Are state-of-the-art machine learning and/or large language models (LLMs) able to predict infection risks after TKA using only pre-operative variables?

Farnaz Rezvani, MS[1], Michael R. Kann, BE[2], Palak Shah, BA[2], Nicole Myers, MSHI, RN[3], Shyam Visweswaran, MD, PhD[4], Sara Tafti [6], Kyle S. Yoon, BS[6], Lauren K. Gross, BS[6], Ahmad P. Tafti, PhD, FAMIA[5], Kenneth L. Urish, MD, PhD[6], Johannes F. Plate, MD, PhD[6]

[1] IEEE Member, London, UK; [2] School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; [3] School of Nursing, University of Pittsburgh, Pittsburgh, USA; [4] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA; [5] School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, USA; [6] Department of Orthopaedic Surgery, University of Pittsburgh, Pittsburgh, PA, USA

Introduction

- Accurate preoperative risk stratification of postoperative infection following total knee arthroplasty (TKA) is essential for clinical decision making
- Machine learning approaches may help predict periprosthetic joint infection (PJI) **[1-3]**
- Performance of machine learning using only preoperative data is unclear
- Role of large language models (LLMs) is unexplored

Aims: Evaluate whether advanced ML models, LLM-based approaches using GPT-4o-mini (OpenAI), and their combinations can reliably predict postoperative infection risk following TKA using preoperative data alone.

Methods

- Used American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) to identify TKA cases from 2021-2023 using Current Procedural Terminology (CPT) codes
- Final cohort included ~196,700 patients with overall infection rate of ~2.6%, reflecting substantial class imbalance **(Figure 1A)**.
- Preoperative variables for model development: demographics, comorbidities, laboratory values, and limited perioperative proxy variables (e.g., case type, transfusion, ventilation, oxygen support)
- Evaluated logistic regression, random forest, gradient boosting, XGBoost, CatBoost, and few-shot large language model (LLM) approaches with undersampling and ensemble methods used to address class imbalance
- Assessed performance using accuracy, recall, precision, F1-score, and receiver operating characteristic area under the curve (ROC-AUC)

Results

- XGBoost: High accuracy (0.92) but poor infection detection (recall 0.13).
- GPT-4o-mini: High infection detection (recall 0.91) but lower accuracy (0.52).
- Hybrid model: Best overall balance, with 0.67 accuracy, 0.72 recall, and 0.68 F1-score **(Table 1).**
- Among 196,700 total knee arthroplasty patients, infection was uncommon; superficial incisional surgical site infection was most common **(Figure 1B)**.

Table 1. Performance Comparison of XGBoost, GPT-4o-mini, and Hybrid Model

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.92	0.07	0.13	0.08
GPT-4o-mini	0.52	0.51	0.91	0.65
Hybrid	0.67	0.65	0.72	0.68

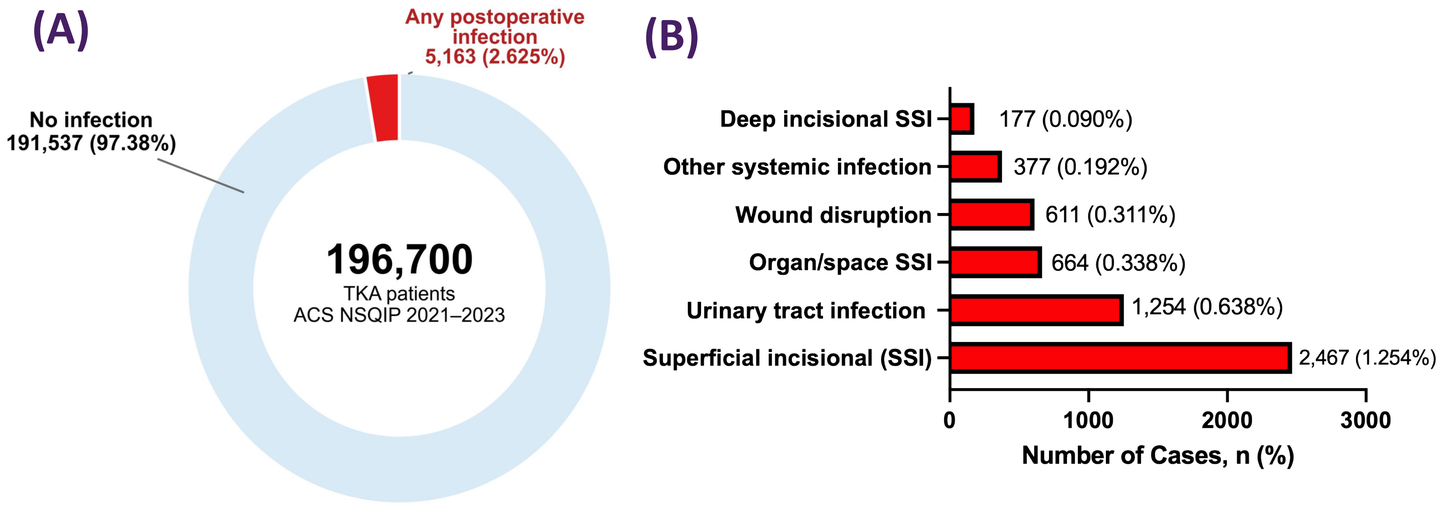

Figure 1. (A) Study cohort and (B) postoperative infection characterization

Conclusions

- Despite advanced ML, LLM, and hybrid approaches, infection prediction after TKA using preoperative variables alone remains limited and unsuitable for clinical use.
- **Clinical Significance:** Clinically meaningful prediction will likely require incorporation of intraoperative and postoperative factors, including surgical technique, contamination risk, and postoperative care.
- **Limitations:** Preoperative data and selected perioperative proxy variables may not adequately capture these important infection-related factors, limiting model performance and generalizability.

References

[1] Kuo et al., J Arthroplasty, 2021 [2] Chong et al., BMC Musculoskeletal Disorders, 2025 [3] Biavardi et al., J Clin Med, 2026

UPMC LIFE CHANGING MEDICINE

Department of Orthopaedic Surgery

University of Pittsburgh

EBJIS-MSIS 2026 | 17-19 SEPTEMBER · PORTO · PORTUGAL