Supplemental Online Content

Strömblad CT, Baxter-King RG, Meisami A, et al. Effect of a predictive model on planned surgical duration accuracy, patient wait time, and use of presurgical resources: a randomized clinical trial. *JAMA Surg*. Published online January 27, 2021. doi:10.1001/jamasurg.2020.6361

This supplemental material has been provided by the authors to give readers additional information about their work.

eMethods.

Current process

The current process relies on a combination of surgeon and scheduler estimates, as well as the Epic software procedure time averaging calculations. The Epic process currently represents the institution's best effort to provide accurate estimates and uses the median of similar cases, identified with an inflexible decision rule based on exact procedure matches, scheduled surgeons, and scheduled surgical platforms. This algorithm only works well when it identifies previous cases that match the CPT code combination of the current case; when this is not so, it provides unreliable estimates or simply provides a duration of zero, forcing the [human] scheduler to determine duration. Thus, this process relies on oversight from the scheduler, who may overwrite the duration (at times, at the request of the surgeon). In 2017 the Epic estimates were overwritten nearly 50% of the time. The error in case duration estimates was far larger if the human input was ignored–perhaps due to the heterogeneity of case mix, i.e. the vast number of unique CPT code combinations present at a cancer center. In this study, we chose to compare the predictive accuracy of the machine learning model with the current process, a stricter criterion that reflects the current state.

Modeling

The notes were streamlined by: 1) removing stop-words such as "the", "is", "are", etc.; 2) words were converted into binary (presence or absence) variables; 3) words were categorized into unigrams, bigrams and trigrams; 4) words were run through a LASSO regression so that only words with predictive value across many cases were included. We then trained a model using these features and the Random Forest algorithm [1] as implemented in the R package "ranger" [2]. We chose the Random Forest model because, in this instance it outperformed other Machine Learning models we tested, including gradient boosting machine and linear regression. The training process used an 80/20 train/test split as well as a 10-fold cross validation, and minimized Root Mean Square Error; this helped reduce the risk of overfitting a model to a sample set that only represents part of the population and extends the model's ability to work in a real-world setting.

We performed testing to ensure that the majority of each input was available prior to surgery and excluded inputs that were often missing. For example, the ASA score was often not available until the day of surgery; thus, we excluded this potentially valuable variable from the model. If there were only a few missing inputs, we automatically imputed values using simple assumptions and averaging. For example, if a categorical variable such as race was missing, we would calculate the Mode of this field from that service to impute, and if a continuous variable such as BMI was missing, we would impute it using the average value for that surgical service.

Outcome measures

The primary outcome measure was case duration accuracy, primarily measured as the MAE in the scheduled case duration. Error was defined as planned case duration minus the actual case duration, as recorded in the Electronic Health Record. Cases performed in less time than planned were considered "over-predicted" and noted to have positive scheduling error. This impacts operations differently from "under-predicting" (i.e. when cases take longer than planned), which has a negative scheduling error. By averaging the magnitude of the individual errors, MAE measures the inaccuracy of an average case and prevents individual errors from "canceling out." By contrast, the mean error (ME)–a secondary accuracy metric–measures whether an average case is performed in less or more time than expected.

As the primary outcome metric, MAE was used for the power calculation that determined the sample size of the trial. The nonparametric Mann-Whitney U test of two means was chosen to test the differences in MAE, due to the non-normal and strictly positive distribution of the MAE.

The study protocol also provided for subgroup analyses to assess the efficacy of the predictive model across the different surgical services and locations included in the study. However, the sample size derived from the power analysis was designed only to detect statistical significance for the combined sample. The protocol also included secondary accuracy metrics, including the ME and percent of cases scheduled accurately within 60 minutes. The two-sided t-test of means for two samples was used to test for differences in ME after the conclusion of the study.

Finally, the study assessed the impact of accurate prediction of case duration on two important operational metrics: patient wait time (the difference between planned and actual “toes in” times); and surgeon wait time (the amount of time that surgeons wait between cases). These secondary analyses were also recorded in the IRB protocol.

eReferences 1.

1. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

2. Wright MN, Zeigler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J Statist Software*. 2017;77(1):1-17.

eDiscussion.

General findings

There are many examples of health centers utilizing retrospective datasets including data, statistics, and machine learning to improve accuracy in predicting surgical case duration. The early literature includes the 1996 study by Dexter [1] in which the focus was to assess the upper boundaries of surgical case duration for a limited set of procedures, in order to ascertain the possibility of adding more cases to the schedule without extending the day significantly. Since then, several studies have focused on fitting statistical models, including log-normal distribution models, multiple linear regression models and machine learning models [2-10]. Each of these studies utilized retrospective data to test the validity and performance of their predictive modeling approach. They showed that accuracy may be improved using larger datasets leveraging the Electronic Health Record, surgeon estimate as an input, and different modeling approaches. However, we found no examples of studies that actually implemented case duration predictions into daily operations prior to the day of surgery. Dexter [11] described implementation of a model providing the remaining expected duration of an ongoing surgical case in the OR when cases are delayed. Our study provides the full case duration estimate prior to the day of surgery, to assist with planning for the entire day.

Data availability prior to surgery

We forced an added level of rigor to ensure the availability of the input data prior to the day of surgery. Most inputs are dependent on health systems' nightly database extract/transfer/load processes, and some inputs are not recorded digitally until the day of surgery. Notably, the ASA score was used in several earlier studies; we also found that this input was a strong predictor in our early models. However, this data is rarely available digitally before the day of surgery. Consequently, a model that includes this datapoint would not provide a prediction prior to surgery without performance loss, compared with results derived from a retrospective dataset. Health systems wishing to leverage this data point must ensure that the ASA score is consistently available digitally, in advance of surgery.

Procedure combinations

Published case duration prediction studies are generally limited to a subset of procedures within a single surgical service or exclude CPT code combinations that occur less frequently. For example, Stepaniak et al. [2] excluded procedures that occurred <10 times in their dataset. They estimated this exclusion to be minimal, but it still comprised 14% of the total population. In the current study, we did not omit any procedure combination groups. Although it may be challenging to develop accurate predictions for every possible CPT code combination, we deem it necessary to successful implementation of a predictive model into the scheduling workflow of a busy operating suite.

Hosseini et al.[7] used a classification algorithm to cluster over 2,000 procedures into 49 categories. Instead of using surgical procedures in the model directly, we aggregated the RVU for each case as model inputs. We also explored other ways to group procedures, including identifying similar words in the procedure code descriptions. It might also be possible to develop unsupervised models that cluster procedure code combinations of similar duration and variability. While these approaches hold promise for future model iterations, our RVU-based approach enabled us to develop a model for each service that included all possible procedure code combinations.

Operational impact

As we implemented the model in a controlled setting, we were able to measure operational outcome metrics that have not been published before. We were able to demonstrate the benefits of such a model for patient, staff, and perioperative surgical resources (pre- and post-surgical areas). As the control group tended to under-predict case duration, and the intervention group limited over- or under-predictions, we were able to reduce patient wait time by 33 minutes on average. This was achieved without increasing time between cases (turnover time) that might otherwise increase the surgeon's wait time. The improvement we demonstrated is not a simple tradeoff of "who is waiting for whom", but rather a net improvement.

It is important to note that we designed the study to allow our surgeons to overwrite scheduled durations at their discretion. However, we received no such demands and the surgeons involved in the study did not report any

disruptions in their surgical days due to implementation of the study. Similarly, the anesthesiologists and surgical staff did not report any complaints related to implementation of the study.

In addition to improvement in patient wait time and turnover time between cases on the day of surgery, operational impact included greater efficiency in the use of pre-surgical resources. Reduced patient wait time in the pre-surgical area correlates strongly with a reduced need for pre-surgical beds. We also showed a reduction in the use of pre-surgical beds, which could lead to reduced wait time for registered patients waiting for those beds. More importantly, a proportion of pre-surgical beds are also used as PACU beds; therefore, any reduction in use of pre-surgical beds could improve the flow of patients from OR into PACU after surgery.

For the sake of simplicity, we decided to weigh (i.e. penalize) under-predictions and over-predictions evenly and optimize the model to produce errors centered around zero. The ideal is to minimize both under- and over-predictions. However, one could imagine a scenario in which an organization had more than enough pre-surgical rooms and staff and might be inclined to sacrifice a marginal amount of patient wait time in order to minimize the risk of an unoccupied OR. Conversely, another organization with an especially high focus on reducing the time patients spend in the system, or an organization with limited preoperative resources, might choose to reduce patient wait time, minimizing patients' time in the facility at the cost of reduced OR utilization. If there is deviation from the safe choice of centering the error around zero, it would be necessary to quantify the cost of patient wait time compared with the cost of time between cases (i.e. turnover time or surgeon wait time).

eReferences 2.

1. Dexter F. Application of prediction levels to OR scheduling. *AORN J.* 1996;63(3):607-615

2. Stepaniak PS, Heij C, Mannaerts GH, 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

3. 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

4. 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

5. 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

6. 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

7. Hosseini N, Sir MY, Jankowski CJ, Pasupathy KS. Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study. *AMIA Annu Symp Proc.* 2015:640-648

8. Kayis E, Wang H, Patel M, et al. Improving prediction of surgery duration using operational and temporal factors. *AMIA Annu Symp Proc*. 2012:456-46

9. 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

10. 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

11. Dexter F, Epstein RH, Lee JD, Ledolter J. Automatic updating of times remaining in surgical cases using bayesian analysis of historical case duration data and "instant messaging" updates from anesthesia providers. *Anesth Analg*. 2009;108(3):929-940