Predicting surgery times with machine learning versus traditional methods

Richard Park, MS, BSN, RN, CNOR
Christopher Stucky, PhD, RN, CNOR, CSSM, CNAMB, NEA-BC, FAORN, FAAN
Chandler Moser, PhD, BSN, RN, CNOR

DHA Defense Health Agency

Eleven studies showed a 25.7% average improvement in prediction accuracy with machine learning, which could optimize scheduling

Introduction

- Operating room expenditures are projected to reach $912 billion by 2025, with costs averaging approximately $46 per minute.
- Current estimation methods relying on surgeon intuition and historical averages frequently result in scheduling inefficiencies and lost productivity.
- Inaccurate duration predictions lead to increased overtime, idle room time, and secondary impacts on staff burnout and patient satisfaction.
- Despite potential for cost-saving improvements, the actual performance and factors influencing the success of ML models are not well understood.

Aims

- To compare the predictive accuracy of machine learning models against traditional surgical duration estimation methods.
- To identify the specific model architectures and dataset characteristics that influence predictive performance.
- To examine the practical challenges and methodological limitations affecting real-world clinical implementation.

Methods

- Conducted a systematic review following PRISMA 2020 guidelines and registered via PROSPERO.
- Searched MEDLINE, Embase, and CINAHL for peer-reviewed studies published between 1/2019 and 10/2024.
- 11 studies met the inclusion criteria
- Assessed risk of bias and evidence certainty using the PROBAST tool

Inclusion	At least one ML model for surgical duration, traditional estimate comparison with quantitative performance data, published in English
Exclusion	Non-human studies, no comparator, lacking quantitative measurement of error, predictive methods not based on ML

Results

Model performance

SCAN ME

- Machine learning models outperformed traditional estimation methods across 10 of 11 studies, showing an average accuracy improvement of 25.7% (reduced error between predicted and actual surgery minutes).
- The most significant performance gains (up to 51% error reduction) were observed in models utilizing tree-based algorithms like XGBoost and neural networks.
- Narrowly trained models focusing on specific specialties or elective versus acute cohorts consistently achieved higher precision than broad, hospital-wide models.

Common predictors used in machine learning models	
Historical Data	Average duration of the latest ten surgeries (at procedure, surgeon, or sub-procedure levels).
Case Specifics	Scheduled time, type of surgical procedure, and surgical diagnosis.
Patient Factors	Age (specifically < 45), BMI (> 40 kg/m^3), and ASA (American Society of Anesthesiologists) scores.

- Statistical analysis revealed no correlation ($r = -0.01$) between the number of predictor variables and the percentage improvement in prediction accuracy.
- A weak negative correlation ($r = -0.18$) was found between training set size and performance improvement, indicating that clinical relevance and cohort are more critical than raw data volume.

Discussion

- Predictive accuracy is heavily influenced by model specialization; models tailored to specific surgical populations yield higher utility than generalized institutional models.
- Inconsistent definitions of "surgical duration" across the literature, ranging from anesthesia start to skin closure, complicate the aggregation of data and cross-site comparisons.
- Methodological limitations like inadequate handling of missing data and a lack of external validation suggest that many models may be over-optimized for their specific retrospective datasets.
- Latent factors such as team familiarity and sociodynamics significantly influence procedural efficiency but are not often studied – many potential combinations of OR staff make model fitting challenging.

Implications for Perioperative Leaders

- Improved scheduling precision can stabilize the perioperative continuum, leading to more predictable overtime planning.
- Better duration predictions could reduce patient wait times and help mitigate the risks associated with rushed room turnovers.
- This study did not find evidence supporting any specific commercial ML solutions.
- Clinical leaders should advocate for the integration of site-specific ML tools into existing electronic health records to ensure scheduling reflects the reality of local workflows.
- Site-specific retraining suggests that widespread adoption will require significant investment in local data infrastructure and skilled personnel.

REFERENCES

Read our published work in *Perioperative Care and Operating Room Management!*

PUBLICATION

Primary Colors:

DHA Dark Blue
RGB: 36/43/100
CMYK: 100/95/28/23
HEX: #092068

DHA Blue
RGB: 90/146/202
CMYK: 65/34/2/0
HEX: #5A92CA

DHA Maroon
RGB: 88/40/49
CMYK: 45/82/61/52
HEX: #582831

DHA Background Blue
RGB: 158/189/219
CMYK: 37/16/4/0
HEX: #9EBDDB

Secondary Colors:

DHA Green
RGB: 90/172/69
CMYK: 69/8/100/0
HEX: #5AAC45

DHA Yellow
RGB: 255/208/65
CMYK: 69/8/100/0
HEX: #FFD041

DHA Orange
RGB: 234/116/37
CMYK: 4/67/99/0
HEX: #EA7425

DHA Gray
RGB: 65/64/66
CMYK: 0/0/0/90
HEX: #414042