both).

Finally, we contribute to the literature of operating room scheduling. Operating rooms are big cost centers and revenue generators of the hospital. The literature on operating room scheduling is huge (see Keskinocak and Savva 2020, May et al. 2011). Different objectives are considered in operating room scheduling, including minimizing costs, maximizing profit and utilization, as well as smoothing downstream census (e.g., Freeman et al. 2016 and Zenteno et al. 2016). Staff planning in the operating room environment is also widely studied (Rath and Rajaram 2022). However, most of the existing literature assumes the surgery duration to be exogenous and independent of surgeon workload. To our best knowledge, we are the first to develop a scheduling model that incorporates the effects of surgeon daily workload. A recent example of endogenous surgery duration is Pourghannad and Wang (2025), in which the surgery duration is affected by the surgeon's past volume. Different from their work, we focus on the impact of surgeon daily workload and apply an IV method to address the endogeneity bias. In addition, we consider the assignment of operations to available days, instead of matching patients and surgeons as in their work.

2. Data and Clinical Setting

2.1. Data Selection

In this study, we use cardiac surgery data from a large academic hospital over the period of July 2015 to July 2019. The data is collected from the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database. This study was approved by the Columbia University Institutional Review Board, protocol number AAAT7723. The STS data contains detailed information of patient demographics, risk factors, preoperative status, operative procedures and timelines, as well as postoperative events for all cardiac operations occurred in the sample period in our partner hospital. This comprehensive data set allows us to control for the severity of patients and complexity of operations when analyzing the impact of daily workload, as measured by the total number of surgical operations per day. We also have the surgeon's identifier for each case, which controls for surgeon-specific differences in outcomes (e.g., Pourghannad and Wang 2025).

We start from 5,604 cases from the STS data in the four year horizon. We first drop 20 cases that are canceled before or during the operation. We then drop 232 cases from seven "infrequent" surgeons in our sample. These surgeons performed a very small number of cases during the four year sample period. They are dropped for the following two reasons. First, these surgeons are more likely to only perform unusual procedures that require special expertise. Second, the small sample size of these surgeons does not allow us to effectively control for surgeon fixed effects. Thus, we focus on the cases from the other eight surgeons, each of which performed at least 200 cases in the sample period. In addition, we find that there are no specific procedure types that are performed exclusively by a particular surgeon. This leaves us with a sample of 5,352 cases in total, which consists of 95.5% of the initial sample.

2.2. Patient Risk Factors and Surgery Metrics

The STS data set provides a comprehensive set of characteristics of each patient, which allows us to control for the patient's severity and condition. It contains basic patient demographics such as gender, age, and race. It also includes risk factors such as a patient's status for liver illness, lung disease, diabetes control, and renal failure, as well as preoperative conditions such as whether the patient experienced heart failure, cardiogenic shock, or myocardial infarction (MI) before the operation. Table 1 reports the summary statistics of patients' gender, age, and critical status. Specifically, a patient is classified as critical if he or she experiences a cardiogenic shock or syncope before the operation. In Section EC.1 of the Electronic Companion, we provide a detailed description of other risk factors in our econometric framework and their summary statistics.

Table 1 Summary Statistics of Patients for Full, Elective, and Non-elective Samples (Full: N = 5,352, Elective: N = 2,480, Non-elective: N = 2,872)

	Full Sample			Elective Sample			Non-elective Sample		
	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
Gender: Male	0.675	-	-	0.655	-	-	0.692	-	-
Age	64.73	66.00	12.56	66.08	68.00	12.12	63.59	65.00	12.79
Critical	0.103	-	-	0.031	-	-	0.166	-	-

Cardiac operations are divided into four main risk categories (surgery status) in increasing order of patient severity and urgency in need of operation: elective, urgent, emergent, and salvage. Elective cases are operations that can be deferred without increased risk; urgent cases are supposed to be performed during the same clinical stay to reduce further risk; emergent and salvage cases refer to the situation that requires emergent operations with no delay upon the outbreak.[1] While the hospital has relatively high flexibility in scheduling elective cases, the schedules of urgent cases should ideally be completed within 72 hours (e.g. Karim et al. 2016), and the hospital has less control over the timing of emergent and salvage cases. In our data, a significant proportion of the operations are urgent or emergent cases, which consists of 53.5% of the sample. In addition to surgery status, we also obtain the procedure information for each case from the STS data. We classify the operations to different types to control for the differences in their procedures. The classification of surgery type is described in Section EC.1 of Electronic Companion.

For each operation, we can determine its operating room (OR) time and incision time from the STS data using the timestamps (hour, minute, and second) of its OR entry and exit, as well as skin incision start and end. We calculate the OR time of each case as the time elapsed between its OR entry and OR exit using the exact timestamps. The OR time can be decomposed to three stages: pre-incision time, incision time, and post-incision time. The incision stage corresponds to the time between skin incision start and end, and the pre-incision (resp. post-incision) stage refers to the time before (resp. after) it. In cardiac operations, the tasks in the pre- and post-incision stages (e.g., pre-operative tests, cleaning up after surgery) can be largely performed by medical staff or surgical fellows without the presence of the focal surgeon. On the other hand, the incision stage requires relatively high level of participation of the surgeon. Thus, the incision time is the most relevant measure for a surgeon's working time than the OR time.

We present the summary statistics for OR time and its three stages by the four surgery status in Table 2. We see fairly consistent pre-incision time spent in the OR across elective, urgent, and emergent patients. This is likely due to the fact that, for these patients, the pre-incision stage is very protocol-driven where the patient goes through standard preparation before the surgeon actually cuts the patient. On the other hand, the pre-incision time for salvage patients is shorter. Although the sample size is very small, this may be indicative of the highly time-sensitive nature of these procedures. We also find that the average incision and OR time are longer for the urgent and emergent cases than that for the elective cases. This is not surprising as the non-elective cases tend to be more complicated and thus take longer. On average, the incision stage (4.78 hours) consists of 67% of the OR time (7.11 hours).

Finally, we obtain from the STS data each patient's hospital admission date, surgery date, and discharge date. Thus, we can compute the patient's LOS before and after the operation. We calculate the pre-surgery LOS (pre-LOS) for each patient as the number of days between the hospital admission and the operation, and the post-surgery LOS (post-LOS) as that between the OR exit and hospital discharge. We also have the total time a patient spends in ICU after the operation, including both the initial ICU visit and the potential revisits. Their summary statistics are reported in Table 2. First, we can see that the elective cases have relatively short pre-LOS. This is because most elective patients are admitted one day before or on the same

[1] See page 154 in the training manual: https://www.sts.org/sites/default/files/Training%20Manual%20V2-9%20June%202020.pdf. Accessed on April 15th, 2024.

Table 2 Summary Statistics of Surgery Metrics and Patient Outcomes: mean (std. dev.)

	Elective	Urgent	Emergent	Salvage	All
Pre-incision time (hours)	1.48 (0.28)	1.51 (0.31)	1.48 (0.45)	1.18 (0.51)	1.49 (0.31)
Incision time (hours)	4.52 (1.56)	4.88 (1.76)	5.80 (2.31)	5.96 (2.27)	4.78 (1.75)
Post-incision time (hours)	0.73 (0.38)	0.78 (0.40)	0.87 (0.47)	0.95 (0.49)	0.76 (0.40)
OR time (hours)	6.79 (1.79)	7.24 (1.99)	8.31 (2.59)	8.28 (2.60)	7.11 (1.99)
Pre-surgery LOS (days)	1.16 (2.92)	4.90 (9.83)	15.32 (30.11)	8.09 (7.54)	3.90 (11.19)
Post-surgery LOS (days)	8.73 (7.85)	13.22 (18.53)	25.60 (22.27)	20.45 (12.15)	12.02 (15.55)
Total ICU time (days)	3.61 (5.55)	5.88 (11.36)	13.29 (17.35)	15.59 (13.03)	5.37 (10.09)
Number	2479	2488	374	11	5352

day of their operation, reflecting the flexibility in their schedule. Second, the elective cases have the shortest post-LOS and total ICU time, while the emergent cases have the longest. This reflects the fact that the patients with elective cases are generally less severe than those of urgent and emergent cases and usually follow typical post-surgery protocols. Notice that the post-LOS includes the ICU time by definition, and the two metrics have a high correlation of 0.80 in the sample.

3. Empirical Framework

We now develop the econometric framework for identifying the effect of surgeon daily workload on surgery duration and outcomes. For each case i, denote its surgeon's daily workload on i's surgery date by $Workload_i$. We measure $Workload_i$ by the number of cases in addition to i performed by its surgeon on the same day. Note that if the surgeon performs only one case in a day, $Workload_i$ equals to zero by definition. In addition to the operations recorded in the STS data, the surgeon may have other duties which may also contribute to their overall daily workload. These duties are not included in the data, and thus are not accounted in our study. That said, we expect our measure, i.e., the number of operations performed, captures the primary workload of surgeons, as cardiac surgeries are extremely complicated procedures and take a long time to complete.

In the medical literature, it is common to measure surgeon workload based on the number of cases performed (e.g., Reijmerink et al. 2024). In Section EC.2 of the Electronic Companion, we further explain why we consider the number of cases as a valid measure for surgeon's daily workload. First, we find that the distributions of surgery duration (OR and incision times) are similar for cases with different surgeon workload levels (e.g., those on single-case versus. two-case days). Thus, the number of cases largely reflects the actual working time of the surgeon. Second, as a robustness check, we also measure $Workload_i$ by the total incision time of other cases performed by case i's surgeon on its surgery day, which takes continuous values. The two workload measures have a high correlation of 0.93 in our sample and lead to very similar estimation results (see Section 4.3). We also show that the patient's predicted risk is not correlated with surgeon daily workload. In addition, we check the distributions of surgeon daily workload for ten common procedure types in our sample (97% of all observations). The distributions are generally similar across

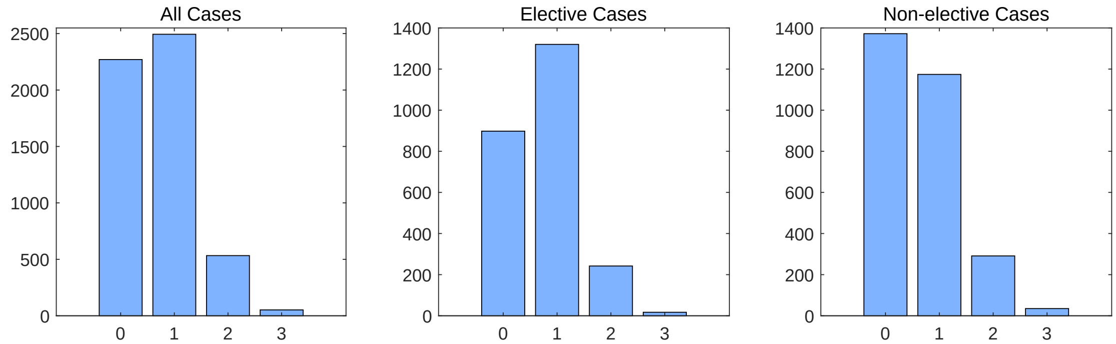


Figure 1 Frequency distributions of $Workload_i$ (number of additional cases excluding focal case i) for full (left), elective (middle), and non-elective (right) samples

different procedures. We do not find certain procedure types exclusively scheduled on single-case or multi-case days. These results are discussed in Section EC.2 of the Electronic Companion.

Figure 1 shows the distribution of $Workload_i$ (number of additional cases) for the full sample, the elective sample, and the non-elective sample. We find that it is very common for a surgeon to perform multiple cases – i.e., $Workload_i \geq 1$, in a day. For at least half of the cases, the focal surgeon performs at least two cases on the focal surgery date. This observation holds for the full sample as well as the elective and non-elective samples. Moreover, for about 11% of the cases, the surgeon's total daily workload is three or four cases.[2] Such high daily workload of surgeons may affect surgery duration and patient outcomes. We also see that non-elective cases are more likely to be scheduled on a single-case day than elective cases, although the difference is limited (37% for elective and 47% for non-elective). Surgeons have more flexibility in scheduling the elective cases; thus, they tend to schedule elective cases on their common working days, which usually have more than one case. We find that our results are robust when we drop the emergent and salvage cases, which are more likely to happen on single-case days (see Section 4.3).

We control for an extensive set of demographic, medical, and operative factors as explanatory variables in our estimation as described in Section 2.2. We also include five operational variables: dummy variables for the day of week, month, and year of the operation, the pre-LOS of patient, and the cardiac patient census in the hospital. The *cardiac patient census* is calculated for each day t as the number of patients in our sample that have been admitted before day t, but have not been discharged from the hospital by day t. We use the cardiac patient census to measure the system congestion level, which has been shown to affect healthcare outcomes (e.g., Kim et al. 2015). We note that the cardiac department in our study operates in a relatively independent manner. Thus, the cardiac patient census serves as a good measure of the congestion level faced by the cardiac department. Specifically, the cardiac surgery ICU is not available to non-cardiac

[2] In cardiac operations, surgeon only needs to participate in key steps of the surgery (Zhang et al. 2023). Thus, paralleling of operations is common in practice. This explains why it is possible to schedule multiple long cases in a single day.

surgery patients. As most of the patients (99.5%) go to the cardiac surgery ICU after their operation, we also expect the cardiac patient census to be highly correlated with the downstream cardiac ICU congestion. The cardiac patient census has a mean of 62.3 patients with a standard deviation of 9.3. As an additional check, we also construct a cardiac surgery ICU census using the patients in our sample, under a reasonable assumption that they visit the ICU once after exiting the OR. We find that including the cardiac surgery ICU census in the analysis does not affect our estimation results.

In total, we have 82 covariates in our model, including demographic, medical, and operative factors, as well as system congestion level, and dummies for time and surgeon fixed effects. We provide a detailed description of these independent variables in Section EC.1 of the Electronic Companion. We represent these variables plus a constant by X_i for case i. To estimate the effect of daily workload, we employ the linear model for a dependent variable y_i:

$$y_i = \beta X_i + \gamma Workload_i + \varepsilon_i, \tag{1}$$

where the error term ε_i follows a normal distribution. We use equation (1) to estimate γ, which is the effect of daily workload on y_i averaged across all cases by the surgeon in a day. We consider four dependent variables related to surgical metrics and patient outcomes: OR time and incision time of the surgery, as well as the post-LOS and total ICU time of the patient. We do not estimate the effect of surgeon daily workload on common binary outcomes of an operation such as mortality and reoperation. This is due to the size of our data set and the extremely low occurrence of these adverse outcomes. For example, the mortality and reoperation probability in our sample are 2.7% and 5.9%, respectively. In addition, since our model includes many categorical variables for patient's risk and operative factors, some of them can lead to perfect prediction for the binary outcomes, thus prohibiting estimation of the effect. We believe that accurately estimating the effect on the binary outcomes would require a much larger sample which provides more variation in the outcomes.

As a naive approach, we can estimate the coefficients in (1) by ordinary least squares (OLS) and interpret the estimated γ as the effect of daily workload on the dependent variable y_i. However, this approach ignores the endogeneity in the daily workload of surgeons. That is, the surgeon daily workload can be affected by patients' severity factors that are unobserved in the data but are considered by the surgeon (e.g., a patient's cognitive state). For example, the surgeon may schedule more cases in a day if the unobserved severity levels are lower and imply shorter OR times. Consequently, both the dependent variable (e.g., OR time) and the daily workload (e.g., number of additional cases) are affected by regressor X_i as well as the unobserved severity factors. If we ignore this endogeneity problem, the estimated coefficients will be biased. In the example described above, ignoring the unobserved severity factor introduces a negative bias to the estimate of γ, as the unobserved severity level is negatively correlated with the daily workload and positively correlated with OR time. The opposite bias may also be observed if the unobserved severity factors are positively correlated with the surgeon's daily workload, e.g., when surgeons need to pack in more cases due to the deterioration of sicker patients. To address the endogeneity bias, we employ an instrumental variable (IV) method to obtain consistent estimates of the coefficients. We construct two IVs using the operational data from the cardiac department and discuss their validity in the next section.

We acknowledge the effect of workload in (1) is likely to vary across cases within the same day, e.g., the first and last case by the surgeon in a day. However, as we do not have information on the decision-making process for how cases are scheduled within a day, we cannot effectively control for the potential endogeneity in the sequencing of cases. More importantly, given the limited observations where a surgeon performs three or four cases in a day, it is not possible to estimate the heterogeneous effects for cases at each sequencing position in a day. Thus, in our main specification, we choose to estimate the effect averaged across all cases performed by the surgeon in a day. As a robustness check, we find consistent results when estimating the effect for being the non-first case of the surgeon as well as the impact of surgeon's prior incision time in a day. We discuss these analyses in Section 4.3.

3.1. Instrumental Variables

We propose two novel IVs and discuss their validity. The two IVs are defined (and vary) on the surgeon-day level. This is inline with our estimation of the average effect in (1) for all cases by the surgeon in a day.

3.1.1. Total Cases by Other Surgeons The first IV we consider is $TotOther_i$, the total number of cases performed by *other* surgeons on the same day as case i. It satisfies the relevance condition because the number of cases by other surgeons can affect the daily workload of the focal surgeon through *resource sharing* across surgeons. In our study hospital, although the surgeons' schedules are typically fixed well in advance in terms of the day of the week they are assigned an OR, the exact number of cases they will perform on a given day is usually not finalized until shortly before the day starts. In addition, while the hospital uses a block booking system, it only serves as a loose guidance for how surgeons schedule their cases,[3] and many resources are still shared by the surgeons within a day (e.g., staff, medications, equipment). For example, a perfusionist has to work for multiple surgeons in a day when the department workload is high. If the other surgeons are performing many operations whose patients must stay in the ICU post-surgery, this could impact the ICU bed-availability for the focal surgeon's patients. Thus, more cases performed by other surgeons in a given day could translate into fewer resources available for the focal surgeon. This tends to limit the workload of the focal surgeon on the same day. The resource sharing phenomenon may be why such a large focus of the literature on workload has been at the system/unit level rather than at the individual level as we study. We aim to pick up such variation using the IV. By the above discussion, we expect $TotOther_i$ to be negatively correlated with the focal surgeon's daily workload.

[3] Using the periods with block information, we find that roughly 20% of cases are performed out of the focal surgeon's blocks.

We next consider whether $TotOther_i$ satisfies the exclusion restriction. The surgeons in the cardiac department at our study hospital have substantial ownership of their patients and schedules. They usually do not coordinate with other surgeons beyond whether there is available OR time and relevant resources when scheduling their own cases, and it is entirely up to the discretion of the focal surgeon which operations to prioritize amongst his/her own patients. Thus, an individual surgeon has little control over other surgeons' patients and schedules. This suggests that $TotOther_i$ should not be directly correlated with the unobserved severity factors of the focal surgeon's patients. A remaining concern is that the other surgeons' workload may be correlated with the downstream congestion, which can affect the surgical outcomes. We emphasize that the IV we use is the same-day workload, while the downstream congestion is based on cumulative prior workload. In addition, we have controlled for cardiac patient census in our model (1) to account for the downstream congestion.[4] Similar approaches of controlling for downstream hospital/unit congestion have been used in a number of studies when congestion has been used as an IV (e.g. Kim et al. 2015, Freeman et al. 2021, Soltani et al. 2022).

The statistics of $TotOther_i$ is reported in Table EC.13 of the Electronic Companion. We find that there are on average about four cases performed by other surgeons in a day. There is substantial variation in $TotOther_i$, as shown by its standard deviation of 1.79 cases. This holds for both the full sample as well as the elective and non-elective samples.

3.1.2. Ratio of Elective Cases on Current Weekday We construct a second IV that captures the frequency a surgeon performs elective cases on the focal weekday. This is based on the following operational feature within the cardiac surgery service. Cardiac surgeons usually have multiple responsibilities in addition to performing operations, including seeing patients in the office, teaching, and attending conferences. Thus, the hospital tends to allocate OR time for each surgeon on specific weekdays to reduce the uncertainty in his/her schedule. These block schedules are often set weeks and even months in advance. Such approach is especially significant for elective cases, which are typically scheduled at least two weeks in advance. Thus, if a given weekday (e.g., Tuesday) is associated with a large proportion of surgeon s's recent elective cases, it is likely to be one of the typical days surgeon s is allocated in the block schedule. This leads to higher expected workload for surgeon s on the days that fall on the same weekday.

To capture the long-term working pattern of surgeons, we construct a second IV as the ratio of focal surgeon's elective cases that fall on the current weekday, $ElecRatioWD_i$. For case i, it is calculated as:

$$ElecRatioWD_i = \frac{\text{Num. of elective cases by } s \text{ in } [t-L, t+L] \text{ performed on } WD_i}{\text{Total number of elective cases by } s \text{ in } [t-L, t+L]}, \tag{2}$$

[4] As stated in Woodridge (2010), the exclusion restriction only requires the IV to be uncorrelated with the residual in the model, after all other (exogenous) variables are controlled. Thus, it is permissible for the IV to be correlated with one of the control variables (downstream congestion in our model).

where WD_i denotes the weekday on which case i is performed; t and s are the surgery date and surgeon of case i; L is a parameter determining the horizon around current date, i.e., L days before and after day t. We exclude the elective cases done in the current week in both the numerator and denominator of (2). [5] IVs based on long-term time average of workload have been widely used in OM literature (see, e.g., Tan and Netessine 2014, Freeman et al. 2017, Soltani et al. 2022). Our second IV follows a similar idea.

The operational factor captured by $ElecRatioWD_i$ is reflected by the block booking system used in the hospital, in which a surgeon is assigned fixed time slots (blocks) and resources to perform their operations. We obtained the surgeons' block schedule for 22 out of the 48 months in our sample; the remaining months were unavailable due an administrative-leave of the scheduling staff. We find that surgeons' blocks indeed tend to fall on specific weekdays. For example, 71% of Surgeon 3's blocks fall on Wednesday and Friday. On the other hand, 92% of elective cases are performed within a surgeon's blocks. These factors suggest that the elective ratio in (2) is an effective indicator for the surgeon's working pattern among weekdays. As the surgeons' schedules are adjusted infrequently (e.g., twice a year), we use $L = 180$ days in (2) to calculate the IV. In robustness checks, we try other values of L and obtain similar results (see Section 4.3). The summary statistics of $ElecRatioWD_i$ are reported in Table EC.13. The mean of $ElecRatioWD_i$ is 22% for the full sample. The standard deviation of the IV is 12%, suggesting there is substantial variation that can be leveraged for identification.

In summary, we expect $ElecRatioWD_i$ to be positively correlated with the surgeon daily workload, satisfying the relevance condition as a valid IV. We also expect $ElecRatioWD_i$ to satisfy the exclusion restriction. First, in the calculation of $ElecRatioWD_i$, we use a relatively long horizon and have dropped the focal surgeon's cases (including the focal case) in the current week. Second, a surgeon's block schedule is usually fixed well in advance and adjusted very infrequently. Thus, $ElecRatioWD_i$ tends to reflect the surgeon's long-term working pattern over weekdays and is unlikely to be correlated with the severity factors of patients treated by the surgeon in the current day.[6] Section 3.3 provides further empirical evidence for the relevance and exogeneity conditions of this IV.

One might be concerned that a case performed on a non-regular working day may be more complex and time sensitive, and that these unobserved factors may jeopardize the validity of our IV. While one can never fully rule out all potential threats to identification, the potential bias introduced by such unobservables would only attenuate our estimated negative effects of daily workload. This is because the surgeon's daily workload tends to be high on the regular working days with large $ElecRatioWD_i$. If the cases on a surgeon's non-regular working day (low $ElecRatioWD_i$) were on average more severe, the days with low

[5] In rare cases, we set $ElecRatioWD_i$ to be zero if the number of elective cases in the denominator is smaller than ten, as it suggests that the surgeon was not working regularly for the period considered. Our results are robust if we drop these cases entirely.

[6] The correlation between $ElecRatioWD_i$ and the focal surgeon daily workload is 0.19. Thus, the IV is not mechanically correlated with surgeon's workload on current day.

surgeon workload would be associated with more unobserved risk. This relationship implies that the true (negative) effects of surgeon daily workload might be larger in magnitude compared with our estimation results. Thus, such potential violation of the exclusion criteria would make it the estimated effect of surgeon daily workload less negative, and our results should be interpreted as conservative estimates.

3.2. Estimation Methods

We estimate the effect of daily workload in model (1) using the two IVs introduced above. We estimate the linear model (1) using the two-stage least squares (TSLS) regression (Woodridge 2010). The TSLS estimation is conducted as follows. In the first stage, we regress the daily workload on the exogenous variables X_i and the two IVs *together* using OLS:

$$Workload_i = \rho X_i + \eta_1 TotOther_i + \eta_2 ElecRatioWD_i + \xi_i. \tag{3}$$

The first stage regression measures the impact of the two IVs on a surgeon's daily workload. For the two IVs to affect the daily workload (i.e., the relevance condition), at least one of η_1 and η_2 should be statistically different from zero. In the second stage, we replace $Workload_i$ in (1) with its fitted values from (3) and estimate γ by OLS. The standard error of the coefficient γ are adjusted to incorporate the estimation error from estimating $Workload_i$ in the first stage (Woodridge 2010).

We find that the distributions of OR time, incision time, post-LOS, and total ICU time have long tails on the right end; thus we winsorize them by their 97.5th percentiles to mitigate the impact of extreme values. This corresponds to 11.9 hours for OR time, 9.2 hours for incision time, 50 days for post-LOS, and 29.7 days for total ICU time. These winsorization choices are quite conservative. As discussed in Section 4.3, our estimation results are robust to other winsorization choice as well as a logarithm transformation of dependent variables. In model (1), we cluster the standard errors by the surgeon's identifier to account for the potential correlation across cases of the same surgeon.

3.3. Validity of the IVs

In this section, we discuss in further detail the relevance and exogeneity conditions for the two IVs. To test the relevance condition, we run the first stage regression (3) to check how the two IVs affect the surgeon daily workload. The estimated coefficients are reported in Table 3. We run regression (3) for the full, elective, and non-elective samples, respectively. The controls X_i in (1) are included in the regression.

We see the two IVs are statistically significant with expected signs for both workload measures and the three samples. For the full sample, the coefficient of $TotOther_i$ and $ElecRatioWD_i$ is -0.075 and 0.993 respectively, both of which are significant at the 0.1% level. This shows the two IVs indeed explain variation in the focal surgeon's daily workload, thus satisfying the relevance condition. Specifically, increasing the $TotOther_i$ (decreasing $ElecRatioWD_i$) by one standard deviation is associated with 0.13 (resp. 0.12) fewer cases for the focal surgeon in a day. This translates to a nearly 20% standard deviation change in the

Table 3 First Stage Regression: Estimated Effects of IVs on Surgeon Daily Workload

	Full	Elective	Non-elective
$TotOther_i$	-0.075^{***}	-0.065^{***}	-0.084^{***}
	(0.006)	(0.008)	(0.008)
$ElecRatioWD_i$	0.993^{***}	0.871^{***}	1.105^{***}
	(0.087)	(0.139)	(0.115)
Num. Obs	5345	2474	2871

The estimated effects of the two IVs on surgeon daily workload. Controls X_i described in Section 3 are included in the regression. We report the estimated coefficients η_1 and η_2 in (3). Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

average surgeon daily workload, which is a substantial effect. Similar observations hold for the elective and non-elective samples. Weak instrument tests for the hypotheses $\eta_1 = 0$ and $\eta_2 = 0$ are strongly rejected at the significant level of 0.1% for all the three samples.

Next, we consider the exogeneity conditions of the two IVs. Although it is impossible to verify the exogeneity condition directly, we provide some empirical evidence for why this is likely to hold. First, we check whether the two IVs are correlated with *observed* measures of patient severity. Table EC.14 of the Electronic Companion summarizes the correlation between the two IVs and 21 observed severity factors. The correlations are generally small. For $TotOther_i$ (resp. $ElecRatioWD_i$) , the correlation is smaller than 0.1 for all the 21 (resp. 20 of the 21) patient's risk factors. The average absolute correlation is only 0.019 and 0.044 for $TotOther_i$ and $ElecRatioWD_i$, respectively. Thus, we conclude that the two IVs are unlikely to be correlated with the patient severity factors. We then conduct the Sargan's overidentification test for the exogeneity of the two IVs (see Woodridge 2010). This is possible because we have two IVs but only one endogenous variable ($Workload_i$). The p-values from the Sargan's test are reported in Table EC.15 of the Electronic Companion for all the four dependent variables (OR time, incision time, post-LOS, total ICU time) and three samples (full, electives, non-electives) in our study. The p-values all are greater than 0.1. Thus, we cannot reject the null hypothesis that the two IVs are valid.

As we have conducted an IV analysis, our results only provide insight into cases that *comply* with the IVs, which are based on operational factors. There are some operations that must happen, regardless of shared resources or surgeon's schedule. The estimated effects from our analysis may not apply to such cases. In Section 4.3, we conduct extensive additional tests to show the robustness of our main findings.

4. Estimation Results

This section provides the main estimation results regarding the effects of surgeon daily workload (number of additional cases) on surgery duration and outcomes. We first show the results estimated from the full sample. Then we analyze the heterogeneity in the effects for elective and non-elective patients.

4.1. Effect of Daily Workload on Surgery Duration and Patient Outcomes

In this section, we report the estimated effects of surgeon daily workload in (1). For the four dependent variables (OR time, incision time, post-LOS, and total ICU time), we provide the estimated γ and its standard errors for the full sample in Table 4. By model (1), it captures the effect averaged across all the cases performed by the surgeon in a day. Panel A shows the estimated effects from the TSLS with the two IVs. For comparison, Panel B reports the results when we ignore the endogeneity problem and estimate (1) by a simple OLS. As described in Section 3, we include a comprehensive set of control covariates X_i in the regression, including patient demographics, risk factors, surgeon and time fixed effects. The estimated coefficients of select control covariates are reported in Table EC.20 of the Electronic Companion.

By the "OR time" column in Panel A, we see that higher daily workload tends to increase the surgery duration of the cases performed by the focal surgeon. In particular, adding one more case increases the OR (resp. incision) time of each case performed by the surgeon by 27 (resp. 28) minutes on average. This translates to a 6.5% (resp. 9.7%) relative increase of the average OR (resp. incision) time. The effects are statistically significant at the 0.1% level. In contrast, if we ignore the endogeneity in daily workload and estimate the model by OLS, the effects become the opposite as shown in Panel B. Thus, it is essential to address the endogeneity in the daily workload as surgeons may schedule more cases if the unobserved severity factors imply shorter OR times, resulting in a negatively biased estimate of the effect. Under high daily workload, surgeons may take more time to complete their tasks. Possible reasons include surgeon fatigue due to long working hours (e.g., Zhang et al. 2023), as well as operational constraints in the related resources when more patients are serviced. Our empirical results suggest that these factors outweigh potential channels for speedup and cause longer surgery duration on average when more cases are performed.

Table 4 Estimated Effects of Surgeon Daily Workload (Number of Additional Cases) on Surgery Duration and Patient Outcomes: Full Sample

	Panel A: With IV				Panel B: Without IV			
	OR time (in hours)	Incision time (in hours)	Post-LOS (in days)	ICU time (in days)	OR time (in hours)	Incision Time (in hours)	Post-LOS (in days)	ICU time (in days)
Estimated γ	0.455^{***}	0.458^{***}	1.109^{***}	1.166^{***}	-0.149^{***}	-0.100^{**}	-0.077	-0.018
	(0.137)	(0.101)	(0.335)	(0.240)	(0.036)	(0.038)	(0.129)	(0.086)
Num. Obs	5345	5345	5344	5319	5345	5345	5344	5319

The estimated effects of surgeon daily workload (number of additional cases) on surgery duration and patient outcomes for the full sample. Controls X_i described in Section 3 are included in the regression. Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

We now consider the effect of daily workload on patient outcomes, including post-LOS and total ICU time. By the "Post-LOS" and "Total ICU" columns in Panel A, we find that higher daily workload increases the post-LOS and total ICU time when estimating the model using TSLS. Specifically, adding one more case increases the total ICU time and post-LOS by 1.17 and 1.11 days, respectively, for the cases performed by the surgeon on the same day. This is equivalent to a 9.3% increase for post-LOS and a 21.7% increase for total ICU time. Without using the IVs, the effect is insignificant for both outcomes, as shown in Panel B. The results here suggest that increased surgeon daily workload is associated with longer post-surgery recovery time for patients. As we discussed above, the negative effect can be attributed to multiple potential factors, including surgeon fatigue due to long working hours, as well as other constraints due to operating later in the day. We discuss the possible mechanisms for the estimated effects in Section 4.4.

The effect on total ICU time and post-LOS is important to consider when managing patient flow. Longer post-surgery LOS will result in increased demand for downstream units and resources and reduce the system throughput. This can lead to overcrowding in the perioperative environment and delay in operations (Zenteno et al. 2016). Additionally, the ICU is often congested and extremely expensive to operate (Kim et al. 2015). Given almost all patients (>99%) in our sample are sent to the ICU after operation, understanding the factors that impact their ICU recovery time provides a potential solution for managing ICU congestion.

We also estimate the effects of surgeon daily workload for two auxiliary dependent variables. The first one is the total pre and post-incision time of the focal case, i.e., the difference between its OR time and incision time. The second one is the patient's post-surgery recovery time out of the ICU, which is the difference between the post-LOS and ICU time. We find no significant effect of surgeon daily workload for both of them. Thus, the surgeon daily workload increases the surgery duration mainly through the incision stage, in which the surgeon is most involved. On the other hand, the increase in post-surgery recovery time is mainly driven by the patient's ICU stay, which is the more complex portion of the patient's post-survery recovery. These results are consistent with the findings from Bavafa and Jónasson (2024) in the ambulance paramedics setting: the effect of workload is more significant for the sub-processes that require complex knowledge work.

4.2. Heterogeneous Effects of Surgeon Daily Workload: Elective and Non-elective Patients

We further investigate the impacts of workload for elective and non-elective cases. The non-elective cases include urgent, emergent, and salvage cases, and account for more than half (53%) of the full sample. By Table 3, the two IVs still impact the surgeons daily workload with expected signs in the two subsamples. We add interaction terms for the workload measure $NumCases_i$ and all exogenous control variables to allow their effects to differ for elective and non-elective cases.[7] We estimate the model

$$y_i = \beta X_i + \beta_{int} X_i \times \mathbf{1}\{NonElec_i\} + \gamma_1 \times NumCases_i + \gamma_2 \times NumCases_i \times \mathbf{1}\{NonElec_i\} + \varepsilon_i, \quad (4)$$

[7] For categorical variables, we define new dummies by both the original covariate and the non-elective status, e.g., "Male×Elective", "Male× Non-elective", "Female×Elective", and "Female × Non-elective". For continuous variables, we add a new column as the product of the original variable and an indicator for non-elective status, e.g., "Age $\times \mathbf{1}\{NonElec\}$".

where $\mathbf{1}\{NonElec_i\}$ is the indicator for case i being a non-elective case. We double-cluster the standard errors by the surgeon and elective/non-elective status. We also interact each of the two IVs with the non-elective status indicator. Then, the original IVs and their interaction terms are used as new IVs in the first stage regression (see Woodridge 2010). This allows the IVs to have different impacts for the two subsamples.

Table 5 Estimated Effects of Surgeon Daily Workload (Number of Additional Cases) from Interaction Regression (4)

	OR Time (in hours)	Incision Time (in hours)	Post-LOS (in days)	ICU time (in days)
Elective: γ_1	0.567^{*}	0.531^{**}	-0.155	0.101
	(0.247)	(0.181)	(0.837)	(0.599)
Interaction term: γ_2	-0.190	-0.135	1.760	1.546^{*}
	(0.291)	(0.245)	(1.144)	(0.696)
Non-elective: $\gamma_1 + \gamma_2$	0.377^{*}	0.395^{*}	1.605^{*}	1.647^{***}
	(0.153)	(0.163)	(0.780)	(0.355)
Num. Obs	5345	5345	5344	5319

The estimated effects of surgeon daily workload (number of additional cases) on surgery duration and patient outcomes for elective and non-elective samples. Controls X_i described in Section 3 and their interaction terms are included in the regression. Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

The estimation results for (4) are reported in Table 5 for the workload-related variables. The coefficient γ_1 measures the workload effect on elective cases; γ_2 measures the heterogeneity between elective and non-elective cases; and $\gamma_1 + \gamma_2$ is the effect for non-elective cases. By the “OR time” and “Incision Time” columns, we see the surgeon daily workload significantly increases the OR and incision time for both elective and non-elective cases. Both effects are significant at the 5% level. In particular, performing one more case in a day increases the OR time (resp. incision time) of each elective and non-elective case by 34 and 22 (resp. 30 and 24) minutes, respectively. The results suggest that the surgeon daily workload has a larger impact on surgery duration for elective cases, though the difference γ_2 is not statistically significant. Such heterogeneity may be explained by the fact that urgent and emergent cases are generally more time sensitive than elective ones, so their surgery duration are less impacted by surgeon daily workload.

By the last two columns of Table 5, we find that for post-LOS and ICU time, the effects of surgeon daily workload is statistically significant for the non-elective cases, but insignificant for the elective ones. This reveals substantial heterogeneity in the effects of surgeon daily workload on patient’s outcomes. Moreover, the magnitudes of the effects are larger for the non-elective cases than those for the full sample: Adding one more case leads to 1.61 more days in the post-LOS of non-elective patients, which is 45% larger than that for the full sample (1.11 days). Similar heterogeneity is also observed for the ICU time (1.65 days for non-electives vs. 1.17 days for the full sample). The difference γ between the elective and non-elective samples is significant at the 5% level for total ICU time but insignificant for the post-LOS. This may be explained by the large standard errors of the post-LOS estimates due to sample size.

One potential explanation for the heterogeneous effects for patient outcomes is that non-elective cases are generally more urgent and complicated than elective ones. Thus, the outcomes of non-elective cases may be more sensitive to surgeon fatigue or operational constraints due to high daily workload. This can be seen in Table 2: non-elective patients have longer post-LOS and total ICU time on average. The standard deviations of total ICU time and post-LOS are also much larger for non-elective patients, implying there is more variation in their surgical outcomes.

Our results demonstrate the consistent negative impact of surgeon daily workload on surgical outcome. Such effects are more significant for the more severe, non-elective patients. This provides new empirical evidence for the link between workload level and patient outcomes (see, e.g., Kc and Terwiesch 2009 and Kuntz et al. 2015) – specifically at the cumulative and individual level. From a managerial perspective, it suggests that when hospitals design their surgery schedules, they should take into account the effects of surgeon daily workload to improve patient flow and outcomes. We explore this direction in Section 5.

4.3. Robustness Checks on Surgeon Workload Effect

We perform extensive robustness analyses to examine the validity of our main findings. The results are documented in Section EC.5 of the Electronic Companion. We briefly discuss them below.

In our main specification, we use the number of other cases ($NumCases_i$) as the workload measure for a surgeon. As robustness checks, we consider three alternative surgeon workload measures. The first is the total incision time ($TotInc_i$) of other cases performed by a surgeon in a day. The second one is an indicator for being a non-first case ($NonFirst_i$) of the focal surgeon in a day. The third one is the prior incision time ($PriorInc_i$) of focal surgeon before the start of focal case. The four workload measures are highly correlated because most surgeons perform one or two cases in a day. In particular, the correlation between the number of cases ($NumCases_i$) and total incision time ($TotInc_i$) is 0.93. Using TSLS with the two IVs, we find that the estimated effects are significant and robust for the four workload measures. The magnitudes of effects are also consistent. The model set-up and estimation results are discussed in Section EC.3. Because our IVs vary at the daily level, instead of the intra-day level, we cannot fully distinguish whether our estimated effects are truly driven by surgeon fatigue versus being the later cases in the day. We discuss the potential mechanism in more details in Section 4.4.

We additionally control for the OR entry time of each case by including it as an exogenous variable in our model (1). We consider two choices: including a dummy variable for each OR entry hour (in total 24 levels) or using a categorical variable for morning (6AM to 12PM), afternoon (1PM to 6PM), or night (7PM to 5AM). The regression results are shown in Panel A of Table EC.16. We still observe statistically significant effects of surgeon daily workload, and the heterogeneity between elective and non-elective cases also holds. The effects on surgery time become mildly smaller (from 27 to 21 minutes for the incision time) after controlling for the OR entry time.

In our main model, we winsorize the four dependent variables at their 97.5th percentiles, respectively. As a robustness check, we have run our regression using the logarithm transformations of the four dependent variables. The results are given in Panel B of Table EC.16. We still find statistically significant effects of surgeon daily workload with expected directions. Compared with our main specification, the relative effect drops from 9.2% to 8.0% for post-LOS and 21% to 12% for total ICU time under the logarithm transformation.[8] We also try the winsorization level of 95th percentile of the variables. The effects are still statistically significant but with smaller magnitudes. The differences in magnitudes can be explained as the dependent variables have long tails on the right-hand side, which is common in medical data. We note that the effects are statistically significant in all these settings, but their estimated magnitudes can depend on the winsorization choice and/or logarithm transformation.

Next, we run regression (1) after dropping the emergent and salvage cases, which account for 7.2% of the sample. We also estimate the model without the cases performed after 3PM in a day (12.5% of sample). It is possible that these patients are more likely to have special medical conditions or requests, making the IVs less applicable. The estimation results are reported in Panel C of Table EC.16. They are largely similar to those under our main specification. In another test, we include an additional control variable which measures the focal surgeon's number of cases in the previous three days before the focal day. Panel D of Table EC.16 shows that this has no impact on our estimation results.

We perform multiple checks regarding the two IVs used in our identification. In our main set-up, we use two IVs to improve the statistical efficiency of identification (e.g., Mogstad et al. 2021). As a robustness check, we run regression (1) by TSLS using only $TotOther_i$ or $ElecRatioWD_i$ as the IV, which allows us to estimate their effectiveness separately. As shown in Panel E of Table EC.16, the estimated workload effects with one IV are qualitatively similar to those in our main specifications.[9] In addition, we experiment with alternative specifications to construct the second IV, $ElecRatioWD_i$. First, we calculate $ElecRatioWD_i$ using only the period prior to the current day, i.e., using interval $[t-L, t)$ in (2). Second, we change the window length L in (2) from 180 days to 90 days. The results are reported in Panel F of Table EC.16. The direction and magnitude of the estimated effects are again similar to our main findings.

In our sample, paralleling surgery is common when a surgeon performs multiple cases in a day. Thus, there may be overlaps in the incision time of these cases. In Panel A of Table EC.17, we show that our

[8] Given the long duration and significant burden of cardiac surgery, the effects of increasing surgeon workload have been observed in many studies with magnitudes comparable to ours (see, e.g., Kc and Terwiesch 2009 and Zhang et al. 2023).

[9] The effects become insignificant for post-LOS when we use $ElecWDRatio_i$ alone. This can be explained as using one IV reduces the exogenous variation that can be leveraged for identification.

main results remain robust after controlling for the overlapping incision time in the model. In our main specification, we use surgeon fixed effects in (1) to control for surgeon heterogeneity. As robustness checks, we further include surgeons' past volumes in the model and separately use more granular surgeon-procedure type fixed effects. The estimation results are reported in Panel B of Table EC.17. They remain similar to our main setting.

Lastly, we perform an additional robustness check for post-LOS. Unlike total ICU time which is measured in hours, post-LOS can only be computed on a daily basis as we do not have the exact discharge time in the STS data. To account for patients' actual recovery time, we use several conservative measures of post-LOS by adjusting the original measure using the OR entry or exit time of each case (e.g., 12PM versus 6PM). We find that the estimated effect are very similar when using these alternative measures. This is discussed in Section EC.5 with results given in Table EC.19.

4.4. Mechanism of Estimated Workload Effects

Our empirical results provide clear evidence for the negative impact of high surgeon daily workload. With the IV method and a large set of control variables, our identification is expected to address the confounding effects from the observable or unobservable factors that are *not* related to surgeon daily workload. However, we cannot disentangle the effects from potential factors that arise from performing multiple cases in a day. Such factors include surgeon fatigue due to long working hours, operational constraints for non-first cases in a day (e.g., shared staff or medical equipment), delay in bed flow when more patients are serviced, and parallel of surgery. Therefore, we cannot fully identify the exact channel for the observed workload effects among these factors. This would likely require more granular data such as direct measures of surgeon fatigue, shift data of staff and nurses, and details of the current scheduling system. In addition, our IV-based identification method hinges on the variation in the two IVs. As both IVs vary on the surgeon-day level, their values are the same for all cases performed by a surgeon in a day. This further hinders our identification of potential confounding factors associated with case sequencing.

In the follows, we provide some suggestive evidence that surgeon fatigue is one possible channel for the observed workload effect, among others. First, cardiac operations usually take long time to complete, meaning that performing multiple cases in a day can lead to surgeon fatigue both mentally and physically. In the medical literature, surgeon fatigue is commonly considered as a driving factor for the negative effects associated with high surgeon workload or late start of operations (e.g., Thomas et al. 2012, Zhang et al. 2023, Reijmerink et al. 2024). In addition, we find that the surgeon daily workload significantly increases the time of the incision stage of a surgery, but not the pre and post-incision stages in which the surgeon is less involved. These results are consistent with the surgeon fatigue channel.

In our study, we have controlled for a comprehensive set of patient demographic, risk, and operative factors, as well as the time and surgeon fixed effects (see the descriptions in Section EC.1). In addition, we identify the causal effect of surgeon daily workload based on the exogenous variation from the two IVs, which are based on other surgeons' workload and the focal surgeon's long-term working pattern. As discussed in Section 3.1, the two IVs are unlikely to correlate with the patient's severity factors. In Section EC.2 of the Electronic Companion, we find that the cases with different surgeon daily workload levels and/or sequencing positions in a day are not systematically different in terms of their procedure types, surgery duration, and predicted risks. These analyses mitigate the concern that the estimated effects are mainly driven by unobserved risk factors. Moreover, in Section EC.3, we show that being a non-first case in a day and/or having longer prior incision time are also associated with longer surgery duration and recovery time. This reveals the negative effects of surgeon cumulative workload before the focal case.

As mentioned in Section 4.3, we perform several additional tests – some of which shed light on the potential channel. First, we find that the effects of surgeon daily workload remain robust even after excluding the emergent/salvage cases in our sample, which are more likely to have unobserved risk factors. The same conclusion holds when we drop the late cases performed after 3PM in a day. Additionally, we add the OR entry hour of focal patient as a covariate in our model. This further controls for potential intraday periodic factors that may affect the surgery outcomes, such as patient circadian rhythm, staff shifts, and resource availability patterns. We still find similar and statistically significant effects of surgeon daily workload. Finally, in Section EC.4 of the Electronic Companion, we provide more empirical tests on the non-first case effects. In general, we do not find evidence that the estimated workload effects are primarily driven by the status of being a later (e.g., third/fourth) cases in a day.

In summary, we find consistent empirical evidence that high surgeon daily workload is associated with adverse surgical outcomes. Such effect can be driven by surgeon fatigue, operational constraints, or a combination of the two. As a limitation of our study, we cannot identify the exact channels for the observed workload effects among these factors. This would likely require more direct measures on surgeon fatigue levels (e.g., via questionnaires or biometric signals) and/or new IVs that can address the endogeneity in intraday case sequencing. These are deferred to future research.

5. A Surgery Scheduling Model with Impact from Daily Workload

In this section, we propose a surgical scheduling model that accounts for the effect of surgeon daily workload. While there is a rich literature on surgery scheduling, most of it assumes exogenous distributions for the surgery duration and patient outcomes (e.g., post-LOS). As shown in our empirical analyses, increased surgeon daily workload can lead to longer surgery duration and post-surgery recovery time. However, to the best of our knowledge, such impact is largely ignored in current scheduling literature.

We briefly introduce the surgical scheduling model in below and include its details in Electronic Companion Section EC.6. We solve the scheduling model for each week (Monday to Friday) in our four year sample. Weekends are excluded in the model, as surgeons rarely work on weekends except for emergent and salvage cases. Our model considers the reassignment of the operations in our sample to different days within the week they were actually performed in the data. By changing the surgery dates, we aim to capture the potential benefit of decreasing high workload days, e.g., by smoothing workload across days. On the other hand, we keep the patient cohort and the surgeon assigned to each patient unchanged. Thus, the changes from our model are relatively small.

The objective in our model is to minimize the total expected OR time, post-LOS, or ICU time of all cases in our sample, as estimated using our econometric model (1). To focus on the impact of daily workload, we assume the term βX_i, which primarily depends on the patient's risk and operative factors, remains unchanged in the new schedule. In Section EC.6, we show that the objective function can be written as

$$\min \sum_s \sum_t \tilde{n}_{s,t} \cdot \left(\gamma^{(el)} \tilde{n}_{s,t}^{(el)} + \gamma^{(ne)} \tilde{n}_{s,t}^{(ne)}\right). \tag{5}$$

Here we index the day in the week by t and the surgeon by s; $\tilde{n}_{s,t}^{(el)}$ and $\tilde{n}_{s,t}^{(ne)}$ denote the number of elective and non-elective cases performed by surgeon s on day t in the new schedule respectively.

In (5), the coefficients $\gamma^{(el)}$ and $\gamma^{(ne)}$ are the estimated effect of surgeon daily workload effect on OR time, post-LOS, or total ICU time – depending on which one we are to minimize – for the elective and non-elective cases respectively. They are reported in Table 5. We set the coefficient to be zero if it is not statistically significant at the 10% level. When we ignore the heterogeneity in the impacts of daily workload, we use the average treatment effects in Table 4 with $\gamma^{(el)} = \gamma^{(ne)} = \gamma^{(avg)}$. Choosing which effect to account for in scheduling depends on the concrete challenges and management goals of hospitals, which can vary in different situations. For example, for a hospital that believes its elective cases and non-elective cases are more likely to exhibit heterogeneous effects, it may be more proper to account for the different impact of surgeon daily workload on elective and non-elective cases when scheduling.

We consider multiple feasibility constraints for the new schedule. We allow elective cases to be assigned to any weekday of the week. On the other hand, we impose that urgent cases can only be assigned to the original date or the adjacent days[10], while the emergent and salvage cases can only be scheduled on the original dates. These constraints reflect the varying levels of time sensitivity for different types of patients. In addition, we assume the surgeon can perform at most $\bar{n}^{(c)} = 3$ cases in a day in the new schedule, which is reasonable according to the data. Finally, given surgeons have other responsibilities in their work days, we assume the maximum number of working days for a surgeon in a week is $\bar{n}^{(d)} = 3$, unless the surgeon works for more days in the original schedule. This is a realistic assumption for the following reasons. In the original schedule from our data, the number of working days in a week is equal or greater than three for 80% of surgeon-week combinations (773 out of 972) if a surgeon performs at least three cases in a

[10] By the STS dataset, urgent cases should be performed within the same hospital stay, which usually takes several days. According to medical literature (e.g., Karim et al. 2016), it is generally suggested that urgent cases should be performed in a 72-hour window.

week.The ratio increases to 92% (resp. 98%) for surgeons who perform at least four (resp. five) cases in a week. We show in Electronic Companion Section EC.6 that the optimization problem can be formulated as a MIQP with binary decision variables, quadratic objective, and linear constraints.

We note that the goal of our surgical scheduling model is to illustrate the potential benefits of incorporating the surgeon's workload effects in scheduling. Thus, we take a retrospective approach and make several simplifications in the model. For example, in our main set-up, we assume that the hospital knows the surgery date of each case in the current week. Another assumption is that we evaluate the performance of schedules based on the average effects in (1) for all cases in a day. Thus, the estimates from our model represent an optimistic "upper bound" on the achievable benefits under these assumptions; the actual gains in practice may be smaller. In Section EC.7, we conduct multiple tests to show the robustness of our findings. We acknowledge that in order to implement the surgery schedules in practice, the hospital needs to consider other important factors, such as stochastic patient arrivals, matching between patients and surgeons, and intraday case sequencing. These are beyond the scope of this study. Therefore, our scheduling model below should be viewed as a stylized illustration rather than prescriptive.

Our empirical findings may motivate hospitals to make incremental changes in their scheduling system. For example, hospitals may try to facilitate the matching between patients with currently "underloaded" surgeons. Hospitals may also increase the available weekly working days of surgeons when they have many cases to perform. It is also helpful to increase flexibility in the scheduling procedures. For instance, elective cases can still be scheduled to specific week well in advance, while deferring the determination of the exact surgery day closer to the time of surgery in a rolling manner. Additionally, hospitals may encourage more block sharing across surgeons, which is already observed in the data. There may be practical obstacles associated with these changes, but our empirical evidence provides the potential direction of improvements such changes may facilitate.

5.1. Results

In this section, we summarize the results from our scheduling model, which demonstrate the benefit of incorporating the effects of surgeon daily workload in surgical scheduling. We show our main results in Table 6. The first two columns show the variable we are optimizing ("Objective") and the estimated effects ("Effect") we use in our objective function (5), with which we solve the surgical scheduling model. The third and fourth columns ("Obj orig" and "Obj new") report the objective values (5) under the original and new schedules, respectively. The fifth and sixth columns ("ΔObj" and "Rel. ΔObj") report the absolute and relative reduction in the objective function,[11] which demonstrate the benefit of applying our surgical scheduling model. The last column ("Number of improved week") reports the number of weeks (out of the 209 weeks in our sample) that we can achieve reduction in the objective function.

[11] The objectives and reductions in Table 6 are computed based on the sum of workload term of the dependent variable, i.e., $\gamma Workload_i$, in (1). They do not account for the term βX_i, which is assumed to be unaffected by the schedule.

Table 6 Estimated Effects of the Surgical Scheduling Model

Objective	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Number of improved weeks
OR time (in hours)	Avg	4124.58	3764.22	360.36	8.7%	184
	Het	4240.78	3804.25	436.52	10.3%	205
Post-LOS (in days)	Avg	10053.09	9174.76	878.33	8.7%	184
	Het	7594.86	6273.95	1320.92	17.4%	203
ICU Time (in days)	Avg	10569.79	9646.32	923.47	8.7%	184
	Het	7793.60	6438.12	1355.48	17.4%	203

In Table 6, we see that our new schedule leads to substantial improvement for all three outcomes using both average and heterogeneous effects of daily workload. For total OR time, the new schedule with average (resp. heterogeneous) effect leads to a 360 (resp. 437) hour decrease in the four-year horizon, which is a 8.7% (resp. 10.3%) relative reduction. The OR is an extremely expensive resource with cost estimated to be up to \$37 per minute (Childers and Maggard-Gibbons 2018). Thus, the reduction in the OR time could save the hospital up to \$242,535 each year. In addition, given the average OR time is 7.1 hours in our sample, the hospital may be able to add 16 new cases each year due to the reduction in OR time from the new schedule. With some cardiac operations netting margins of over \$21,000 per case on average (Robinson 2011), this has the potential to translate to an additional \$336,000 in profit for this service. We find that the new schedule leads to improvement for most (205) of the 209 weeks in our sample. This shows that the benefit of our scheduling model is not limited to a small number of weeks, and the original schedule can be substantially improved.

Our scheduling model also substantially reduces the total expected post-LOS and ICU time when optimized to do so. Using the average effect, the new schedule decreases the total post-LOS by 878 days and the total ICU time by 923 days, translating to a 8.7% relative drop. The benefit is larger with the heterogeneous effect, with the reduction for post-LOS (resp. ICU time) increasing to 1321 (resp. 1355) days. In addition, we find similar reduction in the average occupancy level in the downstream unit from the new schedule. These results highlight the potential benefits of our surgical scheduling model in reducing downstream congestion, which can be a bottleneck in perioperative environment (e.g., Zenteno et al. 2016).

We note that the resulting schedules can lead to improvement for multiple objectives simultaneously, instead of improving one at the cost of another. For example, when considering the average effects, the resulting schedules are identical (see Section EC.6), so all the three objectives are improved by the amounts reported in Table 6. On the other hand, when considering the heterogeneous effects, the resulting schedule is the same when optimizing for total post-LOS and ICU time. Thus, both post-surgery metrics are reduced simultaneously. This further highlights the benefits from the new schedules.

5.2. Managerial Insights and Robustness Checks

In this section, we take a closer look at the new schedule to investigate the mechanisms that lead to the improvement. We consider two schedules from our model. The first is the one using the average effect with $\gamma^{(el)} = \gamma^{(ne)} = \gamma^{(avg)}$ in (5). The second is the one using the heterogeneous effect for post-LOS and ICU time, where the surgeon daily workload impacts the non-elective cases but not the elective ones. In Table 7, we provide some summary statistics of the two schedules (first two rows) as well as the original one (last row). The columns $\tilde{n}_{s,t} = i$ report the number of surgeon-day pairs for which the surgeon performs i cases in a day. The next column "$\bar{n}_{\text{day}}$" shows the average number of weekly working days by a surgeon, given the surgeon appears at least once in the schedule. The columns $\tilde{n}^{(ne)}_{s,t} = i$ report the number of surgeon-day pairs that a surgeon performs i non-elective cases a day. The last column shows the average number of elective cases performed by a surgeon in a day, given the surgeon performs at least one non-elective case in that day. It thus measures the co-occurrence of elective and non-elective cases in the schedule.

Table 7 Summary Statistics of Schedules

Effect	$\tilde{n}_{s,t}=1$	$\tilde{n}_{s,t}=2$	$\tilde{n}_{s,t}=3$	$\tilde{n}_{s,t}=4$	$\bar{n}_{day}$	$\tilde{n}^{(ne)}_{s,t}=1$	$\tilde{n}^{(ne)}_{s,t}=2$	$\tilde{n}^{(ne)}_{s,t}=3$	$\mathsf{E}(\tilde{n}^{(el)}_{s,t} \mid \tilde{n}^{(ne)}_{s,t}>0)$
Avg	2628	1261	65	1	3.02	2003	412	15	0.22
Het (LOS)	2594	868	337	2	2.91	2133	353	11	0.10
Orig	2268	1249	177	13	2.83	1872	424	48	0.27

Under the average effect, the new schedule should smooth surgeon workload across days, i.e., reducing the number of days with multiple cases. This is indeed seen by the first row of Table 7: the number of surgeon-day pairs with three (resp. four) cases in a day decreases from 177 (resp. 13) in the original schedule to 65 (resp. 1) in the new schedule. The reduction in these high workload days is mostly made up by the single-case days ($\tilde{n}_{s,t} = 1$), which increases from 2,268 to 2,628 in the new schedule. This result shows that the new schedule effectively smooths surgeon workload across days. Note that the number of two-case days does not change much under the new schedule. This is because we impose a relatively strict limit on the number of weekly working days by a surgeon. Thus, the model will primarily optimize for the surgeon-day with three cases or more. In Section EC.7, we perform a robustness check under smaller workload effects for three- and four-case days., which provides a conservative estimate for the benefits.

Under the heterogeneous effect for post-LOS or ICU time, the minimization of the objective suggests that the hospital should reduce the workload on the days with non-elective cases, as we have $\gamma^{(el)} = 0$ and $\gamma^{(ne)} > 0$ in (5) by Table 5. This is achieved in the new schedule by the second row of Table 7. First, in the new schedule, the number of surgeon-day pairs with multiple non-elective cases decreases significantly from 472 (424+48) to 364 (353+11). Second, the new schedule decreases the co-occurrence of elective and non-elective cases: given the surgeon performs at least one non-elective case, the average number of elective cases in that day drops from 0.27 in the original schedule to 0.10 in the new schedule.

The above analysis leads to useful managerial insights from a new aspect in surgical scheduling — the cumulative daily workload of individual surgeons. First, when feasible, hospitals should smooth surgeon's workload to reduce the frequency of very busy days, e.g., with three or four cases. Second, hospitals should try to control the workload of surgeons on the days when they have to perform more complicated non-elective cases. In Section EC.6.3 of the Electronic Companion, we use a concrete example to illustrate how the scheduling model works (see Figure EC.4). We believe such insights can be useful for setting surgical schedules to improve patient flow, OR time, and patient's post-surgery outcomes.

We check the feasibility of the new schedule as follows. First, by the "$\bar{n}_{day}$" column in Table 7, the average weekly working days of a surgeon remain very similar under the original schedule (2.83 days) and new schedules (3.02 days for average effect and 2.91 days for heterogeneous effect). Thus, our new schedules do not require surgeons to work for significantly more days in a week, though this naturally smooths the workload. Next, we show that our new schedules do not lead to significant increase in the peak OR usage of the cardiac department. This is important as the OR capacity of the department is limited. Figure EC.3 of the Electronic Companion plots the frequency distributions of department daily workload (number of cases by all surgeons) from the original and new schedules. We see that the distributions of department workload are similar under the new schedules, and the number of days with very high department workload (e.g., more than ten cases) remains small. These results support the feasibility of the new schedules.

In Section EC.7 of the Electronic Companion, we perform several robustness checks for our surgical scheduling model. First, we fix the surgery date of all non-elective cases and do not allow them to be moved. Thus, only elective cases can be rescheduled within a week. Second, we fix the number of working days of surgeons as in the original schedule. The results for the two scenarios are reported in Panels A and B of Table EC.21. We also solve the scheduling model after fixing both the number of working days and the time of non-elective cases (Panel C of Table EC.21). This baseline result reflects the "pure" benefit from workload smoothing. In above scenarios, we still find economically meaningful benefits from our scheduling model, although the magnitudes become smaller. On the other hand, we experiment with a more relaxed constraint by allowing a surgeon to work for a maximum of four days in a week. By Panel D of Table EC.21, this leads to significantly larger benefits from the scheduling model, though the average number of weekly working days increases mildly from 3.02 to 3.28. Finally, we show in Section EC.7.1 that under heterogeneous effect, the scheduling model can achieve Pareto improvement for both surgery duration and post-LOS.

We further develop a heuristic scheduling policy to show the robustness of our findings. The heuristic policy determines the schedules for each surgeon in a given week, and imposes the same set of feasibility constraints as our main model. In the heuristic policy, we first smooth the number of non-elective cases across days. Then, we schedule the elective cases to smooth the total number of cases (both elective and urgent). Thus, the heuristic policy is relatively easy to implement and does not involve complex optimization model. The effects from the heuristic scheduling policy are reported in Panel A of Table EC.23. We see that this simple policy still achieves substantial improvements. The benefits are similar to those in our main model (Table 6) under average effects, and become mildly smaller under the heterogeneous effects.

Furthermore, we account for the uncertainty in the surgery date of emergent/salvage cases in the heuristic policy. In particular, we first use the heuristic policy to schedule the elective and urgent cases based on their information. Then, the emergent/salvage cases occur as in the original data and are assigned to surgeons. As another test, we additionally assume the surgery date is uncertain for some of the urgent cases when making the schedule. In both scenarios, the benefits from the heuristic policy become smaller in magnitude, but are still economically meaningful (see Table EC.23). The details of the heuristic policy and the results are described in Section EC.7.2 of the Electronic Companion.

6. Conclusion

In many human-run service systems, service time and quality can be endogenously affected by the level of workload. In this work, we focus on such relationship in the context of cardiac surgery. Specifically, we study how surgery duration and patient outcomes are impacted by individual surgeon daily workload. Using a detailed data set, we find that higher surgeon daily workload leads to longer surgery duration and post-surgery recovery time. We develop two novel IVs using the operational factors in the cardiac department. The IVs effectively addresses the endogeneity problem due to unobserved risk factors. This provides new evidence for the negative impact of surgeon fatigue or operational constraints due to high daily workload.

Based on our findings, we develop an illustrative surgery scheduling model that incorporates the effect of surgeon workload. We find that by simply rescheduling operations within a week, with practical restrictions on how much non-elective operations can be moved, substantial improvements could be achieved for both surgery duration and patient outcomes. Our model demonstrates the potential improvements in patient flow in the OR (via OR time) and post-surgery (via post-LOS and total ICU time) by accounting for the impact of surgeon workload when scheduling surgery. As such, our results suggest that it is important for hospital managers and surgeons to consider the impact of surgeon workload when managing their ORs.

There are several important limitations of our study. First, due to data limitations, we cannot identify the exact mechanism driving the effects of surgeon daily workload (e.g., surgeon fatigue or operational resource constraints). Thus, it would be interesting to explore more granular data on the operational factors of the surgery department (e.g., nurse shifts) and/or more direct measures on surgeon fatigue (e.g., sleep duration, questionnaire for fatigue level, biometric signals). Second, as we have conducted an IV analysis, our results only provide insight into cases that *comply* with the IVs. Other hospitals may have different scheduling procedures which may make the IVs more or less appropriate. It would also be valuable to develop new IVs that can adequately address the potential endogeneity in intraday sequencing. Third, our empirical results are based on the particular clinical setting of cardiac operations in a large hospital, in which the surgery duration is very long and patients are severe (e.g., more than half are non-elective cases). The results may not apply to other settings with less burdensome operations and less severe patients. Finally, more realistic and ready-to-use, prescriptive surgical scheduling models (instead of the illustrative example in this work) can be developed for the practical use of hospitals and practitioners. These are deferred to future research.

Acknowledgment

The authors would like to thank the CUIMC IT team for their help with managing the data infrastructure system. We would also like to acknowledge the Editors and anonymous reviewers for their constructive feedback during the review process.

References

Bashir M, Harky A, Fok M, Shaw M, Hickey GL, Grant SW, Uppal R, Oo A (2017) Acute type a aortic dissection in the united kingdom: surgeon volume-outcome relation. *Journal of Thoracic and Cardiovascular Surgery* 154(2):398–406.

Batt RJ, Terwiesch C (2016) Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* 63(11):3531–3551.

Bavafa H, Jónasson JO (2024) The distributional impact of fatigue on performance. *Management Science* 70(5):3319–3337.

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4):1042–1062.

Childers CP, Maggard-Gibbons M (2018) Understanding costs of care in the operating room. *JAMA Surgery* 153(4):e176233–e176233.

Delasay M, Ingolfsson A, Kolfal B (2016) Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research* 64(4):867–885.

Delasay M, Ingolfsson A, Kolfal B, Schultz K (2019) Load effect on service times. *European Journal of Operational Research* 279(3):673–686.

Freeman M, Robinson S, Scholtes S (2021) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* 67(7):4209–4232.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* 63(10):3147–3167.

Freeman NK, Melouk SH, Mittenthal J (2016) A scenario-based approach for operating theater scheduling under uncertainty. *Manufacturing & Service Operations Management* 18(2):245–261.

Govindarajan A, Urbach DR, Kumar M, Li Q, Murray BJ, Juurlink D, Kennedy E, Gagliardi A, Sutradhar R, Baxter NN (2015) Outcomes of daytime procedures performed by attending surgeons after night work. *New England Journal of Medicine* 373(9):845–853.

Green LV, Nguyen V (2001) Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* 36(2):421.

Hockey R (2013) *The psychology of fatigue: Work, effort and control* (Cambridge University Press).

Janhofer DE, Lakhiani C, Song DH (2019) Addressing surgeon fatigue: current understanding and strategies for mitigation. *Plastic and Reconstructive Surgery* 144(4):693e–699e.

Karim MN, Reid CM, Cochrane A, Tran L, Billah B (2016) When is ‘urgent’really urgent and does it matter? misclassification of procedural status and implications for risk assessment in cardiac surgery. *Heart, Lung and Circulation* 25(2):196–203.

Kc DS (2014) Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* 16(2):168–183.

Kc DS, Staats BR (2012) Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* 14(4):618–633.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

Kc DS, Terwiesch C (2011) The effects of focus on performance: Evidence from california hospitals. *Management Science* 57(11):1897–1912.

Kc DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Keskinocak P, Savva N (2020) A review of the healthcare-management (modeling) literature published in manufacturing & service operations management. *Manufacturing & Service Operations Management* 22(1):59–72.

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.

Kuntz L, Mennicken R, Scholtes S (2015) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* 61(4):754–771.

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.

McDermott KW, Freeman WJ, Elixhauser A (2017) Overview of operating room procedures during inpatient stays in us hospitals, 2014: statistical brief# 233. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville: Agency for Healthcare Research and Quality* .

Mogstad M, Torgovitsky A, Walters CR (2021) The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review* 111(11):3663–3698.

Needleman J, Buerhaus P, Pankratz VS, Leibson CL, Stevens SR, Harris M (2011) Nurse staffing and inpatient hospital mortality. *New England Journal of Medicine* 364(11):1037–1045.

Neuraz A, Guérin C, Payet C, Polazzi S, Aubrun F, Dailler F, Lehot JJ, Piriou V, Neidecker J, Rimmelé T, et al. (2015) Patient mortality is associated with staff resources and workload in the ICU: a multicenter observational study. *Critical Care Medicine* 43(8):1587–1594.

Oliva R, Sterman JD (2001) Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* 47(7):894–914.

Pourghannad B, Wang G (2025) Matching patients with surgeons: Heterogeneous effects of surgical volume on surgery duration. *Manufacturing & Service Operations Management* .

Rath S, Rajaram K (2022) Staff planning for hospitals with implicit cost estimation and stochastic optimization. *Production and Operations Management* 31(3):1271–1289.

Reijmerink IM, van der Laan MJ, Wietasch JG, Hooft L, Cnossen F (2024) Impact of fatigue in surgeons on performance and patient outcome: systematic review. *British Journal of Surgery* 111(1):397.

Robinson JC (2011) Variation in hospital costs, payments, and profitabilty for cardiac valve replacement surgery. *Health Services Research* 46(6pt1):1928–1945.

Rothstein DH, Raval MV (2018) Operating room efficiency. *Seminars in Pediatric Surgery*, volume 27, 79–85.

Soltani M, Batt RJ, Bavafa H, Patterson BW (2022) Does what happens in the ed stay in the ed? the effects of emergency department physician workload on post-ed care use. *Manufacturing & Service Operations Management* 24(6):3079–3098.

Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management Science* 58(6):1141–1159.

Tan TF, Netessine S (2014) When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* 60(6):1574–1593.

Thomas M, Allen MS, Wigle DA, Shen KR, Cassivi SD, Nichols III FC, Deschamps C (2012) Does surgeon workload per day affect outcomes after pulmonary lobectomies? *Annals of Thoracic Surgery* 94(3):966–972.

Woodridge JM (2010) *Econometric Analysis of Cross Section and Panel Data, 2nd Edtion* (The MIT Press).

Zenteno AC, Carnes T, Levi R, Daily BJ, Dunn PF (2016) Systematic or block allocation at a large academic medical center. *Annals of Surgery* 264(6):973–981.

Zhang D, Gu D, Rao C, Zhang H, Su X, Chen S, Ma H, Zhao Y, Feng W, Sun H, et al. (2023) Outcome differences between surgeons performing first and subsequent coronary artery bypass grafting procedures in a day: a retrospective comparative cohort study. *BMJ Quality & Safety* 32(4):192–201.

Electronic-Companion for “The Impact of Surgeon Daily Workload and its Implications for Operating Room Scheduling”

Yiwen Shen
Information Systems, Business Statistics and Operations Management,
Hong Kong University of Science and Technology, yiwenshen@ust.hk

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School, cwchan@columbia.edu

Fanyin Zheng
Decision, Risk, and Operations, Columbia Business School, fanyin.zheng@columbia.edu

Michael Argenziano
Department of Surgery, New York Presbyterian/Columbia University Irving Medical Center, ma66@cumc.columbia.edu

Paul Kurlansky
Department of Surgery, New York Presbyterian/Columbia University Irving Medical Center, pk2245@cumc.columbia.edu

EC.1. Description and Summary Statistics of Independent Variables in (1)

To control for the effects of patient characteristics and severity levels, we include a comprehensive list of demographic, risk, operative, and operational factors as independent variables in our estimation models. Some of these factors are already discussed in Section 3. We now provide the description and summary statistics for other independent variables included in X_i for our models (1).

In Table EC.1, we document the descriptions, types, and summary statistics of the independent variables. We also provide their locations in the STS data collection form. We handle missing values in the binary and categorical variables as follows: if the number of missing observations is smaller than 100 (1.8% of the sample), we impute their values using the majority from the cases in the same New York Heart Association (NYHA) class. Otherwise, we add a new category “Unknown” to represent the missing values. Summary statistics of the categorical variables are reported in Table EC.2. Note that the NYHA classification is not available (N/A) if the patient has not experienced heart failure. The Pulmonary Artery (PA) pressure is coded as “High” if it is higher than 55mg and “Low” otherwise. We also include the patient’s admission type, which refers to the channels for the patient to be admitted to the hospital: elective (3,777), emergency department (448), transfer-in (1,117), and other (10).

We classify the cases to different surgery types to control for the procedures performed by the surgeons. First, we have eight standard surgery types from the STS data: coronary artery bypass graft (CABG), aortic valve replacement (AVR), mitral valve replacement (MVR), mitral valve repair (MVr), and their combinations CABG+AVR, CABG+MVR, CABG+MVr, and AVR+MVR. For the cases that do not fall into the standard types, we classify their surgery types by the following heuristic rule. We collect from the STS data which of the following four procedures are performed in the operation: coronary artery bypass, valve, other cardiac procedure, and other non-cardiac procedure. If only one of the four procedures is performed, we classify the case as a non-standard isolated type, e.g., "non-standard isolated valve" if only the valve procedure is conducted. If more than one of the procedures are performed, we classify the case as the "non-standard multiple" type. Finally, if none of the four procedures is performed, we classify it as "others". In total, we have six types for the non-standard procedures, i.e., four non-standard isolated ones, non-standard multiple, and others. The numbers of cases of each type (both standard and non-standard ones) are summarized in Table EC.3.

Table EC.1 Description and Summary Statistics of Other Independent Variables in Model (1)

Variable	Description	Section in STS	Type	Mean
Race	Patient's race	Demographics	Categorical	-
Endocard	Endocarditis	Risk factor	Binary	0.053
PeriAD	Peripheral arterial disease	Risk factor	Binary	0.088
Lung	Lung disease with severity $\geq$ mild	Risk factor	Binary	0.192
Hypertension	Hypertension	Risk factor	Binary	0.777
CaroStenosis	Carotid Stenosis	Risk factor	Binary	0.054
Syncope	Syncope	Risk factor	Binary	0.031
Dialysis	Dialysis for renal failure	Risk factor	Binary	0.030
Diabetes	Insulin control for diabetes	Risk factor	Binary	0.111
Liver	Liver disease	Risk factor	Binary	0.022
Cancer	Cancer within five years	Risk factor	Binary	0.062
Thoracic	Thoracic aorta disease	Risk factor	Binary	0.094
DrugUse	Recent or remote drug use	Risk factor	Binary	0.088
Smoke	Smoke status of patient	Risk factor	Categorical	-
PrevCI	Previous cardiac intervention	Previous Intervention	Binary	0.431
CardShock	Cardiogenic shock	Preoperative	Binary	0.076
MI	Prior MI	Preoperative	Binary	0.120
NYHA	NYHA classification	Preoperative	Categorical	-
Aorta	Aorta procedure performed	Operative	Binary	0.123
Incidence	Non-initial cardiovascular surgery	Operative	Binary	0.188
PAPressure	Systolic pressure	Hemodynamics	Categorical	-
TotCABG	Number of arteries bypassed	Coronary Bypass	Continuous	1.36

In summary, the independent variable X_i in (1) includes the factors in Table EC.1, patient's gender and age, surgery status, patient's admission type, procedure type in Table EC.3, surgeon's identifier, patient's pre-LOS, cardiac patient census, and dummies for weekday, month, and year of the operation.

Table EC.2 Summary Statistics of Categorical Variables in Table EC.1

Variable	Category	Num Obs.	Ratio
NYHA	N/A	1933	36.1%
	I	516	9.6%
	II	998	18.6%
	III	991	18.5%
	IV	663	12.4%
	Unknown	251	4.7%
Race	White	4273	79.8%
	Asian	590	11.0%
	Black	274	5.1%
	Other	215	4.0%
Smoke	FALSE	2694	50.3%
	TRUE	2429	45.4%
	Unknown	229	4.3%
PA Pressure	High	376	7.0%
	Low	2247	42.0%
	Unknown	2729	51.0%

Table EC.3 Numbers of Cases by Surgery Types

Surgery Type	Num Obs.	Ratio
Standard (N = 3420)		
CABG	1718	32.1%
AVR	683	12.8%
MVR	225	4.2%
MVr	254	4.7%
CABG + AVR	318	5.9%
CABG + MVR	57	1.1%
CABG + MVr	58	1.1%
AVR + MVR	107	2.0%
Non-standard (N = 1932)		
Isolated Valve	574	10.7%
Isolated CAB	28	0.5%
Isolated cardiac	369	6.9%
Isolated non-cardiac	15	0.3%
Multiple	690	12.9%
Others	256	4.8%

EC.2. Validity of Number of Cases as Workload Measure

In this section, we discuss the validity of using the number of (other) cases, $NumCases_i$, as the surgeon daily workload measure in our main model. One potential concern is that the cases scheduled on single-case versus multi-case days may be very different in their surgery duration and patient risks. If this were the case, the number of cases may not be a proper measure for surgeon daily workload and the observed effects may be driven by unobserved factors. We address this concern as follows.

First, in Table EC.4, we report the summary statistics of the surgery duration (OR time and incision time) for cases with different (1) total number of cases by its surgeon and (2) sequencing position in their surgery day. The sequencing position is determined by the OR entry time of all cases by the focal surgeon in a day (e.g., first case is the one with the earliest OR entry time). In addition, Figure EC.1 shows the box plots for the surgery duration distributions. We see that the distributions of surgery duration are generally similar for the cases with different surgeon daily workload levels and sequencing positions. The only visible difference happens for the cases scheduled as the fourth one in a day, which only account for 13 observations. Thus, we expect the number of cases to be a valid measure for the surgeon daily workload in terms of the time they spend on operation.

In addition, we check the distributions of surgeon daily workload (number of cases) and the sequencing position by the ten common procedure types in our sample (see Section EC.1). Each of the ten common procedure types has more than 100 cases in our sample (Table EC.3). They account for 97% of all observations. The distributions are plotted in Figure EC.2, with the left (resp. right) panel for surgeon daily workload

Table EC.4 Summary Statistics of Surgery Duration by Surgeon Daily Workload and Sequencing Position

	Incision time				OR time			
Total Num. Cases	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
1	3.717	4.650	4.862	5.767	5.933	7.000	7.251	8.233
2	3.533	4.417	4.617	5.450	5.667	6.633	6.876	7.800
3	3.717	4.583	4.763	5.633	5.900	6.867	7.117	8.150
4	3.354	4.817	4.856	5.612	5.692	7.092	7.215	8.137

	Incision time				OR time			
Seq. Position	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
1	3.600	4.517	4.733	5.600	5.817	6.833	7.094	8.050
2	3.667	4.533	4.738	5.600	5.800	6.733	6.975	8.000
3	3.642	4.517	4.765	5.683	5.817	6.733	7.024	7.992
4	3.767	5.533	5.765	7.650	6.083	8.133	8.333	10.550

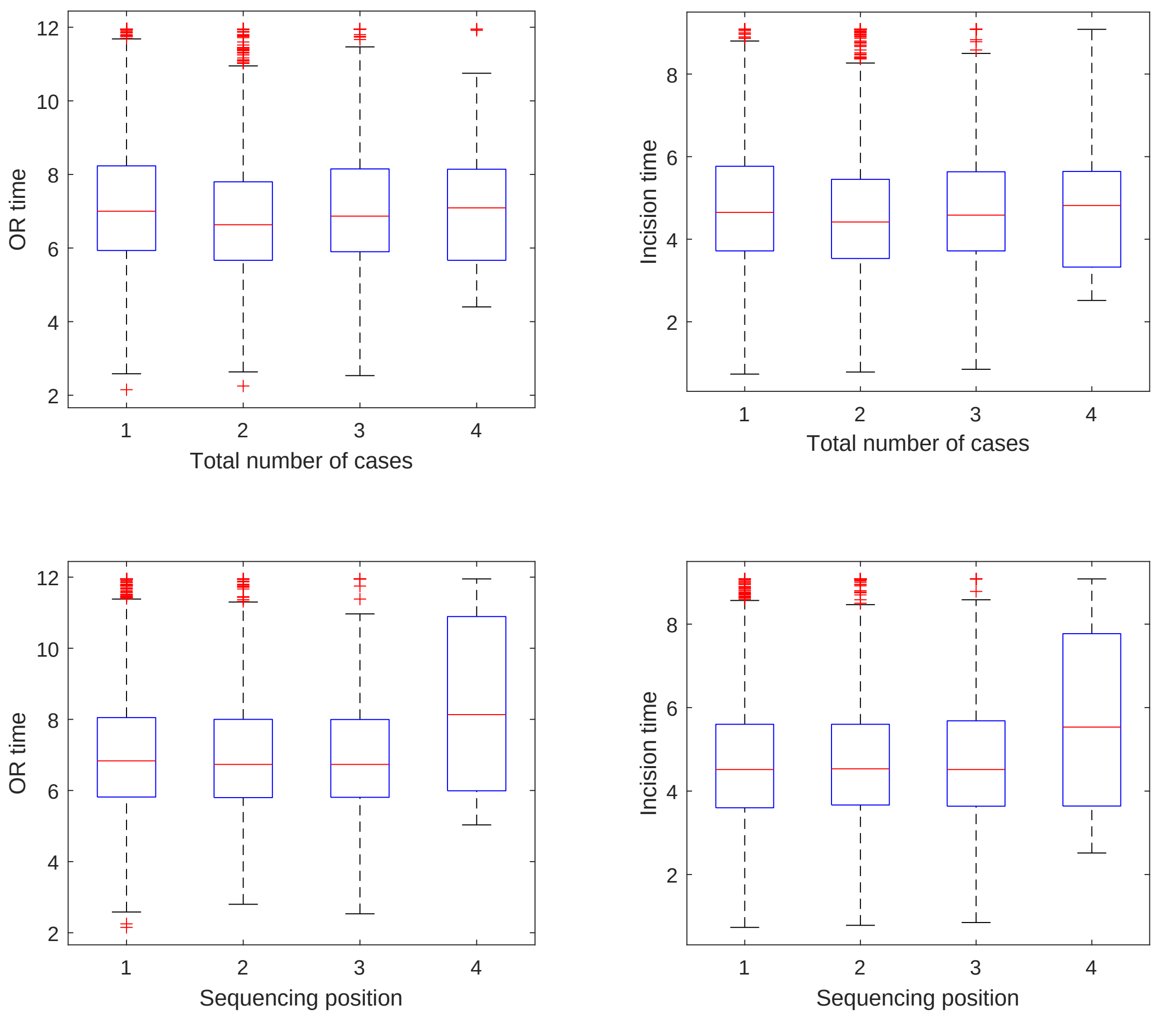
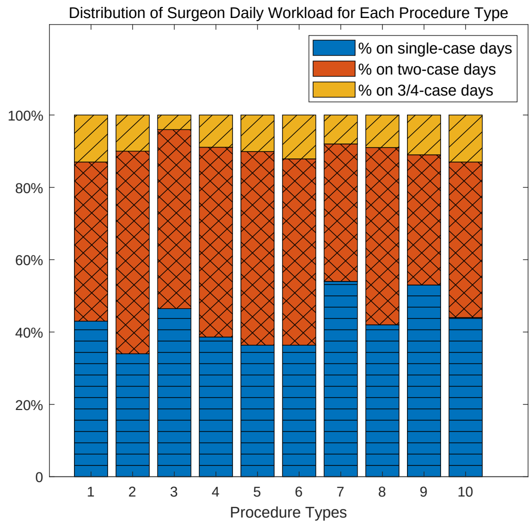


Figure EC.1 Distribution of surgeon daily workload (left) and sequencing position (right) by procedure types.



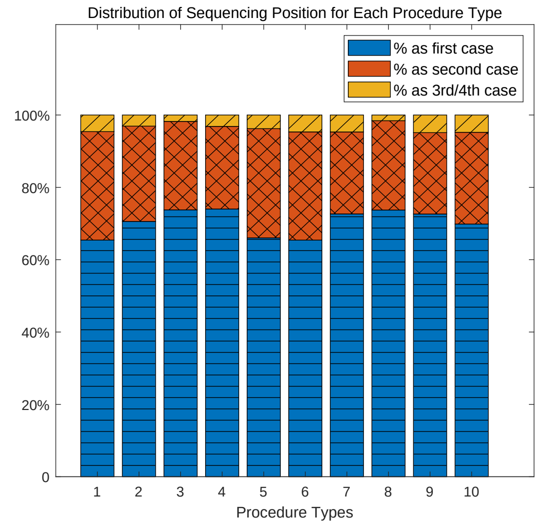


Figure EC.2 Distribution of surgeon daily workload (left) and sequencing position (right) by procedure types.

(resp. sequencing position). In the figure, each bar represents a procedure type. In the left panel, the blue, red, and yellow areas represent the proportion of cases scheduled on single-case, two-case, and three/four-case days, respectively. In the right panel, the blue, red, and yellow areas represent the proportion of cases scheduled as the first, second, and third/fourth case in a day, respectively. We see that the distributions are generally consistent across the procedure types. In particular, we do not find that certain procedure types are exclusively scheduled on single-case or multi-case days, or as the first or later cases in a day.

Finally, we compute the correlations between three main risk scores (i.e., predicted rates for mortality, reoperation, readmission) and the surgeon daily workload or the sequencing position of cases. The predicted scores for eight standard procedure types (62%) are provided by the STS model. The predicted scores for the other six procedure types (38%) are estimated using our sample. The results are shown in Table EC.5. We see that the correlations are relatively low in all cases.

Table EC.5 Correlations Between Predicted Risk and Surgeon Daily Workload/Sequencing Position of Cases

	PredictMort	PredictReOp	PredictReadm
Num. cases	−0.062	−0.103	−0.075
Seq position	−0.010	−0.026	−0.009

The above results suggest that the surgery duration, procedure type, and predicted risks are largely consistent across cases with different number of operations by their surgeons in a day and sequencing positions. These analyses support the validity of using the number of cases as the surgeon daily workload measure. They also mitigate the concern that the observed effects are driven by unobserved risk factors.

EC.3. Estimated Effects of Alternative Workload Measures

In the main model, we use the number of other cases by a surgeon in a day ($NumCases_i$) as the measure for surgeon workload. In this section, we consider three alternative workload-related measures to examine the robustness of our main findings. First, we measure a surgeon's workload by the total incision time of other cases in a day, $TotInc_i$, which takes continuous values. Second, we compute a surgeon's total prior incision time in a day before the start of the focal case, $PriorInc_i$. It includes the incision time for all cases by the focal surgeon that started before the OR entry time of case i. Finally, we introduce a dummy for being a non-first case of focal surgeon, $NonFirst_i$. The summary statistics of the three alternative workload measures are reported in Table EC.6. Here the total incision and prior incision time are measured in hours.

Table EC.6 Summary Statistics of Alternative Workload Measures

	1st Qu.	Mean	Median	3rd Qu.	Std
$TotInc_i$	0.00	3.20	3.17	5.17	3.35
$PriorInc_i$	0.00	1.25	0.00	2.45	2.19
$NonFirst_i$	-	0.31	-	-	-

Similar to $NumCases_i$ in our main setting, the total incision time $TotInc_i$ also captures the average effect of surgeon daily workload. On the other hand, the prior incision time $PriorInc_i$ and the non-first indicator $NonFirst_i$ distinguish between the first and later cases: If case i is the first case of its surgeon in a day, then we have $PriorInc_i$ and $NonFirst_i$ equal zero by definition. The four workload measures are highly related in our sample, with their correlations reported in Table EC.7. This is because in our data, a surgeon mostly has one or two cases in a day (90% of sample). Thus, the state of being a non-first case (or prior incision time) is highly correlated with surgeon daily workload. This is common in the cardiac operations literature (e.g., Zhang et al. 2023) as cardiac surgery usually takes long time to complete. Thus, it is relatively rare for surgeons to perform more than two cases in a day. Moreover, we note that the correlation between the total incision time $TotInc_i$ and number of cases $NumCases_i$ is extremely high (0.93). This can be explained as the average surgery duration is relatively similar for cases with different $NumCases_i$, as shown in previous section.

Table EC.7 Correlations Between Four Workload-related Measures

	$NumCases_i$	$TotInc_i$	$PriorInc_i$	$NonFirst_i$
$NumCases_i$	1.00	0.93	0.52	0.55
$TotInc_i$	0.93	1.00	0.53	0.50
$PriorInc_i$	0.52	0.53	1.00	0.85
$NonFirst_i$	0.55	0.50	0.85	1.00

For all the three alternative measures, we estimate model (1) using TSLS with the two IVs described in Section 3.1. The control covariates X_i (e.g., patient demographic, operative factors) are included in all models. Since our two IVs are constructed at the daily level, the variation in the workload measures picked up by the IVs in the first-stage is also at the daily level. That is, we cannot differentiate the treatment effect between the second case versus the third or fourth case. The two IVs still significantly impacts the alternative workload measures with expected signs. The coefficients of the two IVs in the first-stage regression are reported in Table EC.8. They are statistically significant at the 0.1% level in all cases. This shows that the relevance condition of the two IVs is still satisfied.

Table EC.8 First Stage Regression: Estimated Effects of IVs on Surgeon Daily Workload Measures

	$TotInc_i$	$PriorInc_i$	$NonFirst_i$
$TotOthers_i$	-0.372^{***}	-0.059^{***}	-0.027^{***}
	(0.051)	(0.010)	(0.002)
$ElecWDRatio_i$	4.609^{***}	2.413^{***}	0.505^{***}
	(0.878)	(0.403)	(0.058)

Robust standard error is reported in parenthesis; $^{\dagger}p<0.1$, $^{*}p<0.05$, $^{**}p<0.01$, and $^{***}p<0.001$.

The estimated effects of the three alternative workload measures are reported in the Panels B, C, and D of Table EC.9. For comparison, Panel A shows the estimation results of $NumCases_i$ in our main specification (same as Tables 4 and 5). We see that the effects are still statistically significant for the alternative workload measures. In particular, being a non-first case in a day or having a longer total/prior incision time are associated with longer surgery duration and post-surgery recovery time. This is consistent with the findings in our main specification based on the number of cases as surgeon daily workload measure (see Panel A of Table EC.9). The heterogeneity in the treatment effects largely holds: the effects on surgery duration are statistically significant for both electives and non-electives, but the effects on post-LOS and total ICU time are significant only for non-elective cases.

The magnitudes of treatment effects are also consistent for different workload measures. To see this, first consider the total incision time $TotInc_i$, which is measured in hours. Note that the coefficients of $TotInc_i$ are about one fifth of those in in Panel A for the number of cases $NumCases_i$. This is consistent with the fact that the average incision time of a case is 4.7 hours in our sample. We notice that the magnitudes of effects are larger for the $NonFirst_i$ dummy (in Panel D) than for the $NumCases_i$ (in Panel A) as workload measure. This is because the coefficients of $NumCases_i$ represent the average effect for all cases performed by a surgeon in a day, while the coefficients of $NonFirst_i$ are the effects only for the non-first cases. The magnitudes of effects for $NonFirst_i$ and $PriorInc_i$ are also consistent given the average prior incision time for non-first cases is roughly four hours.

Table EC.9 Estimated Effects of Different Measures for Surgeon Daily Workload

Panel A: Effects of $NumCases_i$

	OR Time (in hours)	Incision Time (in hours)	Post-LOS (in days)	ICU time (in days)
Full Sample (N=5345)	0.455^{***} (0.137)	0.458^{***} (0.101)	1.109^{***} (0.335)	1.166^{***} (0.240)
Electives (N=2474)	0.567^{*} (0.256)	0.531^{**} (0.187)	-0.155 (0.867)	0.101 (0.620)
Non-electives (N=2871)	0.377^{*} (0.158)	0.395^{*} (0.170)	1.605^{*} (0.807)	1.647^{***} (0.367)

Panel B: Effects of $TotInc_i$

	OR Time (in hours)	Incision Time (in hours)	Post-LOS (in days)	ICU time (in days)
Full Sample (N=5345)	0.094^{***} (0.027)	0.095^{***} (0.019)	0.232^{***} (0.067)	0.243^{***} (0.048)
Electives (N=2474)	0.112^{*} (0.048)	0.105^{**} (0.034)	-0.031 (0.176)	0.023 (0.126)
Non-electives (N=2871)	0.080^{**} (0.031)	0.084^{*} (0.033)	0.341^{*} (0.165)	0.349^{***} (0.068)

Panel C: Effects of $PriorInc_i$

	OR Time (in hours)	Incision Time (in hours)	Post-LOS (in days)	ICU time (in days)
Full Sample (N=5345)	0.284^{***} (0.043)	0.260^{***} (0.042)	$0.424^{\dagger}$ (0.230)	0.464^{***} (0.088)
Electives (N=2474)	$0.448^{\dagger}$ (0.233)	0.432^{*} (0.216)	0.066 (0.539)	-0.021 (0.399)
Non-electives (N=2871)	0.196^{***} (0.052)	0.169^{**} (0.055)	0.364 (0.367)	0.508^{***} (0.121)

Panel D: Effects of $NonFirst_i$

	OR Time (in hours)	Incision Time (in hours)	Post-LOS (in days)	ICU time (in days)
Full Sample (N=5345)	1.169^{***} (0.230)	1.055^{***} (0.177)	1.948^{*} (0.857)	1.861^{***} (0.378)
Electives (N=2474)	1.661^{*} (0.755)	1.581^{*} (0.659)	0.525 (2.090)	0.347 (1.327)
Non-electives (N=2871)	0.874^{*} (0.341)	0.752^{*} (0.306)	1.976 (1.337)	2.087^{***} (0.477)

Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

The above results show that our estimated effects are robust for different surgeon workload measures. In addition, the results on prior incision time and non-first case status provide suggestive evidence for the surgeon fatigue channel.

EC.4. Additional Tests for Non-first Case Effects

As discussed in Section 4.4, the estimated workload effects can be driven by multiple channels, including surgeon fatigue and operational constraints related to surgeon daily workload. Due to data limitation, we cannot fully separate the effects from these factors to identify the exact channel. In this section, we further investigate this issue with additional empirical tests.

We check the effects of being a non-first case of the focal-surgeon using the sample in which a surgeon performs multiple cases in a day (multi-case day sample). The variable of interest is the dummy $NonFirst_i$ described in Section EC.3. However, there is a fundamental limitation in our IV-based identification method for the multi-case day sample. As shown in Section 3.1, our two IVs are constructed at the surgeon-day level, thus their values do not vary for the cases performed by the same surgeon in a day. This limits the variation that can be used for identification in the multi-case day sample. To see this, Table EC.10 reports the estimation results of the first-stage regression for $NonFirst_i$ using the multi-case day sample, as well as the elective and non-elective subsamples of it. Both the two IVs are statistically insignificant for the multi-case day sample as well as the elective case subsample in it. For the non-elective subsample, only the

$ElecRatioWD_i$ is weakly significant at the 10% level. This suggests that the two IVs do not satisfy the relevance condition necessary for identification. We then run the TSLS regression using $ElecRatioWD_i$ as the IV on the multi-case day sample. The estimated coefficients are statistically insignificant for all the four dependent variables (OR time, incision time, post-LOS, and ICU time) and three subsamples (all cases, electives, and non-electives).

Table EC.10 First Stage Regression: Estimated Effects of IVs on $NonFirst_i$ in Multi-case Day Sample

	Full	Elective	Non-elective
$TotOther_i$	−0.004	−0.001	−0.009
	(0.006)	(0.008)	(0.009)
$ElecRatioWD_i$	0.122	0.053	$0.243^{\dagger}$
	(0.107)	(0.164)	(0.142)
Num. Obs	3,075	1,576	1,499

Note: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Next, we perform a propensity score-based analysis to further investigate the non-first case effect in the multi-case day sample. The propensity score-based approaches can reduce the bias due to confounding variables in the estimation of treatment effects. We first estimate the propensity score using a logistic model as

$$ln\left[\frac{\Pr\left(NonFirst_i = 1|X_i\right)}{1-\Pr\left(NonFirst_i = 1|X_i\right)}\right] = \beta X_i + \varepsilon_i,$$

where X_i is the list of demographic, risk, operative, and operational factors considered in our model (see Section 3). The logistic model estimates the probability for case i to be a non-first case in the multi-case day sample. Then, we estimate the non-first case effects by both propensity score matching and propensity score weighting (see, e.g., Guo and Fraser 2014).

For propensity score matching, we match case i in the treatment group ($NonFirst_i = 1$) with another case in the control group ($NonFirst_i = 0$) that has similar propensity score. We apply the optimal pair (one-to-one) matching algorithm with a very small tolerance. By the above procedure, we create a sample that is balanced in the estimated propensity score after matching. We then perform the paired t-test in the matched sample for each of the four dependent variables, i.e., OR time, incision time, post-LOS, and total ICU time. For propensity score weighting, we perform a weighted least squared (WLS) regression to estimate the treatment effect for each dependent variable. The weight of each observation is determined by its estimated propensity score and treatment status: Each case in the treatment group receives weight $1/\Pr\left(NonFirst_i = 1|\,X_i\right)$, and each case in the control group receives weight $1/(1 - \Pr\left(NonFirst_i = 1|\,X_i\right))$. We trim the propensity score weights to be between their 2.5th and 97.5th percentiles to mitigate the impact of extreme values.

The estimated effects from propensity score matching and propensity score weighting are reported in Table EC.11 for the multi-case day sample. The estimated effects are mostly statistically insignificant, suggesting the non-first case effect is limited. The only exception is the coefficient for OR time from propensity score weighting, which is statistically significant at the 5% level. The effect has the opposite sign compared with that of our main specification and has a much smaller magnitude.

Table EC.11 Estimated Effects of $NonFirst_i$ in the Multi-case Day Sample: Propensity Score Matching and Weighting (N=3075)

	OR Time (in hours)	Incision Time (in hours)	ICU time (in days)	Post-LOS (in days)
Propensity score matching	-0.142 (0.101)	-0.043 (0.062)	0.234 (0.215)	0.402 (0.373)
Propensity score weighting	-0.115* (0.048)	0.012 (0.042)	0.111 (0.168)	0.348 (0.347)

Note: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Finally, we conduct an additional analysis for the effects of being a third/fourth case by a surgeon in a day. In our data, the sample sizes of the third and fourth cases are very small (in total 204 cases, 3.8% of sample). Therefore, we cannot estimate the workload effects for them separately. That said, we investigate their impacts on the outcomes by estimating the model that includes an interaction term between the surgeon daily workload and an indicator of being a third or fourth case in a day. To address endogeneity, we use the fitted workload levels estimated from the first stage regression with the two IVs, like in our main specification. In particular, in the second stage of the TSLS, we estimate:

$$y_i = \beta X_i + \gamma_1 \widehat{NumCases}_i + \gamma_2 \widehat{NumCases}_i \times \mathbf{1}\{ThirdFourthCase\} + \varepsilon_i, \tag{EC.1}$$

where $\widehat{NumCases}_i$ is the fitted workload level from the first stage regression (3). To ensure we have sufficient variations in the IVs, we estimate the two stages of regressions using the full sample.

Table EC.12 Estimated Effects with Interaction Term for Workload and Third/Fourth Case in a Day (N = 5345)

	OR time	Incision time	Post-LOS	ICU time
Workload (γ_1)	0.499** (0.155)	0.481*** (0.138)	1.265* (0.558)	1.297*** (0.330)
Interaction for 3rd/4th cases (γ_2)	-0.107 (0.077)	-0.056 (0.075)	0.026 (0.141)	0.046 (0.250)

The estimation results for (EC.1) are reported in Table EC.12. The estimated coefficients of surgeon daily workload (γ_1) are all statistically significant and very similar to our main specification in Table 4. In addition, the coefficients of the interaction term (between workload and third/fourth case) are all statistically insignificant and small in magnitudes.

By the above analysis, we do not find evidence that the estimated workload effects are primarily driven by the status of being a later (e.g., third/fourth) cases in a day. Nor do we find evidence that the marginal effect of an additional case is different for the third/fourth case compared with the second case. However, as discussed above, we acknowledge that our IV-based identification strategy in our main analysis cannot fully identify the root causes for the estimated workload effects, which can be driven by surgeon fatigue or operational factors (e.g., parallel surgery) related to scheduling multiple cases in a day. In addition, the interaction analysis (EC.1) has the limitation of the small sample size of third/fourth cases in our data.

EC.5. Supplementary Tables

This section includes the supplementary tables. Table EC.13 reports the summary statistics of the two IVs. Table EC.14 summarizes the correlation between the two IVs and 21 observable severity factors of patients. Table EC.15 reports the p-values from the Sargan's test for the validity of the IVs. Table EC.16 reports the estimation results for various robustness checks. In Table EC.16, we report the estimated effects on OR time, post-LOS, and total ICU time. The results for incision time are similar to those for OR time, and are reported separately in Table EC.18. For ease of comparison, we include in the first two rows ("Base Model") the estimation results from our main specification in Tables 4 and 5. Panel A shows the results when we control for the OR entry hour of each case in our model. Panel B examines the impact when we use different winsorization level as well as a log transformation for the dependent variables. In Panel C, we estimate our model by excluding the emergent/salvage cases or cases after 3PM. In Panel D, we additionally control for surgeon's workload in previous three days. Panel E reports the results when we only use one of the two IVs in our identification. Finally, in Panel F, we use different specifications for calculating the second IV $ElecRatioWD_i$. The estimated effects under these set-ups are qualitatively similar to our main analysis and are discussed in Section 4.3.

In Table EC.17, we report the estimation results for two additional robustness checks. First, we compute for each case its incision time that overlaps with another case from the same surgeon in the day. Then, we include the overlapping incision time as an exogenous control in our model (1). The results from this robustness check are reported in Panel A of Table EC.17. Next, we address potential confounding effects from surgeon heterogeneity via two alternative tests. In the first one, we add surgeons' past volume in our sample as a control variable in our model. It is defined as the total number of cases performed by the surgeon prior to the focal case. Our results are very similar when we control for the past volume of cases with the same procedure type. In the second one, we replace the surgeon and procedure type fixed effects by the more granular surgeon-procedure type fixed effects. We include a dummy for each surgeon and procedure type combination. For the surgeon-procedure type pairs with fewer than 50 observations, we assign one dummy to them for each surgeon (e.g., Surg1_Other) to obtain reliable estimates This yields in total 42 dummies for the surgeon-procedure type fixed effects. The estimation results are included in Panel B of Table EC.17. As shown in the table, the estimated effects of surgeon daily workload s remain robust under these robustness checks.

In addition, as a robustness check for the estimation results of post-LOS, we use several more conservative measures of post-LOS by adjusting the original measure using the OR entry or exit time of each case. The estimated coefficients of $Workload_i$ by TSLS are given in Table EC.19. The first column shows the original results in Tables 4 and 5, in which we compute the post-LOS as the number of days between the OR exit and discharge dates. In the second column ("Entry $\geq$ 3PM"), we subtract a day from the post-LOS if the OR entry time of the case is later than 3PM as there is some evidence that late surgery start times are associated with an increase of LOS by one day. In the third column ("OR Exit Hour"), we subtract the hours elapsed before OR exit on the day of OR exit. In the last two columns, we further subtract a day from the post-LOS if the patient leaves the OR after 12PM and 4PM, respectively. We see from Table EC.19 that our estimated effects for post-LOS remain similar and robust in all these conservative measures.

Finally, we report in Table EC.20 the estimated coefficients of selected patient demographic, risk, operative, and operational factors in our main regression. The coefficients are estimated by TSLS using the two IVs on the full sample.

Table EC.13 Summary Statistics of IV

	Full Sample			Elective Sample			Non-elective Sample		
IV	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
$TotOther_i$	4.11	4.00	1.79	4.30	4.00	1.67	3.95	4.00	1.87
$ElecRatioWD_i$	0.22	0.23	0.12	0.25	0.25	0.11	0.20	0.21	0.13

Table EC.14 Correlation between IVs and Observable Severity Factors

	$TotOther_i$	$ElecRatioWD_i$
Gender: Male	−0.019	−0.026
Age	0.054	0.072
Smoke	−0.006	−0.003
Drug use	−0.017	−0.015
NYHA: III or IV	0.002	−0.102
Endocarditis	−0.017	−0.060
Peripheral arterial disease	0.002	−0.042
Incidence	−0.030	−0.084
Lung disease	0.023	0.015
Liver disease	0.020	−0.011
Thoracic aorta disease	−0.020	0.005
Dialysis	−0.013	−0.017
Diabetes	0.005	−0.040
Cancer	0.008	0.011
Hypertension	0.034	0.043
Previous intervention	−0.031	−0.077
Carotid stenosis	0.027	0.013
Cardiogenic shock	−0.052	−0.185
Syncope	0.001	−0.007
MI	−0.007	−0.068
Systolic pressure: high	0.009	−0.027

Table EC.15 p-values from Sargan's Test for the two IV's Validity

OR time			Incision time			Post-LOS			Total ICU time		
Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec
0.751	0.186	0.405	0.535	0.134	0.544	0.632	0.951	0.207	0.827	0.566	0.887

Table EC.16 Robustness Checks: Estimated Effects of Daily Workload (Number of Other Cases)

Panel A: Controlling for OR entry time

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec
Base model	0.455^{***} (0.137)	0.567^{*} (0.247)	0.377^{*} (0.153)	1.109^{***} (0.335)	-0.155 (0.837)	1.605^{*} (0.780)	1.166^{***} (0.240)	0.101 (0.599)	1.647^{***} (0.355)
Dummy for OR entry hour	0.301^{*} (0.123)	0.367^{*} (0.165)	0.253 (0.161)	1.011^{**} (0.335)	0.080 (0.816)	$1.337^{\dagger}$ (0.712)	1.058^{***} (0.173)	0.226 (0.573)	1.440^{***} (0.220)
Dummy for morning, afternoon, night	0.314^{*} (0.130)	0.389^{*} (0.188)	0.270 (0.167)	1.095^{***} (0.324)	0.037 (0.813)	1.511^{*} (0.724)	1.060^{***} (0.223)	0.155 (0.569)	1.484^{***} (0.324)

Panel B: Other winsorization level and log transformation

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec
Winsorize: 95th pct	0.491^{***} (0.109)	0.556^{**} (0.204)	0.434^{**} (0.181)	0.863^{*} (0.336)	0.238 (0.758)	$1.065^{\dagger}$ (0.611)	0.768^{***} (0.165)	0.218 (0.448)	0.984^{***} (0.227)
Log transform	0.081^{**} (0.026)	0.106^{*} (0.045)	0.064^{*} (0.027)	0.080^{*} (0.033)	0.062 (0.060)	0.082 (0.062)	0.119^{*} (0.060)	0.042 (0.102)	$0.153^{\dagger}$ (0.091)

Panel C: Estimation with subsample

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non−elec	Full	Elec	Non−elec	Full	Elec	Non−elec
Drop emergent & salvage cases	0.448^{**} (0.161)	0.567^{*} (0.256)	0.347^{*} (0.169)	1.134^{**} (0.410)	-0.155 (0.867)	1.824^{*} (0.800)	1.205^{***} (0.190)	0.101 (0.620)	1.858^{***} (0.268)
Drop cases after 3PM	0.322^{*} (-0.133)	0.543^{*} (-0.253)	0.177 (-0.154)	1.584^{***} (0.269)	0.253 (1.123)	2.129^{**} (0.779)	1.272^{***} (0.246)	0.024 (0.614)	1.876^{***} (0.296)

Panel D: Controlling for surgeon workload in previous days

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non−elec	Full	Elec	Non−elec	Full	Elec	Non−elec
Add num. of cases in prior three days	0.445^{**} (0.134)	0.557^{*} (0.256)	0.366^{*} (0.156)	1.079^{**} (0.318)	-0.206 (0.885)	1.578^{*} (0.782)	1.155^{***} (0.234)	0.065 (0.626)	1.649^{***} (0.354)

Panel E: Estimation with one IV

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non−elec	Full	Elec	Non−elec	Full	Elec	Non−elec
$TotOther_i$	0.424^{*} (0.181)	0.356^{***} (0.090)	0.477 (0.291)	1.370^{**} (0.528)	-0.110 (1.152)	$2.559^{\dagger}$ (1.347)	1.090^{**} (0.413)	0.371 (0.800)	1.716^{*} (0.846)
$ElecRatioWD_i$	0.498^{*} (0.254)	$0.902^{\dagger}$ (0.529)	0.246 (0.184)	0.738 (0.547)	-0.225 (1.239)	0.350 (1.170)	1.272^{***} (0.193)	-0.314 (0.930)	1.559^{***} (0.468)

Panel F: Other specifications for $ElecRatioWD_i$

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec
Use prior cases in $[t-L,t)$	0.423^{***} (0.127)	0.560^{**} (0.192)	0.367^{*} (0.175)	1.160^{*} (0.514)	0.470 (0.876)	$1.516^{\dagger}$ (0.789)	1.193^{***} (0.300)	0.330 (0.673)	1.630^{***} (0.327)
$L = 90$ days	0.448^{**} (0.152)	$0.479^{\dagger}$ (0.280)	0.419^{*} (0.176)	1.281^{***} (0.376)	0.249 (0.864)	1.730^{*} (0.759)	1.091^{***} (0.232)	0.317 (0.660)	1.437^{***} (0.324)

Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table EC.17 Robustness Checks (cont'd): Estimated Effects of Daily Workload (Number of Other Cases)

Panel A: Controlling for overlapping incision time

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec
Base model	0.455*** (0.137)	0.567* (0.247)	0.377* (0.153)	1.109*** (0.335)	-0.155 (0.837)	1.605* (0.780)	1.166*** (0.240)	0.101 (0.599)	1.647*** (0.355)
Add overlapping incision time	0.429† (0.259)	0.736 (0.541)	0.252 (0.230)	1.474*** (0.312)	-0.306 (0.945)	2.090*** (0.485)	1.158* (0.502)	-0.388 (1.242)	1.491 (1.115)

Panel B: Controlling for surgeon past volume and surgeon-type fixed effects

	OR time (in hours)			Post-LOS (in days)			Total ICU (in days)		
	Full	Elec	Non-elec	Full	Elec	Non-elec	Full	Elec	Non-elec
Add surgeon past volume	0.446** (0.154)	0.576* (0.263)	0.372* (0.158)	1.202*** (0.200)	0.122 (0.596)	1.693*** (0.317)	1.103*** (0.328)	-0.145 (0.841)	1.585* (0.781)
Surgeon-procedure type fixed effects	0.351** (0.114)	0.458* (0.231)	0.270† (0.150)	1.242*** (0.273)	-0.082 (0.645)	1.849*** (0.376)	1.146*** (0.334)	-0.335 (0.927)	1.829* (0.798)

Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table EC.18 Robustness Checks: Estimated Effects for Incision Time

	Full	Elec	Non-elec
Base model	0.458*** (0.101)	0.531** (0.181)	0.395* (0.163)
Dummy for OR entry hour	0.356*** (0.095)	0.391** (0.136)	0.321* (0.159)
Dummy for morning, afternoon, night	0.358*** (0.100)	0.396** (0.139)	0.329† (0.170)
Winsorization 95th pct	0.441*** (0.091)	0.525** (0.168)	0.373* (0.163)
Log transform	0.131*** (0.032)	0.172** (0.058)	0.102* (0.043)
Drop emergent & salvage cases	0.454*** (0.086)	0.531** (0.187)	0.372* (0.146)
Add overlapping incision time	0.448** (0.148)	0.711† (0.404)	0.276 (0.200)
Surgeon-procedure type fixed effects	0.355*** (0.097)	0.417* (0.166)	0.289† (0.172)

Full	Elec	Non-elect	
Drop cases after 3PM	0.319** (0.100)	0.523** (0.197)	0.176 (0.153)
Add num. of cases in prior three days	0.455*** (0.098)	0.544** (0.193)	0.380* (0.168)
One IV $TotOther_i$	0.405* (0.195)	0.322* (0.147)	0.460 (0.308)
One IV $ElecRatioWD_i$	0.532*** (0.155)	0.863* (0.419)	0.311** (0.117)
Using prior cases in $[t-L,t)$	0.411** (0.128)	0.503* (0.210)	0.359* (0.173)
$L = 90$ days	0.477*** (0.110)	0.484* (0.199)	0.415* (0.174)
Add surgeon past volume	0.458*** (0.108)	0.541** (0.187)	0.405* (0.165)

Robust standard error is reported in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table EC.19 Estimated Effects of Daily Workload (Number of Other Cases) on Post-LOS with Different Adjustments

	Original	Entry $\geq$ 3PM	OR Exit Hour	Exit $\geq$ 12PM	Exit $\geq$ 4PM
Full Sample	1.109*** (0.335)	1.275*** (0.316)	1.140*** (0.325)	1.077** (0.333)	1.205*** (0.296)
Non-elective Sample	1.605* (0.807)	1.789* (0.797)	1.635* (0.801)	1.589* (0.803)	1.678* (0.769)

Robust standard errors in parenthesis; $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Table EC.20 Coefficients of Control Variables in Model (1): Full Sample

	OR time	Post-LOS	Tot ICU time
Gender: Male	0.136*	−0.051	0.098
	(0.067)	(0.423)	(0.225)
Age	−0.005	0.050***	0.040***
	(0.003)	(0.015)	(0.007)
Race: Asian	0.097	1.247*	0.179
	(0.122)	(0.520)	(0.259)
Race: Black	0.148	0.401	0.15
	(0.103)	(0.529)	(0.310)
Race: Other	0.158	−0.314	−0.097
	(0.135)	(0.383)	(0.350)
Endocarditis	0.363***	1.431*	0.147
	(0.076)	(0.622)	(0.390)
Lung disease	0.238***	1.487***	0.779***
	(0.062)	(0.247)	(0.190)
Peripheral arterial disease	0.177	1.077***	0.605*
	(0.119)	(0.269)	(0.261)
Liver disease	0.207	2.167†	1.190*
	(0.131)	(1.108)	(0.483)
Thoracic aorta disease	0.346**	0.406	0.491
	(0.123)	(0.678)	(0.540)
Cancer	0.06	0.515*	0.485
	(0.099)	(0.237)	(0.340)
Hypertension	0.095*	0.396	0.374†
	(0.037)	(0.393)	(0.216)
Recent drug use	−0.016	0.990†	0.384
	(0.089)	(0.576)	(0.300)
Dialysis/renal failure	0.398**	6.119***	4.875***
	(0.148)	(0.888)	(0.402)
Diabetes	0.175**	2.167***	1.007***
	(0.065)	(0.246)	(0.255)
Smoke	−0.016	−0.041	0.116
	(0.026)	(0.172)	(0.080)
Carotid stenosis	0.068	0.001	−0.242
	(0.058)	(0.489)	(0.424)
Syncope	−0.176†	−0.265	−0.629
	(0.105)	(0.538)	(0.413)
Previous intervention	0.104*	1.128*	0.844***
	(0.046)	(0.458)	(0.211)
Cardiogenic shock	0.350	4.132***	2.011***
	(0.231)	(0.623)	(0.424)
Prior MI	0.023	0.350	0.477†
	(0.057)	(0.418)	(0.268)
Aorta procedure	1.407***	−0.229	−0.143
	(0.337)	(1.276)	(0.783)

	OR time	Post-LOS	Tot ICU time
Status: Urgent	0.096*	0.976***	0.511**
	(0.046)	(0.284)	(0.188)
Status: Emergent	0.512*	2.515	1.820**
	(0.241)	(1.748)	(0.700)
Status: Salvage	0.253	1.52	5.105***
	(0.675)	(2.933)	(1.142)
Adm Type: ED	0.031	1.160*	0.473†
	(0.142)	(0.531)	(0.246)
Adm Type: Transfer in	−0.053	1.751***	1.118**
	(0.086)	(0.391)	(0.346)
Adm Type: Other	−0.875***	1.394	0.441
	(0.263)	(1.781)	(1.210)
Pre-LOS	0.002	0.040**	0.004
	(0.002)	(0.015)	(0.007)
Department census	0.007***	0.001	0.007
	(0.001)	(0.008)	(0.008)
Num. of arteries bypassed	0.418***	−0.122	0.167
	(0.044)	(0.297)	(0.173)
NYHA: N/A	0.077	-0.001	-0.104
	(0.050)	(0.153)	(0.129)
NYHA: II	0.038	0.037	−0.236
	(0.073)	(0.186)	(0.185)
NYHA: III	0.214***	0.988*	0.683*
	(0.056)	(0.386)	(0.281)
NYHA: IV	0.354***	5.491***	2.643***
	(0.105)	(0.956)	(0.544)
NYHA: Unknown	0.214	1.735**	0.52
	(0.130)	(0.656)	(0.456)
Non-initial surgery	1.486***	0.582	0.312
	(0.132)	(0.468)	(0.252)
Systolic pressure: Low	−0.072	−1.303***	−0.853**
	(0.127)	(0.323)	(0.280)
Systolic pressure: Unknown	−0.108	−1.634***	−1.003**
	(0.123)	(0.319)	(0.307)
Num Cases	0.455***	1.109***	1.166***
	(0.137)	(0.335)	(0.240)
Constant	4.276***	3.583†	−2.075*
	(0.479)	(1.841)	(1.021)
Procedure type	Y	Y	Y
Surgeon fixed effect	Y	Y	Y
Weekday fixed effect	Y	Y	Y
Monthly fixed effect	Y	Y	Y
Year fixed effect	Y	Y	Y
Num. obs	5,345	5,344	5,319

EC.6. Surgical Scheduling MIQP Formulation

In the following, we formulate the surgical scheduling model used in Section 5. We introduce the notation and decision variables, feasibility constraints, and objective functions. We show the optimization model can be formulated as a Mixed-Integer Quadratic Programming (MIQP) problem.

EC.6.1. Decision Variables and Feasibility Constraints

We solve the scheduling model for each calendar week (Sunday to Saturday) in the four year horizon of our sample. For a given week, we index each case by $i \in C$, where C is the set of all cases performed on the weekdays in the given week. We exclude the operations on the weekends (2.8% of sample) in the model, as their times are generally hard to change. According to the surgery status, the set C can be divided into three exclusive subsets C_{el}, C_{ug}, and C_{es}, which represent the elective, urgent, as well as emergent and salvage cases respectively. For each case i, we denote its surgeon by $\tilde{s}(i)$ and original surgery date by $\tilde{t}(i)$.

We index the day in the week by $t \in T$ and the surgeon by $s \in S$, with T and S being the sets of surgery dates and surgeons for cases $i \in C$. We use A_s to denote the set of cases performed by surgeon s. Our optimization model considers which cases to assign to each day. Thus, our decision variables are $x_{i,t}$ for $i \in C$ and $t \in T$. Each $x_{i,t}$ is a binary variable; it takes value one if case i is assigned to day t, and zero otherwise. We formulate the set of constraints to ensure the feasibility of the resulting schedule. First, every case should be assigned one and only one date in the final schedule. This translates to

$$\sum_{t \in T} x_{i,t} = 1, \; \forall i \in C. \tag{EC.2}$$

For each case, we specify its feasible set of surgery dates according to its status. For elective cases, we allow them to be assigned to any day of the week of the original date. On the other hand, we impose that urgent cases can only be scheduled on the original date or the adjacent days, while the emergent and salvage cases can only be scheduled on the original date. These constraints reflect the reality that urgent cases are more time sensitive than elective ones as their patients are more severe. In addition, the hospital has little control over the arrival time of emergent and salvage patients. We formulate these constraints as

$$x_{i,t} = 0, \text{ if } i \in C_{ug} \text{ and } |t - \tilde{t}(i)| > 1, \tag{EC.3}$$

and

$$x_{i,t} = 0, \text{ if } i \in C_{es} \text{ and } t \neq \tilde{t}(i). \tag{EC.4}$$

We impose an upper bound on surgeon daily workload, i.e., the number of cases performed by each surgeon in a day, to reflect a physical limit on how much time a surgeon can spend operating. In line with our data, we set the upper bound to be three cases. As we keep the surgeon assigned to each case unchanged, the number of cases by surgeon s on day t in the new schedule can be expressed as,

$$\tilde{n}_{s,t} = \sum_{i \in A_s} x_{i,t}. \tag{EC.5}$$

The summation on the right-hand side includes all the cases by surgeon s. Then, the constraint on surgeon daily workload can be formulated as

$$\tilde{n}_{s,t} \leq \max\{\bar{n}^{(c)}, n_{s,t}^{(c)}\},\ \forall t \in T \text{ and } \forall s \in S, \tag{EC.6}$$

where $n_{s,t}^{(c)}$ is the number of cases performed by surgeon s on day t in the original schedule; $\bar{n}^{(c)}$ is a model parameter to be specified. It denotes the maximum daily workload of a surgeon in the new schedule, unless the surgeon already performs more cases in the original schedule.

Finally, we set an upper bound on the number of days worked by each surgeon in a week. Although asking the surgeons to work for more days naturally smooths their daily workload, it would be difficult to implement in reality given their other responsibilities. Note that the surgeon s works on day t in the new schedule if at least one case is performed, i.e., $\tilde{n}_{s,t} > 0$. Thus, we can formulate the constraint as

$$\sum_{t \in T} \mathbf{1}\{\tilde{n}_{s,t} > 0\} \leq \max\{\bar{n}^{(d)}, n_s^{(d)}\},\ \forall s \in S. \tag{EC.7}$$

where $n_s^{(d)}$ is the number of working days by surgeon s in the original schedule; $\bar{n}^d$ is the model parameter denoting the maximum number of days worked by a surgeon, unless the surgeon works for more days in the original schedule.

The constraint (EC.7) is inconvenient to implement as the indicator function is non-linear. We circumvent this difficulty by proposing the following linear formulation. Let the binary variable $z_{s,t}$ denote whether surgeon s works on day t in the new schedule. We bound it by

$$z_{s,t} \leq M \cdot \tilde{n}_{s,t} \text{ and } z_{s,t} \geq m \cdot \tilde{n}_{s,t}, \tag{EC.8}$$

where M (resp. m) is a sufficiently large (resp. small) constant. In our study, we can set them as $M = 100$ and $m = 0.01$. It is easy to verify by (EC.8) that $z_{s,t}$ takes value one if $\tilde{n}_{s,t} > 0$ and zero if $\tilde{n}_{s,t} = 0$. Thus, it always equals to the indicator function $\mathbf{1}\{\tilde{n}_{s,t} > 0\}$. Then, we can rewrite the constraint (EC.7) in the following linear form as

$$\sum_{t \in T} z_{s,t} \leq \max\{\bar{n}^d, n_s^{(d)}\},\ \forall s \in S. \tag{EC.9}$$

In summary, our model includes the constraints (EC.2), (EC.3) – (EC.4), (EC.5) – (EC.6), and (EC.8) – (EC.9), all of which are formulated in linear form.

EC.6.2. Objective Functions and MIQP Formulation

We now introduce the objective function for our model and show how to formulate the surgical scheduling model as an MIQP problem. We consider three alternative objective functions: minimizing the total expected OR time, post-LOS, or total ICU time. Following our econometric model (1), the expected value of the three variables can be expressed as

$$\hat{y}_i = X_i \beta + \gamma Workload_i. \tag{EC.10}$$

Here we measure workload as the number of other cases performed by the focal surgeon on the day. The variable $\hat{y}_i$ is specified as OR time, post-LOS, and total ICU time, respectively.

By (EC.10), the expected value $\hat{y}_i$ can be decomposed to two parts

$$l_i = X_i\beta \text{ and } d_i = \gamma Workload_i.$$

To focus on the impact of daily workload, we assume the first part l_i, which primarily depends on the patient's risk and operative factors, remains unchanged in the new schedule. However, the second part d_i will be affected by the surgeon's workload in the new schedule. Our objective is to minimize the total expected value $\hat{y}_i$, i.e., $\min \sum_{i\in C} \hat{y}_i$. As we assume l_i does not change, this is equivalent to minimizing the sum of the d_i terms. We explicitly express the objective $\min \sum_{i\in C} d_i$ under the new schedule. The total daily workload term for surgeon s on day t is

$$d'_{s,t} = \sum_{i\in A_s} d_i x_{i,t}.$$

It is straightforward to see the summation over all cases in C is equal to that over all surgeons and days:

$$\sum_{i\in C} d_i = \sum_{s\in S}\sum_{t\in T} d'_{s,t}. \tag{EC.11}$$

As such, it is sufficient to write out the objective function using $d'_{s,t}$ instead of d_i.

The number of cases performed by surgeon s on day t is given by $\tilde{n}_{s,t}$ in (EC.5). To account for the heterogeneous effects as discussed in Section 4.2, we need to further obtain the number of elective and non-elective cases. Similar to (EC.5), they are given by

$$\tilde{n}^{(el)}_{s,t} = \sum_{i\in A_s} x_{i,t}\mathbf{1}\{i \in C_{el}\} \text{ and } \tilde{n}^{(ne)}_{s,t} = \sum_{i\in A_s} x_{i,t}\mathbf{1}\{i \notin C_{el}\}. \tag{EC.12}$$

Then, the total impact from daily workload for surgeon s on day t can be expressed as

$$d'_{s,t} = (\tilde{n}_{s,t} - 1)\cdot\left(\gamma^{(el)}\tilde{n}^{(el)}_{s,t} + \gamma^{(ne)}\tilde{n}^{(ne)}_{s,t}\right).$$

Here $\gamma^{(el)}$ and $\gamma^{(ne)}$ are the estimated coefficient γ for the daily workload effect for the elective and non-elective cases respectively. We set the coefficient to be zero if it is not statistically significant at the 10% level. When we ignore the heterogeneity in the impacts of daily workload, we use the average treatment effects in Table 4 with $\gamma^{(el)} = \gamma^{(ne)} = \gamma^{(avg)}$. Plugging $d'_{s,t}$ into (EC.11), the objective function is given by

$$\min \sum_{s\in S}\sum_{t\in T} (\tilde{n}_{s,t} - 1)\cdot\left(\gamma^{(el)}\tilde{n}^{(el)}_{s,t} + \gamma^{(ne)}\tilde{n}^{(ne)}_{s,t}\right).$$

Note that the summation terms $\sum_{t\in T}\sum_{s\in S}\tilde{n}^{(el)}_{s,t}$ and $\sum_{t\in T}\sum_{s\in S}\tilde{n}^{(ne)}_{s,t}$ represent the total numbers of elective and non-elective cases from all surgeons in the week, thus they remain unchanged in the new schedule. Then the above objective can be simplified as

$$\min \sum_{s\in S}\sum_{t\in T} \tilde{n}_{s,t}\cdot\left(\gamma^{(el)}\tilde{n}^{(el)}_{s,t} + \gamma^{(ne)}\tilde{n}^{(ne)}_{s,t}\right). \tag{EC.13}$$

It is easy to verify that the objective function is quadratic in the decision variables $x_{i,t}$. Thus, our model is formulated as an MIQP with quadratic objective (EC.13) and linear constraints (EC.2), (EC.3) – (EC.4), (EC.5) – (EC.6), and (EC.8) – (EC.9). The decision variables are $x_{i,t}$ for $i \in C$ and $t \in T$, as well as $z_{s,t}$ for $s \in S$ and $t \in T$ as introduced in (EC.8). All the decision variables $x_{i,j}$ and $z_{s,t}$ are binary. The final MIQP formulation to minimize total OR time, post-LOS, or ICU time is given with the objective (EC.13), constraints (EC.2), (EC.3) – (EC.4), (EC.5) – (EC.6), (EC.8) – (EC.9), and (EC.12). The model parameters $\bar{n}^{(c)}$ and $\bar{n}^{(d)}$ denote the upper bound on surgeon's number of cases performed in a day and number of days worked in a week, respectively.

Note that when we use the average effect of daily workload for elective and non-elective cases, i.e., $\gamma^{(el)} = \gamma^{(ne)} = \gamma^{(avg)}$, the optimal schedules from the model are the same for the three variables. To see this, by (5), the objective with average effect can be expressed as

$$\min \gamma^{(avg)} \cdot \sum_{s \in S} \sum_{t \in T} \left(\tilde{n}_{s,t}\right)^2 . \tag{EC.14}$$

Thus, the coefficient $\gamma^{(avg)}$ does not impact the solution and can be dropped from (EC.14) (although it affects the objective value). Similarly, we obtain the same optimal schedule for post-LOS and ICU time when we use the heterogeneous effect of daily workload. In this case, the objective function can be written as

$$\min \gamma^{(ne)} \cdot \sum_{s \in S} \sum_{t \in T} \tilde{n}^{(ne)}_{s,t} \left(\tilde{n}^{(ne)}_{s,t} + \tilde{n}^{(el)}_{s,t}\right). \tag{EC.15}$$

This follows by plugging $\tilde{n}_{s,t} = \tilde{n}^{(el)}_{s,t} + \tilde{n}^{(ne)}_{s,t}$ in (5) and using the fact $\gamma^{(el)} = 0$ for post-LOS and ICU time, i.e., surgeon daily workload does not impact the two outcomes of elective cases.

EC.6.3. Distribution of Department Workload and An Example of Smoothing Surgeon Workload

First, Figure EC.3 plots the frequency distributions of department daily workload, measured by number of cases by all surgeons, from the original and new schedules. It shows that the distributions of department workload are similar under the original and new schedules

Then, we use a concrete example to show how our scheduling model works. We look at the schedule for Surgeon 2 in Week 39 of our sample. This is an extremely busy week for the surgeon, with a total of eleven cases performed across four weekdays (four elective, six urgent, and one emergent cases). The original schedule is shown in the left panel of Figure EC.4. We see surgeon daily workload is highly unbalanced across days: two and one cases on Monday and Tuesday, but four cases on both Thursday and Friday. The middle and right panels show the new schedules under the average and heterogeneous effects, respectively. The surgeon's workload is smoothed under both schedules, and the maximum daily workload is now three cases. Moreover, we see the schedule under the heterogeneous effect further smooths the number of non-elective cases across days: the maximum number of non-elective cases in a day decreases from three to two. This is because only the non-elective cases are negatively affected by high daily workload when we use heterogeneous effect for post-LOS or ICU time.

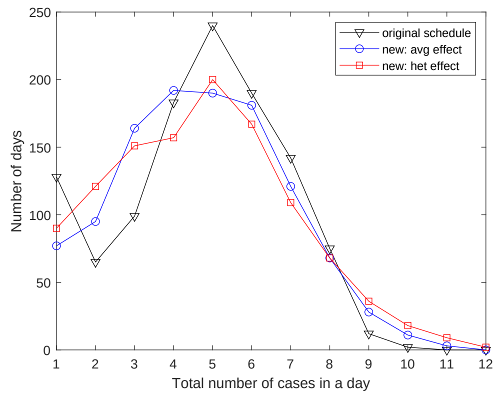


Figure EC.3 **Frequency distributions of total number of cases by all surgeons in a day**

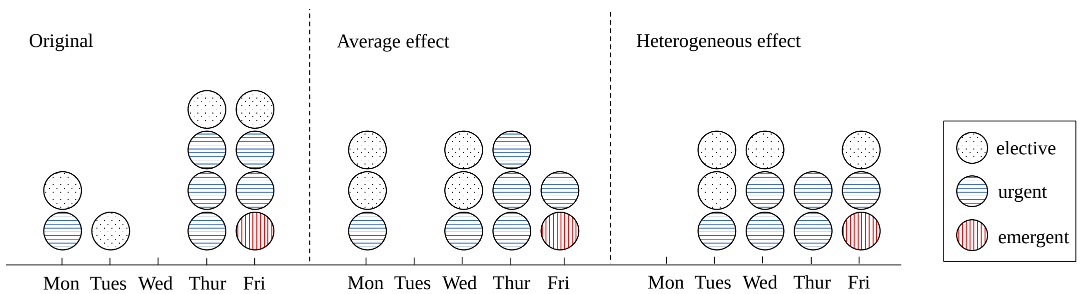


Figure EC.4 **Example surgical schedules from our optimization model: Surgeon 2, Week 39**

EC.7. Robustness Checks on Surgical Scheduling Model

In this section, we perform several robustness checks for our surgical scheduling model, which are discussed in Section 5.2 of the main text. First, we fully fix the time of all non-elective cases (including urgent, emergent, and salvage) and do not allow them to be moved to other days. In this case, only elective cases can be moved within a week. Second, we fully fix the number of working days of surgeons as in the original schedule. The benefits from the scheduling model under the two scenarios are shown in Table EC.21. We see that the new scheduling models still lead to economically meaningful improvements, although the magnitudes become smaller than in our main setting. For example, with time fixed for all non-elective cases (Panel A), the new schedule reduces the OR time by 273 hours, post-LOS by 655 days, and total ICU time by 699 days under the average estimated effects. If we fully fix the number of working days (Panel B), the improvements become smaller under average effects (193 hours for OR time, 470 days for post-LOS, and 494 days for ICU time), but larger under heterogeneous effects for post-LOS and total ICU time (1056 days for post-LOS and 1084 days for ICU time). In both cases, the average numbers of weekly working days are close to that in our original schedule (2.83).

Next, we implement the surgical scheduling model by fixing both the number of surgeon working days per week and the time of non-elective cases. This baseline result can be viewed as the "pure" benefit from workload balancing under restrictive constraints. The estimated benefits are reported in Panel C of Table EC.21. The optimized schedules still lead to meaningful improvements for the outcomes. For example, under the average effects, the new schedule decreases the workload-related terms of OR time, post-LOS, and ICU time by 3.5% simultaneously. The magnitudes of such benefits significantly increase under the heterogeneous effects: the optimized schedule for post-LOS and ICU time can decrease the workload term of the two metrics by 8.4%.

In Panel D of Table EC.21, we report the estimated effects when we allow a surgeon to work for maximum four days in a week (instead of three as in the main setting). We see that the benefits from the scheduling model become significantly larger compared with our main specification. For example, under average effects, the reduction in total OR time increases from 360 hours (in Table 6) to 687 hours. On the other hand, the average weekly working days of a surgeon increases mildly from 2.83 in the original schedule to 3.28 days. This reveals a potential trade-off that can be leveraged by the hospital in its surgery scheduling.

In our optimized schedules in Table 7, the improvement is largely driven by the reductions in the surgeon-days with three and four cases. We further perform a robustness check under smaller workload effects for three-case and four-case days. This provides a conservative estimate for the benefits from the scheduling model. In the first test, we assume the workload effects are 20% smaller for the days with three or four cases by a surgeon. That is, in the objective (EC.13), we set

$$\gamma^{(el)} = 0.8 \times \gamma^{(el)}_{est} \text{ and } \gamma^{(ne)} = 0.8 \times \gamma^{(ne)}_{est}, \text{ if } \tilde{n}_{s,t} = 3, 4.$$

Here $\gamma^{(el)}_{est}$ and $\gamma^{(ne)}_{est}$ denote the original estimated effects from the model (Tables 4 and 5), which are used for one- and two-case days. In the second test, we further decrease the workload effects for three- and four-case days as:

$$\gamma^{(el)} = 2/3 \times \gamma^{(el)}_{est}, \;\; \gamma^{(ne)} = 2/3 \times \gamma^{(ne)}_{est}, \;\; \text{if } \tilde{n}_{s,t} = 3$$

and

$$\gamma^{(el)} = 1/2 \times \gamma^{(el)}_{est}, \; \gamma^{(ne)} = 1/2 \times \gamma^{(ne)}_{est}, \;\; \text{if } \tilde{n}_{s,t} = 4.$$

In this scenario, the workload effect for each case, $\gamma \times \tilde{n}_{s,t}$, is the same for two-case, three-case, and four-case days, which is a conservative assumption.

We evaluate the benefits from the schedules solved in our main model (Table 6) under the two scenarios described above. Note that we cannot easily solve the optimal schedules under the new workload effects, as the model is no longer a MIQP problem. Thus, the estimated benefits from the original schedules provide lower bounds for the maximum benefits that can be achieved under the new workload effects. The results are reported in Table EC.22. Under smaller workload effects for three- and four-case days, the benefits from the scheduling models decrease in magnitude. However, they are still economically significant. In Panel A, the schedule under average treatment effects achieves a relative reduction of 6.3% for all the three outcomes

Table EC.21 Robustness Checks for Surgical Scheduling Model

Panel A: Fixing Time of All Non-elective Cases

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3851.6	273.0	6.6%	156	2.98
(in hours)	Het	4240.8	3908.8	332.0	7.8%	178	2.98
Post-LOS	Avg	10053.1	9387.7	665.4	6.6%	156	2.98
(in days)	Het	7594.9	6861.4	733.5	9.7%	184	2.80
ICU Time	Avg	10569.8	9870.2	699.6	6.6%	156	2.98
(in days)	Het	7793.6	7040.9	752.7	9.7%	184	2.80

Panel B: Fixing Number of Weekly Working Days

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3931.7	192.9	4.7%	134	2.83
(in hours)	Het	4240.8	3979.1	261.7	6.2%	200	2.83
Post-LOS	Avg	10053.1	9582.9	470.2	4.7%	134	2.83
(in days)	Het	7594.9	6538.8	1056.1	13.9%	196	2.79
ICU Time	Avg	10569.8	10075.4	494.4	4.7%	134	2.83
(in days)	Het	7793.6	6709.9	1083.7	13.9%	196	2.79

Panel C: Fixing Number of Weekly Working Days and Time of All Non-elective Cases

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3980.8	143.8	3.5%	105	2.83
(in hours)	Het	4240.8	4053.6	187.1	4.4%	158	2.83
Post-LOS	Avg	10053.1	9702.6	350.4	3.5%	105	2.83
(in days)	Het	7594.9	6959.3	635.6	8.4%	175	2.81
ICU Time	Avg	10569.8	10201.3	368.4	3.5%	105	2.83
(in days)	Het	7793.6	7141.4	652.2	8.4%	175	2.81

Panel D: Allowing Four Working Days in a Week ($\bar{n}^{(d)} = 4$)

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3437.5	687.1	16.7%	205	3.28
(in hours)	Het	4240.8	3473.0	767.8	18.1%	208	3.28
Post-LOS	Avg	10053.1	8378.5	1674.6	16.7%	205	3.28
(in days)	Het	7594.9	5837.4	1757.5	23.1%	204	3.06
ICU Time	Avg	10569.8	8809.1	1760.7	16.7%	205	3.28
(in days)	Het	7793.6	5990.1	1803.5	23.1%	204	3.06

simultaneously, translating to 251 fewer hours for the OR time, 612 fewer days for post-LOS, and 643 fewer days for total ICU time. The benefits are larger under heterogeneous treatment effects for post-LOS and ICU time. Similar observations hold in Panel B when we assume even smaller workload effects for three- and four-case days. These results support the robustness of our main findings.

Table EC.22 Estimated Benefits of Surgical Scheduling Model with Smaller Workload Effects for Three-case and Four-case Days

Panel A: $\gamma_3 = 80\% \times \gamma_{(est)}$; $\gamma_4 = 80\% \times \gamma_{(est)}$

Objective	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Number of improved weeks
OR time (in hours)	Avg	3960.68	3709.52	251.16	6.3%	184
	Het	4074.92	3752.89	322.03	7.9%	205
Post-LOS (in days)	Avg	9653.62	9041.46	612.17	6.3%	184
	Het	7270.65	6194.02	1076.63	14.8%	203
ICU Time (in days)	Avg	10149.80	9506.16	643.63	6.3%	184
	Het	7460.91	6355.11	1105.80	14.8%	203

Panel B: $\gamma_3 = 66.7\% \times \gamma_{(est)}$; $\gamma_4 = 50\% \times \gamma_{(est)}$

Objective	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Number of improved weeks
OR time (in hours)	Avg	3835.65	3671.85	163.80	4.5%	184
	Het	3949.13	3717.65	231.48	6.2%	205
Post-LOS (in days)	Avg	9348.87	8949.63	399.24	4.5%	184
	Het	7017.06	6140.73	876.33	14.3%	203
ICU Time (in days)	Avg	9829.38	9409.62	419.76	4.5%	184
	Het	7200.68	6299.77	900.91	14.3%	203

EC.7.1. Trade-off Between Objectives Under Heterogenous Effects

In this section, we explore potential trade-off in different objectives in our surgical scheduling model. Under the average effects, the resulting optimal schedule is the same for all the three objectives (OR time, post-LOS, total ICU time). Thus they are improved simultaneously. Under the heterogeneous effects, the surgeon daily workload affects the OR time for both elective and non-elective cases, but the ICU time/post-LOS are impacted only for the non-elective cases. Thus, the optimal schedule for OR time is different from that for post-LOS/ICU time. This introduces a potential trade-off between surgery duration and post-surgery recovery time.

For OR time, the term related to surgeon workload is given by:

$$Obj_1 := \sum_{s \in S} \sum_{t \in T} \tilde{n}_{s,t} \cdot \left(\gamma^{(el:or)} \tilde{n}_{s,t}^{(el)} + \gamma^{(ne:or)} \tilde{n}_{s,t}^{(ne)} \right), \tag{EC.16}$$

where $\gamma^{(el:or)} = 0.567$ and $\gamma^{(ne:or)} = 0.377$ hours denote the estimated effects on OR time for elective and non-electives, respectively. For post-LOS time, the objective is given by

$$Obj_2 := \sum_{s \in S} \sum_{t \in T} \gamma^{(ne:los)} \tilde{n}_{s,t} \cdot \tilde{n}_{s,t}^{(ne)}, \tag{EC.17}$$

where $\gamma^{(ne:los)} = 1.605$ days is the estimated effect for non-electives.

To explore potential trade-offs related to the two measures, we construct the following objective

$$\max\left\{w\times\left(1-\frac{Obj1}{WorkLoad_1}\right)+(1-w)\times\left(1-\frac{Obj2}{WorkLoad_2}\right)\right\}, \tag{EC.18}$$

where $w\in[0,1]$ controls the weight of Obj_1 in the optimization problem; $WorkLoad_1$ and $WorkLoad_2$ denote the total workload related term for the OR time and post-LOS in the original schedule, respectively. Thus, $(1-\frac{Obj1}{WorkLoad_1})$ and $(1-\frac{Obj2}{WorkLoad_2})$ represent the relative reduction in the two measures. The weight w controls how the decision maker balances them in the scheduling model.

We vary the value of w from zero to one with a step length of 0.1. For each w, we solve the new schedule using objective function (EC.18). Then we evaluate the reduction in the total OR time and post-LOS from the optimized schedule. The results are shown in Figure EC.5. In the figure, each point corresponds to an optimized schedule (Note that different values of w in (EC.18) may lead to a same optimized schedule). The x-axis shows the reduction in OR time (in hours), and the y-axis shows the reduction in post-LOS (in days). We see that when the hospital fully focuses on post-LOS (with $w=0$ in (EC.18)), the resulting schedule actually leads to an increase in total OR time, as shown by the leftmost point in the figure. This is because the model would schedule multiple elective cases on a single day. According to the estimated heterogeneous effects, this does not hurt the total post-LOS, but can increase the total OR time. However, there is a range of weights w such that the hospital can achieve substantial reductions in both measures (as well as total ICU times). For example, with $w=0.6$, the new schedule is expected to reduce total OR time by 316 hours and total post-LOS by 1122 days. This shows the ability of our scheduling model to achieve pareto improvements under heterogeneous effects.

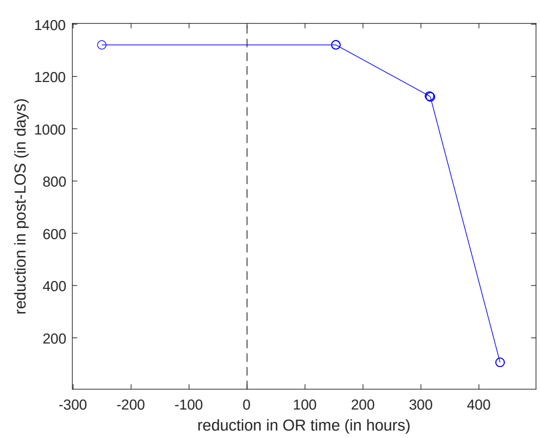


Figure EC.5 Trade-off between post-LOS and OR time under heterogeneous effects scheduling model

EC.7.2. Heuristic Scheduling Policy

In this section, we further develop a heuristic model to show the robustness of our main insights. In the heuristic policy, we still reschedule the cases in each week. As in our main set-up, we allow elective cases to be scheduled within the week, and urgent cases to be moved to adjacent weekdays. The dates of emergent/salvage cases are fixed. Each surgeon works for three days in a week (or more as in the original schedule) if he or she performs at least three cases; this happens for 80% of time in original schedule. The cases are scheduled using the following simple heuristic rules. For each surgeon, we first smooth the number of non-elective cases across days, then we further schedule the elective cases to smooth the total number of cases (both elective and urgent) across days. No complex optimization model (e.g., the MIQP model) is needed for this.

As an additional check, we further assume the time of emergent/salvage cases are not known in advance at the beginning of each week. In this case, the cases are scheduled based only on the elective and urgent patients' information following the above heuristic rule. Then, the times of emergent and salvage cases are realized as in the original data. In line with the data, we assume an emergent/salvage case has a 65% probability to be assigned to a surgeon that does not work on the same day, and a 35% probability to be assigned to a surgeon that already works on the same day. (In the data, 65% of emergent/salvage cases are performed as the only case of their focal surgeons on the day). As we take a retrospective approach, we compare the new and the original schedules for each week and take the one that leads to a lower workload measure. We run simulations for 100 trials to evaluate the average performance of the heuristic schedule.

The results for the heuristic scheduling policy are reported in Table EC.23. Panel A reports the effects under the baseline setting when we assume the surgery dates are known in advance. Panel B further reports the effects with uncertain timing for emergent/salvage cases, as described in above. We see that the simple heuristic scheduling model still leads to economically meaningful improvements. The magnitudes of improvements are comparable to our main MIQP model, although they are smaller in some cases. This holds even we assume uncertain surgery dates for the emergent/salvage cases. For example, under the average estimated effects, the heuristic scheduling model leads to a 345 hour reduction in total OR time and 841 day reduction in total post-LOS. The numbers change to 265 hours and 646 days under uncertain timing of emergent/salvage cases. The results from the heuristic scheduling model further support the robustness of our main managerial insights — smoothing surgeon workload across days can improve the surgical outcomes.

As explained in our main text, the hospital usually has some information for the timing of the cases classified as "urgent". This can be seen by the high proportion (46%) of urgent cases in the sample. We also check this by the time series of the number of urgent cases in each week in our sample, denoted by $NumUrgent_t$ for week t. It has a mean of 12.01 cases and a standard deviation of 5.2 cases. We run a regression for $NumUrgent_t$ as follows:

$$NumUrgent_t = \beta_0 + \beta_1 NumUrgent_{t-1} + Month_t + Year_t + \varepsilon_t,$$

Table EC.23 Estimated Effects from Heuristic Scheduling Policy

Panel A: With Surgery Time Known in Advance

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3797.8	344.9	8.4%	185	3.04
(in hours)	Het	4240.8	3911.0	329.7	7.8%	185	3.04
Post-LOS	Avg	10053.1	9256.6	840.7	8.4%	185	3.04
(in days)	Het	7594.9	6558.0	1036.8	13.7%	185	3.04
ICU Time	Avg	10569.8	9732.4	883.9	8.4%	185	3.04
(in days)	Het	7793.6	6729.6	1064.0	13.7%	185	3.04

Panel B: Uncertain Time for Emergent/Salvage Cases

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3859.5	265.1	6.4%	150	2.94
(in hours)	Het	4240.8	4003.7	237.1	5.6%	150	2.94
Post-LOS	Avg	10053.1	9407.0	646.1	6.4%	150	2.94
(in days)	Het	7594.9	6807.4	787.5	10.4%	150	2.94
ICU Time	Avg	10569.8	9890.5	679.3	6.4%	150	2.94
(in days)	Het	7793.6	6985.5	808.1	10.4%	150	2.94

Panel C: Uncertain Time for Emergent/Salvage Cases and 20% Urgent Cases

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3919.2	205.4	5.0%	131	2.87
(in hours)	Het	4240.8	4041.7	199.1	4.7%	131	2.87
Post-LOS	Avg	10053.1	9552.4	500.6	5.0%	131	2.87
(in days)	Het	7594.9	7069.3	525.6	6.9%	132	2.87
ICU Time	Avg	10569.8	10043.4	526.4	5.0%	131	2.87
(in days)	Het	7793.6	7254.3	539.3	6.9%	132	2.87

Panel D: Uncertain Time for Emergent/Salvage Cases and 40% Urgent Cases

	Effect	Obj orig	Obj new	ΔObj	Rel. ΔObj	Improved week	Avg work day
OR time	Avg	4124.6	3986.3	138.3	3.4%	105	2.81
(in hours)	Het	4240.8	4094.8	146.0	3.4%	105	2.81
Post-LOS	Avg	10053.1	9716.1	337.0	3.4%	105	2.81
(in days)	Het	7594.9	7306.1	288.8	3.8%	107	2.81
ICU Time	Avg	10569.8	10215.5	354.3	3.4%	105	2.81
(in days)	Het	7793.6	7497.2	296.4	3.8%	107	2.81

where $NumUrgent_{t-1}$ is the number of urgent cases in the previous week; $Month_t$ and $Year_t$ denote the month and year of week t. The R-squared from this simple regression is 50.3%, which is relatively high. It suggests that there is much predictability for the number of urgent cases in each week, even with very limited information as included in the regression. In practice, the hospital is likely to have much more information (e.g., surgeon's schedules and patient's admission status) about the cases to be performed in next week.

We perform an additional robustness check as follows. In addition to the emergent and salvage cases, we now randomly choose 20% or 40% of the urgent cases in each week and assume their timing is uncertain and, thus, fixed for the purposes of our scheduling algorithm. We treat these urgent cases as emergent or salvage ones in our implementation of heuristic policy, which is described above. In this test, we have in total about 20% or 30% of cases with uncertain timing ($46.5\% \times 20\% + 10.5\% = 19.8\%$ and $46.5\% \times 40\% + 10.5\% = 29.4\%$). In some ways, this is conservative because even for these cases, we may have some flexibility to move them forward or backward by a day once the elective schedule has been set and as the week evolves.

The estimated effects from the new robustness check are reported in Panels C and D of Table EC.23. The benefits from the heuristic policy indeed decreases after we assume uncertain timing for some urgent cases. However, we still find meaningful reduction from the new schedule. Under the average effects, the new schedule with uncertain timing for 20% of urgent cases can reduce the workload-related term by 5.0% for the OR time, post-LOS, and ICU time simultaneously. Under heterogeneous effects, the new schedule achieves a relative reduction of 6.9% for post-LOS and ICU time. These results further support the robustness of our main managerial insights.

References

Guo, Shenyang, Mark W Fraser. 2014. *Propensity score analysis: Statistical methods and applications*, vol. 11. SAGE publications.

Zhang, Danwei, Dachuan Gu, Chenfei Rao, Heng Zhang, Xiaoting Su, Sipeng Chen, Hanping Ma, Yan Zhao, Wei Feng, Hansong Sun, et al. 2023. Outcome differences between surgeons performing first and subsequent coronary artery bypass grafting procedures in a day: a retrospective comparative cohort study. *BMJ Quality & Safety* **32**(4) 192–201.