

Machine learning prediction of postoperative adverse outcomes after infrainguinal bypass surgery, with benchmark comparison against the VQI cardiac risk index

Sahar Hanifi MS¹, Vasant Honavar Ph.D.², Faisal Aziz MD³, and Lawrence Sinoway MD⁴

^{1,2} Department of Information Sciences and Technology, The Pennsylvania State University, State College, PA, USA.

^{3,4} College of Medicine, The Pennsylvania State University, Hershey, PA, USA.

Background

Despite significant postoperative morbidity, infrainguinal bypass surgery is still an important treatment option for individuals with advanced peripheral arterial disease (PAD), particularly those with chronic limb-threatening ischemia (CLTI).^{1, 2} Accurate risk prediction can help identify high-risk patients prior to surgery, improve perioperative counseling, and for more effective monitoring and preventative intervention allocation. To the best of our knowledge, limited research has examined at machine learning-based prediction of early postoperative adverse outcomes after infrainguinal bypass surgery, especially for major adverse cardiovascular events (MACE) and major adverse limb events (MALE). This poses a significant analytic gap because these events are clinically important, occur within a well-defined postoperative time window, and may be mostly prevented through earlier risk assessment, closer monitoring, and focused perioperative interventions. While previous research on vascular surgery prediction has mostly concentrated on specific outcomes or conventional statistical models, contemporary machine learning techniques provide the chance to evaluate several clinically significant outcomes concurrently and to take advantage of more comprehensive preoperative data.³

In our current work, using the Society for Vascular Surgery Vascular Quality Initiative (SVS-VQI) registry data, we developed machine learning models to predict multiple 30-day postoperative outcomes after infrainguinal bypass surgery. In addition to this comprehensive outcome modeling study, we conducted a targeted benchmark analysis for perioperative myocardial infarction (POMI) against the published Vascular Quality Initiative Cardiac Risk Index (VQI CRI), a well-known calculation specialized to vascular surgery.⁴ This work uses standardized preoperative observational health data from multiple organizations to estimate short-term postoperative risk, which is a transportable patient-level prediction use case from an OHDSI standpoint. In order to assess transportability, reproducibility, and external performance across heterogeneous observational data sources, we see these findings as the basis for a future OHDSI network study.

Methods

Dataset included 59,880 deidentified SVS-VQI infrainguinal bypass surgery records from 2003 to 2020. After excluding duplicate follow-up records, mismatch between the treated and symptomatic limb, and missing outcome information, we obtained a final sample of 51,527 patients. Additionally, the dataset contained 146 candidate variables comprising demographics, comorbidities, medication use, disease severity, physiologic measurements, and procedural context. To avoid information leaking, only preoperative data were used in the primary study's models; postoperative variables were only used to define outcomes.

30-day MALE, MACE, mortality, reintervention, procedural failure, and a composite of MALE or MACE were among the outcomes of interest. We evaluated XGBoost, Random Forest, Naive Bayes, support vector machine, multilayer perceptron, and Logistic Regression using stratified training/test splits, AUROC,

precision-recall metrics, calibration, and Brier score.

As a targeted secondary analysis, we compared the VQI CRI to perioperative myocardial infarction prediction. Using the published infrainguinal bypass predictor set ⁴, we reconstructed an offline VQI-style comparator and matched the evaluation to a temporal holdout cohort of nonemergent infrainguinal bypass procedures from January 2015 to February 2016. Next, we contrasted this VQI-style comparator with a more comprehensive preoperative machine learning model that was created using a larger preoperative feature set on earlier years. We conducted threshold-matched comparisons in addition to standard discrimination metrics, where the machine learning threshold was chosen to match the VQI-style comparator's false-positive rate. This allowed us to directly evaluate whether machine learning could detect more true POMI cases without raising false positives.

Results

As shown in Table 1, XGBoost performed the best or among the best overall across the entire postoperative outcome analysis. Compared to limb-specific outcomes, systemic outcomes were more predictable. While MALE remained significantly more difficult to predict with an AUROC of roughly 0.65, XGBoost produced AUROC values of roughly 0.80 for 30-day mortality and MACE, 0.76 for procedural failure, 0.73 for the composite adverse outcome, and 0.70 for reintervention. Cardiovascular comorbidities, age, dialysis, functional status, and baseline physiologic severity were significant predictors of systemic outcomes, according to feature importance analyses, while limb-specific outcomes were driven more strongly by disease severity and perfusion measures.

Table 1. Best model performance for 30-day postoperative outcomes after infrainguinal bypass surgery

Outcome	Best model	AUROC	AUPRC
Death	XGBoost	0.80	0.53
MACE	XGBoost	0.80	0.51
MALE	XGBoost	0.65	0.12
Reintervention-only	XGBoost	0.70	0.24
Procedural failure	XGBoost	0.76	0.60
Composite (MALE or MACE)	XGBoost	0.73	0.50

Table 2. Benchmark comparison for perioperative myocardial infarction prediction

Model	Cohort	AUROC	AUPRC	Brier score
Published VQI INFRA model	Published external validation cohort	0.72	Not reported	Not reported
Reconstructed VQI-style comparator	Jan 2015-Feb 2016 holdout	0.71	0.058	0.0248
Richer preoperative ML model	Jan 2015-Feb 2016 holdout	0.76	0.097	0.0258

The published VQI infrainguinal bypass model reported an external-validation AUROC of approximately 0.72 in the targeted POMI benchmark analysis.⁴ The reconstructed VQI-style comparator obtained an AUROC of 0.71, an AUPRC of 0.058, and a Brier score of 0.0248 on our own holdout cohort. With an AUROC of 0.76, an AUPRC of 0.097, and a Brier score of 0.0258, our richer XGBoost model improved discrimination on the same holdout cohort, which is shown in Table 2. The XGBoost model consistently detected more true POMI instances at the same false-positive load at moderate and higher comparator thresholds, according to threshold-matched comparison, which demonstrated that the added value of machine learning depended on the operating point.

For instance, at a VQI-style comparator threshold of 0.05, XGBoost found 33 true positives with 333 false

positives, while the comparator found 22 true positives with the same number of false positives. Similarly, at comparator thresholds of 0.04 and 0.03, XGBoost found 45 vs. 37 and 62 vs. 57 true positives, respectively, at matching false-positive counts. These findings suggest that, particularly in more selective alerting regimes, a richer preoperative model can enhance case capture over a small conventional calculator (Figure 1).

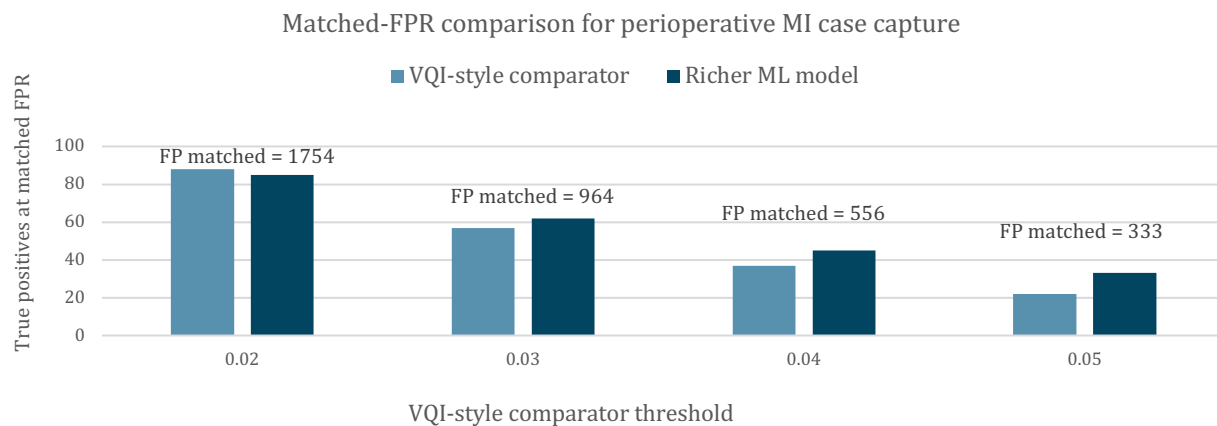


Figure 1. True positives identified by the reconstructed VQI-style comparator and the richer preoperative ML model at matched false-positive rates across clinically relevant calculator thresholds

Conclusion

Overall, these findings demonstrate that machine learning models trained on a large vascular surgery registry can offer clinically meaningful risk prediction for a variety of postoperative outcomes following infrainguinal bypass surgery. These models perform best for systemic outcomes like mortality and MACE, while they perform not as effectively for limb-specific outcomes like MALE. According to the POMI benchmark analysis, richer preoperative machine learning models might improve an established vascular surgery-specific calculator by recognizing more true myocardial infarction patients without raising false positives at clinically meaningful thresholds.

The present findings, however, come from a single registry setting. In order to validate, transport, recalibrate, and stress-test these models across diverse observational databases and healthcare settings, we see the next step as an OHDSI network study. Stronger proof of generalizability, reproducibility, and real-world utility would be provided by such a study, which would also assist in determining whether these prediction models can support more comprehensive perioperative risk stratification across OHDSI data partners.

References

1. Minc SD, Hendricks B, Misra R, Ren Y, Thibault D, Marone L, Smith GS. Geographic variation in amputation rates among patients with diabetes and/or peripheral arterial disease in the rural state of West Virginia identifies areas for improved care. *Journal of vascular surgery*. 2020 May 1;71(5):1708-17.
2. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Medical decision making*. 1993 Dec;13(4):322-38.
3. Dritsas E, Trigka M. Supervised machine learning models for liver disease risk prediction. *Computers* 12 (1): 19 [Internet]. 2023.
4. Bertges DJ, Neal D, Schanzer A, Scali ST, Goodney PP, Eldrup-Jorgensen J, Cronenwett JL, Vascular Quality Initiative. The Vascular Quality Initiative Cardiac Risk Index for prediction of myocardial infarction after vascular surgery. *Journal of vascular surgery*. 2016 Nov 1;64(5):1411-21.