International Journal of Medical Informatics 214 (2026) 106416

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Research paper

Benchmarking text encoding strategies in multimodal clinical data for surgical case duration prediction

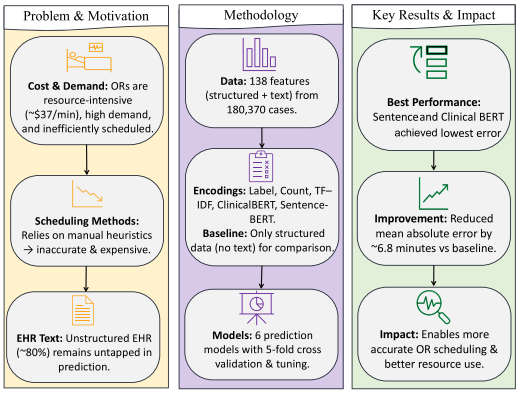
Mohammad Noorchenarboo [a], Michelle Kwong [b,c], Ahmad Elnahas [d,e], Jeff Hawel [d], Christopher M. Schlachta [d], Katarina Grolinger [a,*]

[a] Department of Electrical and Computer Engineering, Western University, London, Canada
[b] Department of Anesthesiology and Pain Medicine, University of Alberta, Edmonton, Canada
[c] Department of Medicine, Western University, London, Canada
[d] Department of Surgery, Western University, London, Canada
[e] London Health Sciences Centre, Western University, 339 Windermere Rd, London, ON, N6A 5A5, Canada

HIGHLIGHTS

- We present a multimodal pipeline for surgical case duration prediction.
- Structured and unstructured data are integrated, and five text encodings are compared.
- Six ML models were evaluated on 180,370 cases across three tertiary care hospitals.
- Sentence-BERT and ClinicalBERT achieved comparable MAE ≈ 26.4 min and SMAPE ≈ 21.6%.
- Clinical text data significantly reduced prediction error compared to structured data alone.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Surgical case duration prediction
Multimodal clinical data
Text encoding strategies
ClinicalBERT
Sentence-BERT
Operating room scheduling
Machine learning

ABSTRACT

Background: Operating rooms (ORs) are highly resource-intensive, yet surgical case duration is often estimated using heuristics that are prone to errors. While machine learning models based on structured perioperative data improve accuracy, unstructured clinical text remains underutilized despite containing valuable contextual details.
Objective: To benchmark classical and contextual text encoding strategies, combined with structured perioperative data, for predicting surgical case durations.
Methods: We retrospectively analyzed 180,370 elective surgical cases from three tertiary care hospitals (2015–2020). Structured variables such as age, sex, BMI, ASA score, and case service were combined with unstructured text features (procedure descriptions), which were encoded using five different methods (label encoding, count vectorization, TF–IDF, ClinicalBERT, Sentence-BERT). We trained a diverse set of machine learning models including linear regression, tree-based ensembles, and neural networks and evaluated predictive accuracy using standard error metrics with cross-validation.

* Corresponding author.
Email address: kgroling@uwo.ca (K. Grolinger).

https://doi.org/10.1016/j.ijmedinf.2026.106416
Received 3 September 2025; Received in revised form 20 March 2026; Accepted 29 March 2026
Available online 1 April 2026

Results: Adding unstructured clinical text to structured perioperative variables improved prediction accuracy across all models. Contextual embeddings consistently outperformed structured-only and traditional text encodings. Sentence-BERT and ClinicalBERT achieved comparable best performance, reducing MAE to approximately 26.4 minutes and SMAPE to 21.6%, with R^2 of 0.86; neither encoder was statistically superior to the other ($p > 0.82$). Improvements over structured-only baselines were statistically significant ($p < 0.01$), corresponding to up to 16% reduction in prediction error. Traditional encodings (label, count, TF–IDF) provided limited benefit.
conclusion: Integrating semantically rich clinical text with structured perioperative data substantially improves surgical duration prediction. Our multimodal approach which combines structured and unstructured data with contextual embeddings, directly improves prediction accuracy, which in turn supports more reliable OR scheduling, better resource utilization, and improved patient care. Future work should incorporate additional narrative sources and interpretability techniques to support clinical adoption.

Glossary of Abbreviations

AI	Artificial Intelligence
ASA	American Society of Anesthesiologists (physical status classification)
BMI	Body Mass Index
BERT	Bidirectional Encoder Representations from Transformers
CI	Confidence Interval
CV	Cross-Validation
EHR	Electronic Health Record
EMR	Electronic Medical Record
GPU	Graphics Processing Unit
ICU	Intensive Care Unit
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
OR	Operating Room
RF	Random Forest
SMAPE	Symmetric Mean Absolute Percentage Error
TPE	Tree-structured Parzen Estimator
TF–IDF	Term Frequency–Inverse Document Frequency
XGBoost	eXtreme Gradient Boosting

1. Introduction

Operating Rooms (ORs) are among the most resource-intensive environments in healthcare, with direct costs estimated at $36–$46 per minute [1]. Since surgical services account for more than half of hospital revenue, even small inefficiencies in OR use carry substantial financial and clinical consequences [2,3]. Yet many hospitals continue to rely on experience-based scheduling practices that are difficult to optimize in the face of constrained capacity, equipment demands, and staff interdependencies [4,5].

Machine learning (ML) has been increasingly applied to address these challenges by predicting surgical durations more accurately than manual heuristics [2,6]. Models based on structured perioperative data such as demographics, procedure codes, and anesthesia times have achieved substantial accuracy gains, with ensemble and deep learning methods often outperforming traditional statistical approaches [7–9]. This advantage is grounded in the nonlinear interactions among patient physiology, procedural characteristics, and scheduling factors that linear models cannot capture—as demonstrated empirically by an artificial neural network reducing mean prediction error by more than 18 minutes relative to conventional linear approaches [6]. Real-world implementations have also shown reductions in scheduling errors [6,10,11]. However, most existing approaches rely primarily on structured variables.

This reliance overlooks the large proportion of information in Electronic Medical Records (EMRs), where nearly 80% of content consists of unstructured text [12]. Clinical narratives capture nuanced reasoning and patient-specific details that are rarely represented in structured fields, yet remain underutilized in predictive modeling [13]. Recent advances in Natural Language Processing (NLP), particularly transformer-based models such as ClinicalBERT and Sentence-BERT [14,15], generate semantically meaningful embeddings that capture contextual nuance in clinical language. These methods have demonstrated strong performance in diverse applications, including disease prediction [16], risk stratification [17], and hospital outcome forecasting [18]. In perioperative contexts, incorporating textual features from procedure descriptions and diagnoses has improved case-duration prediction, especially for rare or complex surgeries [19–21]. This advantage stems from a fundamental architectural difference between classical and contextual encoding methods. Classical approaches such as count vectorization and TF–IDF (Term Frequency–Inverse Document Frequency) operate under a bag-of-words assumption, representing text as token frequency vectors that discard word order and contextual dependencies [22]. As a result, they cannot distinguish distinct descriptions that share vocabulary, nor capture qualifiers such as ‘bilateral,’ ‘revision,’ or ‘laparoscopic’ that substantially alter operative complexity. Transformer-based models, by contrast, encode each token with respect to its full surrounding context through multi-head self-attention [23], enabling them to represent the nuanced meaning of clinical phrases. ClinicalBERT extends this by pretraining on clinical notes [14], while Sentence-BERT applies a contrastive training objective that places semantically similar sentences in proximity within the embedding space [15]—a property directly beneficial for grouping procedurally related descriptions.

Despite these advances, little is known about how different text encoding strategies and ML algorithms compare when applied to perioperative case duration forecasting. Existing studies often rely on either traditional representations such as TF–IDF or a single contextual embedding, without controlled benchmarking across approaches. Table 1 summarizes representative prior studies on surgical case duration prediction along the dimensions most relevant to our contribution. Three observations emerge from this table. First, the majority of studies rely exclusively on structured perioperative variables, with no textual features incorporated [6–11]. Second, among the studies that do incorporate clinical text [19–21], each employs a single encoding strategy without any controlled comparison of how encoding choice affects predictive accuracy. Third, dataset scales vary considerably—from under 1,000 to over 160,000 cases—and validation designs range from simple hold-out splits to prospective evaluations, making direct performance comparisons across studies unreliable. No prior study has benchmarked multiple encoding strategies—from classical bag-of-words methods to domain-specific transformer embeddings—within a single controlled pipeline applied to a large, multi-site cohort. To address this gap, we evaluate

Table 1
Summary of representative prior surgical case duration prediction studies. "Text Source" refers to unstructured clinical text used as a feature input; "–" denotes structured data only.

Study	*N*	Text Source	Encoding	Model Family	Validation
Bartek et al. [7]	46,986	–	–	Linear, XGBoost	80/20 hold-out
Tuwatananurak et al. [11]	990	–	–	GBM, RF	Prospective (3 mo.)
Jiao et al. [19]	53,783	Procedure name, diagnosis	Bag-of-words / NLP	MDN, GBRT	Hold-out + external
Yeo et al. [9]	10,021	–	–	ANN, RF, KNN	Hold-out
Adams et al. [20]	34,457	Procedure descriptions	Term counts, one-hot	Ridge	60/20/20 split
Zaribafzadeh et al. [10]	33,815	–	–	GBM	Hold-out + prospective
Kwong et al. [6]	17,246	–	–	Linear, RF, XGBoost, ANN	Hold-out
Park et al. [8]	161,176	–	–	RF, XGBoost, LightGBM, CatBoost	80/20 + 5-fold CV
Ramamurthi et al. [21]	125,493	Preoperative clinical notes	Fine-tuned LLM	GPT-4, GPT-3.5	Hold-out + external
This study	**180,370**	**Preoperative clinical notes**	**Label, Count, TF–IDF, ClinicalBERT, SentenceBERT**	**OLS, Ridge, Lasso, RF, XGBoost, ANN**	**5-fold CV + Temporal**

RF: Random Forest; GBM: Gradient Boosting Machine; MDN: Mixture Density Network; GBRT: Gradient Boosted Regression Trees; ANN: Artificial Neural Network; KNN: K-Nearest Neighbors; LLM: Large Language Model; OLS: Ordinary Least Squares; TF–IDF: Term Frequency–Inverse Document Frequency; CV: Cross-Validation; NLP: Natural Language Processing; Preop.: Preoperative; ext.: external; prosp.: prospective.

five encoding strategies—label encoding, count vectorization, TF–IDF, ClinicalBERT, and Sentence-BERT—applied to unstructured text from procedure descriptions. These features are integrated with structured perioperative variables to form multimodal inputs. Using data from over 180,000 elective surgical cases across three tertiary care hospitals, we benchmark these representations across multiple ML algorithms with rigorous statistical testing. Our contributions are threefold: (1) a standardized pipeline for combining structured and unstructured perioperative data, (2) the first large-scale comparison of traditional and transformer-based encoding strategies for surgical duration prediction, and (3) evidence that the choice of text encoding strategy and ML algorithm significantly improves surgical case duration prediction, establishing a foundation for reducing inefficiencies in operating room scheduling and ultimately supporting improved patient care delivery.

2. Methodology

This study employed a multi-stage pipeline to build and evaluate predictive models for surgical case duration prediction, incorporating both structured clinical variables and unstructured textual information. The methodology consisted of four main stages: data preprocessing, text encoding, predictive model development, and evaluation strategy. An overview of this multi-stage pipeline is shown in Fig. 1. Each component is described in the following subsections.

2.1. Data preprocessing

Existing de-identified EMR (Cerner, North Kansas City, USA) data were retrospectively collected for surgical cases that took place across three tertiary care academic hospitals in the London Health Sciences Center (London, ON, Canada) over a 5-year period. All data were housed on a secure university-based server for analysis, model development, and internal validation. The study cohort included adult patients undergoing elective surgeries between January 1, 2015, and January 1, 2020; urgent, emergent, and cancelled cases were excluded. Cohort demographics are summarized in Table 2.

The initial dataset contained 194,661 cases. After harmonizing missing-value placeholders and normalizing free-text fields, we excluded 13,445 records with incomplete timestamps and 268 with implausible durations outside the 0–2,880 minute range. Additional removals included 480 records with missing values in non-imputable fields (ASA score, operative diagnosis, sex, encounter type, and case service), 66 invalid sex entries, 30 unmapped service specialties, and 1 unmapped operating room location. In total, 14,291 cases were removed. The text field used as model input — the scheduled procedure description — is sourced from the Cerner EMR and documented by the treating clinician at the time of surgical booking, before the case begins. This field is therefore strictly preoperative and available at the point when OR duration must be estimated, making the prediction task operationally feasible in

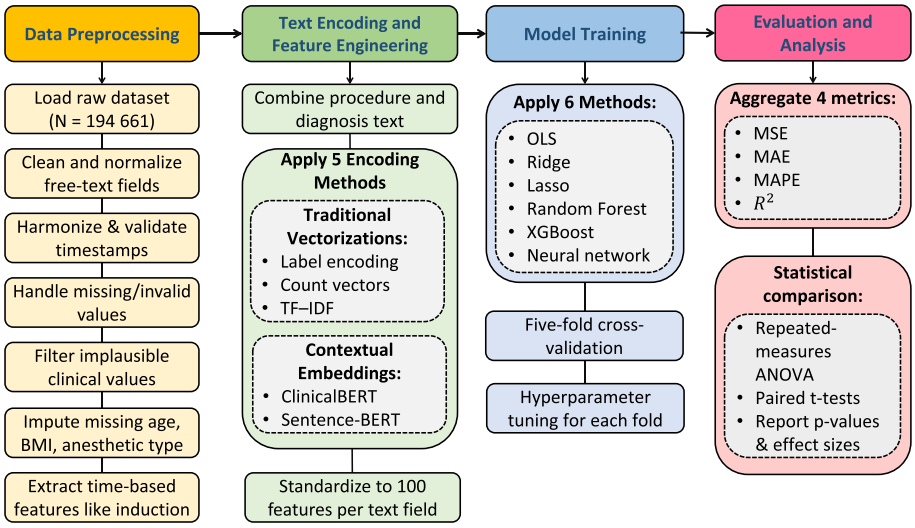


Fig. 1. Main flowchart of the methodology.

Table 2
Cohort demographics.

Characteristic	Cohort (N = 180,370)
Age	53.39 ± 18.69
Sex (M)	97,055 (53.77%)
BMI (kg/m^2)	27.91 ± 7.68
ASA Score	
I	30,394 (16.85%)
II	55,582 (30.82%)
III	68,983 (38.25%)
IV	25,424 (14.10%)
V	84 (0.05%)
Case Service	
Orthopedics	48,728 (27.02%)
General Surgery	26,418 (14.65%)
Abdominal Hernia Repair	7926 (30.00%)
Cholecystectomy	3832 (14.50%)
Ileostomy Closure	963 (3.64%)
Gastric Bypass	797 (3.02%)
Liver Resection	608 (23.01%)
Appendectomy	295 (1.12%)
Obstetrics & Gynecology	24,548 (13.61%)
Otolaryngology	16,956 (9.40%)
Urology	10,894 (6.04%)
Plastic Surgery	10,540 (5.84%)
Neurosurgery	10,333 (5.73%)
Cardiac Surgery	10,058 (5.58%)
Vascular Surgery	6253 (3.47%)
Dental Surgery	5937 (3.29%)
Thoracic Surgery	4873 (2.70%)
Ophthalmology	3531 (1.96%)
Surgical Oncology	1301 (0.72%)
Surgical Encounter Type	
Outpatient	169,348 (93.83%)
Inpatient	11,129 (6.17%)

a prospective scheduling setting. It captures the planned operative intervention as entered by the booking clinician and is consistently populated at booking across all three study hospitals, ensuring that the textual input used in this study is representative of information routinely available to schedulers.

The prediction target is the total OR occupancy time per case, defined as the elapsed time from room entry to room exit and recorded directly in the EMR. This definition mirrors the quantity that schedulers must estimate: the full duration for which an operating room is occupied by a single case, encompassing patient positioning, anesthetic induction, the surgical procedure itself, emergence, and patient transport. The scheduled duration field in the EMR is derived from the same two reference points—scheduled room entry and exit—making the two fields definitionally aligned and the prediction task operationally meaningful. The upper exclusion bound of 48 hours was established through consensus with the clinical team as the longest reasonable duration for any single elective procedure; values exceeding this limit were considered data entry errors or system artifacts rather than genuine surgical encounters. This threshold is empirically supported: the 99.9th percentile of OR occupancy time in the pre-filtered cohort was 769 minutes, and only 30 of 181,216 cases (0.02%) exceeded 2880 minutes, confirming that the filter removed extreme outliers without materially affecting the cohort. The post-filter target distribution was right-skewed, with a median of 101 minutes (IQR: 62–172 min), a mean of 135.5 ± 109.7 minutes, a 90th percentile of 283 minutes, a 95th percentile of 349 minutes, a 99th percentile of 537 minutes, and a maximum of 1810 minutes.

Residual missingness in key variables (BMI, age, anesthetic type) was imputed using a machine learning–based approach. Specifically, we trained an XGBoost regressor/classifier on cases without missing values and used the fitted models to predict missing entries. Rare categorical levels were consolidated, and temporal features (weekday, month, scheduled start hour, case order) were derived. After preprocessing, the analytic cohort comprised 180,370 elective cases across 13 surgical services. The structured feature set consisted of 17 variables (14 numeric or binary predictors and 3 categorical variables), which expanded to 38 features after fold-wise one-hot encoding of anesthetic type, case service, and surgical location.

2.2. Text encoding and feature engineering

We generated a set of 100 text features from scheduled procedure descriptions and concatenated them with 38 structured perioperative variables to form a 138-dimensional input vector, with actual surgical duration in minutes as the prediction target. Three classical encodings were applied to obtain these 100 text features: (1) label encoding, where the 99 most frequent phrases were retained as separate one-hot columns and all remaining phrases were grouped into a single “Other” column, yielding exactly 100 features; (2) count vectorization, producing unigram and bigram term-frequency matrices and selecting the top 100 tokens by raw frequencies; and (3) TF–IDF, re-weighting term counts by inverse document frequency and retaining the 100 tokens with highest average scores. Two transformer-based approaches were also included: ClinicalBERT, extracting [CLS]-token embeddings, and Sentence-BERT, generating sentence-level embeddings optimized for semantic similarity; the embedding matrix was reduced to 100 features. In all cases, text representations were concatenated with the structured features to yield the 138-dimensional inputs. Five-fold train/validation splits were generated with a fixed random seed, with encoding and dimensionality reduction performed independently within each fold.

For the transformer-based encoders, ClinicalBERT was implemented using the `emilyalsentzer/Bio_ClinicalBERT` checkpoint with tokenization truncated to a maximum of 64 tokens; the [CLS] token representation from the final hidden layer was used as the sentence embedding, yielding a 768-dimensional vector for the procedure description. Sentence-BERT was implemented using the `all-MiniLM-L6-v2` checkpoint with its default mean pooling strategy, yielding a 384-dimensional vector for the procedure description. Both models were used in inference mode with weights fully frozen; no fine-tuning was performed on the surgical dataset. To enforce a uniform feature budget across all five encoding strategies and ensure that no method received a dimensionality advantage, all encoders were evaluated at four feature dimensionality settings: 10, 50, 100, and 200 features. For classical encoders this was achieved by retaining the top n tokens by frequency or TF–IDF score; for transformer encoders, PCA was applied fold-wise (fit exclusively on training-fold embeddings and applied to the validation fold without re-fitting) reducing each embedding to n dimensions. The 100-feature setting was selected as the primary reported configuration based on the sensitivity analysis reported in Section 3.4.

2.3. Model training

We evaluated the five text-encoding strategies across six supervised regression algorithms spanning three families: (1) linear models (ordinary least squares, ridge, and lasso regression), providing interpretable baselines; (2) tree-based ensembles (Random Forest and XGBoost), which handle feature selection, nonlinear interactions, and scale invariance; and (3) a neural network (feedforward network with depth tunable from 1 to 3 layers via hyperparameter search) capable of capturing complex relationships and leveraging contextual embeddings.

Each of the 30 model–encoding combinations (six algorithms × five encodings) was trained and evaluated across five folds, yielding 150 independent runs. To eliminate any risk of data leakage, all data-driven preprocessing steps were estimated exclusively on the training fold and subsequently applied to the held-out validation fold without re-fitting. Specifically, the following steps were performed inside each fold: (1) imputation of missing values for age, BMI, and anesthetic type using XGBoost regressors and classifiers trained on complete training-fold cases only; (2) one-hot encoding of anesthetic type, case service, and

surgical location, with category sets determined solely from training-fold values; (3) vocabulary construction and vectorizer fitting for label encoding, TF–IDF, and count vectorization, each fit on training-fold procedure descriptions only and then applied to the validation fold; (4) PCA dimensionality reduction for transformer-based encodings, fit exclusively on training-fold embeddings and applied to the validation fold without re-fitting; and (5) feature scaling via min–max normalization, fit on the training fold and applied to the validation fold. Steps performed globally prior to fold splitting were entirely rule-based and deterministic, including timestamp arithmetic, fixed-dictionary categorical mappings, and outlier removal; none of these involved any statistical estimation from the data and therefore carried no leakage risk. Transformer-based embeddings (ClinicalBERT and Sentence-BERT) were computed globally as a precomputation step because the underlying models are pretrained and were not fine-tuned on this dataset: each text is encoded independently through a fixed forward pass, so the embedding of any validation case is not influenced by training cases.

Except for unregularized linear regression, all models underwent hyperparameter tuning via a Tree-structured Parzen Estimator (TPE) within each fold: tuning λ for ridge/lasso; tree depth, number of estimators, and learning rate η for ensemble methods; and hidden-layer width n, dropout rate γ, and learning rate α for the neural network. Independent searches across folds and encodings ensured fair optimization and comparison. The TPE search was run for 50 trials per model–encoding–fold combination, with 10 startup trials before the TPE sampler activated. The neural network was trained using the AdamW optimizer with a batch size of 512, weight decay tuned via Optuna, and gradient clipping with a maximum norm of 1.0. During hyperparameter search trials, each candidate configuration was trained for 30 epochs; the final model selected after tuning was trained for up to 200 epochs with early stopping (patience = 15 epochs on validation loss) and learning rate reduction on plateau (factor = 0.5, patience = 5, minimum learning rate = 10^{-6}). XGBoost used early stopping with a patience of 20 rounds in both Optuna trials and the final fit. All experiments were executed on a workstation running Windows 11 with an NVIDIA GeForce RTX 3080 GPU (CUDA 11.8). Key software versions were: Python 3.13.5, scikit-learn 1.6.1, XGBoost 3.2.0, TensorFlow 2.20.0, PyTorch 2.7.1, Transformers 4.57.3, Sentence-Transformers 5.2.3, Optuna 4.5.0, pandas 2.2.3, and NumPy 2.1.3.

2.4. Evaluation and analysis

Model performance was assessed using six regression metrics that capture complementary aspects of error and fit: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), and the coefficient of determination (R^2). Each metric was computed within a five-fold cross-validation framework, with all preprocessing confined to the training folds. Final performance values were reported as averages across folds, reducing variance due to data partitioning and providing more generalizable estimates.

The regression metrics are defined as follows, where y_i and $\hat{y}_i$ denote the observed and predicted durations for case i, and n is the number of cases:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

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

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

$$\mathrm{SMAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

To assess whether observed differences in performance across encoding strategies are statistically significant, we used a two-stage hypothesis testing framework. First, a repeated-measures ANOVA was applied for each metric to test the null hypothesis that all encodings perform equally, against the alternative that at least one differs. The observational unit is the (model, fold) combination, yielding 30 observations per encoding strategy (6 models × 5 folds); this repeated-measures structure accounts for algorithm variability by pairing observations across encodings within the same (model, fold) combination. A one-way design was chosen because the primary research question concerns the effect of encoding strategy on predictive performance; the six regression models serve as replications to obtain stable, algorithm-agnostic estimates rather than as a factor of scientific interest, and algorithm variability is controlled for through the repeated-measures pairing rather than modelled explicitly. When significance was found, we conducted pairwise paired t-tests, pairing observations by (model, fold) combination to isolate the encoding effect. All pairwise p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR-BH) procedure across all 15 comparisons. Cohen's d effect sizes are reported alongside each adjusted p-value to convey practical significance. All p-values were evaluated at the 0.05 significance level, and results are reported in Table 4 and Appendix Tables A.2–A.4.

To assess robustness to temporal distribution shift and to simulate prospective deployment conditions, we additionally evaluated all encoding strategies using an expanding-window time-series split ($k = 5$). Cases were sorted chronologically by scheduled start date and divided into five sequential folds; in each fold the training set comprised all cases up to a given cut-off date and the validation set comprised the immediately following cases, with the training window expanding and the validation window advancing chronologically at each fold. This ensures that no future case was ever used during training, directly simulating prospective deployment. Results are reported in Appendix Table A.1.

3. Results

This section presents the outcomes of our experiments, organized into three parts. We begin with an overview of hyperparameter optimization across model families and text encoding strategies. Next, we compare model performance using multiple regression metrics to assess the predictive value of each encoding method. Finally, we report results from statistical tests that quantify the significance of observed differences, highlighting the relative strengths of contextual language models in forecasting surgical case durations.

3.1. Hyperparameter optimization results

Hyperparameter search patterns varied across model families. Ridge regression typically converged on α values between 0.002–7.3, while lasso favored smaller values (0.002–0.019) for count-based encodings and slightly larger ones (up to 0.049) for transformer-based inputs. For tree-based ensembles, Random Forests generally selected 100–190 estimators with depths of 3–10, and XGBoost favored similar estimator counts with learning rates of 0.03–0.10 and depths of 3–6. These ranges illustrate how different encodings influenced the complexity of models chosen during tuning, rather than their predictive performance.

Neural networks displayed the greatest variability across encodings. Hidden layer sizes ranged from ~35 to 130 units per layer, with dropout rates spanning near-zero to ~0.25 depending on the embedding type. Transformer-based inputs often required higher regularization through dropout and narrower learning rate ranges (0.0002–0.005), whereas simpler encodings tolerated wider ranges (up to 0.007). These findings

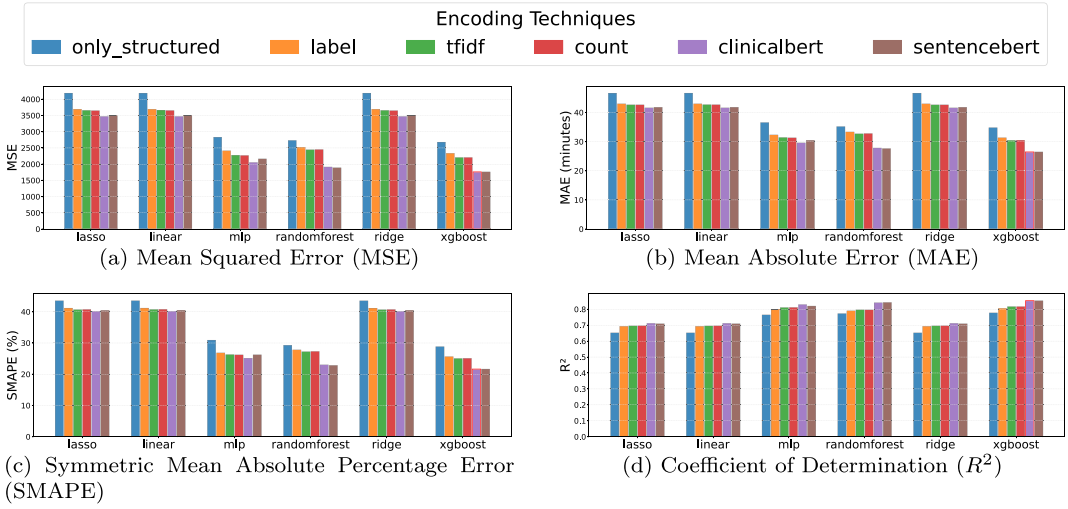


Fig. 2. Performance comparison of different models across four evaluation metrics (MSE, MAE, SMAPE, and R^2). Each subplot shows average results across folds for various encoding strategies, with a shared legend at the top. RMSE and MAPE are omitted for conciseness; complete results for all six metrics are reported in Appendix Table A.5.

suggest that as input features become semantically richer and higher dimensional, optimal configurations for both traditional and neural models shift toward deeper structures, tighter regularization, and finer-tuned learning rates. Overall, hyperparameter optimization was essential to ensure fair comparisons across encodings, as each representation imposed distinct demands on model capacity and regularization.

3.2. Model performance across encoding strategies

Fig. 2 summarizes model performance across all encoding strategies and metrics (MSE, MAE, SMAPE, and R^2), with values averaged across five cross-validation folds per model–encoding pair. RMSE and MAPE are omitted from the figure for conciseness; full results for all six metrics are provided in Appendix Table A.5. Text embeddings, particularly sentence-level representations, achieved the best overall results: both contextual encoders consistently yielded lower errors and higher R^2 scores, with the Sentence-BERT–XGBoost combination reaching the lowest MSE (1751), lowest MAE (26.4 minutes), and lowest SMAPE (21.6%), and ClinicalBERT–XGBoost achieving nearly identical values, reflecting strong predictive accuracy and stability for both contextual methods. In contrast, traditional encodings such as label encoding, count vectors, and TF–IDF underperformed, especially in linear or shallow models, while structured-only baselines consistently lagged behind multimodal models that incorporated textual features. These findings highlight the critical role of text embeddings and demonstrate that aligning richer encodings with flexible model architectures substantially improves surgical duration prediction.

3.3. Subgroup analysis by surgical service and case duration

Table 3 reports MAE (mean ± SD across five folds) for the two best-performing contextual encodings and the structured-only baseline, stratified by surgical service, case duration band, and hospital site. Prediction error varied substantially across services, ranging from 14.78 min (Ophthalmology) to 43.33 min (Neurosurgery), reflecting the higher procedural variability of complex specialties. The structured-only baseline consistently underperformed, with the largest absolute reductions from text encoding observed in Urology (24.31 vs. 37.95 min), ENT (28.67 vs. 37.39 min), and Thoracic Surgery (39.02 vs. 50.69 min). Both contextual encoders produced nearly identical service-level errors across all 13 specialties, suggesting they capture the same procedural signal at this level of aggregation.

Table 3
Subgroup analysis of prediction error (MAE, mean ± SD across five cross-validation folds) for the two best-performing contextual encodings and the structured-only baseline, stratified by surgical service, case duration band, and hospital site.

Level	SentenceBERT	ClinicalBERT	Structured Only
Surgical Service			
Cardiac Surgery	40.35 ± 0.84	40.36 ± 0.95	50.20 ± 1.03
Dental Surgery	22.03 ± 0.68	21.93 ± 0.79	26.19 ± 0.82
ENT	28.67 ± 0.40	28.81 ± 0.46	37.39 ± 0.29
General Surgery	27.78 ± 0.72	27.75 ± 0.62	36.63 ± 0.76
Neurosurgery	43.33 ± 0.79	43.51 ± 0.95	54.56 ± 0.68
OB/GYN	20.49 ± 0.44	20.57 ± 0.49	29.66 ± 0.57
Ophthalmology	14.78 ± 1.19	14.78 ± 1.35	19.73 ± 1.16
Orthopedic	20.31 ± 0.39	20.34 ± 0.34	26.94 ± 0.21
Plastic Surgery	30.42 ± 0.53	30.51 ± 0.61	36.48 ± 0.73
Surgical Oncology	19.10 ± 1.51	18.85 ± 1.41	23.59 ± 1.12
Thoracic Surgery	39.02 ± 0.74	38.98 ± 1.00	50.69 ± 0.65
Urology	24.31 ± 0.65	24.42 ± 0.73	37.95 ± 0.87
Vascular Surgery	31.77 ± 0.61	31.68 ± 0.55	40.90 ± 0.87
Duration Band			
Short (0–60 min)	14.47 ± 0.30	14.49 ± 0.12	25.27 ± 0.19
Medium (61–180 min)	21.25 ± 0.16	21.29 ± 0.16	27.02 ± 0.26
Long (>180 min)	49.91 ± 0.48	49.98 ± 0.31	61.99 ± 0.54
Hospital Site			
Hospital A	25.50 ± 0.17	25.56 ± 0.11	34.48 ± 0.24
Hospital B	28.25 ± 0.19	28.28 ± 0.18	36.10 ± 0.12
Hospital C	15.54 ± 0.35	15.53 ± 0.44	20.37 ± 0.48

All values are MAE in minutes (mean ± SD across five held-out validation folds). Results are shown for the best-performing model configuration at n = 100 text features. Hospital labels are anonymized; the six OR location codes in the EMR were consolidated into three institutional groups prior to analysis.

Stratification by duration band revealed a clear pattern of increasing absolute error with case length. Short cases (≤60 min) were predicted with the highest precision (contextual: ≈14.5 min; structured only: 25.3 min), while long cases (>180 min) yielded the largest absolute errors (contextual: ≈50 min; structured only: 62.0 min). The benefit of text encoding was consistent across all three bands, with contextual models reducing MAE by approximately 10–13 min relative to the structured-only baseline regardless of case length.

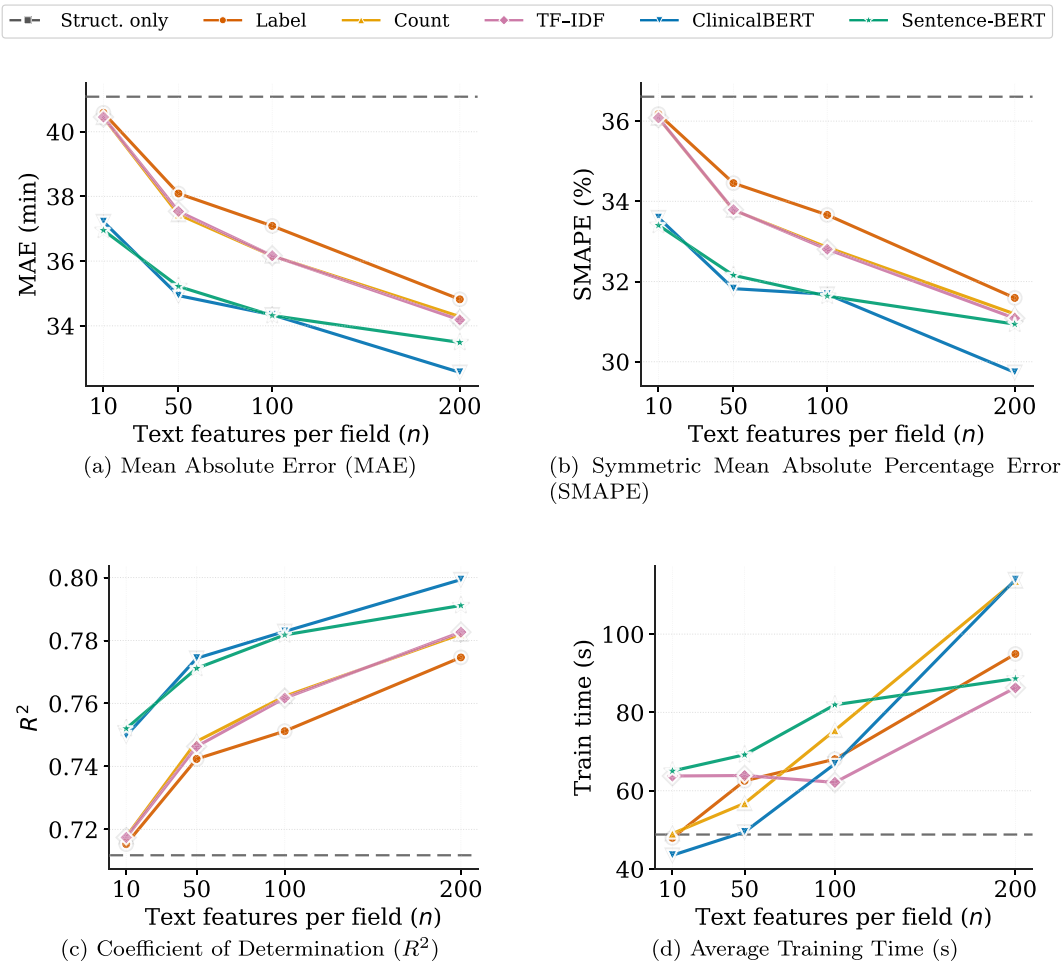


Fig. 3. Sensitivity of encoding-method ranking to text feature dimensionality $n \in \{10, 50, 100, 200\}$, averaged across six models and five folds. Dashed line: structured features only baseline.

3.4. Sensitivity to text feature dimensionality

To assess whether the benchmark conclusions depend on the specific text feature budget n, we evaluated all encoding–model combinations at four dimensionalities: $n \in \{10, 50, 100, 200\}$ features. Fig. 3(a)–(d) report MAE, SMAPE, R^2, and average training time as a function of n, averaged across all six regression models and five cross-validation folds, with the structured-only baseline (MAE = 41.1 min, SMAPE = 36.6%, R^2 = 0.712) shown as a dashed reference. The relative ordering of encoding strategies remained consistent across all four values of n: at $n = 10$, ClinicalBERT and Sentence-BERT already achieved MAE values of 37.2 and 37.0 min respectively, versus 40.4–40.6 min for classical encodings, and this gap widened steadily with increasing n, reaching 32.6 and 33.5 min for contextual methods compared to 34.2–34.8 min at $n = 200$. The same trends were observed for SMAPE and R^2, with contextual encodings consistently achieving lower error and higher explained variance at every dimensionality.

Regarding training cost (Fig. 3(d)), all encodings incurred comparable overhead at $n = 10$ (44–65 s), but diverged at higher dimensionalities, with count vectorization and ClinicalBERT reaching 114 s at $n = 200$ while TF–IDF remained the most stable (62–86 s). These timing differences are attributable to the density of the input feature matrix, as count and transformer-based encoders produce denser matrices at higher n, increasing computation time, rather than to any difference in encoding quality. Importantly, once the model is trained, inference time for all models is under one second, well within practical limits for OR scheduling pipelines. Taken together, these results confirm that the superiority of contextual embeddings over classical encodings is not an artifact of the 100-feature budget used in the primary analysis, and that the ranking is robust from $n = 10$ onward.

3.5. Statistical comparison of encoding methods

Repeated-measures ANOVA was applied to formally test whether the choice of text encoding method significantly affects predictive performance. This analysis confirms that differences among encodings are not due to chance, but reflect systematic effects of encodings on model accuracy. Beyond statistical significance, we report F-statistics and effect sizes (η^2) to show how much variance in outcomes is explained by encoding strategy.

As shown in Fig. 4, the largest effects were observed for MSE ($F = 67.65$, $\eta^2 = 0.93$) and R^2 ($F = 68.06$, $\eta^2 = 0.93$), with higher R^2 values

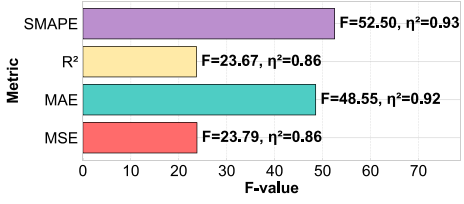


Fig. 4. Repeated-measures ANOVA comparing encoding strategies across evaluation metrics. Bars represent F-values, with effect sizes (η^2) annotated above each bar. Larger values indicate a stronger influence of encoding strategy on model performance.

Table 4
Pairwise comparison of encoding methods based on MAE across all models using paired t-tests with FDR-BH correction. Cohen's d effect size: ∘ negligible (< 0.2), ⊙ small (≥ 0.2), • medium (≥ 0.5), ■ large (≥ 0.8). Significance: * $p_{adj} < 0.05$, ** $p_{adj} < 0.01$.

Encoding 1	Encoding 2	Mean 1	Mean 2	Better	p-value	Cohen's d
Sentence-BERT	ClinicalBERT	34.32	34.34	Sentence-BERT	0.9679	0.019 ∘
Sentence-BERT	TF–IDF	34.32	36.17	Sentence-BERT	0.0155*	1.890 ■
Sentence-BERT	Count	34.32	36.16	Sentence-BERT	0.0146*	1.970 ■
Sentence-BERT	Label	34.32	37.09	Sentence-BERT	0.0036**	2.989 ■
Sentence-BERT	Structured Only	34.32	41.08	Sentence-BERT	< 0.001**	7.394 ■
ClinicalBERT	TF–IDF	34.34	36.17	ClinicalBERT	0.0033**	3.128 ■
ClinicalBERT	Count	34.34	36.16	ClinicalBERT	0.0029**	3.324 ■
ClinicalBERT	Label	34.34	37.09	ClinicalBERT	< 0.001**	4.829 ■
ClinicalBERT	Structured Only	34.34	41.08	ClinicalBERT	< 0.001**	11.461 ■
TF–IDF	Count	36.17	36.16	Count	0.9473	0.069 ∘
TF–IDF	Label	36.17	37.09	TF–IDF	< 0.001**	11.837 ■
TF–IDF	Structured Only	36.17	41.08	TF–IDF	< 0.001**	58.758 ■
Count	Label	36.16	37.09	Count	< 0.001**	20.172 ■
Count	Structured Only	36.16	41.08	Count	< 0.001**	79.739 ■
Label	Structured Only	37.09	41.08	Label	< 0.001**	102.480 ■

indicating that contextual models explain more of the variability in case duration—a key factor for anticipating longer or complex procedures. MAE ($F = 43.73$, $\eta^2 = 0.90$) and SMAPE ($F = 17.21$, $\eta^2 = 0.77$) also showed strong effects, confirming that contextual embeddings reduce both absolute error in minutes and proportional error, leading to more reliable surgical schedules with fewer unexpected deviations. Taken together, these findings demonstrate that encoding strategy is not only statistically significant but also clinically relevant, directly impacting operating room efficiency and planning.

Because MAE directly corresponds to average prediction error in minutes, it provides the most intuitive clinical interpretation. Pairwise comparisons with FDR-BH correction and Cohen's d effect sizes (Table 4) show that both Sentence-BERT and ClinicalBERT reduced scheduling error by approximately 2–7 minutes compared to traditional encodings, with all such differences reflecting large effect sizes. No significant differences were observed between Sentence-BERT and ClinicalBERT, suggesting both contextual approaches capture clinically relevant nuances with comparable fidelity. Traditional methods (label, count, TF–IDF) showed no significant differences among themselves, reinforcing their limited capacity to represent operative text.

In practice, these gains translate into fewer delays, improved utilization of operating room time, and reduced downstream disruptions. Supplementary comparisons for MSE, SMAPE, and R^2 are provided in Appendix Tables A.2–A.4.

4. Conclusion

This study provides the first comparison of five distinct text encoding strategies—label encoding, count vectorization, TF–IDF, ClinicalBERT, and Sentence-BERT—integrated with structured perioperative features for surgical case duration prediction. Evaluated over 180,370 cases and six regression models (ordinary least squares, ridge, lasso, Random Forest, XGBoost, and a feedforward neural network with depth tunable from 1 to 3 layers) using rigorous five-fold cross-validation and fold-wise hyperparameter tuning, our results show that contextual embeddings (Sentence-BERT and ClinicalBERT) yield significant improvements across all metrics (MSE, MAE, SMAPE, and R^2; $p < 0.05$). These semantically rich representations not only outperform traditional encoding techniques but also provide substantial gains over using structured data alone, reducing average prediction error by up to 16% and increasing explained variance, highlighting the added predictive value of clinical text when appropriately encoded.

Structured variables such as ASA score, BMI, and surgical service provide reliable but coarse categorical signals—a single service label may encompass hundreds of distinct operations differing by an hour or more in typical duration [12]. Unstructured procedure descriptions carry the procedural specificity that structured fields cannot capture: whether a case involves revision, bilateral involvement, or a minimally invasive approach are details routinely documented in clinical text and known to substantially affect operative time [24]. When encoded with contextually aware methods, these textual features supply a complementary predictive signal, which explains the consistent performance gains observed across all six regression algorithms evaluated in this study.

Future research should explore several important directions. First, integrating additional narrative sources—such as anesthesia records, nursing notes, and surgeon dictations—could further enrich the textual signal available at scheduling time. Second, the rapid emergence of Large Language Models (LLMs) opens a compelling new avenue for surgical duration prediction. Recent work has demonstrated that fine-tuned LLMs can predict surgical case length with accuracy comparable to or exceeding current institutional scheduling methods [21], while broader evaluations in perioperative medicine suggest that general-purpose LLMs remain limited for numerical regression tasks without fine-tuning [25,26]. Benchmarking LLMs against the contextual embedding pipeline established in this work represents a natural and high-priority next step. Third, real-time model adaptation to case-mix shifts and the development of explainability techniques that translate embedding features into clinician-interpretable insights remain critical prerequisites for broad clinical adoption.

Summary table

What was already known?

- Operating room (OR) scheduling often relies on inaccurate heuristics for surgical durations.
- Machine learning using structured perioperative data improves predictions compared to manual methods.
- Most prior models ignored unstructured clinical text, which contains rich contextual details.
- Few studies systematically compared classical vs. contextual text embeddings in perioperative forecasting.

What does this study add to our knowledge?

- Provides the first large-scale, head-to-head comparison of five text encoding strategies (label, count, TF–IDF, ClinicalBERT, Sentence-BERT) for surgical duration prediction.
- Demonstrates that contextual embeddings (Sentence-BERT and ClinicalBERT) significantly reduce prediction error compared to structured-only and traditional encodings.
- Shows that multimodal models integrating structured + text features improve accuracy across six algorithms and multiple error metrics.

- Offers a standardized, reproducible pipeline for hospitals to leverage EMR text to improve OR scheduling efficiency and patient care.

CRediT authorship contribution statement

Mohammad Noorchenarboo: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Michelle Kwong:** Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Ahmad Elnahas:** Supervision, Resources, Project administration, Funding acquisition. **Jeff Hawel:** Resources, Project administration, Funding acquisition. **Christopher M. Schlachta:** Supervision, Resources, Project administration. **Katarina Grolinger:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT (OpenAI) to improve readability and language fluency. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

This work was supported by the Academic Medical Organization of Southwestern Ontario (AMOSO) Innovation Fund (Grant Number INN23-015).

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the Academic Medical Organization of Southwestern Ontario (AMOSO)'s Innovation Fund (Grant Number INN23-015) for partial support of this study.

Appendix A. Additional results

This appendix provides two types of supplementary results. First, Table A.1 reports temporal cross-validation results as a robustness check against distribution shifts over time. Second, the remaining tables provide pairwise comparisons of encoding strategies across evaluation metrics; the main text presents the MAE-based comparisons (Table 4), and for completeness the MSE, SMAPE, and R^2 comparisons are included here.

Table A.1
Temporal cross-validation results (expanding-window time-series split, $k = 5$). Values are mean ± std across folds, averaged over all six models at $n = 100$ features. The encoding ranking is consistent with the main analysis.

Encoding	MAE (min)	SMAPE (%)	R^2	RMSE (min)
Structured only	41.69 ± 5.86	37.66 ± 8.04	0.699 ± 0.064	59.23 ± 6.17
Label	37.66 ± 5.63	34.65 ± 8.43	0.742 ± 0.055	54.83 ± 5.83
Count	37.04 ± 5.97	34.21 ± 8.60	0.750 ± 0.058	53.89 ± 6.32
TF-IDF	37.16 ± 5.93	34.26 ± 8.51	0.748 ± 0.059	54.06 ± 6.37
ClinicalBERT	34.66 ± 6.71	32.28 ± 9.40	0.777 ± 0.061	50.82 ± 7.04
Sentence-BERT	34.82 ± 7.01	32.41 ± 9.68	0.775 ± 0.064	50.92 ± 7.35

Table A.2
Pairwise comparison of encoding methods based on MSE across all models using paired t-tests with FDR-BH correction. Cohen's d effect size: ∘ negligible (< 0.2), ⊙ small (≥ 0.2), • medium (≥ 0.5), ■ large (≥ 0.8). Significance: * $p_{adj} < 0.05$, ** $p_{adj} < 0.01$.

Encoding 1	Encoding 2	Mean 1	Mean 2	Better	p-value	Cohen's d
Sentence-BERT	ClinicalBERT	2627.6	2614.6	ClinicalBERT	0.8293	0.103 ∘
Sentence-BERT	TF–IDF	2627.6	2869.0	Sentence-BERT	0.0109*	2.092 ■
Sentence-BERT	Count	2627.6	2862.1	Sentence-BERT	0.0109*	2.128 ■
Sentence-BERT	Label	2627.6	2995.4	Sentence-BERT	0.0024**	3.333 ■
Sentence-BERT	Structured Only	2627.6	3469.9	Sentence-BERT	< 0.001**	7.830 ■
ClinicalBERT	TF–IDF	2614.6	2869.0	ClinicalBERT	< 0.001**	5.281 ■
ClinicalBERT	Count	2614.6	2862.1	ClinicalBERT	< 0.001**	5.330 ■
ClinicalBERT	Label	2614.6	2995.4	ClinicalBERT	< 0.001**	8.568 ■
ClinicalBERT	Structured Only	2614.6	3469.9	ClinicalBERT	< 0.001**	20.850 ■
TF–IDF	Count	2869.0	2862.1	Count	0.0988	0.986 ■
TF–IDF	Label	2869.0	2995.4	TF–IDF	< 0.001**	8.353 ■
TF–IDF	Structured Only	2869.0	3469.9	TF–IDF	< 0.001**	36.160 ■
Count	Label	2862.1	2995.4	Count	< 0.001**	13.915 ■
Count	Structured Only	2862.1	3469.9	Count	< 0.001**	57.029 ■
Label	Structured Only	2995.4	3469.9	Label	< 0.001**	87.213 ■

Table A.3
Pairwise comparison of encoding methods based on SMAPE across all models using paired t-tests with FDR-BH correction. Cohen's d effect size: ∘ negligible (< 0.2), ⊙ small (≥ 0.2), • medium (≥ 0.5), ■ large (≥ 0.8). Significance: * $p_{adj} < 0.05$, ** $p_{adj} < 0.01$.

Encoding 1	Encoding 2	Mean 1	Mean 2	Better	p-value	Cohen's d
Sentence-BERT	ClinicalBERT	31.64	31.68	Sentence-BERT	0.9474	0.031 ∘
Sentence-BERT	TF–IDF	31.64	32.80	Sentence-BERT	0.0956	1.019 ■
Sentence-BERT	Count	31.64	32.85	Sentence-BERT	0.0912	1.081 ■
Sentence-BERT	Label	31.64	33.66	Sentence-BERT	0.0255*	1.817 ■
Sentence-BERT	Structured Only	31.64	36.61	Sentence-BERT	0.0012**	4.492 ■
ClinicalBERT	TF–IDF	31.68	32.80	ClinicalBERT	0.0724	1.216 ■
ClinicalBERT	Count	31.68	32.85	ClinicalBERT	0.0618	1.328 ■
ClinicalBERT	Label	31.68	33.66	ClinicalBERT	0.0130*	2.285 ■
ClinicalBERT	Structured Only	31.68	36.61	ClinicalBERT	< 0.001**	5.410 ■
TF–IDF	Count	32.80	32.85	TF–IDF	0.0956	0.999 ■
TF–IDF	Label	32.80	33.66	TF–IDF	< 0.001**	11.900 ■
TF–IDF	Structured Only	32.80	36.61	TF–IDF	< 0.001**	95.139 ■
Count	Label	32.85	33.66	Count	< 0.001**	22.522 ■
Count	Structured Only	32.85	36.61	Count	< 0.001**	107.911 ■
Label	Structured Only	33.66	36.61	Label	< 0.001**	56.686 ■

Table A.4
Pairwise comparison of encoding methods based on R^2 across all models using paired t-tests with FDR-BH correction. Cohen's d effect size: ∘ negligible (< 0.2), ⊙ small (≥ 0.2), • medium (≥ 0.5), ■ large (≥ 0.8). Significance: * $p_{adj} < 0.05$, ** $p_{adj} < 0.01$.

Encoding 1	Encoding 2	Mean 1	Mean 2	Better	p-value	Cohen's d
Sentence-BERT	ClinicalBERT	0.7818	0.7829	ClinicalBERT	0.8277	0.104 ∘
Sentence-BERT	TF–IDF	0.7818	0.7617	Sentence-BERT	0.0110*	2.087 ■
Sentence-BERT	Count	0.7818	0.7623	Sentence-BERT	0.0110*	2.123 ■
Sentence-BERT	Label	0.7818	0.7512	Sentence-BERT	0.0024**	3.323 ■
Sentence-BERT	Structured Only	0.7818	0.7118	Sentence-BERT	< 0.001**	7.765 ■
ClinicalBERT	TF–IDF	0.7829	0.7617	ClinicalBERT	< 0.001**	5.131 ■
ClinicalBERT	Count	0.7829	0.7623	ClinicalBERT	< 0.001**	5.185 ■
ClinicalBERT	Label	0.7829	0.7512	ClinicalBERT	< 0.001**	8.194 ■
ClinicalBERT	Structured Only	0.7829	0.7118	ClinicalBERT	< 0.001**	18.713 ■
TF–IDF	Count	0.7617	0.7623	Count	0.0985	0.987 ■
TF–IDF	Label	0.7617	0.7512	TF–IDF	< 0.001**	8.697 ■
TF–IDF	Structured Only	0.7617	0.7118	TF–IDF	< 0.001**	45.014 ■
Count	Label	0.7623	0.7512	Count	< 0.001**	14.709 ■
Count	Structured Only	0.7623	0.7118	Count	< 0.001**	82.832 ■
Label	Structured Only	0.7512	0.7118	Label	< 0.001**	116.329 ■

Table A.5
Model performance across encoding strategies (n = 100 text features) reported as mean ± standard deviation over five cross-validation folds. MAE and RMSE are in minutes; MSE is in minutes^2; SMAPE and MAPE are in %; R^2 is dimensionless. Structured only uses no text features.

Model	MAE	RMSE	MSE	SMAPE	MAPE	R^2
Sentence-BERT						
MLP	30.47 ± 5.79	46.28 ± 7.26	2183.6 ± 731.8	26.63 ± 7.09	30.34 ± 8.04	0.819 ± 0.060
XGBoost	26.35 ± 0.16	41.84 ± 0.93	1751.0 ± 77.9	21.58 ± 0.18	23.91 ± 0.32	0.855 ± 0.006
Random Forest	27.42 ± 0.09	43.22 ± 0.87	1868.6 ± 75.5	22.69 ± 0.10	25.98 ± 0.21	0.845 ± 0.006
Lasso	40.55 ± 0.15	57.62 ± 0.74	3320.8 ± 85.2	39.62 ± 0.23	44.09 ± 0.25	0.724 ± 0.006
Ridge	40.56 ± 0.15	57.62 ± 0.74	3320.9 ± 85.2	39.65 ± 0.25	44.11 ± 0.26	0.724 ± 0.006
Linear	40.56 ± 0.16	57.62 ± 0.73	3320.7 ± 84.9	39.66 ± 0.26	44.12 ± 0.26	0.724 ± 0.006
ClinicalBERT						
MLP	31.17 ± 3.74	46.88 ± 4.02	2210.3 ± 386.5	27.51 ± 5.51	30.75 ± 5.22	0.817 ± 0.031
XGBoost	26.39 ± 0.11	41.87 ± 0.84	1754.0 ± 70.3	21.61 ± 0.10	23.94 ± 0.13	0.854 ± 0.006
Random Forest	27.55 ± 0.12	43.42 ± 0.88	1886.0 ± 76.7	22.77 ± 0.08	26.10 ± 0.18	0.843 ± 0.006
Lasso	40.31 ± 0.09	57.26 ± 0.79	3279.1 ± 91.0	39.39 ± 0.21	43.62 ± 0.24	0.728 ± 0.007
Ridge	40.31 ± 0.09	57.26 ± 0.79	3279.0 ± 90.7	39.40 ± 0.23	43.63 ± 0.24	0.728 ± 0.007
Linear	40.31 ± 0.09	57.26 ± 0.79	3278.9 ± 90.6	39.41 ± 0.24	43.64 ± 0.25	0.728 ± 0.007
TF–IDF						
MLP	30.13 ± 0.14	46.16 ± 0.86	2131.8 ± 79.1	25.13 ± 0.12	29.40 ± 0.18	0.823 ± 0.006
XGBoost	29.28 ± 0.22	45.51 ± 0.75	2071.9 ± 68.4	24.07 ± 0.18	27.63 ± 0.29	0.828 ± 0.005
Random Forest	32.05 ± 0.17	48.63 ± 0.83	2365.3 ± 80.8	26.72 ± 0.22	32.10 ± 0.44	0.804 ± 0.006
Lasso	41.83 ± 0.14	59.54 ± 0.70	3545.4 ± 83.2	40.28 ± 0.24	45.58 ± 0.26	0.706 ± 0.006
Ridge	41.82 ± 0.14	59.54 ± 0.69	3545.3 ± 82.9	40.27 ± 0.25	45.58 ± 0.26	0.706 ± 0.006
Linear	41.90 ± 0.14	59.62 ± 0.69	3554.6 ± 83.0	40.36 ± 0.24	45.66 ± 0.25	0.705 ± 0.006
Count						
MLP	29.97 ± 0.25	45.97 ± 1.02	2114.0 ± 93.4	25.02 ± 0.24	29.23 ± 0.31	0.824 ± 0.007
XGBoost	29.42 ± 0.15	45.61 ± 0.86	2081.2 ± 78.2	24.22 ± 0.17	27.85 ± 0.32	0.827 ± 0.006
Random Forest	32.12 ± 0.18	48.70 ± 0.88	2372.8 ± 85.6	26.79 ± 0.16	32.20 ± 0.31	0.803 ± 0.006
Lasso	41.80 ± 0.15	59.43 ± 0.71	3531.8 ± 84.5	40.34 ± 0.23	45.71 ± 0.24	0.707 ± 0.006
Ridge	41.79 ± 0.15	59.43 ± 0.71	3531.8 ± 84.3	40.33 ± 0.23	45.71 ± 0.25	0.707 ± 0.006
Linear	41.87 ± 0.15	59.50 ± 0.70	3541.2 ± 83.9	40.41 ± 0.22	45.79 ± 0.24	0.706 ± 0.006
Label						
MLP	31.11 ± 0.19	47.95 ± 0.90	2300.1 ± 86.3	25.62 ± 0.14	30.39 ± 0.27	0.809 ± 0.007
XGBoost	30.49 ± 0.26	47.50 ± 0.95	2257.0 ± 90.0	24.80 ± 0.22	28.95 ± 0.39	0.813 ± 0.007
Random Forest	32.86 ± 0.17	49.63 ± 0.85	2463.8 ± 84.7	27.51 ± 0.19	33.59 ± 0.41	0.795 ± 0.006
Lasso	42.69 ± 0.19	60.42 ± 0.76	3650.6 ± 91.5	41.35 ± 0.33	47.05 ± 0.35	0.697 ± 0.006
Ridge	42.69 ± 0.19	60.42 ± 0.76	3650.6 ± 91.7	41.34 ± 0.34	47.05 ± 0.35	0.697 ± 0.006
Linear	42.69 ± 0.19	60.42 ± 0.76	3650.5 ± 91.5	41.35 ± 0.33	47.06 ± 0.35	0.697 ± 0.006
Structured Only						
MLP	36.56 ± 0.18	53.22 ± 0.86	2833.1 ± 91.4	30.94 ± 0.17	38.87 ± 0.34	0.765 ± 0.007
XGBoost	34.80 ± 0.13	51.77 ± 0.74	2681.0 ± 77.0	28.84 ± 0.13	35.11 ± 0.23	0.777 ± 0.006
Random Forest	35.17 ± 0.16	52.27 ± 0.87	2733.2 ± 90.4	29.30 ± 0.19	36.19 ± 0.36	0.773 ± 0.007
Lasso	46.65 ± 0.22	64.73 ± 0.76	4190.8 ± 98.1	43.52 ± 0.30	51.72 ± 0.32	0.652 ± 0.007
Ridge	46.65 ± 0.22	64.73 ± 0.75	4190.8 ± 97.7	43.52 ± 0.31	51.73 ± 0.33	0.652 ± 0.007
Linear	46.66 ± 0.23	64.73 ± 0.75	4190.7 ± 97.8	43.54 ± 0.31	51.74 ± 0.33	0.652 ± 0.007

Appendix B. Reproducibility

Experimental settings

All models were selected via 5-fold cross-validation. For each fold and configuration (encoding type × number of text features), hyperparameters were tuned independently using Bayesian optimization over the search spaces listed in Table B.1. The best-performing configuration per fold was selected based on validation performance.

Software and hardware environment

All experiments were conducted in the environment detailed in Table B.2, including the operating system, GPU, CUDA version, and key library versions.

Table B.1
Best hyperparameter values selected across all folds and encoding configurations. Ranges reflect the min–max of the best trial per fold; sets reflect all distinct values chosen. "Fixed" indicates no tuning was performed for that parameter.

Model	Hyperparameter	Best Value(s)
Lasso	`alpha`	[0.0010, 0.0171]
	Solver	Fixed: coordinate descent
Ridge	`alpha`	[0.0010, 5.40]
	Solver	Fixed: Cholesky
Random Forest	`n_estimators`	[156, 500]
	`max_depth`	15 (ceiling reached across all folds)
	`max_features`	{0.3, 0.5, 0.7}
	`min_samples_split`	[2, 10]
	`min_samples_leaf`	[1, 4]
XGBoost	`n_estimators`	[258, 500]
	`learning_rate`	[0.025, 0.255]
	`max_depth`	{6, 7, 8}
	`subsample`	[0.61, 0.99]
	`colsample_bytree`	[0.61, 0.99]
	`reg_alpha`	[0.0001, 6.51]
	`reg_lambda`	[0.0001, 9.63]
	Objective / early stopping	Fixed: `reg:squarederror`; early stopping rounds = 20
MLP	`n_layers`	{1, 2, 3} (predominantly 1)
	`activation`	{`relu`, `elu`} (predominantly `elu`)
	`learning_rate`	[0.005, 0.010]
	`weight_decay`	[10^{-6}, 0.010]
	`units` (per layer)	Layer 1: [33, 249]; layer 2: [21, 126]; layer 3: [10, 63]
	`dropout` (per layer)	[0.001, 0.497]
	Optimizer / clipping	Fixed: AdamW; `clipnorm` = 1.0
	`input_dim`	Fixed: $38 + k$, where $k \in \{0, 10, 50, 100, 200\}$
	Epochs / early stopping	Final fit: 200 epochs, patience = 15, `restore_best_weights=True`
	Optuna trials	30 epochs/trial; MedianPruner ($n\_warmup = 5$); subset = 5,000 rows
	Batch size / LR schedule	Fixed: batch = 512; ReduceLROnPlateau: factor = 0.5, patience = 5, min_lr = 10^{-6}
All models — cross-validation		5-fold CV; Optuna TPE sampler ($n\_startup = 10$, multivariate); tune split = 25% of training fold

Table B.2
Software and hardware environment used for all experiments.

Component	Details
Operating System	Windows 11
GPU	NVIDIA GeForce RTX 3080
CUDA Version	11.8
Python	3.13.5
scikit-learn	1.6.1
XGBoost	3.2.0
TensorFlow	2.20.0
PyTorch	2.7.1
Transformers	4.57.3
Sentence-Transformers	5.2.3
Optuna	4.5.0
pandas	2.2.3
NumPy	2.1.3

Data availability

Analytical code and preprocessing scripts are publicly available at https://github.com/mnoorchenarboo/Benchmarking-Text-Encoding-Strategies. The de-identified dataset is not publicly available due to hospital privacy and confidentiality regulations.

References

[1] C.P. Childers, M. Maggard-Gibbons, Understanding costs of care in the operating room, JAMA Surg. 153 (4) (Apr 2018) e176233.

[2] T.G. Smith, H. Norasi, K.M. Herbst, M.L. Kendrick, T.B. Curry, T.P. Grantcharov, V.N. Palter, M.S. Hallbeck, S.P. Cleary, Creating a practical transformational change management model for novel artificial intelligence–enabled technology implementation in the operating room, Mayo Clin. Proc. Innov. Qual. & Outcomes 6(6) (Dec 2022) 584–596.

[3] P.E. Hasvold, J. Scholl, Flexibility in interaction: sociotechnical design of an operating room scheduler, Int. J. Med. Inform. 80 (9) (Sep 2011) 631–645.

[4] O. Martinez, C. Martinez, C.A. Parra, S. Rugeles, D.R. Suarez, Machine learning for surgical time prediction, Comput. Methods Programs Biomed. 208 (Sep 2021) 106220.

[5] S. Zhu, W. Fan, S. Yang, J. Pei, P.M. Pardalos, Operating room planning and surgical case scheduling: a review of literature, J. of Comb. Optim. 37 (3) (Jul 2018) 757–805.

[6] M. Kwong, M. Noorchenarboo, K. Grolinger, J. Hawel, C.M. Schlachta, A. Elnahas, Optimizing surgical efficiency: predicting case duration of common general surgery procedures using machine learning, Surg. endosc. 39 (2025) 5227–5234.

[7] M.A. Bartek, R.C. Saxena, S. Solomon, C.T. Fong, L.D. Behara, R. Venigandla, K. Velagapudi, J.D. Lang, B.G. Nair, Improving operating room efficiency: machine learning approach to predict case-time duration, J. Am. Coll. Surg. 229 (4) (Oct 2019) 346–354e3.

[8] J.-B. Park, G.-H. Roh, K. Kim, H.-S. Kim, Development of predictive model of surgical case durations using machine learning approach, J. Med. Syst. 49 (1) (Jan 2025).

[9] I. Yeo, C. Klemt, C.M. Melnic, M.H. Pattavina, B.M.C. De Oliveira, Y.-M. Kwon, Predicting surgical operative time in primary total knee arthroplasty utilizing machine learning models, Arch. Orthop. Trauma Surg. 143 (6) (Aug 2022) 3299–3307.

[10] H. Zaribafzadeh, W.L. Webster, C.J. Vail, T. Daigle, A.D. Kirk, P.J. Allen, R. Henao, D.M. Buckland, Development, deployment, and implementation of a machine learning surgical case length prediction model and prospective evaluation, Ann. of Surg. 278 (6) (Jun 2023) 890–895.

[11] J.P. Tuwatananurak, S. Zadeh, X. Xinling, J.A. Vacanti, W.R. Fulton, J.M. Ehrenfeld, R.D. Urman, Machine learning can improve estimation of surgical case duration: a pilot study, J. Med. Syst. 43 (3) (Jan 2019).

[12] K.H. Goh, L. Wang, A.Y.K. Yeow, H. Poh, K. Li, J.J.L. Yeow, G.Y.H. Tan, Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare, Nat. Commun. 12 (1) (Jan 2021).

[13] R. Chen, C.H. Joyce, J.-M.S. Lin, Extracting medication information from unstructured public health data: a demonstration on data from population-based and tertiary-based samples, BMC Med. Res. Methodol. 20 (1) (Oct 2020).

[14] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv prepr arXiv:1904.03323, 2019.

[15] N. Reimers, I. Gurevych, Sentence-bert: sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019.

[16] C. Mao, X. Jie, L. Rasmussen, L. Yikuan, P. Adekkanattu, J. Pacheco, B. Bonakdarpour, R. Vassar, L. Shen, G. Jiang, F. Wang, J. Pathak, Y. Luo, Ad-bert: using pre-trained language model to predict the progression from mild cognitive impairment to alzheimer's disease, J. Biomed. Inform. 144 (Aug 2023) 104442.

[17] F. Jaotombo, L. Adorni, B. Ghattas, L. Boyer, Finding the best trade-off between performance and interpretability in predicting hospital length of stay using structured and unstructured data, PLOS ONE 18 (11) (Nov 2023) e0289795.

[18] J. Gatto, P. Seegmiller, G. Johnston, S.M. Preum, Identifying the perceived severity of patient-generated telemedical queries regarding covid: developing and evaluating a transfer learning-based solution, JMIR Med. Inform. 10 (9) (Sep 2022) e37770.

[19] Y. Jiao, A. Sharma, A.B. Abdallah, T.M. Maddox, T. Kannampallil, Probabilistic forecasting of surgical case duration using machine learning: model development and validation, J. Am. Med. Inform. Assoc. 27 (12) (Oct 2020) 1885–1893.

[20] T. Adams, M. O'Sullivan, C. Walker, Surgical procedure prediction using medical ontological information, Comput. Methods Programs Biomed. 235 (Jun 2023) 107541.

[21] A. Ramamurthi, B. Neupane, P. Deshpande, R. Hanson, K.R. Brown, K.K. Christians, D.B. Evans, A.N. Kothari, Development and validation of an artificial intelligence system for surgical case length prediction, Surg. 179 (Mar 2025) 108942.

[22] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) (1988) 513–523.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.

[24] M.JC. Eijkemans, M. van Houdenhoven, T. Nguyen, E. Boersma, E.W. Steyerberg, G. Kazemier, Predicting the unpredictable: a new prediction model for operating room times using individual characteristics and the surgeon's estimate, Anesthesiology 112 (1) (2010) 41–49.

[25] P. Chung, C.T. Fong, A.M. Walters, N. Aghaeepour, M. Yetisgen, V.N. O'Reilly-Shah, Large language model capabilities in perioperative risk prediction and prognostication, JAMA Surg. 159 (8) (2024) 928–937.

[26] Z.A. Nazi, W. Peng, Large language models in healthcare and medical domain: a review, Informatics 11 (2024) 57. MDPI.