

Original Paper

# Predicting Pediatric Urological Surgery Duration Through Multimodal Patient-Physician Feature Fusion: Deep Learning Framework Incorporating Clinical Text Embedding

Yonggen Zhao<sup>1,2\*</sup>, BE; Ruoge Lin<sup>3\*</sup>, MS; Yiyi Sun<sup>1,2\*</sup>, MS; Lingdong Chen<sup>1,2</sup>, MS; Jian Huang<sup>1,2</sup>, PhD; Guangjie Chen<sup>4</sup>, MD; Zhu Zhu<sup>1,2\*</sup>, PhD; Gang Yu<sup>1,2</sup>, PhD

<sup>1</sup>National Clinical Research Center for Children and Adolescents' Health and Diseases, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China

<sup>2</sup>Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou, China

<sup>3</sup>College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, China

<sup>4</sup>Department of Urology, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China

\*these authors contributed equally

## Corresponding Author:

Gang Yu, PhD

National Clinical Research Center for Children and Adolescents' Health and Diseases

Children's Hospital, Zhejiang University School of Medicine

3333 Binsheng Rd

Hangzhou 310052

China

Phone: 86 13588773370

Email: [yugbme@zju.edu.cn](mailto:yugbme@zju.edu.cn)

## Abstract

**Background:** Accurate prediction of surgical duration is critical for optimizing operating room scheduling and resource allocation. Existing models, however, exhibit limited applicability in pediatric urology due to the unique anatomical and developmental characteristics of children.

**Objective:** This study aimed to develop and validate a specialty-tailored prediction framework for estimating the duration of pediatric urological surgeries.

**Methods:** We integrated multisource heterogeneous data, encompassing patient demographics, surgical details, surgeon-specific features, and electronic medical record narratives, to develop a customized prediction system. Large language model techniques were used to extract semantic representations from unstructured clinical text, while a multihead perceptron architecture enabled the efficient fusion of structured and unstructured features. Pediatric-specific clinical variables, such as developmental stage and the severity of urinary tract malformations, were explicitly modeled to capture their impact on surgical duration.

**Results:** The proposed approach achieved a mean absolute error of 11.39 minutes and a root mean square error of 15.58 minutes, markedly outperforming existing methods. Comparative analyses demonstrated that the Qwen-based structured preprocessing with text embeddings provided superior feature representation, surpassing both traditional long short-term memory and direct Embedding-3 approaches. Feature importance analysis identified the primary surgical procedure, surgical plan, and preoperative diagnosis as dominant predictive factors.

**Conclusions:** By combining innovative feature engineering with a tailored model architecture, the proposed framework substantially improves the accuracy of surgical duration prediction in pediatric urology. These findings offer robust technical support for precision operating room scheduling and hold significant clinical value in enhancing the efficiency of surgical resource utilization.

*JMIR Med Inform* 2026;14:e82329; doi: [10.2196/82329](https://doi.org/10.2196/82329)

**Keywords:** surgical duration prediction; operating room management; deep neural network; multihead perceptron; medical informatics

## Introduction

### *Background and Motivation*

As one of the hospital's most resource-intensive departments, the operating room plays a pivotal role in determining the institution's financial health and overall success [1,2]. Optimizing operating room scheduling under constraints of available bed capacity and human resources has become a key lever for improving health care service efficiency. Within this process, accurate prediction of surgical duration is a prerequisite for effective scheduling decisions [3-6]. However, due to the presence of multiple sources of uncertainty, such as patient heterogeneity (eg, disease complexity and anatomical variation), intraoperative contingencies (instrument failure and unexpected bleeding), and dynamic adjustments to the surgical plan [7,8], predicting surgical duration remains a major challenge in perioperative management [9,10]. Traditional estimation methods largely rely on the surgeon's personal experience or simple historical averages [11,12]. These approaches lack standardized classification of surgical procedures and fail to systematically integrate dynamic variables, such as patient-specific characteristics, team collaboration efficiency, and equipment turnover rates, leading to substantial discrepancies between estimated and actual durations [13,14]. Such inaccuracies can trigger a cascade of operational disruptions. The underestimation of duration can lead to staff fatigue, congestion in emergency channels [15], and prolonged patient waiting and fasting times, thereby increasing the risk of adverse events. Overestimation, on the other hand, results in idle resource time. More critically, errors in estimating a single case can be amplified through the scheduling system, ultimately causing mismatches in staffing, cost, and heightened dissatisfaction among patients and their families [16,17]. Therefore, developing more accurate and reliable methods for surgical duration prediction is both important and urgent.

In recent years, machine learning has made significant advances in surgical duration prediction [18,19]. This research can be broadly categorized by the types of features used.

### *Patient- and Procedure-Oriented Studies*

These methods mainly rely on patient characteristics and surgical information, leveraging demographic data, preoperative physiological indicators, and procedural details for predictive modeling. Commonly used algorithms include multilayer perceptrons (MLPs) and Extreme Gradient Boosting, with some studies incorporating real-time intraoperative parameters, such as heart rate and bispectral index, to enable dynamic predictions [20-23]. Key predictive factors identified in these studies include patient BMI, surgical type, and procedural complexity, with certain models achieving prediction errors—the mean absolute error (MAE) within a clinically acceptable margin of 30 minutes.

### *Surgeon- and Procedure-Oriented Studies*

Such studies emphasize surgeon-related and procedural factors, particularly the impact of operator experience, surgical techniques, and case urgency on operative duration. Ensemble methods, such as bagged decision trees, gradient boosting trees, and Extreme Gradient Boosting, have demonstrated strong performance in this domain, especially for single-specialty or emergency procedures, with reported MAEs as low as 14.9 minutes [24-28]. Feature importance analyses consistently highlight surgical complexity and surgeon proficiency as dominant predictors.

### *Comprehensive Studies Integrating Patient, Doctor, and Surgical Information*

The most comprehensive approaches integrate multimodal clinical data, combining patient-specific variables, surgeon expertise, and procedural factors, to optimize predictive accuracy. Random forests and MLPs are frequently adopted in such frameworks, while some studies have further improved performance through specialized department-specific modeling or 2-stage prediction architectures, achieving MAEs below 16 minutes in targeted clinical settings [29-31]. Emerging techniques, including computer vision-based analysis of surgical videos and large language models (LLMs) (eg, GPT-4), are also gaining traction, demonstrating promising capabilities in processing unstructured clinical narratives for duration estimation [32,33].

Despite the diversity of approaches in existing studies, several critical limitations persist. A primary issue lies in feature engineering, where current methods predominantly rely on structured data fields, such as procedure names and diagnostic codes, while largely neglecting the wealth of information contained in unstructured clinical notes. These free-text entries, which include detailed condition descriptions and laboratory results, often capture patient-specific variations that could significantly enhance prediction accuracy. Furthermore, existing models demonstrate limited applicability to pediatric surgical contexts. Adult-derived models fail to account for pediatric-specific physiological considerations, including developmental variations in organ systems and anatomical structures. This oversight is particularly problematic, given the narrower intraoperative safety margins in pediatric patients due to their reduced physiological reserves [12], which demands greater predictive precision.

This study focuses on pediatric urological surgery and develops a novel specialty-oriented prediction method. Compared with adult procedures, pediatric urology presents unique clinical challenges: the patient age range spans from newborns to adolescents, with marked developmental variations in physiological parameters, and differences in urinary tract development and malformation severity influence the level of operative precision required. These

factors make adult-trained models poorly transferable to pediatric cases.

To address these issues, we designed an end-to-end prediction system, which integrates multimodal clinical data, including patient characteristics, surgical information, surgeon attributes, and clinical narratives, to model and accurately predict pediatric urological surgery duration. We innovatively used LLM techniques to automatically process electronic medical record (EMR) text containing detailed case descriptions, examination findings, and physiological features, extracting and embedding unstructured clinical information efficiently. In addition, we proposed a multi-head MLP architecture for heterogeneous data fusion, allowing categorical, numerical, and text feature vectors to be processed independently before integration, outperforming traditional single-head MLP approaches. We further applied permutation importance analysis to identify the most influential predictors, including the lead surgeon, primary surgical procedure, preoperative diagnosis, and pediatric-specific disease characteristics (eg, prepuce characteristics and testicular characteristics). Comparative and ablation experiments demonstrated that our method substantially improved prediction accuracy, achieving an MAE of 11.39 minutes and a root mean square error (RMSE) of 15.58 minutes—significantly better than the existing work. Our model exhibits clear advantages in predicting pediatric urological surgery duration, providing reliable support for precise operating room scheduling.

## Methods

### *Data Acquisition and Preprocessing Flow*

The data used in this study were obtained from surgical cases performed at the Urology Surgery Center of the Children's Hospital, Zhejiang University School of Medicine, between May 22, 2024, and July 10, 2025. The hospital is an independent tertiary pediatric medical center. All surgical information was extracted from the hospital information system, yielding a total of 4526 cases. These encompassed 4 major categories of pediatric urological procedures: penile lengthening (n=2067, 45.7%), orchidopexy (n=937, 20.7%), laparoscopic high ligation of the processus vaginalis (n=1158, 25.6%), and circumcision (n=364, 8%).

### *Feature Extraction*

In this study, the duration of surgery was defined as the time interval commencing from “anesthesia induction” and concluding when the patient left the operating room. This definition encompasses not only the surgical procedure itself but also key stages, such as anesthesia induction, preoperative preparation, and postoperative transfer prior to the patient's departure from the operating room. As a result, it more comprehensively reflects the actual occupancy period of the operating room. When compared with simplified definitions, such as “from skin incision to suture completion” or “from entry to exit of the operating room,” the time frame used in this study more accurately reflects the real-world workflow under multidisciplinary collaboration and holds direct guiding

significance for hospital surgical scheduling and resource allocation [34]. For example, precise prediction of this period facilitates the rational arrangement of consecutive surgeries, optimizes the coordination between the anesthesia team and the recovery room, and mitigates idle time or congestion arising from estimation errors. All duration data were obtained from the hospital information system, guaranteeing the consistency and traceability of the definition. Considering that previous research has suggested that the duration of surgery generally follows a log-normal distribution [35,36], we standardized the target variable (with a mean of 0 and an SD of 1) to improve the stability of model training and the balance of feature scales.

Additional predictor variables extracted from the information system included patient age, patient weight, primary surgical procedure, lead surgeon features, surgical plan, preoperative diagnosis, and case-specific characteristics, including preoperative preparation details. All extracted variables were used by the prediction model.

Patient age and weight are continuous variables. Patient age was recorded in the format “x years x months” and was converted to years, followed by standardization. If the age was missing, it was replaced with the age mentioned in the “case characteristics and preoperative preparation” notes. Patient weight, recorded in kilograms, was also standardized. If missing or abnormal, it was imputed using the average weight of patients in the same age group.

Primary surgical procedure and lead surgeon features are categorical variables. The primary surgical procedure is a single-category variable referring to the main pediatric urological surgery performed for the patient, with only 1 such designation per case. This study primarily covered 4 main procedure types: penile lengthening, orchidopexy, laparoscopic high ligation of the processus vaginalis, and circumcision. This variable was one-hot encoded. The lead surgeon data were expanded into 3 distinct features to capture specific physician characteristics: name, years of practice, and annual surgical volume. The surgeon name is a single-category variable representing the primary attending specialist and was processed using one-hot encoding. Years of practice is a continuous numeric variable representing the duration of the surgeon's professional experience. Annual surgical volume is a discrete numeric variable indicating the total number of surgeries performed by that surgeon in the past year. Both numeric features (years of practice and annual surgical volume) were standardized to normalize their scales. Missing values for these features were imputed using the mean value of the respective column.

Preoperative diagnosis and surgical plan are multilabel categorical variables. The preoperative diagnosis refers to the diagnostic categories assigned to a patient before surgery, and each patient may have 1 or multiple diagnoses. For example, patient A's diagnosis might be “hydrocele,” whereas patient B's diagnosis could be “hydrocele, phimosis, and incomplete testicular descent.” Diagnoses were split by commas, and occurrence counts were tallied for each category. To avoid excessive sparsity, only diagnostic categories appearing  $\geq 10$

times were retained; others were grouped into a “rare type” category. Multilabel encoding was then applied. For example, “hydrocele” = [1, 0, 0], “phimosis” = [0, 1, 0], “incomplete testicular descent” = [0, 0, 1], and “hydrocele, phimosis, incomplete testicular descent” = [1,1,1].

The “surgical plan” refers to the preoperative plan drafted by the surgeon, and each patient may have 1 or more planned procedures. For example, patient A’s plan could be “bilateral laparoscopic high ligation of the processus vaginalis,” whereas patient B’s plan might be “bilateral laparoscopic high ligation of the processus vaginalis + orchidopexy + circumcision.” Due to differences in writing habits, plans often have inconsistent formatting, with procedure types separated by various delimiters, such as plus signs, commas, or phrases like “simultaneously perform.” Naming conventions for the same procedure were also inconsistent; for instance, “right orchidopexy” and “right-sided orchidopexy” refer to the same surgery. For our prediction target, surgical duration, laterality (“left” vs “right”) was assumed to have no effect for unilateral procedures, so these were combined into “unilateral” categories, whereas “bilateral” procedures were treated separately due to their significant impact on duration. Initially, manual tokenization and grouping were performed based on delimiters (plus signs, commas, etc) and standardization of terms (eg, merging “circumcision” and “circumcision [stapler/ligation/manual suture]”). This yielded 170 categorized procedure types, of which only those occurring  $\geq 10$  times were retained. The rest were grouped as “other surgical procedures.” All surgical plans were encoded using the same multilabel approach.

However, due to the high variability in surgical plans, manual tokenization and grouping could not fully resolve inconsistencies, and many equivalent procedures remained uncategorized. To address this, we used an advanced approach using the LLM Qwen3, developed by Alibaba in April 2025. The selection of Qwen over other prominent LLMs (eg, GPT-4 and Llama-3) was driven by its specific alignment with our study’s linguistic and security requirements. Unlike Llama-3, which is primarily optimized for English, Qwen demonstrates superior proficiency in understanding specialized Chinese clinical terminology and semantics relevant to our dataset. Furthermore, in contrast to cloud-based services, such as GPT-4, that necessitate external data transmission, Qwen supports local deployment. This feature allows us to process sensitive patient data entirely within the hospital’s secure intranet, ensuring strict adherence to data privacy regulations while maintaining reasoning and extraction capabilities compared to top-tier proprietary models.

By calling Qwen3’s application programming interface for batch processing, the model summarized all patients’ surgical plans into 23 standardized procedure categories (including “other surgical procedures”). This intelligent consolidation reduced unnecessary feature dimensionality and improved downstream analysis. For example, Qwen3 automatically grouped “penile nevus excision,” “penile cyst excision,” and “penile scar excision” under “penile lesion excision.” Similarly, these standardized categories were then multilabel-encoded for each patient.

Case characteristics and preoperative preparation are unstructured free-text variables describing reasons for admission, degree of organ development, precise lesion locations, and other specifics. Due to similar issues with inconsistent formatting, we again leveraged Qwen3 for structured information extraction. We predefined 4 target features for extraction: prepuce characteristics description, testicular characteristics description, scrotal characteristics description, and penile characteristics description. During extraction, Qwen3 was instructed to output both the structured fields and the corresponding original text segments to guard against hallucinations. Since the extracted text lengths varied, we used the Embedding-3 model (256 dimensions) to produce fixed-length semantic vectors for model input. Embedding-3, developed by ZHIPU AI (Tsinghua University), is a text-embedding model capable of capturing deep semantic relationships and outputting high-quality embeddings. It supports multilingual and long-text inputs, making it well-suited for high-precision semantic representation tasks.

Missing category and text vectors were replaced with null vectors. A detailed description of the surgical feature variables is provided in [Table 1](#).

It is worth noting that while “surgeon name” provides specific identity information, it lacks generalizability, as it is strictly bound to the current roster of physicians and cannot accommodate new surgeons joining the department in the future. To address this limitation, we designed 2 distinct experimental configurations. The first configuration includes “surgeon name” to capture individual-specific variance. The second configuration substitutes “surgeon name” with the quantitative metrics “years of practice” and “annual surgical volume” to construct a more robust and generalizable model capable of predicting surgical duration for new or unobserved surgeons.

**Table 1.** Description of specific surgical feature variables (N=4526).

Patient variables and type	Values/range and statistics
Age (year)	
Numeric	Float (average 5.3, SD 3.7, max 17.0, min 0.0)
Weight (kg)	
Numeric	Float (average 24.6, SD 14.6, max 140, min 5)
Primary surgical procedure	
Categorical	Penile lengthening surgery (2067, 45.6%), laparoscopic high ligation of the processus vaginalis (1158, 25.6%), orchidopexy (937, 20.7%), circumcision (364, 8%)
Lead surgeon	
Categorical	Name: total =31, including “unknown”
Numeric	Years of practice: float (average 16.1, SD 7.1, max 40.6, min 1.5) (y)
Numeric	Annual surgical volume: integer (average 461.2, SD 223.8, max 863, min 2) (cases/y)
Preoperative diagnosis	
Multicategorical	Concealed penis (1839, 40.6%), hydrocele (1202, 26.6%), and cryptorchidism (733, 16.2%).
Surgical plan	
Multicategorical	Penile lengthening (1956, 43.2%), preputioplasty (1649, 36.4%).
Case characteristics and preoperative preparation	
Free text	Prepuce characteristics description: eg, “Phimosis prepuce, can be manually retracted to expose the glans.”
Free text	Testicular characteristics description: eg, “Both testes are located within the scrotum and are well-developed.”
Free text	Scrotal characteristics description: eg, “Scrotum is well-developed.”
Free text	Penile characteristics description: eg, “The penis appears small, with only the distal half of the glans visible on the body surface in the supine position.”
Surgery duration (min)	
Numeric	Float (average 38.6, SD 24.9, max 539, min 4)

## Model Construction

We developed multiple machine-learning methods based on neural networks and traditional models. From the entire dataset of 4526 cases, 15% (n=679 cases) were randomly selected as test data for evaluating the prediction models. The remaining data (referred to as “training data”) were further randomly split into training (n=3168 cases) and validation (n=679 cases) sets at an 85% to 15% ratio. Detailed descriptions of model hyperparameters and validation performance metrics can be found in the Results section. For the finalized model, performance evaluation was conducted only once on the test dataset.

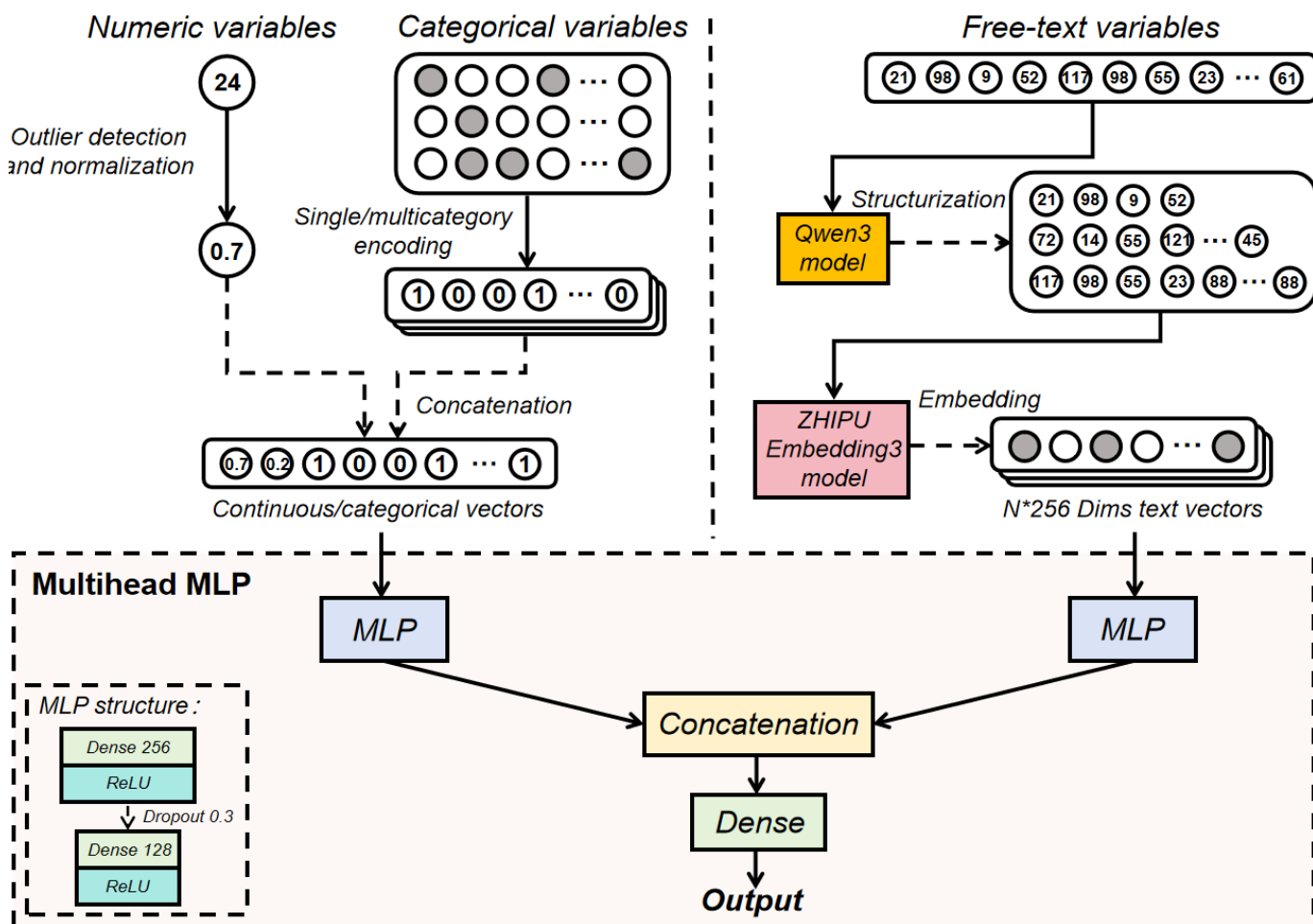
### **Pediatric Urology Surgical Duration Prediction Model Based on Multihead MLP**

In this study, we trained an MLP model with a multihead structure (multihead MLP) to predict surgical duration in pediatric urology. The core idea is to process different types of input features (text vectors and categorical/numerical vectors) separately through independent subnetworks or

branches and then concatenate their outputs and feed them into a shared MLP layer for the final prediction.

Traditional single-head MLP concatenates all heterogeneous features into 1 large input vector, which is then passed through a single MLP for prediction. However, due to the severe dimensional imbalance among features in this study (eg, the primary surgery procedure vector is 4-dimensional, while text vectors have dimensions of 4×256), single-head MLP tends to be dominated by features with larger dimensions, drowning out those with smaller dimensions and leading to a tendency for overfitting. To address this, the multihead MLP uses parallel representation learning by establishing independent subnetworks for each data group. This design enables the model to capture the unique patterns and characteristics of each data type separately, thereby reducing interference and ensuring that heterogeneous data are fused within a balanced latent space. To mitigate overfitting, our multihead MLP was trained with early stopping. The architecture of the proposed model is shown in [Figure 1](#).

**Figure 1.** Architecture of the proposed pediatric urology surgical duration prediction model. MLP: multilayer perceptron; ReLU: rectified linear unit.



The multihead MLP architecture comprises 2 distinct subnetworks designed to process heterogeneous input features separately: one dedicated to textual features and the other to categorical and numerical features. Each subnetwork consists of 2 fully connected (dense) layers with 256 and 128 neurons, respectively, each followed by a dropout layer with a rate of 0.3 to mitigate overfitting. Rectified linear unit activation functions are applied throughout. Outputs from both subnetworks are concatenated in a feature fusion layer, which feeds into a subsequent dense layer of 64 neurons accompanied by another dropout layer (rate 0.3). The final output layer uses a single neuron with a linear activation function to produce continuous predictions of surgical duration. The model is optimized using the Adam optimizer with a learning rate of 0.001, trained over 200 epochs, and uses mean squared error as the loss function.

### Model Training and Experimental Settings

This study included a total of 4526 pediatric urology surgical cases with complete recorded surgical times for model development and evaluation. The training dataset (comprising both training and validation sets) contained 3847 cases, including 1730 cases of penile lengthening surgery, 810 cases of orchiopexy, 992 cases of laparoscopic high ligation of the processus vaginalis, and 315 cases of circumcision, accounting for 85% of the total dataset. The test dataset consisted

of 679 cases, including 337 penile lengthening surgeries, 127 orchioepexies, 166 laparoscopic high ligations, and 49 circumcisions, representing 15% of the total dataset.

Python 3.8.3 (Python Software Foundation) was used for all feature extraction, algorithm implementation, and performance evaluation. The Alibaba Qwen3 LLM was used for all text analysis and structured field extraction. The ZHIPU AI Embedding-3 large model was used for all text embedding. Google’s TensorFlow 2.4.1 (Google Brain) was used to implement the multihead MLP and other neural networks in comparative experiments. The Scikit-Learn library (scikit-learn community and INRIA) was applied for machine-learning methods, such as decision tree regression (DT) in comparative experiments, with default parameters used for all models.

### Evaluation Metrics

Model performance was assessed using 3 metrics: MAE, RMSE, and the coefficient of determination (R-squared,  $R^2$ ).

MAE represents the average of the absolute differences between predicted and true values. It directly reflects the average magnitude of prediction errors and is defined by the following formula:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

RMSE is the square root of the average of the squared differences between predicted and true values. It is a commonly used metric to measure the deviation between model predictions and actual values. The formula is as follows:

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

$R^2$  measures the proportion of variance in the dependent variable (true values) that is explained by the model. It represents the fraction of the total variability in the dependent variable accounted for by the model. An  $R^2$  value of 1 indicates a perfect fit, explaining all variance;  $R^2=0$  means the model explains no more variance than simply predicting the mean; and  $R^2<0$  indicates the model performs worse than predicting the mean. The formula is as follows:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (3)$$

The formulas for the residual sum of squares and the total sum of squares are as follows:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (5)$$

where  $n$  denotes the number of samples,  $y_i$  represents the true surgical duration for the  $i$ th sample,  $\bar{y}$  is the mean of all true values, and  $\hat{y}_i$  is the model's predicted surgical duration for the  $i$ th sample.

### Feature Importance Evaluation

Permutation importance is used to assess the significance of individual features within the model. As a model-agnostic method, its core principle is to measure the impact on model performance by randomly permuting the values of a single feature. Initially, predictions are made on the unshuffled test set to establish a baseline performance score. Then, for each feature in turn, its values in the test set are randomly shuffled while keeping other features unchanged, followed by generating new predictions and recalculating the performance score. The difference between the baseline and the shuffled performance indicates the importance of that feature. A significant drop in performance after shuffling a feature implies that the feature is highly influential to the model's predictive capability.

### Ethical Considerations

From an ethical perspective, this study focuses on the preoperative diagnosis and duration of surgery for patients. The information collection methods adopted by the researchers are designed to ensure that the patients cannot be directly identified, nor can they be indirectly identified through related identifiers. In this study, only the surgery-related content was analyzed, and no personally identifiable information was extracted or associated. Additionally, the identities of the hospital staff involved in the surgery were anonymized and represented only by numerical IDs. This study was performed in line with the principles of the Declaration of Helsinki. All data used in this study were obtained from retrospective surgical cases, and approval was granted by the Academic Ethics Committee of Children's Hospital Zhejiang University School of Medicine (IRB No. 2023-IRB-0181-P-01; granted on July 20, 2023). We have applied for an informed consent waiver for our study. No compensation was provided to participants, as this was a retrospective study using existing surgical records without direct patient involvement or additional data collection.

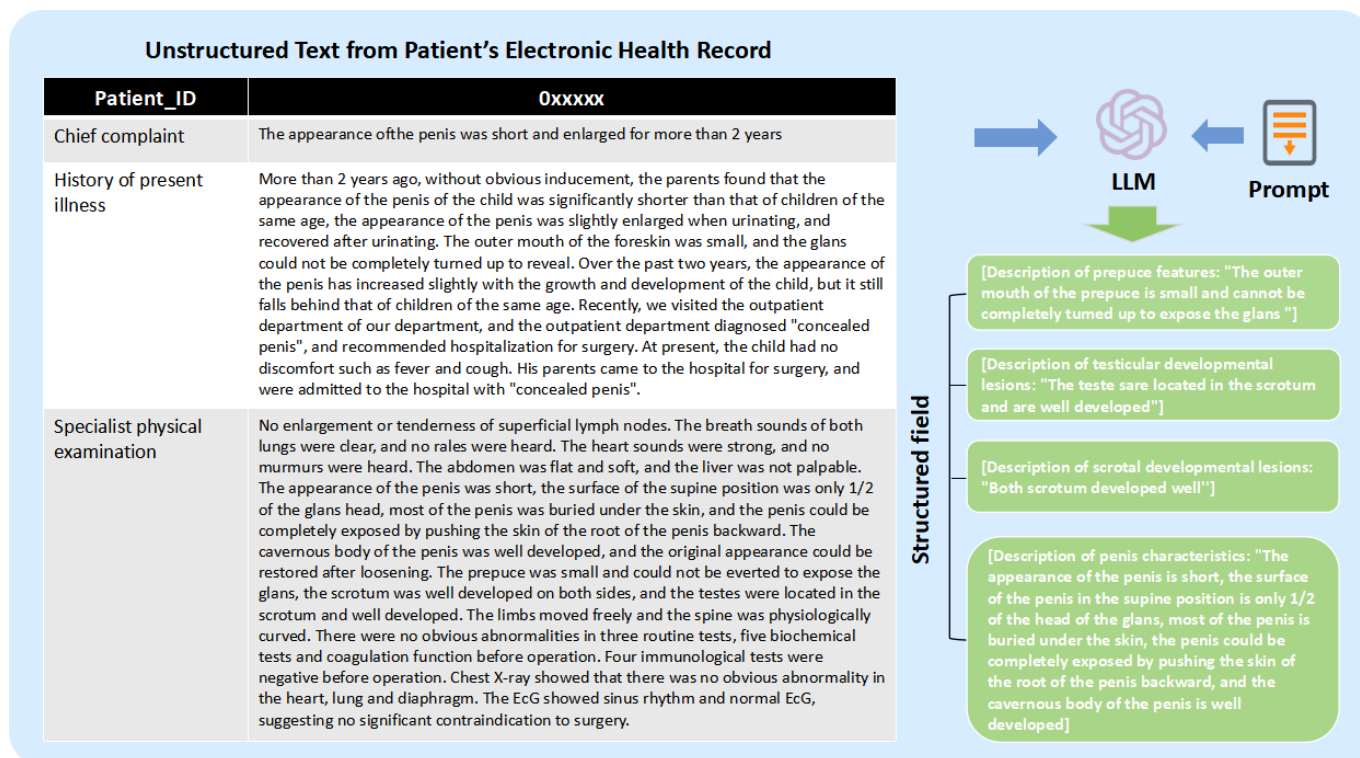
## Results

### Accuracy Verification of LLM-Based Extraction

To quantitatively evaluate the fidelity of the structured data generated by Qwen-3 and to assess the prevalence of hallucinations, we conducted a manual validation on a randomly selected subset of 500 cases. This evaluation focused on 4 key features extracted from the unstructured "case characteristics and preoperative preparation" text: prepuce characteristics description, testicular characteristics description, scrotal characteristics description, and penile characteristics description. The validation was performed by a senior medical informatics engineer with over 10 years of experience in clinical data processing. A strict scoring metric (ranging from 0% to 100%) was applied to measure the consistency between the original text and the extracted structured data. Specifically, to rigorously penalize hallucinations, any extraction containing factually incorrect or fabricated information was assigned a score of 0%. Incomplete extractions were scored proportionally based on the extent of the missing information.

Figure 2 shows a concrete example of extracting structured fields from patient electronic health records using LLMs and prompt text. The evaluation results demonstrated high reliability in the model's extraction capabilities. The average accuracy scores for the 4 features were 99.6% for prepuce characteristics, 98.94% for testicular characteristics, 93.43% for scrotal characteristics, and 99.18% for penile characteristics. These results confirm that Qwen-3 effectively maintained semantic consistency with the source text while minimizing hallucinations.

**Figure 2.** A case of extracting structured fields from patients' electronic health records text using LLM and prompting. EcG: electrocardiogram; LLM: large language model.



## Model Performance Comparison

We compared the proposed method with 6 outstanding surgical duration prediction approaches in the field, including ridge regression (RR) [37], DT [38], support vector regression (SVR) [39], random forest regression (RF) [40], gradient boosting regression tree (GBR) [41], and the machine learning–mixture density network (ML-MDN) model that combines long short-term memory (LSTM) and MLP for surgical duration prediction [23]. All models were trained and tested using the same datasets and feature categories as the proposed model. The RR, DT, SVR, RF, and GBR models used the feature engineering pipeline proposed in this study, whereas the ML-MDN model was implemented using the feature processing strategy described in its original publication. Specifically, the ML-MDN method integrates both structured and unstructured text features using a neural network architecture. In this method, structured variables (categorical and numerical) are encoded or normalized, whereas unstructured text variables (including primary surgical procedure, surgical plan, preoperative diagnoses, and case characteristics) are tokenized and embedded into vectors fed into an LSTM network. The LSTM captures temporal dependencies and contextual information, outputting semantic vectors that are concatenated with structured features and fed into an MDN. The MDN outputs a probability distribution

governed by 3 parameters: mean ( $\mu$ ), SD ( $\sigma$ ), and mixture coefficients ( $\alpha$ ). As the other baseline models in this study are deterministic regressors that output single-point estimates, a direct comparison with a probability distribution is infeasible. To ensure a fair and consistent evaluation framework, we used the expected value (mean,  $\mu$ ) of the MDN's predicted distribution as the representative point prediction. This allows for the standard calculation of MAE, RMSE, and  $R^2$ . Additionally, we compared these metrics against a naive baseline ("average") that simply uses the global mean of the surgical duration from the training set for prediction.

In the first experimental configuration using "surgeon name" as a feature, the proposed method achieved the best performance (MAE=10.74 min, RMSE=14.91,  $R^2=0.58$ ). SVR (MAE=11.74 min), RF (MAE=11.62 min), and GBR (MAE=11.68 min) showed comparable, second-best results. RR (MAE=13.06 min) and DT (MAE=15.07 min) demonstrated relatively lower performance. The ML-MDN model's performance (MAE=16.91 min) was inferior to both our multihead MLP model (MAE=10.74 min) and other traditional machine-learning models. In contrast, the naive "average" baseline, which relied solely on the global mean without utilizing any features or machine-learning algorithms, yielded the poorest performance (MAE=19.06 min). The detailed model evaluation results are shown in Table 2.

**Table 2.** Performance comparison of different surgical duration prediction models and methods on the test dataset utilizing specific surgeon identity “surgeon name”.

Models	MAE <sup>a</sup>	RMSE <sup>b</sup>	R <sup>2c</sup>
Average	19.0579	24.8830	0.0000
RR <sup>d</sup> [37]	13.0554	19.3512	0.2910
DT <sup>e</sup> [38]	15.0692	21.8759	0.0939
SVR <sup>f</sup> [39]	11.7400	16.8024	0.4655
RF <sup>g</sup> [40]	11.6238	16.5873	0.4791
GBR <sup>h</sup> [41]	11.6821	16.3646	0.4930
ML-MDN <sup>i</sup> [23]	16.9117	21.4013	0.1328
Multihead MLP <sup>j</sup>	10.7391 <sup>k</sup>	14.9134 <sup>k</sup>	0.5789 <sup>k</sup>

<sup>a</sup>MAE: mean absolute error.

<sup>b</sup>RMSE: root mean squared error.

<sup>c</sup>R<sup>2</sup>: coefficient of determination.

<sup>d</sup>RR: ridge regression.

<sup>e</sup>DT: decision tree.

<sup>f</sup>SVR: support vector regression.

<sup>g</sup>RF: random forest.

<sup>h</sup>GBR: gradient boosting regression.

<sup>i</sup>ML-MDN: machine learning–mixed dense network.

<sup>j</sup>MLP: multilayer perceptron.

<sup>k</sup>These values in the last row indicate the best performance among all tested models across each evaluation metric.

In the second configuration, where “surgeon name” was substituted with “years of practice” and “annual surgical volume,” all models exhibited a slight increase in overall error rates. However, the overall performance trend remained

consistent: the proposed multihead MLP model still secured the best performance (MAE=11.39 min; RMSE=15.58; R<sup>2</sup>=0.54). Detailed results for this configuration are shown in [Table 3](#).

**Table 3.** Performance comparison of different surgical duration prediction models on the test dataset utilizing generalized surgeon metrics “years of practice” and “annual surgical volume”.

Models	MAE <sup>a</sup>	RMSE <sup>b</sup>	R <sup>2c</sup>
Average	19.0579	24.8830	0.0000
RR <sup>d</sup> [37]	13.6891	20.7138	0.1876
DT <sup>e</sup> [38]	14.6289	21.9035	0.0916
SVR <sup>f</sup> [39]	11.8670	16.9053	0.4589
RF <sup>g</sup> [40]	12.0465	16.8062	0.4652
GBR <sup>h</sup> [41]	11.6706	16.4278	0.4890
ML-MDN <sup>i</sup> [23]	17.2747	21.7401	0.1051
Multihead MLP <sup>j</sup>	11.3889 <sup>k</sup>	15.5808 <sup>k</sup>	0.5404 <sup>k</sup>

<sup>a</sup>MAE: mean absolute error.

<sup>b</sup>RMSE: root mean squared error.

<sup>c</sup>R<sup>2</sup>: coefficient of determination.

<sup>d</sup>RR: ridge regression.

<sup>e</sup>DT: decision tree.

<sup>f</sup>SVR: support vector regression.

<sup>g</sup>RF: random forest.

<sup>h</sup>GBR: gradient boosting regression.

<sup>i</sup>ML-MDN: machine learning–mixed dense network.

<sup>j</sup>MLP: multilayer perceptron.

<sup>k</sup>These values in the last row indicate the best performance among all tested models across each evaluation metric.

As previously discussed, relying on specific surgeon names limits the model’s applicability to the current roster and fails to accommodate unobserved or newly hired surgeons, thereby reducing generalizability. Therefore, despite the marginal decrease in prediction accuracy, substituting specific identities with quantitative professional metrics (years of practice and annual surgical volume) represents a more robust and scalable solution, ensuring that the model remains effective for future clinical applications. Consequently, all

subsequent results and analyses presented in this study utilize “years of practice” and “annual surgical volume” in place of “surgeon name.”

To evaluate the proposed model’s robustness, 5-fold cross-validation was performed. The model demonstrated remarkable stability in point-wise prediction, achieving an average MAE of 11.61 (0.47) minutes, as shown in [Table 4](#). The low SD (0.47 min) indicates that the model’s predictive

accuracy is consistent across different data partitions. It is worth noting that while 4 out of 5 folds yielded  $R^2$  scores exceeding 0.50, the second fold showed a performance dip ( $R^2=0.28$ ; RMSE=24.95). The discrepancy between the relatively stable MAE (12.05 min) and the spiked RMSE

in this fold suggests the presence of extreme outliers (eg, rare complex cases with unexpectedly long durations) in that specific test subset, rather than a systematic failure of the model. Overall, the aggregated results ( $R^2=0.47\pm.11$ ) confirm the model's generalizability in clinical scenarios.

**Table 4.** Performance metrics of the proposed multihead MLP<sup>a</sup> model using 5-fold cross-validation.

Models	1-fold	2-fold	3-fold	4-fold	5-fold	Mean (SD)
MAE <sup>b</sup>	11.9191	12.0483	11.8336	10.9581	11.2793	11.6077 (0.4669)
RMSE <sup>c</sup>	16.0396	24.9463	17.6053	16.0877	15.4815	18.0321 (3.9448)
$R^2$ <sup>d</sup>	0.5291	0.2779	0.5177	0.5396	0.5010	0.4731 (0.1100)

<sup>a</sup>MLP: multilayer perceptron.

<sup>b</sup>MAE: mean absolute error.

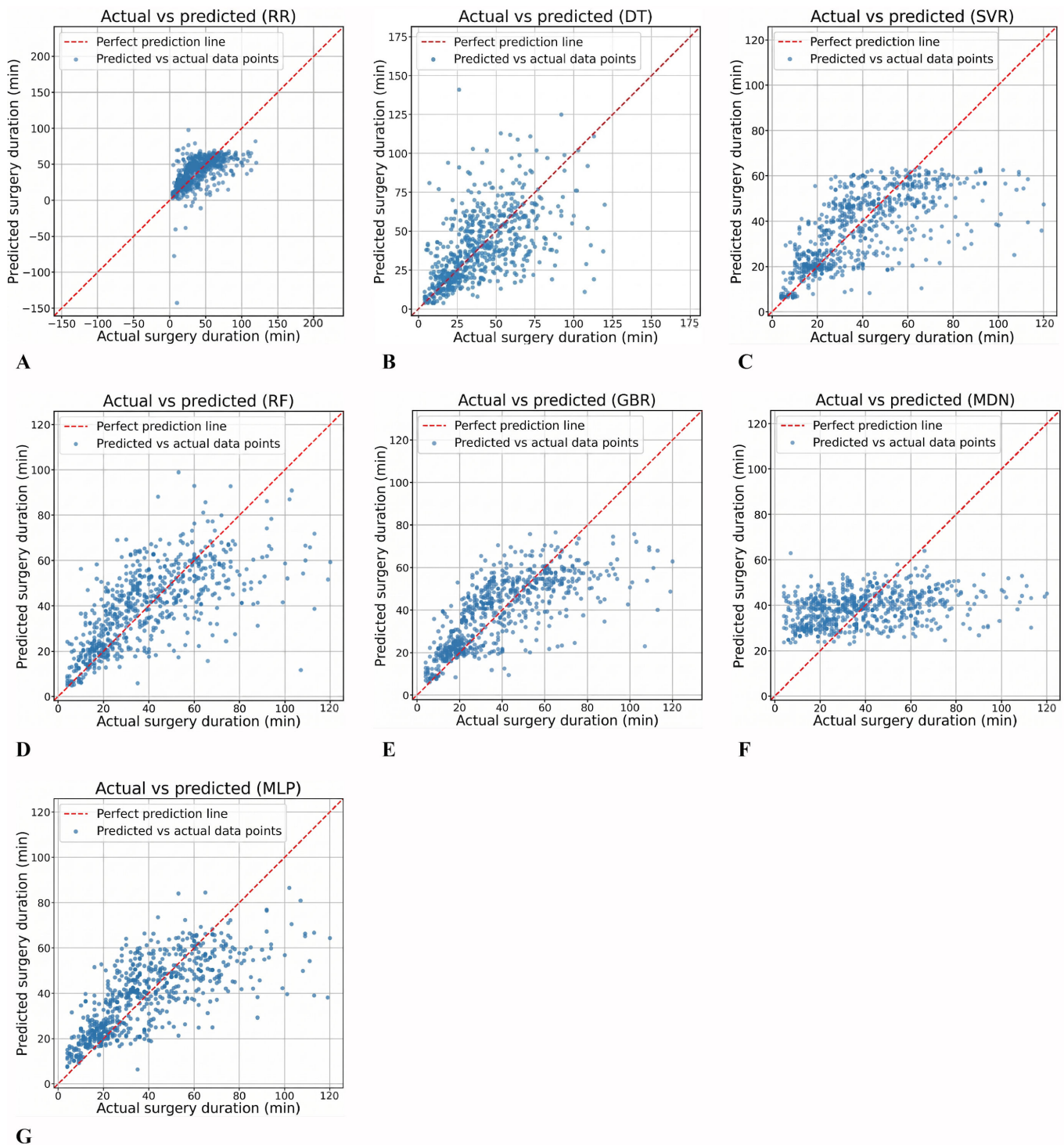
<sup>c</sup>RMSE: root mean squared error.

<sup>d</sup> $R^2$ : coefficient of determination.

The scatterplot further visualizes the predictive performance of the models, where the red-dashed line represents the ideal prediction—that is, the predicted values equal the true values. The blue scatter points show the distribution of predicted versus actual values in the test set; the closer the points are to the red line, the more accurate the model's predictions. Scatterplots for all models are shown in Figure 3. It can be observed that the points predicted by RR and DT

are highly dispersed (with the largest range on the horizontal axis). Additionally, the ML-MDN model's predicted points also deviate notably from the red line, consistent with its poorer performance metrics reported earlier. In contrast, the remaining models show less difference, with the multihead MLP exhibiting the most uniform scatter distribution, indicating more stable performance.

**Figure 3.** Scatterplots of different models on the test dataset. (A) Ridge regression (RR), (B) decision tree (DT), (C) support vector regression (SVR), (D) random forest (RF), (E) gradient boosting regression (GBR), (F) machine learning–mixture density network (ML-MDN), (G) multihead multilayer perceptron (MLP).



## Ablation Study

To demonstrate the effectiveness of the proposed multihead MLP architecture, detailed ablation experiments were conducted on its components. Four different model frameworks were tested separately: single-head MLP with categorical or numerical feature vectors, singlehead MLP with text feature vectors, single-head MLP with concatenated categorical or numerical and text feature vectors, and the multihead MLP. The detailed evaluation results are shown

in Table 5. The experimental results indicate that directly concatenating categorical or numerical feature vectors with text feature vectors causes the single-head MLP to be biased toward the higher-dimensional text features, resulting in degraded performance, which is worse than using only the categorical or numerical features. In contrast, the multihead MLP effectively integrates heterogeneous data, leveraging the strengths of features with different dimensions, and achieves the best performance.

**Table 5.** Ablation study on the proposed model architectures and feature combinations.

Different model architectures	MAE <sup>a</sup>	RMSE <sup>b</sup>	R <sup>2c</sup>
Single-head MLP + categorical/numeric vectors	11.9365	16.4067	0.4903
Single-head MLP <sup>d</sup> + text vectors	12.7383	17.0647	0.4486
Single-head MLP + categorical/numeric vectors + text vectors	12.3874	16.9184	0.4581
Multihead MLP + categorical/numeric vectors + text vectors	11.3889	15.5808	0.5404

<sup>a</sup>MAE: mean absolute error.

<sup>b</sup>RMSE: root mean squared error.

<sup>c</sup>R<sup>2</sup>: coefficient of determination.

<sup>d</sup>MLP: multilayer perceptron.

## Comparison of Performance Using Different Text-Embedding Methods

To further investigate the impact of different text-embedding strategies on model prediction performance, this study designed 2 additional text embedding methods for comparison. The first method uses a traditional tokenization embedding combined with an LSTM network. Specifically, the raw text undergoes preprocessing steps, such as the removal of newline characters and extra whitespace for normalization; the Keras tokenizer layer is then used to tokenize the text and build a vocabulary. The text sequences are padded or truncated to a fixed length and passed through an embedding layer to map tokens into low-dimensional dense vectors. Since the embedding layer only performs initial vector conversion without contextual understanding or complex semantic modeling, the subsequent LSTM layer processes these embedded token vectors to output a semantic vector representing the entire text sequence. The second method directly uses the ZHIPU AI Embedding-3 large model

to embed the entire text segment without prior tokenization or Qwen3-based structured preprocessing. This approach treats the text as a whole, aiming to evaluate the large pretrained model's capacity for text comprehension without explicit preprocessing.

In this experiment, the classification or numerical feature branch was removed, and only the text vectors were used as model input to better assess the influence of different text-embedding strategies on model performance. The results, shown in [Table 6](#), indicate that the method based on the ZHIPU AI Embedding-3 large model outperforms the traditional tokenization + LSTM approach. Furthermore, the Qwen3-based structured preprocessing of the text enhances the embedding model's performance compared to the direct embedding of unprocessed full text. These findings strongly support the superiority of our proposed LLM-based unstructured text extraction and embedding method in predictive tasks.

**Table 6.** Performance comparison of different text-embedding methods.

Text-embedding methods	MAE <sup>a</sup>	RMSE <sup>b</sup>	R <sup>2c</sup>
Tokenizer + embedding layer (Keras) + LSTM <sup>d</sup>	18.7119	22.9938	-0.0010
ZHIPU Embedding-3 model (full text)	12.8774	17.2793	0.4347
ZHIPU Embedding-3 model (structured text)	12.6437 <sup>e</sup>	17.0410 <sup>e</sup>	0.4502 <sup>e</sup>

<sup>a</sup>MAE: mean absolute error.

<sup>b</sup>RMSE: root mean squared error.

<sup>c</sup>R<sup>2</sup>: coefficient of determination.

<sup>d</sup>LSTM: long short-term memory.

<sup>e</sup>These values in the last row indicate the best performance among all tested methods across each evaluation metric.

## Feature Importance Evaluation

Permutation importance was used to assess the contribution of each feature within the multihead MLP model. The ranking of feature importance is presented in [Multimedia Appendix 1](#). It is evident that disrupting the feature order of the primary surgical procedure leads to the largest increase in the model's MAE, indicating that the primary surgical procedure is the most critical feature by a significant margin compared to others. Following this, the surgical plan and preoperative diagnosis rank next in importance, while case characteristics and preoperative preparation, including the unstructured text of the patient's prepuce description, exhibit the lowest importance.

## Discussion

### Principal Findings

This study demonstrates that an end-to-end surgical duration prediction model integrating heterogeneous multisource data, including structured patient and surgeon characteristics alongside unstructured EMR text, can significantly improve prediction accuracy for pediatric urology procedures. The proposed multihead MLP architecture, which processes categorical, numerical, and textual features through dedicated input branches, effectively addresses the challenge of heterogeneous data fusion and achieves state-of-the-art performance. In comprehensive comparisons against the "average" baseline and 6 established surgical duration

prediction methods [23,37-41], our model attains the best results across all metrics (MAE=11.39 min; RMSE=15.58;  $R^2=0.54$ ), confirming its advantage in this specialized clinical setting. This improvement is of clear clinical significance. In the dataset used for this study, each operating room typically handles 5 to 8 surgeries per day. Consequently, on a daily cumulative basis, the proposed multihead MLP model reduces the total prediction error by 38.3 to 61.4 minutes compared to the naive “average” baseline. Given that the mean surgical duration is only 38.6 (SD 24.9) minutes, this improvement is substantial, which implies the potential to schedule 1 to 2 additional surgeries, or crucially, to prevent the cancellation of the last patient due to schedule overruns. Furthermore, compared to the traditional ML-MDN model, our approach reduces the daily cumulative error by 29.4 to 47.1 minutes, which is roughly equivalent to the duration of 1 full surgery. Thus, the proposed model directly enhances operating room turnover efficiency and schedule predictability.

### ***Advantages of Patient-Physician Heterogeneous Feature Fusion***

Experimental results indicate that our multisource feature fusion strategy yields significant accuracy improvements over single-source models. Compared to models using only categorical or numerical features or only textual features, we achieve MAE reductions of 4.6% and 10.6%, respectively. The multihead MLP also outperforms the conventional single-head MLP approach by reducing MAE by 8.1%. This enhancement stems from our architecture’s ability to prevent high-dimensional text vectors from dominating the training process at the expense of lower-dimensional features, such as patient age and weight. By providing separate subnetworks for each feature type, our model enables parallel learning across different dimensions, facilitating deeper feature interactions and synergistic enhancement. This design offers valuable methodological insights for addressing similar heterogeneous data integration challenges in medical applications.

### ***Automated EMR Text Processing Pipeline***

Our second major contribution is an automated EMR text processing pipeline leveraging large-scale pretrained models. This approach demonstrates several advantages over traditional tokenization with LSTM networks, including immediate usability, multilingual support, and superior handling of long text inputs. Furthermore, our structured field extraction preprocessing method yields higher-quality embeddings compared to direct processing of raw text segments, confirming the effectiveness of our automated approach in capturing deep semantic relationships within clinical narratives.

### ***Advantages Over Traditional Methods***

In the Results section, we observed that the traditional ML-MDN model (LSTM + MLP) performed significantly worse than our proposed multihead MLP model, falling behind even the simpler RR baseline. This performance

discrepancy can be attributed to 3 primary factors: first, the difference in text-encoding mechanisms. ML-MDN relies on LSTMs, which process text sequentially and often struggle with long-range dependencies inherent in verbose medical records. In contrast, Qwen uses a self-attention mechanism, allowing the model to capture global semantic relationships simultaneously and ensuring that critical clinical cues are effectively integrated regardless of their position. This superiority of LLM-based encoding is further corroborated by our ablation study in the Results section. The second factor is the divergence in feature processing strategies. ML-MDN applies a uniform LSTM encoding to all unstructured text fields (including surgical plans and case characteristics). Conversely, our approach adopts a hybrid strategy: we reserve LLM-based encoding exclusively for verbose narrative fields (eg, case characteristics) while using precise multilabel encoding for semistructured concepts, such as surgical plans. The fact that the RR model using our proposed hybrid feature set outperformed ML-MDN serves as strong evidence for the effectiveness of this feature engineering strategy. The third factor is architectural differences. While ML-MDN relies on the simple concatenation of all features, our model uses a multihead input structure. As discussed in the Methods section, this design better handles heterogeneous data, further widening the performance gap.

### ***Key Predictors and Clinical Insights***

Permutation importance analysis identifies the primary surgical procedure as the most influential predictor, aligning with clinical experience that the specific type of surgery is the fundamental determinant of procedural duration. The surgical plan and preoperative diagnosis demonstrated secondary importance, reflecting their roles as key indicators of surgical complexity. Furthermore, the lead surgeon’s years of practice and annual surgical volume also exhibited high predictive value; this finding underscores the necessity of incorporating surgeon-specific metrics into operating room scheduling. Interestingly, textual case descriptions displayed relatively lower importance, likely attributable to the clinical tendency to use standardized language when documenting similar conditions. For instance, descriptions of conditions, such as “small penile appearance” typically adhere to consistent phrasing patterns, limiting their discriminative power. This observation suggests that refining clinical documentation practices could offer opportunities to enhance future predictive modeling.

### ***Limitations and Further Research Directions***

A primary limitation of this research stems from its reliance on data obtained from a single tertiary pediatric hospital. While the dataset exhibits strong internal consistency and an adequate sample size, the single-center provenance may nonetheless constrain the model’s generalizability across other health care settings, particularly where disparities exist in clinical workflows, surgical practices, or electronic health record documentation standards. Such limitations are frequently encountered in medical artificial intelligence applications. Future efforts could involve validating model

robustness through multicenter collaborations and exploring approaches, such as domain adaptation or lightweight transfer learning. For instance, fine-tuning models with limited annotated data from target institutions improve applicability across diverse clinical environments.

Moreover, the current predictive model relies exclusively on preoperative data, whereas actual operative duration is frequently influenced by intraoperative factors that cannot be anticipated. For example, the incidental identification of contralateral cryptorchidism may necessitate a shift from a unilateral to a bilateral procedure, while unforeseen events, such as anatomical anomalies or intraoperative bleeding, can extend surgical time. Such dynamic intraoperative developments fall outside the scope of static preoperative variables and may introduce deviations in predictions. To this end, we plan to build a 2-stage prediction framework in future work: the first stage generates baseline estimates using preoperative data, and the second stage dynamically corrects these estimates by integrating real-time information, such as laparoscopic video streams, anesthesia records, and instrument usage logs, after the operation begins. This approach has the potential to maintain the efficiency of preoperative planning while improving the adaptability and precision of resource allocation during surgery.

## Conclusion

This study tackles the clinical challenge of surgical duration prediction in pediatric urology by introducing a specialized predictive model that leverages an innovative multimodal data fusion approach. Our findings demonstrate that combining LLM-based text processing with a multihead MLP

architecture successfully integrates structured clinical data with unstructured medical narratives, leading to a significant improvement in prediction accuracy. The model outperforms existing methods across key performance metrics, including MAE and RMSE, confirming the value of heterogeneous feature collaboration in predictive modeling. Feature importance analysis identifies critical predictors, such as the primary surgical procedure, surgical plan, and preoperative diagnosis, offering evidence-based insights for clinical resource allocation. Additionally, sensitivity analysis underscores the importance of pediatric-specific clinical characteristics, including developmental stage variations and urinary tract malformation severity, reinforcing the need for specialty-tailored modeling approaches. These results provide reliable decision-support for operating room scheduling and establish a methodological framework for surgical duration prediction in other specialized fields.

While the study yields promising results, certain limitations must be acknowledged, particularly the reliance on single-institution data. Future research should prioritize multicenter data collection to enhance model generalizability, investigate dynamic prediction methods that incorporate real-time intraoperative data, and develop more interpretable tools for analyzing feature interactions. Such refinements would further strengthen the model's applicability in complex clinical settings. Overall, this work presents a practical and effective technological solution for intelligent surgical management in pediatric subspecialties, with meaningful implications for optimizing operating room efficiency and health care resource use.

---

## Acknowledgments

The authors have not used generative artificial intelligence tools at any stage in the preparation of this manuscript.

---

## Funding

This study was partially supported by the Key R&D Program of Zhejiang (2023C03101) and the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2024KY1176).

---

## Data Availability

Data that support the findings of the study are available upon request from the corresponding author.

---

## Authors' Contributions

Conceptualization: YZ

Methodology: RL, YS, ZZ

Software: RL, ZZ

Validation: LC, JH

Formal analysis: GC

Visualization: RL, YS

Funding acquisition: YZ, GY

Writing—manuscript: YZ, RL, YS, ZZ

Writing—review and editing: YZ, RL, YS, ZZ

Supervision: ZZ, GY

All authors have read and agreed to the published version of the manuscript.

---

## Conflicts of Interest

None declared.

---

## Multimedia Appendix 1

Feature importance ranking of different variables in the proposed surgical duration prediction model.

[\[PNG File \(Portable Network Graphics File\), 72 KB-Multimedia Appendix 1\]](#)

## References

1. MacMillan L, Madura GM, Elliot M, et al. What affects operating room turnover time? A systematic review and mapping of the evidence. *Surgery*. May 2025;181:109263. [doi: [10.1016/j.surg.2025.109263](https://doi.org/10.1016/j.surg.2025.109263)] [Medline: [40054053](https://pubmed.ncbi.nlm.nih.gov/40054053/)]
2. Wu Q, Xie N, Shao Y. Day surgery appointment scheduling with patient preferences and stochastic operation duration. *Technol Health Care*. 2021;29(4):697-708. [doi: [10.3233/THC-192086](https://doi.org/10.3233/THC-192086)] [Medline: [33386830](https://pubmed.ncbi.nlm.nih.gov/33386830/)]
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*. Apr 2008;106(4):1232-1241. [doi: [10.1213/ane.0b013e318164f0d5](https://doi.org/10.1213/ane.0b013e318164f0d5)] [Medline: [18349199](https://pubmed.ncbi.nlm.nih.gov/18349199/)]
4. Strum DP, May JH, Vargas LG. Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *Anesthesiology*. Apr 2000;92(4):1160-1167. [doi: [10.1097/0000542-200004000-00035](https://doi.org/10.1097/0000542-200004000-00035)] [Medline: [10754637](https://pubmed.ncbi.nlm.nih.gov/10754637/)]
5. 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*. Dec 2010;145(12):1165-1170. [doi: [10.1001/archsurg.2010.255](https://doi.org/10.1001/archsurg.2010.255)] [Medline: [21173290](https://pubmed.ncbi.nlm.nih.gov/21173290/)]
6. Gabriel RA, Harjai B, Simpson S, Goldhaber N, Curran BP, Waterman RS. Machine learning-based models predicting outpatient surgery end time and recovery room discharge at an ambulatory surgery center. *Anesth Analg*. Jul 1, 2022;135(1):159-169. [doi: [10.1213/ANE.0000000000006015](https://doi.org/10.1213/ANE.0000000000006015)] [Medline: [35389380](https://pubmed.ncbi.nlm.nih.gov/35389380/)]
7. Denton B, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Manag Sci*. Feb 2007;10(1):13-24. [doi: [10.1007/s10729-006-9005-4](https://doi.org/10.1007/s10729-006-9005-4)] [Medline: [17323652](https://pubmed.ncbi.nlm.nih.gov/17323652/)]
8. Pandit JJ. Rational planning of operating lists: a prospective comparison of “booking to the mean” vs. “probabilistic case scheduling” in urology. *Anaesthesia*. May 2020;75(5):642-647. [doi: [10.1111/anae.14958](https://doi.org/10.1111/anae.14958)] [Medline: [31867710](https://pubmed.ncbi.nlm.nih.gov/31867710/)]
9. Park HS, Kim SH, Bong MR, et al. Optimization of the operating room scheduling process for improving efficiency in a tertiary hospital. *J Med Syst*. Aug 15, 2020;44(9):1-7. [doi: [10.1007/s10916-020-01644-0](https://doi.org/10.1007/s10916-020-01644-0)] [Medline: [32803733](https://pubmed.ncbi.nlm.nih.gov/32803733/)]
10. Zhang B, Su Q. Dynamic surgery scheduling based on an improved genetic algorithm. *J Healthc Eng*. 2021;2021:1559050. [doi: [10.1155/2021/1559050](https://doi.org/10.1155/2021/1559050)] [Medline: [34853667](https://pubmed.ncbi.nlm.nih.gov/34853667/)]
11. Dexter F, Macario A, Epstein RH, Ledolter J. Validity and usefulness of a method to monitor surgical services' average bias in scheduled case durations. *Can J Anaesth*. Nov 2005;52(9):935-939. [doi: [10.1007/BF03022054](https://doi.org/10.1007/BF03022054)] [Medline: [16251558](https://pubmed.ncbi.nlm.nih.gov/16251558/)]
12. Dexter EU, Dexter F, Masursky D, Kasproicz KA. Prospective trial of thoracic and spine surgeons' updating of their estimated case durations at the start of cases. *Anesth Analg*. Apr 1, 2010;110(4):1164-1168. [doi: [10.1213/ANE.0b013e3181cd6eb9](https://doi.org/10.1213/ANE.0b013e3181cd6eb9)] [Medline: [20145282](https://pubmed.ncbi.nlm.nih.gov/20145282/)]
13. Nasiri MM, Shakouhi F, Jolai F. A fuzzy robust stochastic mathematical programming approach for multi-objective scheduling of the surgical cases. *OPSEARCH*. Sep 2019;56(3):890-910. [doi: [10.1007/s12597-019-00379-y](https://doi.org/10.1007/s12597-019-00379-y)]
14. Vali-Siar MM, Gholami S, Ramezani R. Multi-period and multi-resource operating room scheduling under uncertainty: a case study. *Comput Ind Eng*. Dec 2018;126:549-568. [doi: [10.1016/j.cie.2018.10.014](https://doi.org/10.1016/j.cie.2018.10.014)]
15. Erdem E, Qu X, Shi J. Rescheduling of elective patients upon the arrival of emergency patients. *Decis Support Syst*. Dec 2012;54(1):551-563. [doi: [10.1016/j.dss.2012.08.002](https://doi.org/10.1016/j.dss.2012.08.002)]
16. Tuwatananurak JP, Zadeh S, Xu X, et al. Machine learning can improve estimation of surgical case duration: a pilot study. *J Med Syst*. Jan 17, 2019;43(3):1-7. [doi: [10.1007/s10916-019-1160-5](https://doi.org/10.1007/s10916-019-1160-5)] [Medline: [30656433](https://pubmed.ncbi.nlm.nih.gov/30656433/)]
17. Tankard K, Acciavatti TD, Vacanti JC, et al. Contributors to operating room underutilization and implications for hospital administrators. *Health Care Manag (Frederick)*. 2018;37(2):118-128. [doi: [10.1097/HCM.0000000000000214](https://doi.org/10.1097/HCM.0000000000000214)] [Medline: [29578931](https://pubmed.ncbi.nlm.nih.gov/29578931/)]
18. Abdelrazig Merghani AM, Ahmed Esmail AK, Mubarak Osman AME, Abdelfrag Mohamed NA, Mohamed Ali Shentour SM, Abdelrazig Merghani SM. The role of machine learning in management of operating room: a systematic review. *Cureus*. Feb 2025;17(2):e79400. [doi: [10.7759/cureus.79400](https://doi.org/10.7759/cureus.79400)] [Medline: [40125180](https://pubmed.ncbi.nlm.nih.gov/40125180/)]
19. Macario A. What does one minute of operating room time cost? *J Clin Anesth*. Jun 2010;22(4):233-236. [doi: [10.1016/j.jclinane.2010.02.003](https://doi.org/10.1016/j.jclinane.2010.02.003)] [Medline: [20522350](https://pubmed.ncbi.nlm.nih.gov/20522350/)]
20. Abbas A, Mosseri J, Lex JR, et al. Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty. *Int J Med Inform*. Feb 2022;158:104670. [doi: [10.1016/j.ijmedinf.2021.104670](https://doi.org/10.1016/j.ijmedinf.2021.104670)] [Medline: [34971918](https://pubmed.ncbi.nlm.nih.gov/34971918/)]
21. Gabriel RA, Harjai B, Simpson S, et al. An ensemble learning approach to improving prediction of case duration for spine surgery: algorithm development and validation. *JMIR Perioper Med*. Jan 26, 2023;6:e39650. [doi: [10.2196/39650](https://doi.org/10.2196/39650)] [Medline: [36701181](https://pubmed.ncbi.nlm.nih.gov/36701181/)]

22. DelliCarpini G, Passano B, Yang J, et al. Utilization of machine learning models to more accurately predict case duration in primary total joint arthroplasty. *J Arthroplasty*. May 2025;40(5):1185-1191. [doi: [10.1016/j.arth.2024.10.100](https://doi.org/10.1016/j.arth.2024.10.100)] [Medline: [39477036](https://pubmed.ncbi.nlm.nih.gov/39477036/)]
23. Jiao Y, Xue B, Lu C, Avidan MS, Kannampallil T. Continuous real-time prediction of surgical case duration using a modular artificial neural network. *Br J Anaesth*. May 2022;128(5):829-837. [doi: [10.1016/j.bja.2021.12.039](https://doi.org/10.1016/j.bja.2021.12.039)] [Medline: [35090725](https://pubmed.ncbi.nlm.nih.gov/35090725/)]
24. Martinez O, Martinez C, Parra CA, Rugeles S, Suarez DR. Machine learning for surgical time prediction. *Comput Methods Programs Biomed*. Sep 2021;208:106220. [doi: [10.1016/j.cmpb.2021.106220](https://doi.org/10.1016/j.cmpb.2021.106220)] [Medline: [34161848](https://pubmed.ncbi.nlm.nih.gov/34161848/)]
25. Babayoff O, Shehory O, Shahoha M, Sasportas R, Weiss-Meilik A. Surgery duration: optimized prediction and causality analysis. *PLoS One*. 2022;17(8):e0273831. [doi: [10.1371/journal.pone.0273831](https://doi.org/10.1371/journal.pone.0273831)] [Medline: [36037243](https://pubmed.ncbi.nlm.nih.gov/36037243/)]
26. Miller LE, Goedicke W, Crowson MG, Rathi VK, Naunheim MR, Agarwala AV. Using machine learning to predict operating room case duration: a case study in otolaryngology. *Otolaryngol Head Neck Surg*. Feb 2023;168(2):241-247. [doi: [10.1177/01945998221076480](https://doi.org/10.1177/01945998221076480)] [Medline: [35133897](https://pubmed.ncbi.nlm.nih.gov/35133897/)]
27. Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: algorithm development and validation study. *JMIR AI*. Sep 8, 2023;2(1):e44909. [doi: [10.2196/44909](https://doi.org/10.2196/44909)] [Medline: [38875567](https://pubmed.ncbi.nlm.nih.gov/38875567/)]
28. Riahi V, Hassanzadeh H, Khanna S, et al. Improving preoperative prediction of surgery duration. *BMC Health Serv Res*. Dec 2, 2023;23(1):1343. [doi: [10.1186/s12913-023-10264-6](https://doi.org/10.1186/s12913-023-10264-6)] [Medline: [38042831](https://pubmed.ncbi.nlm.nih.gov/38042831/)]
29. 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*. Apr 1, 2021;156(4):315-321. [doi: [10.1001/jamasurg.2020.6361](https://doi.org/10.1001/jamasurg.2020.6361)] [Medline: [33502448](https://pubmed.ncbi.nlm.nih.gov/33502448/)]
30. Huang L, Chen X, Liu W, Shih PC, Bao J. Automatic surgery and anesthesia emergence duration prediction using artificial neural networks. *J Healthc Eng*. 2022;2022:2921775. [doi: [10.1155/2022/2921775](https://doi.org/10.1155/2022/2921775)] [Medline: [35463687](https://pubmed.ncbi.nlm.nih.gov/35463687/)]
31. Park JB, Roh GH, Kim K, Kim HS. Development of predictive model of surgical case durations using machine learning approach. *J Med Syst*. Jan 14, 2025;49(1):8. [doi: [10.1007/s10916-025-02141-y](https://doi.org/10.1007/s10916-025-02141-y)] [Medline: [39808376](https://pubmed.ncbi.nlm.nih.gov/39808376/)]
32. Frassini E, Vijfinkel TS, Butler RM, van der Elst M, Hendriks BHW, van den Dobbelaars JJ. Deep learning methods for clinical workflow phase-based prediction of procedure duration: a benchmark study. *Comput Assist Surg (Abingdon)*. Dec 2025;30(1):2466426. [doi: [10.1080/24699322.2025.2466426](https://doi.org/10.1080/24699322.2025.2466426)] [Medline: [39992712](https://pubmed.ncbi.nlm.nih.gov/39992712/)]
33. Ramamurthi A, Neupane B, Deshpande P, et al. Applying large language models for surgical case length prediction. *JAMA Surg*. Aug 1, 2025;160(8):894-902. [doi: [10.1001/jamasurg.2025.2154](https://doi.org/10.1001/jamasurg.2025.2154)] [Medline: [40632526](https://pubmed.ncbi.nlm.nih.gov/40632526/)]
34. Dexter F, Epstein RH. Operating room efficiency and scheduling. *Curr Opin Anaesthesiol*. Apr 2005;18(2):195-198. [doi: [10.1097/01.aco.0000162840.02087.15](https://doi.org/10.1097/01.aco.0000162840.02087.15)] [Medline: [16534338](https://pubmed.ncbi.nlm.nih.gov/16534338/)]
35. May JH, Spangler WE, Strum DP, Vargas LG. The surgical scheduling problem: current research and future opportunities. *Prod Oper Manag*. May 2011;20(3):392-405. [doi: [10.1111/j.1937-5956.2011.01221.x](https://doi.org/10.1111/j.1937-5956.2011.01221.x)]
36. van Eijk RPA, van Veen-Berkx E, Kazemier G, Eijkemans MJC. Effect of individual surgeons and anesthesiologists on operating room time. *Anesth Analg*. Aug 2016;123(2):445-451. [doi: [10.1213/ANE.0000000000001430](https://doi.org/10.1213/ANE.0000000000001430)] [Medline: [27308953](https://pubmed.ncbi.nlm.nih.gov/27308953/)]
37. Zhao B, Waterman RS, Urman RD, Gabriel RA. A machine learning approach to predicting case duration for robot-assisted surgery. *J Med Syst*. Jan 5, 2019;43(2):32. [doi: [10.1007/s10916-018-1151-y](https://doi.org/10.1007/s10916-018-1151-y)] [Medline: [30612192](https://pubmed.ncbi.nlm.nih.gov/30612192/)]
38. 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*. Aug 2017;4(1):35-52. [doi: [10.1007/s41060-017-0055-0](https://doi.org/10.1007/s41060-017-0055-0)]
39. Li X, Zhan Y, Ni J, et al. Multimodal feature analysis for surgery duration prediction. Presented at: Proceedings of the 2024 16th International Conference on Bioinformatics and Biomedical Technology (ICBBT '24); May 24-26, 2024:188-195; Chongqing, China. [doi: [10.1145/3674658.3674688](https://doi.org/10.1145/3674658.3674688)]
40. Bellini V, Guzzon M, Bigliardi B, Mordonini M, Filippelli S, Bignami E. Artificial intelligence: a new tool in operating room management. role of machine learning models in operating room optimization. *J Med Syst*. Dec 10, 2019;44(1):20. [doi: [10.1007/s10916-019-1512-1](https://doi.org/10.1007/s10916-019-1512-1)] [Medline: [31823034](https://pubmed.ncbi.nlm.nih.gov/31823034/)]
41. Bartek MA, Saxena RC, Solomon S, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. *J Am Coll Surg*. Oct 2019;229(4):346-354. [doi: [10.1016/j.jamcollsurg.2019.05.029](https://doi.org/10.1016/j.jamcollsurg.2019.05.029)] [Medline: [31310851](https://pubmed.ncbi.nlm.nih.gov/31310851/)]

## Abbreviations

- DT:** decision tree regression
- EMR:** electronic medical record
- GBR:** gradient boosting regression tree
- LLM:** large language model

**LSTM:** long short-term memory  
**MAE:** mean absolute error  
**ML-MDN:** machine learning–mixture density network  
**MLP:** multilayer perceptron  
**RF:** random forest regression  
**RMSE:** root mean square error  
**RR:** ridge regression  
**SVR:** support vector regression

*Edited by Arriel Benis; peer-reviewed by Meng-Hsun Tsai, Takashi Ashihara; submitted 13.Aug.2025; final revised version received 06.Mar.2026; accepted 06.Mar.2026; published 28.Apr.2026*

*Please cite as:*

*Zhao Y, Lin R, Sun Y, Chen L, Huang J, Chen G, Zhu Z, Yu G*

*Predicting Pediatric Urological Surgery Duration Through Multimodal Patient-Physician Feature Fusion: Deep Learning Framework Incorporating Clinical Text Embedding*

*JMIR Med Inform 2026;14:e82329*

*URL: <https://medinform.jmir.org/2026/1/e82329>*

*doi: [10.2196/82329](https://doi.org/10.2196/82329)*

© Yonggen Zhao, Ruoge Lin, Yiyang Sun, Lingdong Chen, Jian Huang, Guangjie Chen, Zhu Zhu, Gang Yu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 28.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.