JAMA Surgery | Original Investigation

Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources
A Randomized Clinical Trial

Christopher T. Strömblad, MS; Ryan G. Baxter-King, BA; Amirhossein Meisami, PhD; Shok-Jean Yee, MA, RN, CNOR; Marcia R. Levine, MSN, RN, NE-BC; Aaron Ostrovsky, MPH; Daniel Stein, MD, PhD; Alexia Iasonos, PhD; Martin R. Weiser, MD; Julio Garcia-Aguilar, MD; Nadeem R. Abu-Rustum, MD; Roger S. Wilson, MD

Invited Commentary page 322

Supplemental content

IMPORTANCE Accurate surgical scheduling affects patients, clinical staff, and use of physical resources. Although numerous retrospective analyses have suggested a potential for improvement, the real-world outcome of implementing a machine learning model to predict surgical case duration appears not to have been studied.

OBJECTIVES To assess accuracy and real-world outcome from implementation of a machine learning model that predicts surgical case duration.

DESIGN, SETTING, AND PARTICIPANTS This randomized clinical trial was conducted on 2 surgical campuses of a cancer specialty center. Patients undergoing colorectal and gynecology surgery at Memorial Sloan Kettering Cancer Center who were scheduled more than 1 day before surgery between April 7, 2018, and June 25, 2018, were included. The randomization process included 29 strata (11 gynecological surgeons at 2 campuses and 7 colorectal surgeons at a single campus) to ensure equal chance of selection for each surgeon and each campus. Patients undergoing more than 1 surgery during the study's timeframe were enrolled only once. Data analyses took place from July 2018 to November 2018.

INTERVENTIONS Cases were assigned to machine learning–assisted surgical predictions 1 day before surgery and compared with a control group.

MAIN OUTCOMES AND MEASURES The primary outcome measure was accurate prediction of the duration of each scheduled surgery, measured by (arithmetic) mean (SD) error and mean absolute error. Effects on patients and systems were measured by start time delay of following cases, the time between cases, and the time patients spent in presurgical area.

RESULTS A total of 683 patients were included (mean [SD] age, 55.8 [13.8] years; 566 women [82.9%]); 72 were excluded. Of the 683 patients included, those assigned to the machine learning algorithm had significantly lower mean (SD) absolute error (control group, 59.3 [72] minutes; intervention group, 49.5 [66] minutes; difference, −9.8 minutes; $P = .03$) compared with the control group. Mean start-time delay for following cases (patient wait time in a presurgical area), dropped significantly: 62.4 minutes (from 70.2 minutes to 7.8 minutes) and 16.7 minutes (from 36.9 minutes to 20.2 minutes) for patients receiving colorectal and gynecology surgery, respectively. The overall mean (SD) reduction of wait time was 33.1 minutes per patient (from 49.4 minutes to 16.3 minutes per patient). Improved accuracy did not adversely inflate time between cases (surgeon wait time). There was marginal improvement (1.5 minutes, from a mean of 70.6 to 69.1 minutes) in time between the end of cases and start of to-follow cases using the predictive model, compared with the control group. Patients spent a mean of 25.2 fewer minutes in the facility before surgery (173.3 minutes vs 148.1 minutes), indicating a potential benefit vis-à-vis available resources for other patients before and after surgery.

CONCLUSIONS AND RELEVANCE Implementing machine learning–generated predictions for surgical case durations may improve case duration accuracy, presurgical resource use, and patient wait time, without increasing surgeon wait time between cases.

TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT03471377

JAMA Surg. 2021;156(4):315-321. doi:10.1001/jamasurg.2020.6361
Published online January 27, 2021.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Roger S. Wilson, MD, Department of Surgery, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065 (wilsonr@mskcc.org).

Surgery is the most common treatment for solid-tumor cancers. More than half of all new patients treated at our institution, Memorial Sloan Kettering Cancer Center (MSKCC), undergo surgical procedures as frontline therapy. In 2019, we were slated to perform more than 27 000 surgeries across 14 surgical services and 3 operating room (OR) campuses.

Planning the OR schedule for a mean of more than 100 cases per day is an endeavor filled with uncertainty. The 4 major areas of uncertainty include on-time starts of the first case; turnover time between cases; cases booked close to the day of surgery; and planned vs actual surgical case duration. Many institutions have undertaken efforts to improve aspects of planning, such as the on-time start of the first case, turnover time, and management of add-on and urgent/emergency cases.[1-7] In studies focusing on efforts to reduce surgical case duration and turnover, Dexter et al,[8] Strum et al,[9] and Stepaniak et al[7] emphasized that improving the reliable time estimate of surgical cases leads to improved efficiency of OR processes. Their published literature[6,8-16] examining accuracy in the estimation of case duration has been limited to retrospective data modeling.

To improve accuracy in predicting surgical case duration, we developed a set of predictive models that leveraged the recorded information (more than 300 data points per case) and knowledge we have regarding the patient, surgeon, and procedure. Relevant patient, procedure, surgeon, and systems data were collected and used to train a machine learning model for 2 surgical services. This was tested in a retrospective analysis and set to run automatically every morning, 1 weekday prior to each patient's surgery, to make predictions of case duration available for application by a scheduler. Typically, patients are given their surgery time 1 day prior to surgery, enabling them to benefit from the predictions generated. Herein we examine whether this predictive model can be implemented and is more accurate than the current process.

Key Points

Question Can the implementation of a predictive model improve accuracy in predicting the duration of scheduled surgical procedures, and how does this potential improvement affect patient wait time, time between cases, and use of presurgical resources?

Findings In this randomized clinical trial of 683 surgical patients, the implementation of a machine learning model significantly improved case duration accuracy, reduced patient wait time, effected no change in time between cases (ie, turnover time or surgeon wait time), and reduced the use of presurgical resources.

Meaning Daily use of machine learning to predict surgical case duration may be warranted.

Methods

This study was reviewed and approved by the MSKCC Institutional Review Board/Privacy Board. The research involves no more than minimal risk to the participants because researchers were only interacting with their protected health information for predicting the case duration and scheduling the OR case. The waiver or alteration did not adversely affect the rights and welfare of the research participants, and therefore they were not notified of these changes and no informed consent procedure was used.

Experimental Design and Sample

This was a single-center, 2-campus, randomized clinical trial performed with 2 surgical services (gynecology and colorectal) at MSKCC. The study compared accuracy in predicting surgical case duration assigned to a machine learning–assisted model, with a control group of cases assigned to the current scheduling-flow system. There are 14 surgical services at our institution; however, study implementation had a manual component, requiring us to limit the number of services to 2. The Gynecology and Colorectal Services in the Department of Surgery were selected because of dissimilarity in the nature of procedures and patients between these 2 services and strong leadership support from the respective service chiefs. In 2017, there were 2416 cases completed and a mean (SD) case duration of 187 (124) minutes for the Gynecology Service, and 1539 cases completed and a mean (SD) case duration of 244 (170) minutes for the Colorectal Service. Study participants were patients in the Gynecology and Colorectal Services undergoing surgery between April 7, 2018, and June 25, 2018, whose cases were scheduled more than 1 weekday before surgery. The size of the date range was designed to ensure the sample size derived from the power analysis, after applying exclusion criteria (trial protocol in Supplement 2).

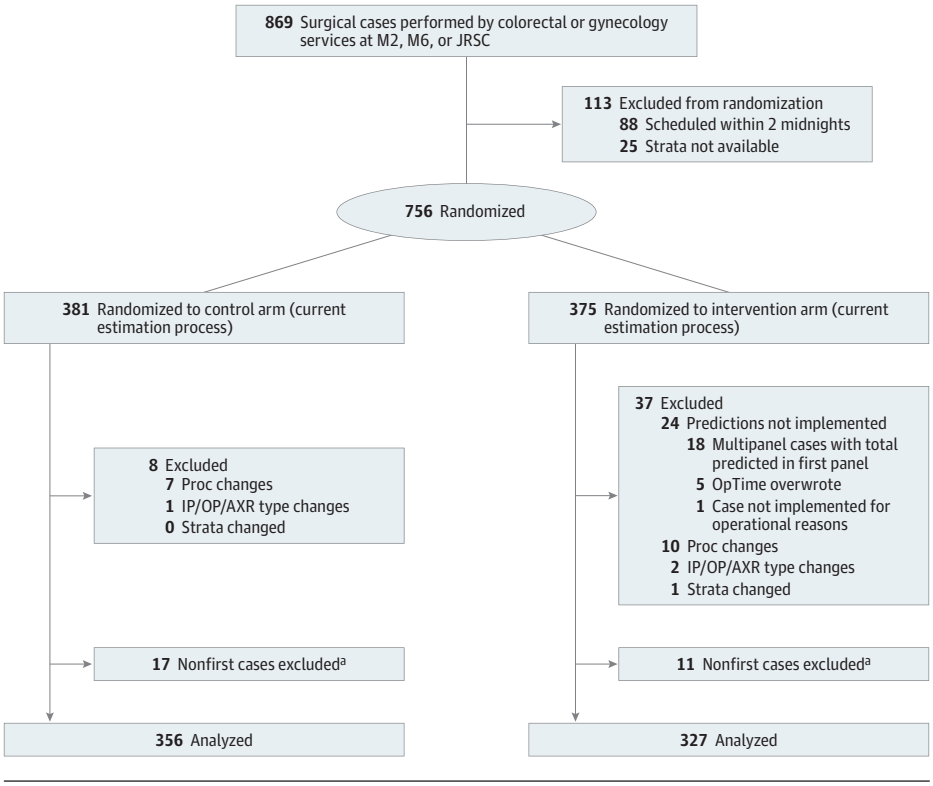
The trial was designed to randomize 50% of included cases to the intervention arm, in which case durations were estimated by a machine learning model. The other 50% were randomized to the control arm, in which case durations were estimated using the current process. The current process is the electronic health record (EHR) standard method for deriving case duration estimates supplemented with the surgeon or scheduler's estimate (eMethods in Supplement 1).

Cases were stratified by surgeon and OR location (30 strata). Randomizations were accomplished by method of random permuted block, pregenerated by the Department of Biostatistics, sequentially numbered, and stored on an undisclosed drive that no surgeons had access to or knew about. The randomization was assigned to cases the day before surgery, ordered by the earliest patient's planned toes-in time (the moment the patient enters the operating room). Cases were excluded if the combination of surgeon and location did not appear in the pregenerated randomizations.

Predicting Case Durations With Machine Learning

The machine learning model predicts OR case durations, measured as patient-in-room to patient-out-of-room time, using data from the institution's EHR. Through discussions with surgeons, anesthesiologists, nurses, and schedulers (including S.-J.Y., M.R.L., M.R.W., J.G.A., N.R.A.-R., and R.S.W.), we identified more than 300 data features that we were able to derive from the EHR, including patient characteristics (such as age,

Figure. Exclusion Flowchart



Surgeries performed by the Colorectal and Gynecology Services at the 2 platforms included in the study (between April 7, 2018, and June 25, 2018). EHR indicates electronic health record; IP/OP/AXR, inpatient/outpatient/patient in ambulatory extended recovery; JRSC, Josie Robertson Surgical Center; M2, second floor surgical rooms on the main campus; M6, sixth floor surgical rooms on the main campus; Proc, procedure.

[a] Control and intervention cases were pooled for the purposes of determining patients' first case in the study period.

sex, race, body mass index, comorbidities), surgeon-associated statistics (surgeon's experience performing specific procedures), procedure groups (eg, using primary procedures, summarized using relative value units), operational factors (such as location and day of the week), and additional factors derived from clinical notes in the patient's medical record.

Implementation

Every morning at 8:30 AM, the model automatically gathered data for the following day's cases from an internal data warehouse, generating predictions for the intervention arm. The predictions were published at a dashboard designed for the OR scheduling office to open and apply. During the study period, every weekday between 8:30 AM and 11:00 AM, the scheduler manually implemented the predictions into the live scheduling system. On Friday mornings, the scheduler would apply the predictions to Saturday, Sunday, and Monday cases.

Power Analysis

A retrospective analysis of the predictive model showed an overall mean absolute error (MAE) (SD) of 50 (56) minutes for the machine learning predictions and 67 (73) minutes for the current state estimates. Historical data showed that the Gynecology Service performed 1.69 times as many cases as the Colorectal Service, with an 18% lower MAE. The sample size was calculated at 568 cases to ensure a 90% power with an α of .10 for a 2-sided test of means for 2 samples or arms (eDiscussion in Supplement 1 and trial protocol in Supplement 2).

The statistical programming language R version 3.5 (R Foundation for Statistical Computing) was used for the data analysis. The data analyses were performed July 2018 to November 2018.

Results

Participants

Between April 7, 2018, and June 25, 2018, 869 surgeries were performed by the Colorectal and Gynecology Services at the 2 platforms included in the study (**Figure**). Of these, 113 cases did not receive randomizations because their strata were not expected (eg, the case was performed by a new surgeon) or the case was scheduled after randomizations were assigned (ie, 11 AM on the morning before the surgery). Of the remaining 756 cases, a set of predetermined exclusion criteria removed 45 cases with substantial input changes and implementation issues, as well as cases already enrolled in the study.

Because the control arm followed the preexisting scheduling process, implementation issues occurred only in the intervention arm. This, along with the fluctuating nature of a randomized study, resulted in 327 of 683 cases (47.9%) belonging to the intervention arm vs 356 of 683 cases (52.1%) in the control arm. Patient characteristics in the 683 included cases did not differ substantially between the 2 groups (**Table 1**). These included 566 women (82.9%), 467 patients with gynecology cases (68.4%), with a mean (SD) age of 55.8 (13.8) years.

Table 1. Study Participant Characteristics

Variable	No. (%)	
	Control (n = 356)	Intervention (n = 327)
Age, median (IQR), y	55 (46-65)	58 (45-67)
Race		
Asian	23 (6.5)	29 (8.9)
Black	21 (5.9)	33 (10)
Other	10 (2.8)	18 (5.5)
Unknown	13 (3.7)	13 (4.0)
White	289 (81)	234 (72)
Sex		
Female	293 (82)	273 (83)
Male	63 (18)	54 (17)
Service		
Colorectal Service	119 (33)	97 (30)
Gynecology Service	237 (77)	230 (70)

Abbreviation: IQR, interquartile range.

Accuracy Outcomes

By implementing a machine learning–prediction model into the scheduling workflow, we demonstrated significantly reduced schedule duration errors for both surgical services (**Table 2**).[17] To ensure reduction in overall error, we designed the study around MAE. We found that MAE (SD) was 59.3 (72) minutes in the control group and 49.5 (66) minutes in the intervention group. This 16.5% improvement was statistically significant (P = .03). The operational value of such an improvement is more evident when one observes how it reduced the number of large error predictions (cases with MAE >60 minutes), in that those are the cases that often disrupt a day in the OR. These metrics were 120 of 356 patients (33.7%) and 83 of 327 patients (25.4%) in the control group and intervention group, respectively (**Table 3**), indicating a reduction of large errors by 8.3% of cases. The secondary outcome metrics also demonstrated substantial improvement in case duration accuracy. The control group had a mean error (ME) (SD) of −33.7 (87) minutes, indicating that error distribution in the control group was centered about a half hour less than the ideal ME of 0 minutes and tended to underpredict case duration. The intervention group had a ME (SD) of −3.6 (83) minutes, indicating that this was still underpredicted slightly but with an error distribution centered closer to 0. With respect to eliminating the tendency to overpredict or underpredict, this represents an improvement of 89% (P < .001).

The results from the Colorectal Service were substantially stronger than the overall findings in MAE (control group, mean [SD], 74.0 [78] minutes; intervention group, mean [SD], 55.7 [87] minutes; difference, 24.8%), ME (control group, mean [SD], −51.7 [94]; intervention group, mean [SD], 5.7 [103]; difference, 89.0%), and percentage of cases scheduled accurately within 60 minutes (72 of 119 patients [60.5%] vs 72 of 97 patients [74.2%]; difference, 13.7%). Although improvements for the Gynecology Service were not as significant on a case-by-case basis (MAE [SD], 51.9 [68] minutes vs 46.9 [55] minutes; difference, 9.6%; ME [SD], −24.6 [82] minutes vs −7.5 [72] minutes; 69.4%; cases scheduled accurately within 60 minutes, 164 of 237 patients [69.2%] vs 172 of 230 patients [74.8%]; difference, 5.6%), it is important to note that the Gynecology Service (mean [SD], 7.1 [3.9] cases) performed a greater number of cases per day than the Colorectal Service (mean [SD], 3.3 [2.1] cases) and these were generally of shorter duration (mean [SD] duration, Colorectal Service: 255.3 [165.0] minutes; Gynecology Service, 190.1 [133.2] minutes), which could lead to a more significant improvement if a mean was calculated over days. However, as the study did not stratify by days, this measure is beyond its current scope. Overall, there was also a reduction in the SD of the error of 8.3% (66 vs 72 cases), although this was not the case for SD in the Colorectal Service. Further investigation revealed that a single case in the Colorectal Service had a duration of 19 hours; no other case had a duration greater than 14 hours. This extremely long case was in the intervention group for the Colorectal Service and initially predicted to take 9 hours and 41 minutes, yielding a large prediction error. For that same case, the current process predicted a duration of only 6 hours. Had we considered this case an outlier and removed it from the study, the SD for the Colorectal Service would have shown an improvement in the intervention arm vs the control arm.

Operational Outcomes

We observed effects on patients in reduced mean wait time for starting surgery as scheduled (from 49.4 minutes to 16.3 minutes; 67.1%; Table 3). By reducing schedule duration error, the time patients spent in the presurgical area was reduced. (The accuracy of an individual case duration estimate can affect the start time of the following patient or the time a surgeon waits for the next patient to be transported into the OR.)

We measured patient wait time as the difference between planned and actual start time for the following patient. To isolate the effect on patients most directly affected by improved case duration accuracy, we measured only the wait time for the patient in the study group who immediately followed the first case in the room (denoted as the *to-follow patient*). By measuring this outcome on to-follow patients, we observed a mean reduction of 33.1 minutes in patient wait time (from 49.4 minutes to 16.3 minutes). Specifically, this involved a mean 16.7-minute wait time reduction (from 36.9 minutes to 20.2 minutes) for the Gynecology Service and a 62.4-minute reduction (from 70.2 minutes to 7.8 minutes) for the Colorectal Service. Finally, we measured the difference in the time from to-follow patients' registration on the day of their surgery to the time they entered the OR, as a way to measure preoperative resource usage, using our predictive model, to-follow patients had a mean reduction of 25.2 minutes (from 173.3 minutes to 148.1 minutes) in presurgical length of stay (Table 3). We reported significantly reduced error and improved wait time and resource usage metrics, without disruption or a negative effect on surgeons' wait time, given the marginal improvement of 1.5 minutes in turnover time (from 70.6 minutes to 69.1 minutes).

Table 2. Accuracy Outcomes[a]

Arm	Control (n = 356), min	Intervention (n = 327), min	Improvement in mean, %	*P* value
Mean absolute error (SD)[b]				
Overall	59.3 (72)	49.5 (66)	16.5	.03
Colorectal Service	74.0 (78)	55.7 (87)	24.8	.004
Gynecology Service	51.9 (68)	46.9 (55)	9.6	.64
Mean error (SD)[b]				
Overall	−33.7 (87)	−3.6 (83)	89.3	<.001
Colorectal Service	−51.7 (94)	5.7 (103)	89.0	<.001
Gynecology Service	−24.6 (82)	−7.5 (72)	69.4	.01

[a] Sample sizes for the service-specific results are listed in Table 1.

[b] In the entire sample, there was 1 case with a duration greater than 1000 minutes: a colorectal service case randomized into the intervention arm. If this case had been considered an outlier and removed, the SD for colorectal service in the intervention arm would have been reduced from 87 to 70 for the mean absolute error SD and 103 to 86 for the mean error SD.[17]

Table 3. Operational Outcomes[a]

Metric	Control (n = 356)	Intervention (n = 327)	Improvement, %
Accurate within 60 min, No./total No. (%)			
Overall	236/356 (66.3)	244/327 (74.6)	8.3
Colorectal Service	72/119 (60.5)	72/97 (74.2)	13.7
Gynecology Service	164/237 (69.2)	172/230 (74.8)	5.6
Patient wait time, mean (SD), min[b]			
Overall	49.4 (70.6)	16.3 (74.6)	67.1
Colorectal Service	70.2 (72.5)	7.8 (82.3)	88.9
Gynecology Service	36.9 (66.8)	20.2 (70.8)	45.3
Turnover time, mean (SD), min[b]			
Overall	70.6 (35.3)	69.1 (42.1)	2.0
Colorectal Service	74.4 (36.6)	74.4 (32.6)	−0.1
Gynecology Service	68.3 (66.7)	66.7 (45.8)	2.4
Patient time in facility (until toes-in time), mean (SD), min[c]			
Overall	173.3 (78.6)	148.1 (62.3)	14.5
Colorectal Service	177.1 (75.4)	146.0 (52.2)	17.6
Gynecology Service	171.0 (80.8)	149.0 (66.6)	12.9

[a] Sample sizes for service-specific results are listed in Table 1.

[b] Sample sizes for patient wait time and turnover time: 91 first cases in the Colorectal Service with a to-follow case (47.3% in the intervention arm); 172 first cases in the Gynecology Service with a to-follow case (53.5% in the intervention arm). Turnover time did not exclude durations greater than a set threshold, nor did we exclude instances in which the following case in a room was performed by a different surgeon, as is common for turnover metrics, making the mean time between cases appear shorter.

[c] Five to-follow cases had a missing time stamp for patient time in facility: 2 in the intervention arm and 2 in the control arm in the Colorectal Service and 1 in the intervention arm in the Gynecology Service.

Discussion

General Findings

Most published literature that we were able to identify to date, examining case duration accuracy, are limited to retrospective data modeling and theoretical improvements.[6,8-16] We found 1 recently published study[18] that performed implementation of a statistical regression model, showing the ability to better predict when the surgical day would end. This study most notably showed improvements in throughput and staff burnout scores. The study was limited to including 4 surgeons from the same service performing exclusively vascular surgical interventions, and their control group was the historical mean duration; however, Zhou et al[19] had already concluded that relying solely on historical surgery times may not be efficient. In contract, this study includes 18 surgeons and 2 campuses, excluding no procedure types within those services. Furthermore, we have chosen a control group representing the institution's current best effort to predict duration—which meant a combination of historical median and surgeon or scheduler estimates—and did not limit it to using only the historical mean. We found no reports in the literature to date describing implementation of a predictive model to more than 1 surgical service and excluding no procedure types, nor any reports providing evidence of the outcome of such a model on patient wait time, surgeon wait time, or preoperative resources.

By implementing a machine learning–prediction model into the scheduling workflow, we identified significant reduction in error (including the primary MAE; $P = .03$) with respect to surgical schedule duration estimates for both services. Although study and sample size were designed to determine whether the effect on the MAE was significant for the entire study population, we also reviewed the results of each of the 2 services to assess the reduced error. Because this was a randomized clinical trial and the predictive model was implemented into daily operations, we were able to measure and observe the effect on operational metrics, yielding significantly reduced error, improved patient wait time, and improved resource usage metrics.

The current process tends to severely underpredict case durations; consequently, there are long patient wait times for the OR but less surgeon wait time or underuse of the OR. With implementation of the predictive model, we were able to not only minimize the underpredicted skew of the estimates' error

distribution but also reduce the absolute prediction error; the result was a substantial reduction in patient wait time, without an increase in surgeon wait time. In direct correlation to decreased patient wait time, we measured a reduction in use of presurgical resources, including beds and staff. Some of the presurgical beds at MSKCC's operating suites serve as so-called swing beds that may also serve as Post Anesthesia Care Unit beds. Thus, the reduction in presurgical beds at certain times of the day may facilitate additional capacity in the Post Anesthesia Care Unit. This suggests that the blockage in flow of surgical patients into this unit could consequently be reduced. To achieve successful implementation, we focused on 3 elements not found in the previous literature, ensuring (1) data availability prior to surgery, (2) avoiding procedure combinations as a limiting factor, and (3) operational effects. Each of these are discussed in the trial protocol in Supplement 2.

Next Steps and Future Opportunity

We have shown significant improvement in case duration accuracy in a controlled setting using our predictive model. We now plan to implement the model for the full day and a new set of surgical services and study the planned end time of each OR procedure compared with the actual end time. This measure might reduce unplanned overtime, providing the opportunity to add cases on appropriate days and resulting in greater staff satisfaction due to fewer last-minute requests to stay longer at the end of the day.

We also plan to work with services that commonly perform multipanel cases. At MSKCC, this includes the Breast Service and the Plastics and Reconstructive Service. Furthermore, there may be value in predicting more detailed time intervals than toes-in time to toes-out time (the moment the patient exits the operating room); for example, toes-in time to the start of the first procedure depends on the difficulty of anesthesia induction and may have other correlated factors than procedure duration (the start of the first procedure to the end of the last procedure). A new version of the model should be able to predict each of these intervals and combine them appropriately to arrive at a deeper level of granularity and a higher level of accuracy.

Next steps also include testing the ability of the model to predict more than 1 day in advance with less available information (ie, less input to the predictive model). This may be challenging, because fewer inputs will be available weeks in advance of surgery. There will likely be a trade-off between information availability and the accuracy of early predictions: the earlier we predict, the less reliable information we may have, and prediction performance may be reduced. Such a model, combined with an implementation workflow that integrates seamlessly into the EHR, is the major work we are pursuing. Finally, we expect this approach to be applicable and relevant for nonsurgical services that use procedure rooms, such as Interventional Radiology, and endoscopic services, such as Gastroenterology and Pulmonology.

Limitations

Although the predictive model will provide a prediction for any case scheduled more than 2 midnights before the day of surgery, it may not provide accurate predictions if the submitted procedure codes deviate significantly from the procedures that are performed. Ensuring proper booking of procedures will improve the model's performance, although there may be times when it is impossible to know exactly which procedure will be performed, because that depends on what is discovered in the OR once the procedure has begun and/or wet laboratory work with unexpected results.

The smaller sample size in the intervention arm could potentially lead to more variability in the intervention sample. While this should not bias interpretations of results, a more even distribution of cases between study arms would provide more robustness to the results.

A less common occurrence were multipanel cases in which multiple surgeons from different services operated on the same patient during the same case. We were not able to conclusively demonstrate improved accuracy results for this subgroup. In services where this is more common, it might be useful to add a set of features around the second surgeon or around common surgeon pairs (surgeons who have worked together in the past) vs less common pairs (surgeons with a limited history of working together).

The model showed improvements in each service. However, when measuring the effect on a typical case, the strongest improvements were in the Colorectal Service. Although the improvements for the typical gynecology case were not as substantial, it is important to note that the Gynecology Service generally performed more cases per day because their cases were generally shorter in duration. A more significant improvement could be demonstrated in this higher volume of cases if the means of the outcome metrics were calculated over entire days. However, because there was no stratification by days, this measure is beyond the scope of the current study.

The prediction was designed to be available 1 day prior to surgery and excludes cases that are scheduled after the predictions were made available. Although this was sufficient for most of the operational gains we report, there will be additional value in providing even earlier predictions and having predictions ready for cases scheduled within 24 hours of surgery. It would make for a more seamless implementation if predictions were available the moment a surgical case was first booked. The challenge is to develop a seamless integration of the predictive model into the EHR without requiring schedulers to change their workflow to incorporate the prediction.

For the past 30 years, the primary obstacle to predictive modeling has been implementation. While work can be done to improve the model, we overcame this obstacle, demonstrating the model's feasibility and its positive effect on real-world practice.

Conclusions

In this randomized clinical trial of 683 surgical patients (after exclusion criteria were applied), the implementation of a machine learning model significantly improved accuracy in predicting case duration and led to reduced patient wait time, no difference in time between cases (ie, turnover time or sur-

geon wait time), and reduced presurgical length of stay. By extension, this decreased the unnecessary use of presurgical resources. While we did not survey all the faculty included in the study, we did correspond before and after the study with both service chiefs and received no complaints or reports of disruption to the surgical day. We demonstrate an outcome on operational metrics that could not have been rigorously tested in a retrospective analysis.

In summary, scheduling of real-world surgical cases using duration estimates from a machine learning model demonstrated that the framework is generalizable beyond specific procedures or services. These findings have led us to pursue systemwide implementation of this model in which all of our institution's 27 000 annual surgical cases will leverage the proposed predictive modeling framework.

ARTICLE INFORMATION

Accepted for Publication: October 25, 2020.

Published Online: January 27, 2021. doi:10.1001/jamasurg.2020.6361

Author Affiliations: Department of Strategy and Innovation, Memorial Sloan Kettering Cancer Center, New York, New York (Strömblad, Baxter-King); Adobe Inc, San Jose, California (Meisami); Department of Nursing, Memorial Sloan Kettering Cancer Center, New York, New York (Yee, Levine); Health Informatics, Memorial Sloan Kettering Cancer Center, New York, New York (Ostrovsky, Stein); Epidemiology-Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York (Iasonos); Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York (Weiser, Garcia-Aguilar, Abu-Rustum, Wilson); Weill Medical College of Cornell University, New York, New York (Weiser, Garcia-Aguilar, Abu-Rustum, Wilson).

Author Contributions: Mr Strömblad had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Strömblad, Meisami, Levine, Ostrovsky, Stein, Abu-Rustum, Wilson.
Acquisition, analysis, or interpretation of data: Strömblad, Baxter-King, Meisami, Yee, Iasonos, Weiser, Garcia-Aguilar, Abu-Rustum, Wilson.
Drafting of the manuscript: Strömblad, Baxter-King, Meisami, Levine, Ostrovsky, Abu-Rustum, Wilson.
Critical revision of the manuscript for important intellectual content: Meisami, Yee, Stein, Iasonos, Weiser, Garcia-Aguilar, Abu-Rustum, Wilson.
Statistical analysis: Strömblad, Baxter-King, Meisami, Iasonos.
Administrative, technical, or material support: Baxter-King, Meisami, Yee, Levine, Ostrovsky, Stein, Garcia-Aguilar, Abu-Rustum, Wilson.
Supervision: Yee, Weiser, Wilson.

Conflict of Interest Disclosures: Dr Iasonos reported personal fees from Mylan outside the submitted work. Dr Garcia-Aguilar reported personal fees from Intuitive Inc, Medtronic, and Johnson & Johnson outside the submitted work. Dr Abu-Rustum reported grants from Stryker and Grail during the conduct of the study. No other disclosures were reported.

Funding/Support: This research was supported in part through the National Institutes of Health/National Cancer Institute (Cancer Center support grant P30 CA008748).

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 3.

REFERENCES

1. Cen dan JC, Good M. Interdisciplinary work flow assessment and redesign decreases operating room turnover time and allows for additional caseload. *Arch Surg*. 2006;141(1):65-69. doi:10.1001/archsurg.141.1.65

2. CIA RR, Brown MJ, Heal JR, et al; Surgical Process Improvement Team, Mayo Clinic, Rochester. Use of lean and six sigma methodology to improve operating room efficiency in a high-volume tertiary-care academic medical center. *J Am Coll Surg*. 2011;213(1):83-92. doi:10.1016/j.jamcollsurg.2011.02.009

3. Nissan J, Campos V, Delgado H, Mazaedial C, Spector S. The automated operating room: a team approach to patient safety and communication. *JAMA Surg*. 2014;149(11):1209-1210. doi:10.1001/jamasurg.2014.1825

4. Scala TM, Carco D, Reece M, Fouche YL, Pollak AN, Nagarkatti SS. Effect of a novel financial incentive program on operating room efficiency. *JAMA Surg*. 2014;149(9):920-924. doi:10.1001/jamasurg.2014.1233

5. Smith CD, Spackman T, Brommer K, et al. Re-engineering the operating room using variability methodology to improve health care value. *J Am Coll Surg*. 2013;216(4):559-568. doi:10.1016/j.jamcollsurg.2012.12.046

6. Stepaniak PS, Heij C, Mannaerts GHH, de Quelerij M, de Vries G. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesth Analg*. 2009;109(4):1232-1245. doi:10.1213/ANE.0b013e3181b5de07

7. Stepaniak PS, Vrijland WW, de Quelerij M, de Vries G, Heij C. Working with a fixed operating room team on consecutive similar cases and the effect on case duration and turnover time. *Arch Surg*. 2010;145(12):1165-1170. doi:10.1001/archsurg.2010.255

8. Dexter F, Dexter EU, Masursky D, Nussmeier NA. Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesth Analg*. 2008;106(4):1232-1241. doi:10.1213/ane.0b013e318164f0d5

9. Strum DP, May JH, Vargas LG. Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *Anesthesiology*. 2000;92(4):1160-1167. doi:10.1097/00000542-200004000-00035

10. Bartek MA, Saxena RC, Solomon S, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. *J Am Coll Surg*. 2019;229(4):346-354.e3. doi:10.1016/j.jamcollsurg.2019.05.029

11. Dravenstott RR, Reich E, Strongwater S, Devapriya P. B1-3: improving surgical case duration accuracy with advanced predictive modeling. *CMR Clin Med Res*, 2014. 12(1-2):93. doi:10.3121/cmr.2014.1250.b1-3

12. Eijkemans MJ, van Houdenhoven M, Nguyen T, Boersma E, Steyerberg EW, Kazemier G. Predicting the unpredictable: a new prediction model for operating room times using individual characteristics and the surgeon's estimate. *Anesthesiology*. 2010;112(1):41-49. doi:10.1097/ALN.0b013e3181c294c2

13. Hosseini N, Sir MY, Jankowski CJ, Pasupathy KS. Surgical duration estimation via data mining and predictive modeling: a case study. *AMIA Annu Symp Proc*. 2015;2015:640-648.

14. Kayis E, Wang H, Patel M, et al. Improving prediction of surgery duration using operational and temporal factors. *AMIA Annu Symp Proc*. 2012;2012:456-462.

15. Master N, Zhou Z, Miller D, Scheinker D, Bambos N, Glynn P. Improving predictions of pediatric surgical durations with supervised learning. *Int J Data Sci Anal*. 2017;4(1):35-52. doi:10.1007/s41060-017-0055-0

16. Strum DP, May JH, Sampson AR, Vargas LG, Spangler WE. Estimating times of surgeries with two component procedures: comparison of the lognormal and normal models. *Anesthesiology*. 2003;98(1):232-240. doi:10.1097/00000542-200301000-00035

17. Dexter F, Abouleish AE, Epstein RH, Whitten CW, Lubarsky DA. Use of operating room information system data to predict the impact of reducing turnover times on staffing costs. *Anesth Analg*. 2003;97(4):1119-1126. doi:10.1213/01.ANE.0000082520.68800.79

18. Kougias P, Tiwari V, Sharath SE, et al. A statistical model-driven surgical case scheduling system improves multiple measures of operative suite efficiency: findings from a single-center, randomized controlled trial. *Ann Surg*. 2019;270(6):1000-1004. doi:10.1097/SLA.0000000000002763

19. Zhou J, Dexter F, Macario A, Lubarsky DA. Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. *J Clin Anesth*. 1999;11(7):601-605. doi:10.1016/S0952-8180(99)00110-5