Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

MACHINE LEARNING ENHANCED DISCRETE EVENT SIMULATION OF A SURGICAL CENTER

Chen He[1], Marianne Sarazin[3], François Dufour[3], Marie Paule Bourbon[3], and Xiaolan Xie[2]

[1]College of Artificial Intelligence, Southwest University, Chongqing, CHINA
[2]Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Saint-Etienne, FRANCE
[3]Clinique Mutualiste Saint-Etienne, AESIO Group, Saint-Etienne, FRANCE

ABSTRACT

In discrete event simulation (DES) of surgical centers, activity durations are inherently stochastic and are typically modeled with standard probability distributions. In reality, however, actual procedure times may not conform to standard distributions, depending instead on the individualized characteristics of patients. Accounting for such covariate effects in DES is challenging, making personalized care process simulation impractical. This shortcoming can compromise model fidelity and constrain the utility of DES for precise service management. Machine learning (ML) techniques offer a promising solution by incorporating patient-specific features into duration estimates. In this study, we introduce an ML-driven simulation framework that integrates advanced predictive algorithms with a DES model of a surgical center. We benchmark our approach against conventional DES methodologies that rely solely on predefined probability distributions to represent time variability. Results show that our approach can significantly reduce the randomness of service durations and lead to a significant improvement in simulation accuracy.

1 INTRODUCTION

The staffing and utilization of surgical centers incur substantial costs, motivating stakeholders to optimize the use of critical resources such as operating rooms (ORs), surgeons, and post-anesthesia care unit (PACU) facilities (Khaniyev et al. 2020) (Van Tunen et al. 2020). A prerequisite for effective resource optimization is the accurate estimation of resource-occupancy durations; however, this task is complicated by the inherent uncertainty in case-time durations (Varmazyar et al. 2020) (Pandit 2020). Moreover, operational decision-making—such as same-day online scheduling—often relies on subjective estimates of activity durations, which may yield suboptimal outcomes. For example, managers may estimate activity times based on a limited number of observations when allocating on-call staff (Khaniyev et al. 2020) (Dexter et al. 2004), scheduling add-on cases (Zhou and Dexter 1998), or determining the need for additional ORs (Dexter 2000). Accurate predictions of case-time durations can enable efficient case scheduling (Varmazyar et al. 2020), optimized resource allocation (Pandit 2020), and improved patient flow (Zheng et al. 2022), whereas inaccurate estimates may increase surgery costs (Childers and Maggard-Gibbons 2018), prolong working hours (Strachota et al. 2003), exacerbate staffing fatigue and medical errors (Strachota et al. 2003) (West et al. 2009), and extend both staff turnover and patient waiting times (Strachota et al. 2003) (Denton et al. 2007). These challenges underscore the need for advanced predictive methodologies in estimating case-time durations.

Traditional DES models predominantly employ statistical inference to characterize activity durations, often overlooking variability introduced by patient-specific factors (Forbus and Berleant 2022). For instance, Zheng et al. (2022) fitted parametric distributions to all activity durations within their DES framework and validated goodness-of-fit via the Kolmogorov-Smirnov test. Likewise, Dexter and Ledolter (2005) developed a Bayesian statistical method that combines surgeons' subjective estimates with historical data to forecast OR occupancy duration. Although effective in certain contexts, these conventional approaches

frequently neglect critical determinants of duration variability. In reality, surgery duration is influenced by patient attributes such as age, procedure type, severity level, and comorbidities, yet these covariates are often omitted from standard statistical models, thereby constraining estimation accuracy.

Machine learning offers a promising alternative by integrating patient-specific factors to uncover latent patterns and enhance prediction precision. Although prior studies have employed ML to forecast OR length of stay (LOS) for preoperative planning, several limitations remain. First, simulation of surgical centers requires the appropriate level of granularity and nonstandardized duration estimates, such as OR occupancy time and post-anesthesia care bed occupancy times, rather than the common practice that estimates anesthesia and surgery durations separately (Bodenstedt et al. 2019) (Jiao et al. 2022) (Strömblad et al. 2021) (Zhao et al. 2019), thereby reducing the applicability of previous ML models for predictive simulation. Second, many investigations focus on narrow subsets of cases (e.g., specific medical departments (Devi et al. 2012) (Fang et al. 2023), reducing generalizability. Third, current ML approaches predominantly rely on structured inputs (categorical and numerical variables) (Elfanagely et al. 2021), despite the high heterogeneity of data across institutions and practitioners. For example, some electronic health records (EHRs) may encode procedure names as single categories, multiple categories, or even free text. Moreover, the categories themselves may be inconsistent (e.g., "Ligature des artères" versus "Ligature des artères hémorroïdaires avec guidage doppler, avec mucopexie, par voie anale" to denote identical procedures). As such, robust ML approaches that can be generalized must be able to semantically represent a common version of heterogeneous data structures, including free-text elements.

In response to these challenges, we propose a novel framework for predicting surgical case durations that leverages both structured variables and unstructured text. We describe the development and evaluation of ML models trained on heterogeneous data sources, demonstrate their integration into DES for precise simulation, and illustrate their utility in guiding day-of-surgery operational decisions. Our approach represents a significant advancement toward ML-driven simulation models for comprehensive surgical process optimization.

2 SIMULATION MODEL

2.1 Process Description

The surgical procedure consists of three standard phases: pre-operative, intraoperative, and post-operative phases. From a process improvement standpoint, it comprises a sequence of discrete events. At the Saint-Etienne Mutualist Clinic, the general surgical workflow entails six key stages (illustrated in Fig. 1, where labels above each box denote the unit in which each event occurs): (i) registration and check-in: upon arrival at the Ambulatory Reception Unit (ARU) or Inpatient Reception Unit (IRU), a registration nurse verifies the patient's identity, confirms demographic and clinical information, and completes all requisite documentation; (ii) pre-operative preparation: the patient is escorted to the pre-operative holding unit (PHU), where a nurse conducts pre-operative preparation, including physical assessment, surgery preparation (i.e., intravenous drips, blood pressure tests, ECG tests, etc.), and anesthesia preparation (if required), etc.; (iii) anesthesia: the patient is then transferred to designated downstream units to receive anesthesia induction. Locoregional anesthesia (LRA) is performed in some dedicated beds located within the PACU, while other types of anesthesia are carried out in the OR pre-assigned to the patient; (iv) surgery: The patient undergoes the planned surgical procedure in the OR; (v) post-anesthesia care: depending on patient categories/pathways, patients may be discharged directly, transferred to the PHU for minimal post-anesthesia care, or sent to PACU for a standard recovery. Patients recover in the OR if PACU/PHU is not available; (vi) departure: Patients who require further post-operative monitoring are transferred to outpatient wards, inpatient wards, or the intensive care unit (ICU) through IRU or ARU based on their clinical needs. Patients not requiring further monitoring exit the surgical center directly through the ARU or IRU.

Surgical patients can be categorized as either outpatients (same-day admission, surgery, and discharge) or inpatients (requiring at least one overnight stay). Further classification of patients is based on admission/discharge date/mode, surgery type, hospitalization duration, and post-anesthesia care needs.

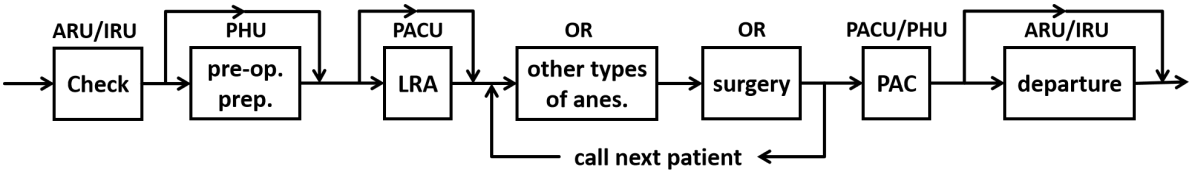

Figure 1: Workflow of the surgical center.

As illustrated in Table 1, the surgical center involves patients from five distinct clinical pathways. Specifically, external patients involve outpatients who undergo minor procedures (e.g., endoscopy, biopsy, colonoscopy, etc.) without pre-operative preparation and post-anesthesia care. Short-stay ambulatory patients are outpatients who undergo minor procedures requiring minimal post-anesthesia care in the PHU and leave the surgical center within a few hours. Long-stay ambulatory patients are outpatients who need extended post-operative monitoring in outpatient wards but are discharged the same day. D0 patients are inpatients who have their admission, surgery, and discharge on the same day. D-n patients are inpatients who are hospitalized $n, n \in \mathbb{N}^+$ days before the surgery for necessary pre-operative preparations, typically for complex procedures or fragile patients. In summary, external, short-stay ambulatory, and long-stay ambulatory patients are classified as outpatients, whereas D0 and D-n patients are considered inpatients.

Table 1: Patient flow at the surgical center.

pathway	arrival	pre-oprative prep.	anesthesia	surgery	post-anesthesia care	departure
External	ARU	-	OR/PACU	OR	-	ARU
Short ambulatory	ARU	PHU	OR/PACU	OR	PHU	ARU
Long ambulatory	ARU	PHU	OR/PACU	OR	PACU	IRU
D0	ARU	PHU	OR/PACU	OR	PACU	IRU
D-n	IRU	-	OR/PACU	OR	PACU	IRU

2.2 Assumptions

The following assumptions are introduced to define the workflow, patient flow, resources, and management policies based on the the current setup of the surgical center and analysis of the collected data: (i) A surgical center in a hospital comprises ORs, PAC beds, and other resources; (ii) Only the processes from patient arrival at the ARU/IRU to discharge from the surgical center are incorporated into the model; (iii) The simulation model incorporates only weekday cases, while weekend cases are excluded due to differing operational configurations (e.g., reduced human resources) and policies (e.g., protocols for managing emergency cases); (iv) Only elective surgeries are considered, with non-elective cases treated as elective. Each patient has a scheduled arrival time, a scheduled surgery start time, a designated surgeon, and an allocated OR; (v) Case rescheduling ("surgery jumps") and cancellations are not considered; (vi) The patient transition times between medical units are short and depend on their movement speed; (vii) Once a patient following a long ambulatory, D0 or D-n pathway completes the post-anesthesia care, he/she will be transferred to inpatient or outpatient wards by a stretcher-bear, which are limited resources; (viii) Each OR accommodates only one patient at a time, with each patient pre-assigned to a specific OR; (ix) Anesthetists administer LRA in the PACU and other anesthesia modalities in the OR on a first-call, first-served basis.

Remark 1: Each surgical team typically comprises a surgeon, an anesthetist, an anesthetist nurse, a circulating nurse, an instrument nurse, and a bandage nurse. Note that the anesthetist oversees multiple ORs, and the PAC nurse is shared among all ORs.

Remark 2: For analytical convenience, the model is initially developed based on probability distributions. In Section 3.2, the discussion is extended to include the application of machine learning prediction algorithms.

2.3 Data and Simulation Model Configuration

A dataset was collected from the surgical center of the Saint-Etienne Mutualist Clinic, part of Group Aesio. It comprises 1,742 surgical cases across 13 medical specialties (e.g., urology, ophthalmology, orthopedics, vascular surgery, etc.) recorded between January 2, 2024, and January 31, 2024. A standardized pre-processing protocol was implemented, involving the following steps: (i) the removal of cases with at least 50% missing values; (ii) correction of incoherent timestamps when sufficient information was available (refer to Chapter V of the thesis (Rifi 2023)); (iii) exclusion of cases with unresolved inconsistencies; (iv) missing start or end timestamps for activities were imputed using the average duration of the respective activity. After these steps, a refined dataset containing 1,664 cases was obtained.

Based on the above assumptions and the dataset formulated, a discrete event simulation model comprising 16 ORs and 24 PAC beds, specifically, 19 beds in the PACU and 5 beds in the PHU, was developed using AnyLogic. The simulation encompasses 22 weekdays over a one-month period, with daily operations simulated from 7:30 AM to 8:00 PM. Patient arrivals are generated sequentially according to a predetermined schedule derived from historical data, thereby naturally incorporating daily variability into each day's schedule. A first-come, first-served policy is applied to all shared resources. Notably, patients are permitted to enter the surgical center before the completion of the preceding surgery. The entry into ORs is contingent upon the completion of OR cleaning from the prior procedure, which subsequently triggers the entry of the next patient. Using the Python `distfit` library, best-fit statistical distributions are selected to model case-time durations.

The registration and check-in process is conducted in the ARU for patients following pathways other than D-n, while D-n patients are processed in the IRU. The IRU comprises a single reception desk staffed by one nurse responsible for patient reception and departure. Given that D-n patients have completed their preoperative preparations in their inpatient wards, they experience no queuing or delays in the IRU, facilitating rapid entry into the surgical center. Conversely, the ARU features two reception windows and is staffed by four nurses, each assigned to a specific pathway (i.e., external, short-stay ambulatory, long-stay ambulatory, and D0). The registration and check-in process in the ARU is modeled using a triangular distribution with a minimum of 3 minutes, a mode of 5 minutes, and a maximum of 8 minutes. Patients processed through the ARU are required to change into surgical gowns before entering the surgical center, utilizing one of four available dressing rooms. The dressing process follows an empirical distribution with a mean duration of 9.615 minutes and a standard deviation (SD) of 4.718 minutes. Subsequently, patients place their belongings into one of 36 locker compartments, a process modeled by a triangular distribution with a minimum of 1 minute, a mode of 1.5 minutes, and a maximum of 2 minutes. It is assumed that locker capacity is sufficient, with scalability as needed, thereby preventing any bottlenecks in luggage placement.

Preoperative preparations, such as intravenous drips and ECG tests, are conducted in the PHU exclusively for ambulatory and D0 patients. The PHU is equipped with 9 rest sofas, 10 beds, and 6 nurses (4 of whom are shared with the ARU) to facilitate these preparations. The duration of preoperative preparation is modeled as an empirical distribution with a mean of 31.137 minutes and a SD of 22.829 minutes. Additionally, the LRA is performed solely in the PACU, which includes 5 dedicated beds for this procedure. The LRA duration follows an empirical distribution with a mean of 20.976 minutes and a SD of 9.21 minutes.

The OR process encompasses several stages: OR entry, initiation of anesthesia, incision, closure, OR exit, and OR cleaning. Based on available data, the length of stay in ORs (i.e., OR LOS) is modeled as an empirical distribution with a mean of 71.645 minutes and a SD of 45.143 minutes. The post-anesthesia care is provided in either the PACU or PHU, depending on patient pathways (refer to Table 1). The PACU contains 19 beds dedicated to the PAC, while the PHU has 5 such beds. The duration of PAC in the PACU is modeled as an empirical distribution with a mean of 100.193 minutes and a SD of 32.785 minutes, whereas in the PHU, it follows an empirical distribution with a mean of 39.035 minutes and a SD of 12.412 minutes. Following PAC in the PACU, patients are transferred to their respective wards via the IRU, contingent upon the availability of stretcher bearers. In the current setting, 6 stretcher bearers are available to meet the demands of PAC nurses. Conversely, after PAC in the PHU, patients retrieve their

belongings and change clothes, processes modeled by a triangular distribution with a minimum of 1 minute, a mode of 1.5 minutes, and a maximum of 2 minutes, and an empirical distribution with a mean of 9.615 minutes and a SD of 4.817 minutes, respectively. All estimated distributions have been validated using the Kolmogorov–Smirnov test to ensure the goodness-of-fit.

2.4 Validation of the Simulation Model

The simulation model has been validated through a comparison with 1,664 data samples in the dataset. Two primary metrics, including the OR LOS and OR utilization, were evaluated. Ten simulation replications are conducted, and the average values of metrics are reported. As shown in Table 2, the mean OR LOS produced by the simulation closely approximates that observed in the collected data, with only minor differences. Comparisons are conducted for OR utilization for each OR. As illustrated in Figure 2, the utilization rates obtained from the simulation replications closely align with those derived from the collected data. OR utilization is calculated as the ratio of the total OR usage time for all surgeries completed before the end of the day (numerator) to the theoretical total available time (denominator), where the theoretical availability is 12.5 hours per OR per day. Note that ORs 13, 14, and 15 are dedicated to orthopedics, ophthalmology, and the digestive system, respectively, with relatively low utilization rates (i.e., below 30%).

Table 2: Validation of OR LOS measured in hours (N=1,664).

	Actual data	Simulation	Deviation
Mean	1.077	1.194	0.117
SD	0.752	0.644	0.108

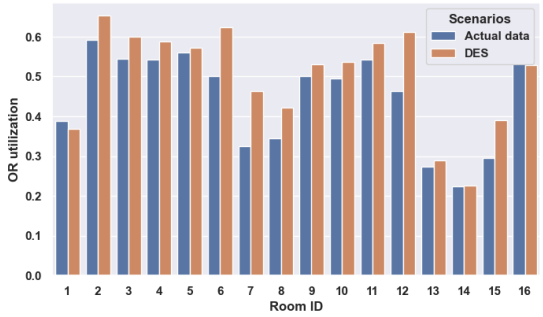

Figure 2: Validation of OR utilization.

The results demonstrate that the simulation model accurately replicates the OR workflow and patient flow, yielding performance outcomes within an acceptable range. Therefore, the developed simulation model can be utilized to predict and evaluate various scenarios, as well as to provide recommendations for operational improvements and decision-making.

3 MACHINE LEARNING–ENHANCED SIMULATION

3.1 Variable Definition and Feature Extraction

To develop ML models for case-time duration prediction, the dataset was chronologically partitioned into training and test subsets so as to preserve its time-series structure. The training set comprised 1,026 cases occurring between 2 January and 19 January (the first three weeks), whereas the test set comprised 638 cases from 22 January to 31 January (the final week and additional days).

The prediction process involves two interdependent clinical services essential to surgical center operation: (i) the intraoperative care in ORs and the post-operative recovery process in ORs, the PACU, or the PHU. The primary target variables for the ML models are the total OR LOS, defined as the time from patient entry into the OR to patient OR exit, rather than skin incision to skin closure, and the PAC LOS, defined as the time difference between the start and end of post-anesthesia care. These durations correspond to the occupancy of key medical resources, i.e., the ORs and PAC beds. Prior research (Van Eijk et al. 2016) shows that surgical case duration typically conforms to a log-normal distribution across most procedure types. Consistent with this, Figure 3(a) illustrates the distribution of surgical case durations within the dataset, which exhibits significant right skewness. Consequently, the surgical case duration is log-transformed (as

shown in Figure 3(b)) and normalized to have a mean of 0 and a standard deviation (SD) of 1 for use in ML regression models. In contrast, the PAC LOS (as shown in Figure 3(c)) is only normalized to have a mean of 0 and an SD of 1.

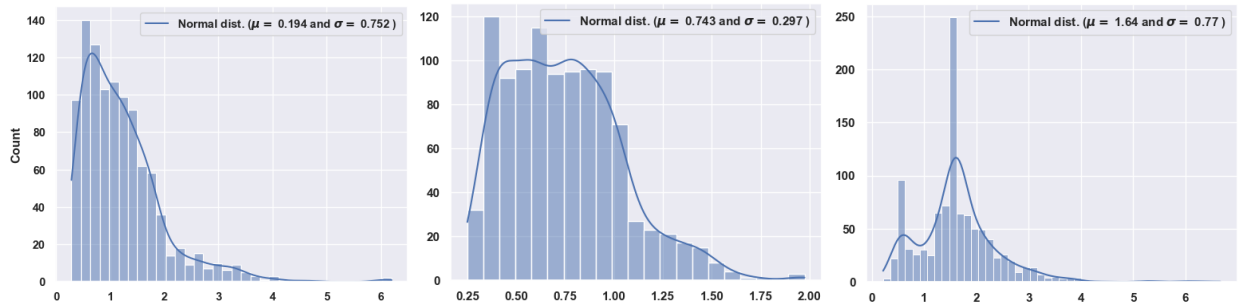

Figure 3: Histogram and density estimates of case time durations (in hours) in the training data: (a) OR LOS (left); (b) Log-transformed OR LOS (middle); (c) PAC LOS (right).

The dataset includes several categorical variables: OR, surgeon, surgeon specialty, primary procedure code, first and second sub-procedure codes (when available), OR nurse, instrument nurse, bandage nurse, anesthesiologist, primary anesthesia type (which comprises six classes, e.g., general anesthesia (GA), local-regional anesthesia (LRA), pure local anesthesia (LA), sedation, GA & LRA, LA & sedation), anesthesia nurse, PAC bed, PAC nurse, robot surgery (a binary indicator of whether the surgery was robot-assisted), clinical pathway (designating five distinct pathways: extern, short ambulatory, long ambulatory, D0, and D-n), day-of-week of surgery (with seven categories: Monday, Tuesday, ..., Sunday), and arriving late (a binary indicator of whether a patient arrives late). For de-identification purposes, the names of the medical staff have been replaced by unique identification codes. Regarding the day of the week, the surgical center operates on a six-day schedule, with only rare emergency cases accepted on Sundays. In January 2024, 1,643 surgeries (98.73%) were performed on weekdays (i.e., Monday through Friday), whereas Saturday experiences a considerably lower volume with only 16 cases (0.96%). On Sundays, the center is exclusively open for unforeseen emergency cases, accounting for merely 5 cases (0.3%). Missing values in these categorical variables have been imputed with a dedicated category "unavailable". Prior to integration into machine learning models, these categorical variables were transformed into binary variables using the one-hot encoding method. Table 3, lists some critical categorical variables included in the dataset and their summary statistics.

Table 3: Statistics of categorical variables in the dataset (N=1,664).

Variable	Unique	Most freq.	Variable	Unique	Most freq.
OR	16	OR16 (251)	Surgeon	43	ID??? (112)
Specialty	7	Orthopedic (452)	Procedure	299	HHQE005 (152)
OR Nurse	61	ID??? (174)	Instrument nurse	50	ID??? (64)
Bandage Nurse	224	ID??? (152)	Anesthesiologist	36	ID??? (115)
Anes. Type	7	GA (866)	Comp. anes. type	27	TAP Bloc (87)
Anes. Nurse	33	ID??? (131)	PAC bed	24	BX06 (126)
PAC Nurse	27	ID??? (162)	Pathway	5	Short Ambu. (518)

The "unique" column specifies the number of distinct categories within the variable. The "freq" column provides the count of the most frequently occurring class.

Regarding numerical variables, the dataset includes age (in years), severity (an ordinal variable with increasing levels 1, 2, 3, and 4), scheduled OR LOS (i.e., the pre-assigned time blocks in the operating room schedule based on anticipated duration, which differ from the actual OR LOS), the start time of the scheduled surgery (in 24-hour format), the end time of the scheduled surgery (in 24-hour format), time of

arrival at the surgical center (in 24-hour format), time of arrival at the OR, and the start of the PAC (in 24-hour format). At our institution, the scheduled OR LOS is primarily based on the surgeons' estimates. It is occasionally adjusted using historical data, although this is not a standardized or consistent process. In rare cases, the scheduled OR LOS is missing, and a "-1" placeholder is entered for emergency cases. The scheduled duration was log-transformed as recommended by (Van Eijk et al. 2016). All timestamps are expressed using the 24-hour convention. Consequently, time variables are converted into numerical values ranging from 1 to 24. For example, 1.5 and 18.25 represent 1:30 AM and 6:15 PM, respectively. Missing values in the numerical variables are encoded as "-1", thereby allowing machine learning models to distinguish between missing data and valid measurements of zero. Prior to incorporation into the predictive models, the numerical variables were normalized using z-score normalization to achieve a mean of 0 and a standard deviation of 1.

Table 4: Summary statistics of the numerical variables in the dataset.

Variable	Traning (N=1,026)					Test (N=638)				
	mean	SD	min	median	max	mean	SD	min	median	max
Age (years)	62.36	15.63	17.13	63.78	99.03	62.34	14.99	20.21	63.26	100.21
Scheduled OR LOS (hours)	1.118	0.726	0.166	1.0	6.75	1.131	0.855	0.333	1.0	6.0

The study also incorporates unstructured data, including the procedure name and the surgeon's note. The procedure name is a short, free-text description of the intended surgery or procedure(s) to be performed. Billing codes (e.g., Homogeneous Groups of Patients (GHM)) were excluded from the analysis, as they are not available at the time of surgery. The surgeon's note comprises a brief free-text description of the operating location and associated precautions. Missing text data was represented with an empty string, i.e., "". Common French stop words (e.g., "le", "de", "ce", etc.) and punctuation are removed from text descriptions. Subsequently, unigrams and bigrams are extracted from these free-text descriptions and encoded using a TF-IDF (Term Frequency–Inverse Document Frequency) sparse vector representation (Havrlant and Kreinovich 2017). Trigrams were excluded as their inclusion yielded no obvious performance improvement.

3.2 Baseline Models and Machine Learning Modeling

We developed two families of predictive models: OR LOS prediction models and PAC LOS prediction models. Model training was performed using a five-fold cross-validation framework, and the final evaluation was conducted on a separate test set. Specifically, the training set was employed within the cross-validation framework to fine-tune model parameters, after which each model was evaluated on the unseen test set. All results presented in this section pertain exclusively to the performance observed on the test set. For baseline comparisons, we constructed simple heuristics that replicate how stakeholders estimate OR LOS and PAC LOS, denoted as $\tilde{t}_{OR}^{Comb}$ and $\tilde{t}_{PAC}^{Comb}$, respectively. Additionally, we considered a widely used Bayesian approach for OR LOS prediction (Dexter and Ledolter 2005), referred to as $\tilde{t}_{OR}^{Bayes}$.

These baseline models leverage historical data to compute average case-time durations as the predicted values, which include (i) **statistical estimation method**, i.e., sampling from the empirical distributions estimated for OR LOS ($\tilde{t}_{OR}^{Stat}$) and PAC LOS ($\tilde{t}_{PAC}^{Stat}$) using training data (refer to Section 2.3); (ii) **surgeon-procedure specific OR LOS** ($\tilde{t}_{OR}^{Comb}$), i.e., the average OR LOS for the last 10 surgical cases (if available) with the same procedure code by a specific surgeon; (iii) **anesthesiologist–anesthesia specific PAC LOS** ($\tilde{t}_{PAC}^{Comb}$), i.e., the average PAC LOS for all cases with the same anesthesia type performed by a specific anesthesiologist; (iv) **Bayesian statistical method** ($\tilde{t}_{OR}^{Bayes}$), i.e., a weighted combination of the surgeon's estimate (i.e., scheduled duration) with historical data to forecast OR LOS. Using the estimated values, we developed baseline models to forecast relevant case-time durations within the test set. The performance of these baseline models serves as a reference, such that any ML approach must yield superior results to be considered effective.

Subsequently, we trained a set of predictive models using a five-fold cross-validation approach. Each method was implemented with the Python `sklearn` library, and hyperparameter tuning was conducted using a random search strategy. The selection of models was guided by a preliminary screening process, during which we initially evaluated various approaches, including linear regression, LASSO regression, and k-nearest neighbors (KNN). ML models that demonstrated inferior performance, measured by MAE, RMSE, R2, and PDw15, compared to at least one of the above-mentioned approaches were excluded. Following this screening, we identified and selected the most promising ML models for further analysis.

3.3 Experimental Results of Machine Learning Predictions

Table 5 compares the results obtained from ML methods and baseline models used to forecast OR LOS and PAC LOS. The first column lists the method names, while the second through fourth columns present the mean absolute error (MAE), root mean squared error (RMSE), and the r^2 value, respectively. The final column, PDw15P, indicates the percentage of forecasted cases for which the absolute deviation from the actual duration is less than 15%. The results demonstrate that all ML methods outperform the baseline models across all four quality metrics. Furthermore, HGBR and GBR exhibit the best performance, with HGBR achieving the lowest MAE. Notably, HGBR is capable of forecasting 51% (70%) of the OR LOS (PAC LOS) with an absolute error not exceeding 15%. According to the literature (Saadouli et al. 2015) and expert opinion in the hospital, a prediction is considered acceptable when the deviation from the actual duration is less than 15%. This figure represents an improvement over the baseline model $\tilde{t}_{OR}^{Stat}$ ($\tilde{t}_{OR}^{Stat}$) by 37.940% (52.664%).

Table 5: Testing results for OR LOS (left) and PAC LOS (right) prediction (measured in minutes).

Method	MAE	RMSE	R^2	PDw15P	Method	MAE	RMSE	R^2	PDw15P
$\tilde{t}_{OR}^{Stat}$	47.306	61.031	-0.836	13.157	$\tilde{t}_{PAC}^{Stat}$	48.883	62.534	-1.222	18.025
$\tilde{t}_{OR}^{Comb}$	21.894	37.683	0.272	26.018	$\tilde{t}_{PAC}^{Comb}$	35.001	50.361	-0.441	25.549
$\tilde{t}_{OR}^{Bayes}$	19.832	33.747	0.133	32.601	MLP	21.028	29.613	0.501	44.201
DT	18.528	29.240	0.578	36.520	DTR	19.296	29.715	0.498	50.783
RF	13.919	22.832	0.742	44.670	SVR	17.983	33.384	0.366	49.843
BR	13.138	22.253	0.755	44.827	RF	15.650	22.156	0.720	53.605
MLP	13.013	21.334	0.775	45.141	BR	14.961	21.503	0.737	55.956
GBR	12.311	20.364	0.795	45.297	GBR	10.733	15.686	0.860	68.808
SVR	11.281	17.788	0.843	46.551	**HGBR**	**10.084**	**15.427**	**0.864**	**70.689**
HGBR	**10.840**	**16.517**	**0.865**	**51.097**					

SVR: Support Vector Regression, DT: Decision Tree regression, RF: Random Forest, BR: Bagging Regression Tree, MLP: Multiple Layer Perceptron, GBR: Gradient Boosting Regression Tree, HGBR: Histogram-based Gradient Boosting Regression Tree.

We present a detailed statistical analysis to determine whether the tested ML models yield statistically different results. In line with the recommendations of Demsar (Demšar 2006), we conducted an analysis aimed at establishing (i) whether the prediction methods produce different forecasts and (ii) whether a ranking among the different methods can be defined. To achieve this, we performed a two-step statistical analysis.

We first applied the Kruskal–Wallis test (Liu and Chen 2012) to evaluate whether differences in mean absolute error (MAE) among the prediction algorithms were statistically significant. This non-parametric test is typically used when comparing more than two independent groups. The null hypothesis of the test posits that the prediction results of these ML models are indistinguishable (i.e., similar MAE values are provided), and thus, any observed ranking arises by chance. Table 6 summarizes the results of the Kruskal–Wallis tests for both OR LOS and PAC LOS on the MAE of different algorithms. It reports the number of runs per method (N), degrees of freedom (df), the χ^2 statistic, the significance level, and the

associated p-value. At a significance level of $\alpha = 1 \times 10^{-4}$, the null hypothesis is rejected, indicating that MAE differs significantly across methods.

Table 6: Results of the Kruskal-Wallis test.

target	N	df	χ^2	α	p-value
OR LOS	100	9	564.605	1e-4	<1e-110
PAC LOS	100	8	627.746	1e-4	<1e-120

Table 7: The setup for Nemenyi tests.

target	k	N	α	q_α	CD
OR LOS	9	100	1e-4	6.721	0.910
PAC LOS	8	100	1e-4	6.660	0.859

Second, given that the null hypothesis of the Kruskal–Wallis test was rejected, a post-hoc analysis was performed using the Nemenyi test (Liu and Chen 2012) to compare each method with the others, thereby identifying which methods differ significantly in their rankings. Two methods are considered significantly different if their average ranks differ by at least the critical difference (CD), computed as $CD = q_\alpha \sqrt{k(k+1)/6N}$, where k is the number of configurations used, N the number of runs per algorithm, and q_α the critical value obtained from (Demšar 2006). Table 7 details the Nemenyi test configuration and the resulting CD. Methods whose mean-rank differences exceed the CD are deemed to perform significantly differently.

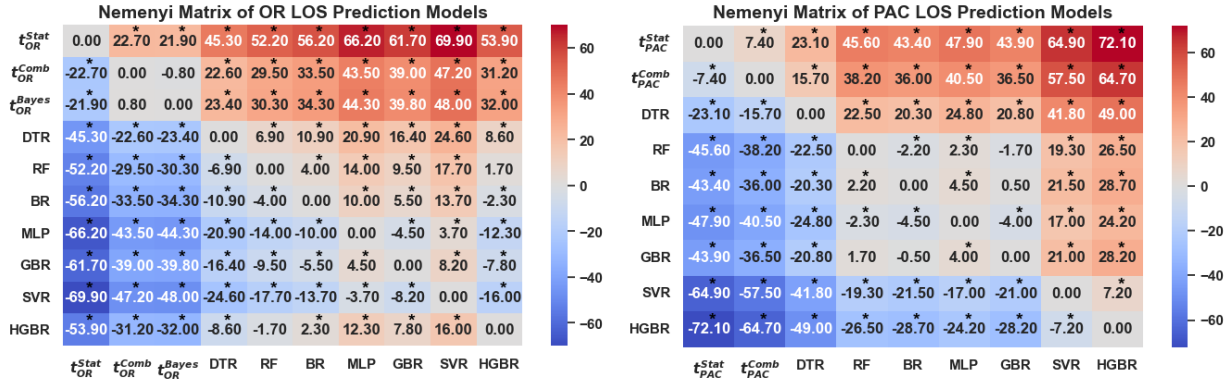

Figure 4: Nemenyi post-hoc comparison matrixs.

Figure 4 shows the Nemenyi matrix of significant results for different prediction methods. Cells marked with * indicate that the difference between the two groups (methods) exceeds the critical value. The number in cells represents the difference in rank mean between the row group and the column group. The results show that HGBR and GBR are significantly different from other methods at the α level of 0.01%. These findings indicate that HGBR and GBR perform significantly better than the other methods based on the evaluated statistical metrics.

4 BENEFITS OF ML-ENHANCED SIMULATION

4.1 Conventional Simulation vs. ML-enhanced Simulation

Utilizing the case-time duration prediction approach described in the previous section, the management of the surgical center becomes both personalized and precise. Specifically, for each request for key resources (i.e., OR and PAC bed) from a patient, the relevant services employ the validated HGBR models to forecast the duration of resource occupancy. The ML-enhanced DES treats the ML-predicted durations as point estimates, directly replacing the stochastic sampling from empirical or parametric distributions. That is, once the characteristics of a patient are known, a single deterministic duration predicted by the ML model is assigned in the DES model. This substitution eliminates residual randomness for these durations, thereby enabling more personalized and precise simulation.

To assess the impact of integrating predictive capabilities of ML algorithms into the simulation model, we propose two distinct management scenarios: (i) precise surgical service management based on the machine learning prediction models (i.e., ML-enhanced DES) and (ii) classical surgical service management (i.e., conventional DES) that relies solely on probability distributions derived from the entire training dataset (as described in Section 2.3) and does not consider personal characteristics. The key performance indicators (KPIs) for evaluation are the patient LOS in key services (e.g., surgeries and post-anesthesia care) and utilization of key resources (e.g., OR and PAC beds).

4.2 Numerical Results

To validate the ML-enhanced DES model, we conducted ten independent replications of each model, i.e., the ML-enhanced DES model versus the conventional DES model. The average values of the four primary OR management metrics, i.e., OR LOS, OR utilization, PAC LOS, and PAC bed utilization, are calculated and compared. As shown in Table 8, the ML-enhanced DES model achieved a 6.46% reduction in mean OR LOS, with a corresponding 5.03% reduction in its standard deviation (SD), relative to the traditional DES model. OR utilization was compared on an individual-OR basis. Figure 5 (left) demonstrates that utilization rates produced by the ML-enhanced DES model deviate from the actual data by at most 0.0677 (6.77%), whereas the conventional DES model exhibits a maximum deviation of 0.1155 (11.55%).

Table 8: Comparison of OR LOS measured in hours (N=638).

	Actual data	DES	ML-based DES
Mean	1.198	1.359	1.272
SD	0.700	0.596	0.563

Table 9: Comparison of PAC LOS measured in hours (N=638).

	Actual data	DES	ML-based DES
Mean	1.337	1.643	1.551
SD	0.843	0.675	0.763

Table 9 reports a 5.59% decrease in average PAC LOS under the ML-enhanced DES framework. In Figure 5 (right), the letters A-E denote the 5 PAC beds located in the PHU, and the numbers 1-19 correspond to the 19 PAC beds located in the PACU. The ML-enhanced DES model's PAC bed utilization predictions incur a maximum deviation of 0.0812 (8.12%) from observed data, in contrast to 0.1290 (12.90%) for the conventional DES model.

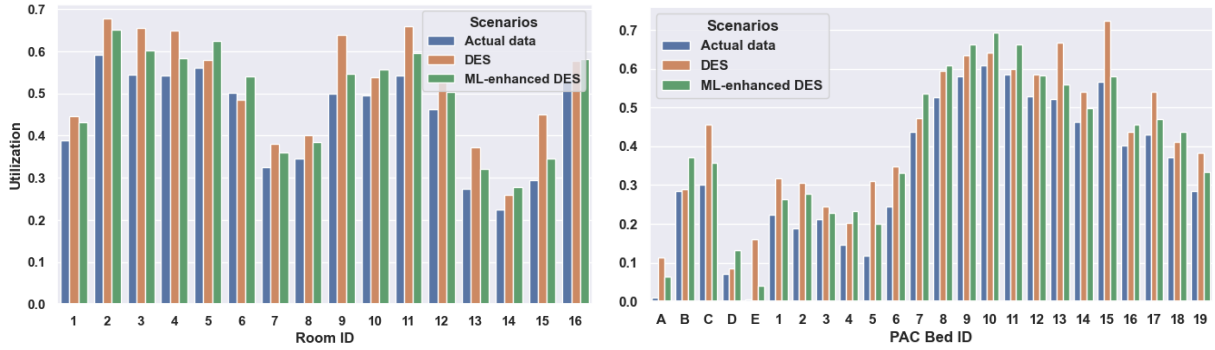

Figure 5: Comparison of OR utilization (left) and PAC bed utilization (right).

To statistically validate these improvements, paired t-tests ($\alpha = 0.01$) were conducted using the results from 10 simulation replications for both OR LOS and PAC LOS. The observed reductions in mean OR LOS (0.087 hours) and mean PAC LOS (0.092 hours) were both statistically significant ($p < 0.001$). Additionally, paired t-tests ($\alpha = 0.05$) were performed on the utilization rates of each OR and PAC bed, also based on 10 replications. The differences in utilization were statistically significant for all 16 ORs and all 19 PAC

beds ($p < 0.01$). Other ancillary metrics were validated against historical data and reviewed with hospital staff.

5 CONCLUSION, DISCUSSION AND FUTURE WORK

This study presents an ML-enhanced simulation approach aimed at optimizing the utilization of critical resources in surgical centers through the integration of ML techniques. The proposed hybrid simulation model, which combines ML and DES, facilitates the incorporation of individual patient characteristics and enables the quantification of service process variability via pre-trained ML models. The results derived from the ML-enhanced simulation highlight the potential advantages of embedding predictive capabilities of ML algorithms within simulation frameworks.

While the ML-enhanced DES framework demonstrated significant improvements in simulation accuracy at the Saint-Etienne Mutualist Clinic, its generalizability to other surgical centers warrants consideration. The transferability of the HGBR models depends on: (i) data availability (e.g., availability of structured patient covariates and unstructured clinical notes), (ii) workflow similarity (e.g., patient pathways, resource configurations), and (iii) operational heterogeneity (e.g., scheduling policies, emergency case protocols). Institutions with divergent data ecosystems or care processes may require model retraining using local datasets. Future multisite validation studies are recommended to establish broader applicability.

In resource-limited settings, such integrated approaches can enable precise resource allocation and reduce workflow disruptions. Future research will pursue two key objectives. First, we aim to develop methodologies for both offline planning and online scheduling of key surgical resources within the simulation framework, with the goal of achieving performance improvements beyond those offered by the current ML-enhanced DES model. Second, for stochastic events such as late arrivals of patients or surgeons and the occurrence of emergency cases, we intend to construct predictive models and design proactive mitigation strategies. These efforts are expected to reduce OR overtime and promote a more balanced and efficient utilization of critical resources.

Acknowledgments: This research was partially supported by the Fundamental Research Funds for the Central Universities (Grant No. SWU-KQ25028) and the Youth Program of the National Natural Science Foundation of China (Grant No. 62306246).

REFERENCES

Bodenstedt, S., M. Wagner, L. Mündermann, H. Kenngott, B. Müller-Stich, M. Breucha, *et al*. 2019. "Prediction of Laparoscopic Procedure Duration Using Unlabeled, Multimodal Sensor Data". *International Journal of Computer Assisted Radiology and Surgery* 14:1089–1095.

Childers, C. P., and M. Maggard-Gibbons. 2018. "Understanding Costs of Care in the Operating Room". *JAMA Surgery* 153(4):e176233–e176233.

Demšar, J. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets". *Journal of Machine Learning Research* 7:1–30.

Denton, B., J. Viapiano, and A. Vogl. 2007. "Optimization of Surgery Sequencing and Scheduling Decisions Under Uncertainty". *Health Care Management Science* 10:13–24.

Devi, S. P., K. S. Rao, and S. S. Sangeetha. 2012. "Prediction of Surgery Times and Scheduling of Operation Theaters in Optholmology Department". *Journal of Medical Systems* 36:415–430.

Dexter, F. 2000. "A Strategy to Decide Whether to Move the Last Case of the Day in an Operating Room to Another Empty Operating Room to Decrease Overtime Labor Costs". *Anesthesia & Analgesia* 91(4):925–928.

Dexter, F., R. H. Epstein, R. D. Traub, Y. Xiao, and D. C. Warltier. 2004. "Making Management Decisions on the Day of Surgery Based on Operating Room Efficiency and Patient Waiting Times". *Anesthesiology* 101(6):1444–1453.

Dexter, F., and J. Ledolter. 2005. "Bayesian Prediction Bounds and Comparisons of Operating Room Times Even for Procedures with Few or No Historic Data". *Anesthesiology* 103(6):1259–1167.

Elfanagely, O., Y. Toyoda, S. Othman, J. A. Mellia, M. Basta, T. Liu, *et al*. 2021. "Machine Learning and Surgical Outcomes Prediction: A Systematic Review". *Journal of Surgical Research* 264:346–361.

Fang, F., T. Liu, J. Li, Y. Yang, W. Hang, D. Yan, *et al*. 2023. "A Novel Nomogram for Predicting the Prolonged Length of Stay in Post-anesthesia Care Unit After Elective Operation". *BMC Anesthesiology* 23(1):404.

Forbus, J. J., and D. Berleant. 2022. "Discrete-event Simulation in Healthcare Settings: A Review". *Modelling* 3(4):417–433.

Havrlant, L., and V. Kreinovich. 2017. "A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (TF-IDF) Heuristic (and Variations Motivated by This Explanation)". *International Journal of General Systems* 46(1):27–36.

Jiao, Y., B. Xue, C. Lu, M. S. Avidan, and T. Kannampallil. 2022. "Continuous Real-time Prediction of Surgical Case Duration Using a Modular Artificial Neural Network". *British Journal of Anaesthesia* 128(5):829–837.

Khaniyev, T., E. Kayış, and R. Güllü. 2020. "Next-day Operating Room Scheduling with Uncertain Surgery Durations: Exact Analysis and Heuristics". *European Journal of Operational Research* 286(1):49–62.

Liu, Y., and W. Chen. 2012. "A SAS Macro for Testing Differences among Three or More Independent Groups Using Kruskal-Wallis and Nemenyi Tests". *Journal of Huazhong University of Science and Technology - Medical Sciences* 32(1):130–134.

Pandit, J. 2020. "Rational Planning of Operating Lists: a Prospective Comparison of 'booking to the Mean' vs. 'probabilistic Case Scheduling' in Urology". *Anaesthesia* 75(5):642–647.

Rifi, L. 2023. *Digital Twin-based Decision Support System for the Prospective and the Retrospective Analysis of an Operating Room Under Uncertainties*. Ph. D. thesis, Ecole des Mines d'Albi-Carmaux.

Saadouli, H., B. Jerbi, A. Dammak, L. Masmoudi, and A. Bouaziz. 2015. "A Stochastic Optimization and Simulation Approach for Scheduling Operating Rooms and Recovery Beds in an Orthopedic Surgery Department". *Computers & Industrial Engineering* 80:72–79.

Strachota, E., P. Normandin, N. O'brien, M. Clary, and B. Krukow. 2003. "Reasons Registered Nurses Leave or Change Employment Status". *The Journal of Nursing Administration* 33(2):111–117.

Strömblad, C. T., R. G. Baxter-King, A. Meisami, S.-J. Yee, M. R. Levine, A. Ostrovsky, *et al*. 2021. "Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources: a Randomized Clinical Trial". *JAMA Surgery* 156(4):315–321.

Van Eijk, R. P., E. Van Veen-Berkx, G. Kazemier, and M. J. Eijkemans. 2016. "Effect of Individual Surgeons and Anesthesiologists on Operating Room Time". *Anesthesia & Analgesia* 123(2):445–451.

Van Tunen, B., M. Klimek, K. Leendertse-Verloop, and R. J. Stolker. 2020. "Efficiency and Efficacy of Planning and Care on a Post-anesthesia Care Unit: a Retrospective Cohort Study". *BMC Health Services Research* 20:1–8.

Varmazyar, M., R. Akhavan-Tabatabaei, N. Salmasi, and M. Modarres. 2020. "Operating Room Scheduling Problem Under Uncertainty: Application of Continuous Phase-type Distributions". *IISE Transactions* 52(2):216–235.

West, C. P., A. D. Tan, T. M. Habermann, J. A. Sloan, and T. D. Shanafelt. 2009. "Association of Resident Fatigue and Distress with Perceived Medical Errors". *JAMA* 302(12):1294–1300.

Zhao, B., R. S. Waterman, R. D. Urman, and R. A. Gabriel. 2019. "A Machine Learning Approach to Predicting Case Duration for Robot-assisted Surgery". *Journal of Medical Systems* 43(2):32.

Zheng, H., Q. Wang, J. Shen, Y. Kong, and J. Li. 2022. "Modeling and Analysis of Operating Room Workflow in a Tertiary a Hospital". *IEEE Robotics and Automation Letters* 7(3):7006–7013.

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

AUTHOR BIOGRAPHIES

CHEN HE is a lecturer at the College of Artificial Intelligence, Southwest University, Chongqing, China. His research interests include machine learning, data mining, and the application of operations research in healthcare systems. His email address is chenhe95@swu.edu.cn and his website is https://hechen95.github.io.

MARIANNE SARAZIN is a public health physician holding a doctorate in life sciences, France. She is the Director of the Medical Information Department at the Saint-Étienne Mutualist Clinic. Her email address is marianne.sarazin@iplesp.upmc.fr.

FRANÇOIS DUFOUR is an anesthesiologist and intensivist at the Saint-Etienne Mutualist Clinic. His research interests lie in operating theater optimization, including surgical planning, scheduling, and critical resource allocation. His email address is FDUFOUR@mutualite-loire.com.

MARIE PAULE BOURBON is the Director of the Operating Center at the Saint-Etienne Mutualist Clinic. Her research interests include the simulation and modeling of surgical centers. Her email address is MBourbon@mutualite-loire.com.

XIAOLAN XIE is a professor of industrial engineering at the École des Mines de Saint Etienne. His research interests include healthcare system engineering, optimization, and data analytics. He is the author of 350+ publications including over 130+ journal articles and six books. Prof. Xie is a fellow of IEEE. He was the founding chair of the Technical Committee on Automation in Health Care Management of the IEEE Robotics & Automation Society. He has been editor/associate editor for various international journals (IEEE TASE, IEEE TAC, IEEE TRA, IJPR) and special issue guest editor. His email address is xie@emse.fr.