Bruins for Business Health

Artificial Intelligence for Predicting Surgical Duration: Operational Value, Workflow Integration, and Scalability in Health Systems

Elizabeth Sautter, Ellie Korman, Jiaa Bartake, Josephine Wong, Kyra Shah, Laasya Balupari, Nitya Jhamb, Rohan Gianchandani, Sahasra Kalluri, Trevor Kuo, Vinay Panchal, Jason Yu

Bruins for Business Health, University of California, Los Angeles

INTRODUCTION

- Operating rooms (ORs) are among the most resource-intensive environments in health systems, often accounting for a significant proportion of hospital revenue and operating costs. Small inefficiencies in surgical scheduling can therefore produce disproportionately large financial and workflow consequences.
- Inaccurate surgical duration estimates contribute to persistent operational challenges, including OR overtime, underutilized block time, delayed case starts, staffing inefficiencies, and downstream bottlenecks affecting postoperative care and bed availability.
- Beyond financial impact, unreliable duration estimates can negatively affect patient experience, contributing to prolonged wait times, last-minute cancellations, and reduced transparency around expected surgery timelines.
- Traditional prediction methods rely heavily on historical averages, surgeon judgment, and static EHR-derived heuristics. These approaches often fail to account for nonlinear interactions between patient characteristics, procedural complexity, provider variability, and institutional workflow patterns.
- Artificial intelligence (AI) and machine learning (ML) models offer the potential to improve predictive precision by leveraging electronic health record (EHR) data and identifying patterns not captured by conventional statistical methods.
- This review evaluates current evidence on AI-based surgical duration prediction, focusing on model accuracy, reduction in extreme errors, operational impact, and scalability within large health systems.

METHODS

- A systems-oriented narrative review was conducted to evaluate published studies assessing artificial intelligence (AI) models for surgical duration prediction, with emphasis on both statistical performance and operational relevance within real-world hospital workflows.
- Inclusion criteria required peer-reviewed studies conducted in clinical settings that evaluated AI-based prediction models, reported quantitative performance metrics (e.g., MAE or accuracy thresholds), and compared results against at least one institutional or statistical baseline. Also studies spanned across specialties.
- Models were compared against heterogeneous baselines, including EHR-derived estimates, institutional OR scheduling systems, Bayesian statistical approaches, and manual estimation methods. This variability was considered when interpreting reported performance gains.
- Evaluated modeling approaches included ensemble tree-based models (e.g., Random Forest, CatBoost), neural networks, and large language models (LLMs), representing distinct strategies for handling structured clinical data and complex feature interactions.
- Ensemble tree-based models were frequently highlighted for their strong performance with structured EHR variables, ability to model nonlinear relationships, and relative ease of integration into existing hospital IT infrastructures.
- Neural networks and LLM-based approaches demonstrated capacity to process high-dimensional data and capture complex patterns, though they often required larger datasets, careful validation, and workflow-specific adaptation to ensure reliability in heterogeneous clinical environments.
- Primary outcomes included mean absolute error (MAE), threshold-based accuracy metrics, reduction in large prediction errors, and reported operational endpoints such as decreased OR overtime and improved utilization.

RESULTS

Table 1: Accuracy Gain Across Models

Model	Model Value	Baseline Value	Accuracy Gain
MANN[1]	89%	80%	9%
Bagged Trees[2]	Not Reported	Not Reported	16%
Random Forest[3]	64.7%	55.8%	8.9%
Leap Rail[4]	41.1%	31.2%	8.9%
CatBoost Regressor[5]	48%	17%	31%
GPT-4-Fine-Tuned[6]	46.12%	45.76% - EHR model 40.92% - OR model	0.36% - EHR model 5.2% - OR model

Table 1 shows that across studies, AI models consistently outperformed institutional baselines, with ensemble tree-based approaches, particularly CatBoost, demonstrating the largest absolute accuracy gains. While improvements varied depending on baseline comparator, most models showed measurable enhancement over traditional EHR-derived or scheduling estimates.

Table 2: Reductions in Time Errors Across Models

Model	Error Metric	Model Value	Baseline Value	Reduction in Prediction Error (Minutes)	Reduction in Prediction Error (Percent)
MANN[1]	CRPS	13.6 min	20.3 min	6.4 min	31.5%
Bagged Tree[2]	RMSE	26.09 min	No pooled value reported	4.75 min compared to linear regression model	Not reported
Leap Rail[3]	MAE	Not reported	Not reported	7 min	Not reported
CatBoost XGBoost[4]	MAE	Not reported	Not reported	9.6 min 8.5 min	Not reported
CatBoost Regressor[5]	MAE	46.4 min	120.0 min (EHR)	73.6 min	61.3%
Fine-tuned GPT-4[6]	MAE RMSE	47.6 min 74.9 min	49.3 min (EHR) 79.8 min	1.7 min 4.9 min	3.44% 6.14%

Table 2 shows that AI-based models reduced mean prediction error in minutes across multiple evaluation metrics, with CatBoost showing the largest absolute and percentage reductions. These reductions in time error are operationally meaningful, as minimizing large deviations directly improves OR efficiency and reduces overtime risk.

INTERPRETATION

- AI models outperform traditional methods because they capture nonlinear relationships among patient demographics, comorbidities, procedural complexity, provider experience, and institutional workflow factors that are not fully represented in heuristic-based systems.
- Ensemble tree-based approaches appear particularly well-suited for structured EHR data due to their robustness, interpretability, and strong performance across heterogeneous datasets.
- Operationally, improved duration prediction enhances scheduling precision, which can increase daily case throughput, reduce overtime expenditures, and optimize resource allocation without requiring fundamental changes to surgical practice.
- Compared to diagnostic AI applications, surgical duration prediction represents a lower regulatory risk use case, as it supports operational decision-making rather than direct clinical diagnosis or treatment selection.

CONCLUSION

- AI-based surgical duration prediction models consistently demonstrate improved accuracy compared to traditional EHR-derived and institutional scheduling estimates across diverse clinical settings.
- Ensemble tree-based models, particularly CatBoost and Random Forest, show the strongest and most consistent operational promise, while neural networks also demonstrate meaningful improvements over conventional statistical and heuristic-based approaches.
- The value of these models extends beyond statistical accuracy to tangible workflow and financial benefits, including reduced OR overtime, improved block utilization, increased daily case throughput, and enhanced scheduling stability.
- Despite consistent performance gains, heterogeneity in baseline comparators and evaluation metrics across studies highlights the need for standardized reporting frameworks to enable clearer cross-institutional comparison.
- Successful real-world implementation will depend not only on predictive accuracy but also on EHR integration, workflow alignment and stakeholder adoption
- Surgical duration prediction represents a pragmatic and scalable entry point for operational AI within health systems, offering potential for measurable return on investment with relatively low regulatory and clinical risk.
- As hospitals expand digital infrastructure, these models may serve as a foundation for broader predictive analytics across perioperative and operational domains.
- Future work should prioritize standardized evaluation metrics, multi-institutional validation, prospective deployment studies, and assessment of long-term financial and operational impact.

REFERENCES

1. Jiao, A., York, J., et al. (2024). Continuous real-time prediction of surgical case duration using a modular artificial neural network. *British Journal of Anaesthesia, 128*(5), 829–837. https://doi.org/10.1016/j.bja.2021.12.039
2. Martinez, O., Martinez, C., Parra, C. A., Rugeles, S., & Suarez, D. R. (2021). Machine learning for surgical time prediction. *Computer Methods and Programs in Biomedicine*, 208, 106220. https://doi.org/10.1016/j.cmpb.2021.106220
3. Tuwatananurak, J. P., Zadeh, S., Xu, X., et al. (2019). Machine learning can improve estimation of surgical case duration: A pilot study. *Journal of Medical Systems, 43*, 44. https://doi.org/10.1007/s10916-019-1160-5
4. Miller, L. E., Goedicke, W., Crowson, M. G., Rathi, V. K., Naunheim, M. R., & Agarwala, A. V. (2023). Using machine learning to predict operating room case duration: A case study in otolaryngology. *Otolaryngology–Head and Neck Surgery*, 168(2), 241–247. https://doi.org/10.1177/01945998221076480
5. Ramamurthi, A., Neupane, B., Deshpande, P., Hanson, R., Brown, K. R., Christians, K. K., Evans, D. B., & Kothari, A. N. (2025). Development and validation of an artificial intelligence system for surgical case length prediction. *Surgery*, 179, 108942. https://doi.org/10.1016/j.surg.2024.09.051
6. Ramamurthi, A., Neupane, B., Deshpande, P., Hanson, R., Vegesna, S., Cray, D., Crotty, B. H., Somai, M., Brown, K. R., Pawar, S. S., Taylor, B., & Kothari, A. N. (2025). Applying large language models for surgical case length prediction. *JAMA Surgery, 160*(8), 894–902. https://doi.org/10.1001/jamasurg.2025.2154