Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 12: 1–12

DOI: 10.1177/20552076261443585
journals.sagepub.com/home/dhj
Sage

Precision vs recall in surgical AI: Performance and limitations of tabular foundation model for perioperative classification and risk prediction

Junqi Cui[1], Ze He[1], Enoch Chi Ngai Lim[2], and Chi Eung Danforn Lim[2,3,4]

Abstract

Background: The Tabular Prior-data Fitted Network (TabPFN) is a recently introduced Transformer-based foundation model designed specifically for structured tabular data. TabPFN enables task inference without the need for hyperparameter tuning or extensive data preprocessing. Despite its disruptive potential, the application of TabPFN in surgical data science remains unexplored. In this study, we evaluate the performance of TabPFN across six surgical classification tasks. **Objective:** To assess TabPFN's performance against benchmark machine learning models in surgical classification tasks and identify optimal application scenarios based on sample size and outcome incidence characteristics. **Methods:** In this study, perioperative data from two independent medical centres were utilized, comprising a large-scale cohort (n=67,134) and a medium-scale cohort (n=6,888). Six clinically relevant classification tasks were developed. The performance of TabPFN was systematically compared to benchmark models including XGBoost, Random Forest, Support vector machine, Logistic regression, and Decision tree, using area under the receiver operating characteristic curve, accuracy, precision, recall, F1-score, and calibration. **Results:** In tasks with large-sample sizes (n > 3,000) and higher outcome incidence (>40%), TabPFN achieved the highest recall and F1-score among all models. For tasks with low outcome incidence (<20%), TabPFN attained the highest precision. Calibration analysis demonstrated that TabPFN provided reliable probability estimates in large-sample tasks (n > 3,000), but its calibration performance declined noticeably in low outcome incidence (<20%). **Conclusions:** TabPFN represents a promising methodological approach for tabular modelling in surgical data science. It also shows considerable promise in tasks with sample sizes exceeding 10,000. However, it is not yet capable of fully replacing established benchmark models. The application scenarios of TabPFN should be selected based on key task-specific characteristics. Further targeted training is necessary in future.

Keywords

tabular data, surgical classification task, learning algorithms, prior-data fitted networks, machine learning

Received: 15 October 2025; Revised: 10 March 2026; Accepted: 3 April 2026

[1]Faculty of Medicine & Health, University of New South Wales, Kensington, NSW, Australia
[2]Translational Research Department, Specialist Medical Services Group, Earlwood, NSW, Australia
[3]School of Life Sciences, University of Technology Sydney, Ultimo, NSW, Australia
[4]NICM Health Research Institute, Western Sydney University, Westmead, NSW, Australia

Corresponding author:
Chi Eung Danforn Lim, NICM Health Research Institute, Western Sydney University, 158 Hawkesbury Rd, Westmead, NSW 2145, Australia.
Email: chi.lim@westernsydney.edu.au

Introduction

Surgical data plays an essential role in improving the quality of clinical decision-making. This type of data integrates diverse information modalities, including continuous physiological signals, categorical clinical variables and text annotations. However, surgical data have long presented significant challenges for modelling, including high dimensionality, multi-modality, heterogeneity sources, frequent missing values and severe class imbalance.[1–4]

For instance, the VitalDB dataset,[5] which contains 6,388 surgical cases, includes only approximately 0.9% in-hospital mortality cases. The performance of conventional models remains suboptimal when handling imbalanced distributions.[1,2,6–9] This bias can result in misleadingly high accuracy metrics, while the models fail to effectively identify minority categories.[6–9]

A substantial proportion of published model-based studies in medical research lack external validation. Models developed using single-centre data often exhibit limited generalizability and frequently suffer from performance degradation when applied across different patient populations.[6,9,10] Chekroud et al.[11] highlighted the phenomenon of "illusory universality" in clinical prediction models. Models trained on large single-centre data, often fail to generalize effectively to smaller patient cohorts from other institutions.[11–13]

Hollmann et al.[13] introduced the Tabular Prior-data Fitted Network (TabPFN), a novel model designed for tabular data processing. TabPFN is a Transformer-based model that is pre-trained on millions of synthetic tabular datasets generated from diverse prior distributions. This approach enables the model to capture generalizable input-output relationships, thereby allowing one-shot inference on a new dataset without requiring task-specific feature engineering, hyperparameter tuning, or gradient-based optimization.[13]

This study aims to evaluate the performance and clinical applicability of TabPFN compared to baseline models across six surgical classification tasks with varying sample sizes and outcome incidences, and to determine optimal application scenarios for TabPFN in surgical data science. It has the potential to contribute to improved modelling strategies for surgical data science, thereby offering more robust tools for clinical decision support systems.

Methods

Study design and setting

This study was designed as a retrospective comparative modelling study evaluating the performance of a tabular foundation model (TabPFN) against conventional machine learning algorithms for surgical outcome classification. The analysis utilised perioperative datasets obtained from two publicly available sources: the Medical Informatics Operating Room Vitals and Events Repository (MOVER) dataset from the University of California, Irvine Medical Centre (United States), and the VitalDB dataset from Seoul National University Hospital (South Korea). The datasets include surgical cases recorded between 2016 and 2022, providing multi-institutional perioperative data for model evaluation.

Statistical analysis

Model performance metrics including AUC, accuracy, precision, recall, and F1-score were calculated on the test dataset. To quantify uncertainty, 95% confidence intervals were estimated using bootstrap resampling (1,000 iterations). Statistical comparisons between AUC values of competing models were conducted using the DeLong test, with significance defined at $p < 0.05$.

Handling of class imbalance

Several tasks exhibited substantial outcome imbalance. To preserve the natural clinical incidence of events, no artificial resampling methods (such as oversampling or undersampling) were applied. Model thresholds were determined using the default probability cut-off of 0.5. In addition to ROC curves, precision–recall performance was evaluated, which is particularly informative in low-incidence classification tasks.

Data sources

In this study, perioperative data from two independent medical centres were utilized. The first cohort, representative of large-scale data, consists of 67,134 surgical procedures performed at the University of California, Irvine Medical Centre between 2017 and 2022.[14] The second cohort, representative of medium-scale data, comprises 6,888 surgical cases recorded at Seoul National University Hospital in South Korea, with the majority of data gathered between 2016 and 2017.[5]

Variables with missing rates exceeding 15% were excluded. The remaining missing data were imputed using the multiple imputation by chained equations.[15]

Task design

We designed six surgical classification tasks, each defined by distinct outcome occurrence rates, sample sizes, and data completeness, to simulate issues commonly encountered in surgical data.[1–4] These tasks represent two clinically significant evaluation orientations: one prioritizing recall maximization, and the other emphasizing precision optimization.[2]

Table 1 provides a comprehensive overview of the six classification tasks, demonstrating the deliberate variation in sample sizes (ranging from 375 to 17,312) and outcome incidences (14.4% to 74.4%). This design enables systematic evaluation of TabPFN's performance across different data characteristics commonly encountered in surgical settings. The tasks span two independent medical centers, ensuring external validity and generalizability assessment.

Task A targeted ICU patients aged 18–75 years with ASA classification ≥2. The intraoperative doses of propofol and fentanyl were defined as treatment variables, with the aim of predicting whether bradypnea would occur within 72 hours following surgery.

Hollmann et al.[13] reported that TabPFN outperformed all baseline models on most tasks when sample sizes did not exceed 10,000. This task aims to evaluate the performance of TabPFN in classification tasks where the sample size is close to 10,000 and the outcome variable is highly imbalanced.

Task B targeted patients aged 65-85 years with ASA classification ≥3. The intraoperative doses of propofol and fentanyl were defined as treatment variables. Aiming to predict whether patients will require ICU admission following surgery. This task aims to evaluate the performance of TabPFN in a classification task, using a moderately sized subset derived from a large-scale dataset with a high incidence of outcomes.

Task C targeted patients aged 18-85 years who underwent anaesthesia for over 60 minutes. The intraoperative doses of propofol and fentanyl were defined as treatment variables. Aiming to predict whether patients will require ICU admission following surgery.

The current default running limit for TabPFN is 10,000 samples. Hollmann et al.[13] proposed prospects for future advancements in handling data with sample sizes beyond 10,000. The sample size for this task has exceeded the threshold, and the task aims to conduct a breakthrough test.

Task D targeted patients who underwent laparoscopic cholecystectomy. Aiming to predict whether patients will require postoperative ICU admission. This task aims to evaluate the performance of TabPFN in classification task with a moderate outcome incidence, using a small-sized subset derived from a large-scaledata.

Task E targeted patients aged 18-85 years with ASA classification ≥3. The intraoperative doses of phenylephrine were defined as treatment variables. Aiming to predict postoperative mortality. This task aims to evaluate the performance of TabPFN in classification task with a low outcome incidence, using a small-sized subset derived from a medium-scaledata.

Task F targeted patients aged 18-65 years who underwent liver transplantation. Aiming to predict whether patients will require postoperative ICU admission. This task aims to evaluate the performance of TabPFN in classification task with a moderate outcome incidence, using a small-sized subset derived from a medium-scaledata.

Comparative evaluation

We conducted a comparative analysis of TabPFN's classification performance against five baseline models: XGBoost, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Decision Tree. Following the recommendations of Muntean et al.[16] for evaluating classification models, we used the following metrics: area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1 score.

Table 1. Overview of task data sources, sample sizes, and outcome incidence.

Task	Data source	Sample size	Outcome	Outcome incidence (%)
A	MOVER	9,876	Bradycardia	19.8
B	MOVER	3,507	ICU admission	74.4
C	MOVER	17,312	ICU admission	64.2
D	MOVER	779	ICU admission	41.1
E	VitalDB	815	In-hospital mortality	14.4
F	VitalDB	375	ICU admission	41.6

Baseline model configuration

Baseline machine learning models were implemented using the scikit-learn framework with standard configurations. Logistic regression used L2 regularization with the default solver. Random forest was implemented with 100 trees and default depth parameters. Support vector machines used the radial basis function kernel. XGBoost was implemented with default learning rate and tree depth parameters. Decision tree models used the Gini impurity criterion without pruning. Hyperparameter tuning was not performed to maintain consistency with the TabPFN approach, which does not require model-specific optimisation.

Calibration evaluation

We plotted the decile average calibration (DAC) curves to evaluate the calibration performance of TabPFN. For each model, the predicted probabilities from the independent test set were divided into 10 equally sized bins. Within each bin, we computed the average predicted probability and the observed event frequency. By comparing these two metrics, the DAC curves provide a visual representation of how well the predicted probabilities align with the actual outcomes across the full probability spectrum.

Experimental environment

All experiments were conducted using Python 3.13.5 on Google Colab. The TabPFN model was implemented using PyTorch 2.7.1 and the HuggingFace Transformers library version 4.53.0. Baseline models, including XGBoost (version 3.0.2), Random Forest, Support Vector Machine, Logistic Regression, and Decision Tree, were implemented using scikit-learn version 1.7.0.

TabPFN configuration

The TabPFN model was implemented using the official pretrained architecture provided in the TabPFN library. The model was applied in inference mode without task-specific hyperparameter tuning, following the standard implementation described by Hollmann et al.[13] The model accepts tabular inputs comprising numerical and categorical variables and performs probabilistic classification using transformer-based attention mechanisms. Default configuration settings were used for all experiments to ensure reproducibility and to reflect typical usage scenarios of the foundation model.

Model training and data Splitting

For each task, the dataset was randomly split into training and testing sets with an 80:20 ratio. Stratified sampling was applied to maintain the original outcome incidence within both subsets. Model performance was evaluated on the independent test set. To reduce variance due to random partitioning, experiments were repeated across five independent random splits, and the average performance metrics were reported. Baseline machine learning models were trained using the training dataset and evaluated on the same test dataset to ensure consistent comparison with TabPFN.

Results

Task A

TabPFN achieved the highest accuracy (0.805) and precision (0.703) among all evaluated models, indicating its strong ability to correctly distinguish cases of postoperative bradypnea. In terms of discriminative performance, TabPFN also attained the highest AUC (0.649), outperforming all benchmark models. However, its recall (0.023) and F1-score (0.044) were substantially lower than those of logistic regression (recall = 0.575, F1-score= 0.361) and support vector machine (recall = 0.488, F1-score= 0.337). This suggests that under conditions of low outcome incidence, TabPFN demonstrates limited ability to identify all true positive cases.

In such clinical scenarios, there is a considerable risk of missing many true positives. When predicting postoperative complications, recall is typically considered more critical because false negatives may lead to more severe clinical outcomes.

Task B

TabPFN achieved an AUC of 0.861, surpassing all benchmark models. It also demonstrated the highest overall accuracy (0.807), F1-score (0.882) and recall (0.971), indicating effectiveness in identifying all patients requiring ICU admission.

However, TabPFN's precision (0.808) was lower than that of the baseline models, suggesting a higher rate of false positives. In the context of predicting ICU admissions, an increased number of false positives may result in unnecessary use of ICU resources.

Task C

TabPFN achieved the highest performance across all evaluation metrics, including AUC, accuracy, precision, recall, and F1-score, outperforming all baseline models. TabPFN demonstrates substantial potential for classification tasks with sample sizes exceeding 10,000.

Task D

TabPFN achieved the highest AUC value of 0.786, significantly outperforming all benchmark models. It also demonstrated superior performance in terms of accuracy (0.713), precision (0.722), and F1 score (0.587). The high precision indicates that TabPFN provides reliable predictions for ICU admission, thereby facilitating more efficient utilization of ICU resources. Its recall (0.494) was lower than that of the support vector machine (0.563) and the logistic regression (0.563). However, in this task, recall may be less critical for this task.

Task E

TabPFN achieved the highest AUC (0.882) and superior accuracy (0.895) and precision (1.000), significantly outperforming all benchmark models. However, its recall (0.266) and F1 score (0.419) were relatively lower than most baseline models. In predicting in-hospital mortality, recall is paramount, as failing to detect true in-hospital mortality may have serious implications.

Task F

TabPFN achieved the highest AUC (0.985), significantly outperforming all benchmark models. However, its accuracy, precision, and recall were not the highest among the evaluated models. XGBoost achieved the highest accuracy (0.920) and precision (0.931), while SVM attained the highest recall (1.000).

Figure 1 presents the ROC curve comparisons across all six tasks, providing visual confirmation of TabPFN's discriminative performance. The curves demonstrate TabPFN's superior AUC performance in most tasks, particularly evident in Tasks C, D, E, and F where TabPFN curves consistently lie above those of benchmark models. The ROC analysis supports the quantitative findings, showing TabPFN's strong discriminative ability across diverse surgical classification scenarios.

Table 2 summarizes the comprehensive performance comparison across all tasks and models. The results reveal distinct performance patterns: TabPFN consistently achieved the highest precision in low-incidence tasks (Tasks A, D, E with <20% incidence), while demonstrating superior recall in high-incidence, large-sample tasks (Tasks B and C). This systematic pattern suggests that TabPFN's performance characteristics are predictably influenced by task-specific data properties, enabling informed model selection based on clinical priorities and data characteristics.

DAC curve analysis

As illustrated in Figure 2, the calibration curves for Tasks A and C closely align with the ideal diagonal line, suggesting that the predicted probabilities are highly reliable in these cases. Task B demonstrates reasonably good calibration, although a mild tendency toward overestimation is evident in the medium-probability range. In contrast, Task D shows clear overestimation, with the calibration curve consistently above the diagonal. Task E shows severe miscalibration. When the predicted probability exceeds 0.4, the actual event rate jumps to 1.0, indicating extreme overestimation. In Task F, the calibration curve shows substantial fluctuations across the probability intervals 0.3-0.7, indicating both over- and underestimation in different regions.

Discussion

Surgical data science seeks to improve the quality of interventional healthcare by systematically collecting, organizing, analysing, and utilizing data through modelling techniques.[1–4,10,12,13] Efficient modelling of structured tabular data is crucial for improving the accuracy of outcome predictions and facilitating optimal resource allocation.[13]

Machine learning tools have been increasingly utilized in perioperative care settings. For example, Zeng et al.[17] developed a XGBoost-based predictive model to forecast complications following congenital heart disease surgery. Barker et al.[18]

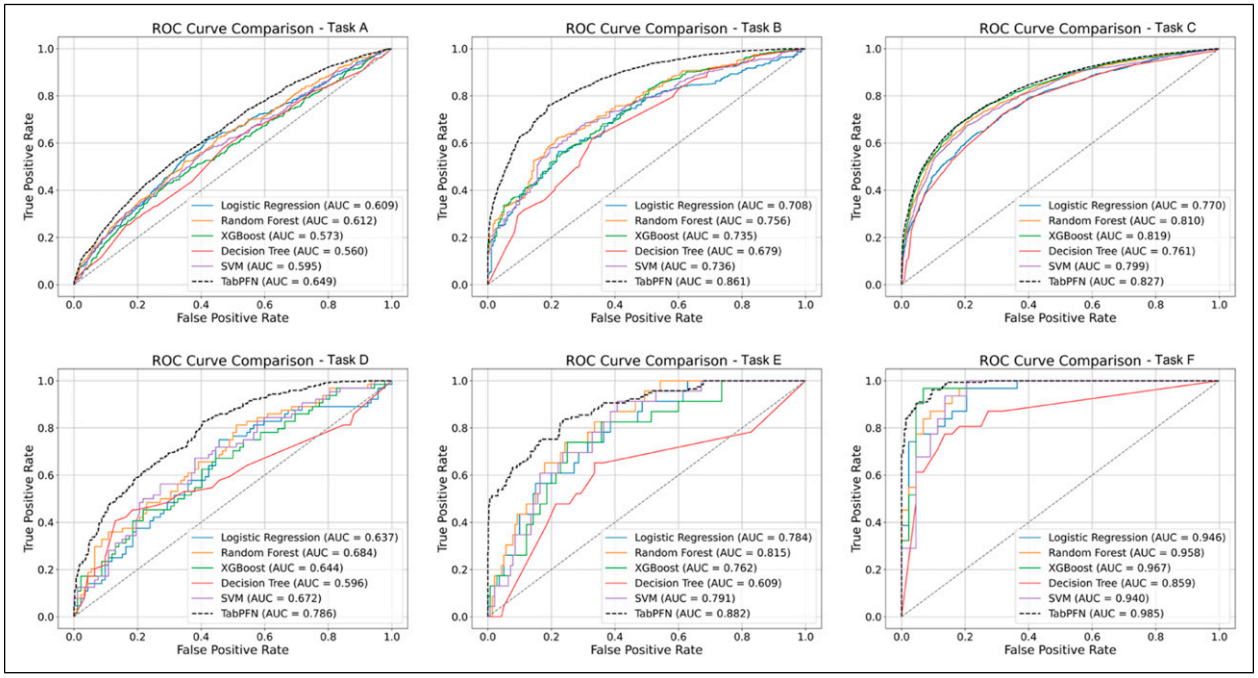

Figure 1. ROC curve comparison of TabPFN and baseline models.

employed an elastic-net regularized logistic regression model to predict the likelihood of unplanned care escalation among surgical patients following their transfer out of the post-anaesthesia care unit. Recent literature has also discussed the potential role of TabPFN-based modelling in broader surgical systems, including applications in healthcare resource optimisation and surgical health economics.[19]

The TabPFN model proposed by Hollmann et al.[13] is expected to become a transformative approach for modelling in the surgical field. They claimed that, on data within 10,000 samples and 500 features, TabPFN outperforms traditional models.

Performance of surgical classification tasks

Our research findings indicate that in task A with a sample size close to 10,000 and highly imbalanced outcome, the recall of TabPFN is lower than that of most baseline models. This phenomenon is also observed in task E. Conversely, in Task B and Task C where the sample size >3,000 and the outcome incidence was relatively high, TabPFN achieved the highest recall.

Hollmann et al.[13] acknowledged that models may systematically underestimate outcomes when those outcomes are underrepresented in training data. Therefore, in classification tasks where recall is prioritized, TabPFN may not be suitable for small samples (n < 1,000) or low outcome incidence (<20%) scenarios. TabPFN may be more suitable for tasks with abundant samples (n > 3,000) and high outcome incidence (>40%).

TabPFN achieved the highest precision in Tasks A, D, and E, all of which are characterized by low outcome incidence (<20%). In contrast, TabPFN did not attain the highest precision in Tasks B, C, and F, which have higher outcome incidence (>40%). These results suggest that for classification tasks where precision is prioritized, TabPFN may be more suitable for tasks with low outcome incidence (<20%).

Model characteristics and generalizability

In our datasets,[5,14] each patient record comprises a distinct set of features, such as demographic variables, intraoperative treatments, and laboratory test results. Given that these data originate from multiple sources and adhere to varying measurement standards, substantial heterogeneity exists across the dataset.

To address this challenge, TabPFN employs a dual attention mechanism: each cell first attends to its row, capturing interactions among all features within an individual patient, and subsequently attends to its column, identifying cross-patient patterns for the same feature. This enables effective modeling of both individual patient characteristics and population-level.

This mechanism endows TabPFN with strong generalization ability and transferability in heterogeneous settings. Hollmann et al.[13] emphasized that strong generalization ability is an important feature of ideal model. With the increasing

Table 2. Comparison of model performance metrics.

Task A				
Model	Accuracy	Precision	Recall	F1-score
Random forest	0.764	0.336	0.199	0.250
Logistic regression	0.597	0.263	0.575	0.361
Support vector machine	0.619	0.257	0.488	0.337
XGBoost	0.777	0.292	0.090	0.137
Decision tree	0.656	0.239	0.338	0.280
TabPFN	**0.805**	**0.703**	**0.023**	**0.044**
Task B				
Model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.635	0.856	0.613	0.714
Decision tree	0.610	0.843	0.584	0.690
Random forest	0.751	0.818	0.854	0.836
XGBoost	0.766	0.808	0.900	0.851
Support vector machine	0.655	0.876	0.625	0.729
TabPFN	**0.807**	**0.808**	**0.971**	**0.882**
Task C				
Model	Accuracy	Precision	Recall	F1-score
XGBoost	0.751	0.790	0.834	0.811
Random forest	0.734	0.821	0.750	0.784
Support vector machine	0.712	0.857	0.663	0.748
Logistic regression	0.693	0.812	0.679	0.740
Decision tree	0.687	0.814	0.664	0.732
TabPFN	**0.758**	**0.798**	**0.834**	**0.816**
Task D				
Model	Accuracy	Precision	Recall	F1-score
Random forest	0.641	0.587	0.422	0.491
XGBoost	0.641	0.583	0.438	0.500
Support vector machine	0.628	0.545	0.563	0.554
Logistic regression	0.603	0.514	0.563	0.537
Decision tree	0.615	0.534	0.484	0.508
TabPFN	**0.713**	**0.722**	**0.494**	**0.587**
Task E				
Model	Accuracy	Precision	Recall	F1-score
Random forest	0.853	0.333	0.043	0.077
XGBoost	0.840	0.364	0.174	0.235
Support vector machine	0.785	0.350	0.609	0.444
Logistic regression	0.663	0.258	0.739	0.382
Decision tree	0.656	0.230	0.609	0.333
TabPFN	**0.895**	**1.000**	**0.266**	**0.419**
Task F				
Model	Accuracy	Precision	Recall	F1-score
Random forest	0.880	0.844	0.871	0.857
XGBoost	0.920	0.931	0.871	0.900
Support vector machine	0.867	0.756	1.000	0.861
Logistic regression	0.840	0.771	0.871	0.818
Decision tree	0.813	0.774	0.774	0.774
TabPFN	**0.915**	**0.878**	**0.923**	**0.900**

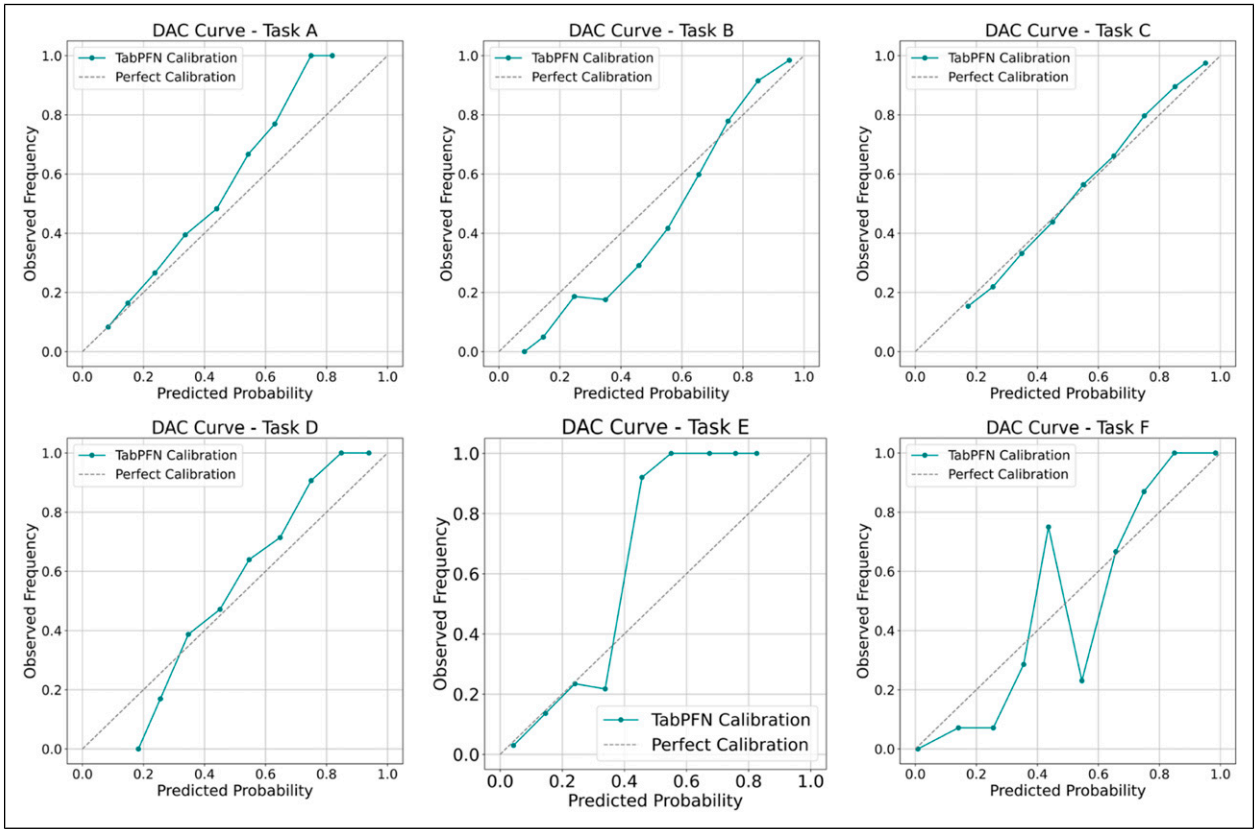

Figure 2. Calibration curves across six tasks.

prevalence of multi-centre collaborations, this generalization ability becomes crucial.[4] However, memory consumption scales linearly with dataset size, rendering it impractical for extremely large datasets.[13]

TabPFN significantly reduces the modelling workload. The model automatically handles missing values, encodes categorical variables, and normalizes numerical features,[13] which is particularly advantageous in perioperative data analysis where missing values and mixed feature types are prevalent.[1–4] In our study, however, we pre-processed the data using MICE imputation.[15] Therefore, we did not evaluate this built-in functionality.

Hollmann et al.[13] did not specify whether surgical domain knowledge was incorporated into the training of these preprocessing capabilities. Although TabPFN can handle missing input features, it cannot infer missing result labels. The performance of classification tasks still fundamentally depends on the quality of the data.

Clinical implications

Resource allocation and healthcare economics

The superior precision of TabPFN in low-incidence scenarios has significant implications for healthcare resource allocation. In settings where false positives are costly (e.g., unnecessary ICU admissions, excessive monitoring), TabPFN's high precision can reduce healthcare expenditure while maintaining patient safety. Conversely, in high-stakes scenarios where missing positive cases have severe consequences (e.g., mortality prediction), the model's lower recall in certain conditions necessitates careful consideration of the clinical cost-benefit ratio.

Implementation in clinical decision support systems

TabPFN's automated preprocessing capabilities and elimination of hyperparameter tuning make it particularly suitable for integration into real-time clinical decision support systems. The model's ability to handle heterogeneous data from multiple sources without extensive preprocessing aligns well with the reality of modern electronic health records, where data

standardization remains challenging. However, the memory scaling limitations suggest that implementation strategies should consider data sampling or distributed computing approaches for large-scale deployments.

Multi-center collaboration and data sharing

The evidence from two different medical centres indicates that TabPFN can support multi-centre collaboration on research projects. It appears that the model's generalisation capabilities could enable the creation of shared prediction models useful across various institutions, thereby addressing a major gap in surgical prediction models. This is especially important for rare surgical procedures, which are not common enough to provide sufficient data for model training and development in a single centre.

Quality improvement and surgical outcomes

The governance of quality improvement processes focused on outcomes is defined by TabPFN's performance attributes. In procedures with a high caseload and a moderate complication rate, increased recall from TabPFN can help with the timely identification of at-risk patients who necessitate early intervention. Focused screening in specialised or low-frequency procedures could benefit from the model's high precision, thereby optimising the trade-off between sensitivity and specificity tailored to the presenting clinical scenario.

Training and education implications

Lowered barriers to advanced predictive modelling granted by TabPFN may influence the design of surgical education and research training. As TabPFN removes technical hurdles, there is potential for faster dissemination of axioms among clinician-researchers, which could accelerate the integration of surgical data science results into practice. Nonetheless, this ease of access must be mitigated by adequate training in interpreting and applying the model meaningfully in clinical contexts.

Regulatory and validation considerations

TabPFN's use of the foundation model approach brings forth significant issues for approval and clinical validation. A model's pre-training on general synthetic datasets, as in TabPFN's case, is likely to improve its generalisability; however, regulatory requirements may necessitate some validation studies focused on specific surgical use cases. As this study's findings on interpretability demonstrated, surgical decision-making needs to incorporate explainable AI techniques.

Future directions

In this study, we designed Task C using 17,312 samples to challenge the default upper limit of the model. The results showed that TabPFN outperformed all benchmark models in most performance metrics. The DAC was highly consistent with the diagonal line, indicating that the predicted probabilities were highly reliable. TabPFN remains highly promising in scenarios with over 10,000 samples.

Model interpretability and explainability

Future research should prioritize the development of interpretability methods specifically tailored for TabPFN in surgical contexts. Given the critical nature of surgical decision-making, explainable AI approaches that can provide clinicians with transparent reasoning for predictions are essential for clinical adoption and regulatory approval.

Domain-specific training and optimization

Applying surgical domain-specific training on TabPFN may improve its performance in medical use cases. Further research should examine whether incorporating surgical domain knowledge into the model's preprocessing and attention mechanisms can improve accuracy and calibration, particularly for rare outcomes.

Real-world clinical validation

Prospective clinical validation studies are necessary to assess the performance and effectiveness of TabPFN in real-world surgical settings. These investigations should determine predictive accuracy alongside other measurements of clinical value, workflow assimilation, impact on patient outcomes, and resource utilisation across the healthcare system.

Multi-modal data integration

Exploratory research should focus on the capability of TabPFN to integrate multi-modal surgical data, including intraoperative monitoring signals, imaging data, and textual clinical notes. Expanding these capabilities would strengthen the predictive and clinical utility of the model in numerous surgical contexts.

Addressing class imbalance and rare events

To address low-incidence TabPFN outcomes, the severe class imbalance in surgically predictive modelling requires further attention. This could involve bespoke class-imbalance solutions, such as ensemble strategies, advanced sampling methods, or hybrids combining TabPFN with more conventional approaches.

Temporal dynamics and longitudinal modeling

Considering the heuristics of TabPFN's capabilities in relation to the temporal aspects of perioperative data is a significant research avenue. Further research should determine how the model can be designed to handle time-series data and forecast results at varying postoperative time intervals predictive modelling.

Federated learning and privacy-preserving approaches

The construction of privacy-preserving federated learning paradigms for TabPFN could support collaborations between multiple institutions while safeguarding patient confidentiality. Acquisition of more sophisticated and more widely applicable surgical prediction models could be achieved through this data-less centralised sharing approach.

Limitations

This study has several limitations. The TabPFN model has inherent constraints, including potential bias reflection from historical data, limited interpretability, and assumptions of data stationarity that may be violated by evolving surgical practices.[4,13] Methodologically, unmeasured confounding factors were not included, and the task design does not necessarily reflect direct clinical value. The evaluation could be improved by incorporating additional performance measures beyond standard classification metrics. Sample sizes for some tasks (<1,000 samples) may have limited statistical power, and the binary classification approach may oversimplify complex clinical outcomes.

Another limitation relates to the engineered design of the six classification tasks. Although the tasks were constructed to represent different combinations of sample size and outcome incidence, they may not fully correspond to clinically validated prediction endpoints. The purpose of this design was methodological benchmarking rather than development of deployable clinical prediction tools. Future studies should evaluate TabPFN using clinically validated prediction targets and prospectively collected datasets. In addition, model comparisons were restricted to conventional machine learning approaches. More recent deep learning architectures specifically designed for tabular data, such as neural decision networks or attention-based tabular models, were not included. Inclusion of these models in future studies would provide a more comprehensive benchmark for evaluating the relative advantages of TabPFN.

Data limitations include the potential impact of missing data patterns despite MICE imputation, temporal bias from different data collection periods (2016-2017 vs 2017-2022), and exclusion of variables with >15% missing data. Generalizability is constrained by the limitation to two academic medical centers, potentially limiting applicability to other healthcare systems. Finally, comparisons were restricted to traditional baseline models, and future studies should include broader model comparisons to better contextualize TabPFN's performance capabilities.

Conclusion

This study evaluated TabPFN performance across six surgical classification tasks using perioperative data from two independent medical centers, comparing it against five benchmark machine learning models. The findings demonstrate that TabPFN's applicability is highly task-dependent, with superior recall in large-sample, high-incidence scenarios (>3,000 samples, >40% incidence) and highest precision in low-incidence conditions (<20%). Cross-institutional validation confirmed strong generalizability, though calibration performance varied significantly with the incidence of the outcome.

Clinical impact

TabPFN represents a transformative advancement in surgical data modeling by eliminating hyperparameter tuning requirements and demonstrating robust cross-institutional performance. However, interpretability limitations and task-dependent performance characteristics indicate it should complement rather than replace established models in clinical practice.

Implementation pathway

Clinical adoption should prioritize high-volume, moderate-to-high incidence scenarios for recall-critical applications, while leveraging its precision advantages for screening in low-incidence conditions. Deployment in small-sample scenarios (<1,000) requires careful consideration, given the superior performance of traditional models in these contexts.

Future directions

Priority research areas include developing clinical interpretability methods, optimizing performance for severely imbalanced datasets, and conducting prospective validation studies to assess real-world clinical utility and workflow integration.

ORCID iDs

Junqi Cui https://orcid.org/0009-0001-6895-5036
Enoch Chi Ngai Lim https://orcid.org/0009-0009-6349-4531
Chi Eung Danforn Lim https://orcid.org/0000-0002-4448-8154

Ethical considerations

This study used only publicly available, fully de-identified datasets obtained from the MOVER dataset (University of California, Irvine Medical Center) and the VitalDB dataset (Seoul National University Hospital). Both datasets are publicly released research resources and were accessed in accordance with their respective data use agreements and repository policies. No institutional data access permissions beyond the published dataset licenses were required. The VitalDB dataset was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. 1703-039-836), with informed consent waived due to the retrospective and anonymized nature of the data. No new patient data were collected for this study.

Author contributions

JC: conceptualisation, data curation, methodology, formal analysis, investigation, project administration, resources, software, writing – original draft, writing – review and editing.
ZH: data curation, resources, software, writing – original draft.
ECNL: formal analysis, resources, visualization, software, validation, writing – review and editing.
CEDL: formal analysis, resources, visualisation, software, validation, supervision, writing – review and editing.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication

Data Availability Statement

All data generated or analysed during this study are included in this published article [and its supplementary information files].

References

1. Kirtac K, Aydin N, Lavanchy JL, et al. Surgical phase recognition: From public datasets to real-world data. *Appl Sci* 2022; 12: 8746. https://doi.org/10.3390/app12178746
2. Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science—from concepts toward clinical translation. *Med Image Anal* 2022; 76: 102306. https://doi.org/10.1016/j.media.2021.102306
3. Levin M, McKechnie T, Kruse CC, et al. Surgical data recording in the operating room: A systematic review of modalities and metrics. *Br J Surg* 2021; 108(6): 613–621. https://doi.org/10.1093/bjs/znab016
4. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng* 2017; 1(9): 691–696. https://doi.org/10.1038/s41551-017-0132-7
5. Lee HC, Park Y, Yoon SB, et al. VitalDB: A high-fidelity multi-parameter vital signs database in surgical patients. *Sci Data* 2022; 9: 279. https://doi.org/10.1038/s41597-022-01411-5
6. Krawczyk B. Learning from imbalanced data: Open challenges and future directions. *Prog Artif Intell* 2016; 5: 221–232. https://doi.org/10.1007/s13748-016-0094-0
7. Kaur H, Pannu HS and Malhi AK. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput Surv* 2019; 52: 1–36. https://doi.org/10.1145/3343440
8. Emmanuel T, Maupong T, Mpoeleng D, et al. A survey on missing data in machine learning. *J Big Data* 2021; 8: 140. https://doi.org/10.1186/s40537-021-00516-9
9. Mirza B, Wang W, Wang J, et al. Machine learning and integrative analysis of biomedical big data. *Genes* 2019; 10: 87. https://doi.org/10.3390/genes10020087
10. Rockenschaub P, Hilbert A, Kossen T, et al. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the ICU. *Crit Care Med* 2024; 52: 1710–1721. https://doi.org/10.1097/CCM.0000000000006359
11. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science* 2024; 383(6679): 164–167. https://doi.org/10.1126/science.adg8538
12. Yang J, Soltan AAS and Clifton DA. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *NPJ Digit Med* 2022; 5: 69. https://doi.org/10.1038/s41746-022-00614-9
13. Hollmann N, Müller S, Purucker L, et al. Accurate predictions on small data with a tabular foundation model. *Nature* 2025; 637: 319–326. https://doi.org/10.1038/s41586-024-08328-6
14. Samad M, Angel M, Rinehart J, et al. Medical Informatics Operating Room Vitals and Events Repository (MOVER): a public-access operating room database. *JAMIA Open* 2023; 6: ooad084. https://doi.org/10.1093/jamiaopen/ooad084
15. White IR, Royston P and Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; 30: 377–399. https://doi.org/10.1002/sim.4067
16. Muntean M and Militaru FD. Metrics for evaluating classification algorithms. In: Ciurea C, Pocatilu P and Filip FG (eds) *Education, Research and Business Technologies. Smart Innovation, Systems and Technologies*. Springer, 2023, vol 321, pp. 307–317. https://doi.org/10.1007/978-981-19-6755-9_24
17. Zeng X, Hu Y, Shu L, et al. Explainable machine-learning predictions for complications after pediatric congenital heart surgery. *Sci Rep* 2021; 11: 17244. https://doi.org/10.1038/s41598-021-96721-w
18. Barker AB, Melvin RL, Godwin RC, et al. Machine learning predicts unplanned care escalations for post-anesthesia care unit patients during the perioperative period: A single-center retrospective study. *J Med Syst* 2024; 48: 69. https://doi.org/10.1007/s10916-024-02085-9
19. Lim E and Lim C. Redefining surgical health economics: The potential of TabPFN for real-time precision modelling. *J Comput Commun* 2025; 13: 30–40. https://doi.org/10.4236/jcc.2025.1311003