Arch Iran Med. March 2026;29(3):148-155

doi:10.34172/aim.36838 http://journalaim.com

ARCHIVES OF IRANIAN MEDICINE

Open Access

Original Article

Practical Limits of Preoperative Registry-based Risk Prediction After Bariatric Surgery: A National MBSAQIP Temporal Validation

Amir Monshizadeh[1,2*], Masoud Rezvani[3,4], Khosrow Najjari[5], Seyed Hossein Hosseini Nourzad[6], Amir Hossein Latif[7]

[1]Department of Surgery, Tehran University of Medical Sciences, Tehran, Iran
[2]Department of Surgery, School of Medicine, Tehran, Iran
[3]Yale School of Medicine, Division of Bariatrics
[4]Assistant Professor of Surgery, Division of Bariatrics, Department of Surgery, Sentara Northern Virginia Medical Center, Virginia, USA
[5]Department of Surgery, Sina Hospital, Tehran University of Medical Sciences. Tehran, Iran
[6]Department of Project Management, University of Tehran, Tehran, Iran
[7]Department of Surgery, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran

Abstract

Introduction: Preoperative registry variables are widely used for bariatric risk counseling; however, the ceiling of patient-level prediction using routinely collected data remains uncertain.

Objective: To benchmark the performance of logistic regression versus gradient-boosted decision trees for prediction of 30-day outcomes after bariatric surgery using contemporary MBSAQIP data and to evaluate achievable discrimination, calibration, and clinical utility when models are restricted to standard preoperative registry variables.

Methods: We analyzed adults undergoing bariatric surgery in the 2020–2023 MBSAQIP Participant Use Files. Models were developed using 2020–2022 data and temporally validated in an independent 2023 cohort. Using 30 fixed preoperative predictors, we assessed discrimination (AUROC), calibration (Brier score; calibration intercept and slope), and decision-curve utility across 1–10% risk thresholds. Performance was examined by procedure type and procedural intent.

Results: Among 702,614 patients, several outcomes were rare (leak 0.2%, mortality 0.03%). Discrimination was modest across outcomes and subgroups (typical AUROC 0.58–0.62), and XGBoost did not consistently outperform logistic regression. For sleeve gastrectomy readmission, both models demonstrated acceptable population-level calibration (Brier 0.022; slopes near 1), but net clinical benefit was limited. For rare outcomes such as leak, calibration was poor, and no meaningful net benefit was observed.

Conclusion: Using standard preoperative MBSAQIP variables, both logistic regression and XGBoost provide limited individual-level prediction of 30-day outcomes, particularly rare events. Clinically actionable improvement will likely require richer perioperative inputs rather than increased algorithmic complexity alone.

Keywords: Bariatric surgery, Machine learning, Obesity, Postoperative complications risk assessment

Cite this article as: Monshizadeh A, Rezvani M, Najjari K, Hosseini Nourzad SH, Latif AH. Practical Limits of Preoperative Registry-based Risk Prediction After Bariatric Surgery: A National MBSAQIP Temporal Validation. Arch Iran Med 2026;29(3):148-155. doi:10.34172/aim.36838

Received: January 9, 2026, **Revised:** January 25, 2026, **Accepted:** February 17, 2026, **ePublished:** March 1, 2026

Introduction

Metabolic and bariatric surgery is the most effective treatment for severe obesity, producing sustained weight loss and improvement or remission of obesity-related comorbidities. Despite ongoing improvements in perioperative safety, clinically meaningful 30-day complications, readmissions, and reinterventions continue to occur, underscoring the importance of accurate preoperative risk assessment for patient counseling, perioperative planning, and expectation setting.[1]

Existing bariatric risk calculators derived from national registries, including the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) risk/benefit calculator, are designed to provide population-level expected risks to support patient counseling and shared decision-making rather than precise individualized prediction.[2,3] In parallel, machine learning techniques have been increasingly applied to surgical risk prediction, driven by the expectation that greater model flexibility may improve predictive performance.[4,5] As proprietary "black-box" perioperative artificial intelligence tools become more common, surgeons require clear, evidence-based benchmarks for what standard preoperative registry data can (and cannot) predict in routine clinical practice.

However, evidence supporting the superiority of machine learning over logistic regression when applied to structured clinical registry data remains inconsistent. Several bariatric studies have reported higher

*Corresponding Author: Amir Monshizadeh, Email: amir_mz1@yahoo.com

discrimination using machine learning approaches, but many are limited by single-center designs, narrow outcome selection, or the absence of temporal or independent external validation, restricting generalizability.[6,7] More broadly, prior methodological work has demonstrated that when prediction relies on routinely collected registry variables alone, machine learning rarely outperforms traditional regression unless substantially richer or more granular data inputs are available.[8]

Clarifying not only whether machine learning improves prediction, but also where and why prediction fails, is essential for setting realistic expectations for clinical use and guiding future model development. Demonstrating the practical limits of preoperative risk prediction using routinely collected registry variables, therefore, represents both a clinically and methodologically relevant objective.

In this study, we used contemporary national MBSAQIP data to evaluate the practical performance limits of preoperative registry-based risk prediction after bariatric surgery. We compared logistic regression with a commonly used machine learning approach across multiple 30-day postoperative outcomes using temporal validation, assessed discrimination, calibration, and clinical utility, and examined performance across procedure types and surgical intent to benchmark achievable prediction when models are restricted to preoperative registry data alone.

Materials and Methods

Study Design and Data Source

We conducted a retrospective cohort study using the 2020–2023 Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) Participant Use Files.[9] MBSAQIP is a national, prospectively collected registry capturing standardized preoperative variables and 30-day postoperative outcomes for bariatric procedures performed at accredited centers in the United States.

Dataset and Cohort

We analyzed MBSAQIP data from January 1, 2020, through December 31, 2023. Patients were excluded for body mass index (BMI) < 30 or > 80 kg/m², age < 18 or > 80 years, lack of 30-day follow-up, or inability to assign a definitive procedure type using CPT codes. Patients meeting these criteria were included in the analytical cohort. Model development was performed using cases from 2020–2022, and temporal validation was conducted in an independent 2023 cohort.

Subgroup Definitions

Model performance was examined across clinically relevant contexts using mutually exclusive subgroup assignments based on CPT mappings:

1. Procedure type: laparoscopic sleeve gastrectomy (SG), laparoscopic Roux-en-Y gastric bypass (LRYGB), biliopancreatic diversion with duodenal switch (BPD/DS), and laparoscopic adjustable gastric band (LAGB).
2. Procedural intent: initial (primary) bariatric operation, revision after prior bariatric surgery, and conversion from one bariatric procedure to another. Robotic clarification: In MBSAQIP, robotic-assisted procedures are coded as laparoscopic with a robotic modifier; therefore, robotic cases were included within their corresponding laparoscopic procedure groups.

Outcomes

All outcomes followed MBSAQIP registry definitions and were recorded within 30 days. Primary outcomes, adapted from the Bariatric Surgical Risk/Benefit Calculator, included all complications, serious complications, bleeding, surgical site infection, mortality, readmission, reintervention, reoperation, and anastomotic/staple-line leak.[2,7] Composite definitions and secondary outcomes are provided in Supplementary File 1 (Panels A/B).

Predictor Variables and Preprocessing

We used a fixed set of 30 preoperative predictors, extending the original calculator to incorporate laboratory measures and updated variable modifications available in later MBSAQIP datasets. Full predictor details are listed in Supplementary File 1 (Panel C). Continuous variables were median-imputed and categorical variables were mode-imputed. Given the benchmarking aim of this study, we used this transparent approach rather than more complex imputation strategies. Laboratory missingness indicators were incorporated into the modeling pipeline for machine-learning models. Comorbidities were encoded as binary indicators (0/1). Categorical variables (e.g. race, ASA class) were retained as recorded or collapsed *a priori* to preserve clinical interpretability and sample-size stability. All preprocessing parameters were learned from the training period (2020–2022) and applied unchanged to the 2023 temporal validation set. CPT verification was used to ensure consistent procedure-type assignment.

Model Development

Logistic regression models were estimated without regularization to preserve interpretability and maintain comparability with the original MBSAQIP calculator structure. XGBoost models incorporated SMOTE applied only within training folds for rare outcomes to avoid information leakage; sensitivity analyses using class weighting yielded similar conclusions.

Statistical Analysis

Model discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) for all outcomes and average precision for imbalanced outcomes (Supplementary File 2). Precision, recall, and F1 were summarized for rare-event tasks. Calibration was evaluated using calibration plots, Brier score, and calibration intercept and slope. Clinical utility was assessed using decision-curve analysis across 1%–10% risk thresholds in 0.5% increments. Model interpretability was examined using odds ratios for logistic regression and SHAP values for XGBoost.

Software

All models were developed in Python using scikit-learn, XGBoost, shap, imbalanced-learn, and matplotlib within reproducible, cross-validated pipelines.

Ethical Approval

This study used the de-identified MBSAQIP Participant Use File and was determined to be exempt from institutional review board review according to institutional policy. The MBSAQIP dataset contains no direct patient identifiers.

Clinical trial Registration

Not applicable.

Results

Cohort Characteristics

Of 828,481 patients recorded in MBSAQIP from 2020–2023, a total of 702,614 met the inclusion criteria and comprised the final analytic cohort. Baseline characteristics and 30-day outcomes are summarized in Table 1. All eligible cases from the registry during this period were included; no formal sample-size calculation was performed.

The mean age was 43.8 ± 11.7 years, and the mean preoperative body mass index was 44.9 ± 7.5 kg/m^2. Most patients were female (82.7%), White (64.9%), and non-Hispanic (83.0%). ASA class III–IV was common (81.8%), and 99.4% of patients were functionally independent preoperatively. The most prevalent comorbidities included hypertension (44.3%), sleep apnea (37.3%), gastroesophageal reflux disease (32.8%), and hyperlipidemia (23.1%).

Procedures consisted predominantly of laparoscopic sleeve gastrectomy (66.6%) and laparoscopic Roux-en-Y gastric bypass (31.3%), with biliopancreatic diversion with duodenal switch (1.6%) and laparoscopic adjustable gastric banding (0.5%) performed less frequently. Most operations were primary procedures (91.1%), nearly all were minimally invasive (99.9%), and robotic assistance was used in 28.1%. The majority of patients were discharged home (99.2%).

The most common postoperative event was 30-day readmission (3.2%), followed by serious complications (2.9%). Bleeding (0.9%), anastomotic or staple-line leak (0.2%), and mortality (< 0.1%) were rare.

Table 1. Baseline characteristics and 30-day outcomes (MBSAQIP 2020–2023; N = 702,614)

Variable		Value (Number, Mean ± SD or %)
Demographics		
Age (years)		43.8 ± 11.7
BMI (kg/m^2)		44.9 ± 7.5
ASA Class (%)	I	0.2%
	II	18.0%
	III	78.5%
	IV	3.3%
	V	< 0.1%
Functional Status (Independent)		698,560 (99.4%)
Female Sex		580,765 (82.7%)
Race (White)		455,806 (64.9%)
Hispanic Ethnicity		119,547 (17.0%)
Comorbidities		
Diabetes		160,832 (22.9%)
Hypertension		311,507 (44.3%)
Sleep Apnea		262,418 (37.3%)
COPD		8,203 (1.2%)
GERD		230,596 (32.8%)
Smoker		44,161 (6.3%)
Renal Insufficiency		3,649 (0.5%)
Hyperlipidemia		162,283 (23.1%)
Immunosuppression		16,825 (2.4%)
IVC Filter		1,512 (0.2%)
On Dialysis		2,077 (0.3%)
Cardiac History[1]		15,986 (2.3%)
Vascular History[2]		31,401 (4.5%)
Lab Values[3]		
Albumin (g/dL)		4.2 ± 0.4
Creatinine (mg/dL)		0.8 ± 0.5
Hematocrit (%)		41.0 ± 3.7
Hemoglobin A1c (%)		5.9 ± 1.0
Surgical Information		
Laparoscopic Approach[4]		702,108 (99.9%)
Robotic Assistance[5]		197,589 (28.1%)
Discharge Destination (Home)[6]		697,101 (99.2%)
Length of Stay (days)		1.3 ± 1.1
Outcomes (30 Days)		
All Complications[7]		105,698 (15.0%)
Serious Complication[8]		20,037 (2.9%)
Bleeding[9]		6,121 (0.9%)
Surgical Site Infection		5,238 (0.7%)
Mortality		206 (0.03%)
Readmission		22,428 (3.2%)
Reintervention		5,565 (0.8%)

Logistic Regression Performance

Logistic regression identified statistically significant associations across several subgroup–outcome pairs. Effect directions and odds ratios with 95% confidence intervals are reported in Supplementary file 3. Selected statistically significant predictors across representative subgroup–outcome pairs are summarized in Supplementary file 4.

Overall predictive performance was modest. Across outcomes, discrimination was limited, with mean AUROC values at approximately 0.60 ± 0.07. Threshold-dependent performance metrics were poor for many outcomes, with

Table 1. Continued.

Variable	Value (Number, Mean ± SD or %)
Reoperation	7,342 (1.0%)
Anastomotic/staple-line leak	1,386 (0.2%)

[1]Cardiac History: composite variable including history of percutaneous transluminal coronary angioplasty, myocardial infarction, or cardiac surgery.
[2]Vascular History: composite variable including history of deep vein thrombosis, therapeutic anticoagulation, or venous stasis.
[3]Laboratory values: missing data were imputed using median imputation as described in methods.
[4-5]In MBSAQIP, robotic procedures are coded as laparoscopic with a robotic-assisted modifier; thus, robotic cases are included within the laparoscopic group.
[6]The remaining 0.8% were discharged to acute care hospitals, other facilities, or left against medical advice.
[7,8,9]All complications, Serious complications, Bleeding: See Supplementary File 1 for details.
ASA, American Society of Anesthesiologists; BMI, Body mass index; COPD, Chronic obstructive pulmonary disease; GERD, Gastroesophageal reflux disease; IVC, Inferior vena cava.

median precision and F1 scores near zero, particularly for rare events.

XGBoost Discrimination

In the independent 2023 temporal validation cohort, XGBoost models demonstrated modest discrimination across procedures, surgical intent strata, and outcomes. Typical AUROC values ranged from approximately 0.58 to 0.62, with no consistent advantage over logistic regression.

Within sleeve gastrectomy, 30-day readmission and anastomotic or staple-line leak showed the highest discrimination among evaluated outcomes. Threshold-dependent performance metrics remained low for imbalanced outcomes (Supplementary file 2). Table 2 summarizes AUROC across subgroups for the XGBoost models, and Figure 1 displays the corresponding heatmap. Because sleeve gastrectomy comprised the largest subgroup, SHAP interpretation and detailed calibration and decision-curve analyses are presented for sleeve gastrectomy as a representative example; discrimination results across all subgroups are summarized in Table 2 and Figure 1.

Feature Attribution (SHAP)

To aid interpretability of XGBoost models in sleeve gastrectomy, SHAP analyses were performed. For 30-day readmission, Figure 2A shows that serum albumin, hypertension, sleep apnea, hemoglobin A1c, and gastroesophageal reflux disease were among the most influential predictors. For anastomotic or staple-line leak (Figure 2B), serum albumin, gastroesophageal reflux disease, sleep apnea, and hypertension again emerged as leading contributors, followed by Hispanic ethnicity, sex, and prior bariatric surgery.

Missingness indicators contributed to model predictions across outcomes.

Calibration and Clinical Utility

For 30-day readmission in sleeve gastrectomy, both modeling approaches demonstrated acceptable population-level calibration in the 2023 validation cohort. The Brier score was 0.0221 for both models. Calibration intercepts and slopes were +0.088 and 0.976 for XGBoost, and +0.080 and 1.021 for logistic regression, respectively. Calibration curves are shown in Figure 3.

Decision-curve analysis across 1–10% risk thresholds in 0.5% increments demonstrated limited net clinical benefit for both modeling approaches. For anastomotic or staple-line leak, calibration was poor for both approaches, and neither model demonstrated meaningful net benefit, consistent with the very low event prevalence. Calibration and decision-curve results are summarized in Table 3.

Table 2. XGBoost Model discrimination (AUROC) across subgroups and outcomes (2023 temporal validation cohort)

Surgical Subgroup	Postoperative Outcome	AUROC
SG	All Complications	0.586
	Serious Complications	0.595
	30-Day Readmission	0.597
	30-Day Reoperation	0.585
	30-Day Intervention	0.617
	Anastomotic or staple-line leak	0.601
LRYGB	All Complications	0.578
	Serious Complications	0.590
	30-Day Readmission	0.582
	30-Day Reoperation	0.579
BPD/DS	All Complications	0.609
	30-Day Reoperation	0.624
LAGB	30-Day Readmission	0.573
	30-Day Reoperation	0.569
Initial Procedure	Serious Complications	0.592
Revision procedure	Surgical site infection	0.576
Conversion procedure	All Complications	0.588

- AUROC values are for XGBoost models trained on MBSAQIP 2020–2022 and evaluated in the independent 2023 temporal validation cohort.
- Sensitivity analysis: subgroup-trained XGBoost models (trained independently within each stratum) are reported in Supplementary File 5 and are not directly comparable to the global models summarized here.

AUROC, Area under the Receiver Operating Characteristic curve; SG, Sleeve gastrectomy; LRYGB, Laparoscopic Roux-en-Y gastric bypass; BPD/DS, Biliopancreatic diversion with duodenal switch; LAGB,Laparoscopic adjustable gastric band; SSI, Surgical site infection.

Discussion

Using national MBSAQIP data from more than 700,000 patients, we found that preoperative risk models, whether based on logistic regression or gradient-boosted decision trees, had limited ability to predict individual 30-day outcomes after bariatric surgery. Despite rigorous temporal validation in an independent 2023 cohort, increased model complexity did not meaningfully improve discrimination, calibration for rare events, or clinical utility. Together, these findings indicate that routinely collected preoperative registry variables impose an intrinsic ceiling on patient-level risk prediction for many short-term outcomes, particularly uncommon events, and that algorithmic changes alone are unlikely to

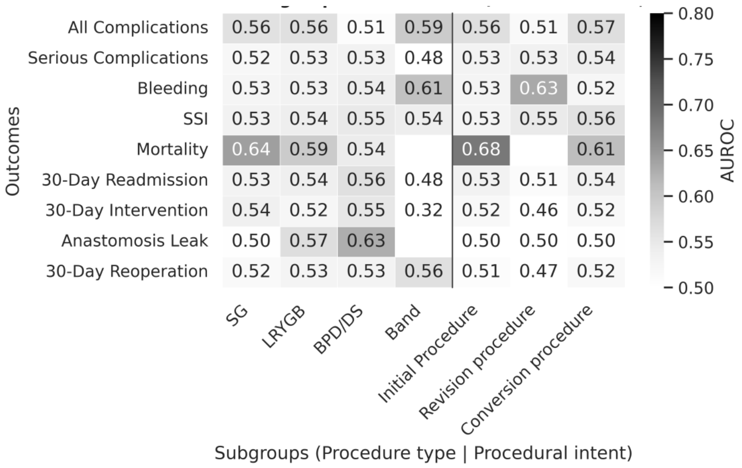

Figure 1. XGBoost discrimination across subgroups and outcomes.
Area under the receiver operating characteristic curve (AUROC) for the XGBoost model, stratified by procedure type (SG, LRYGB, BPD/DS, LAGB) and procedural intent (Initial, Revision, Conversion). Darker shading indicates higher AUROC. AUROC values are from the 2023 temporal validation cohort.

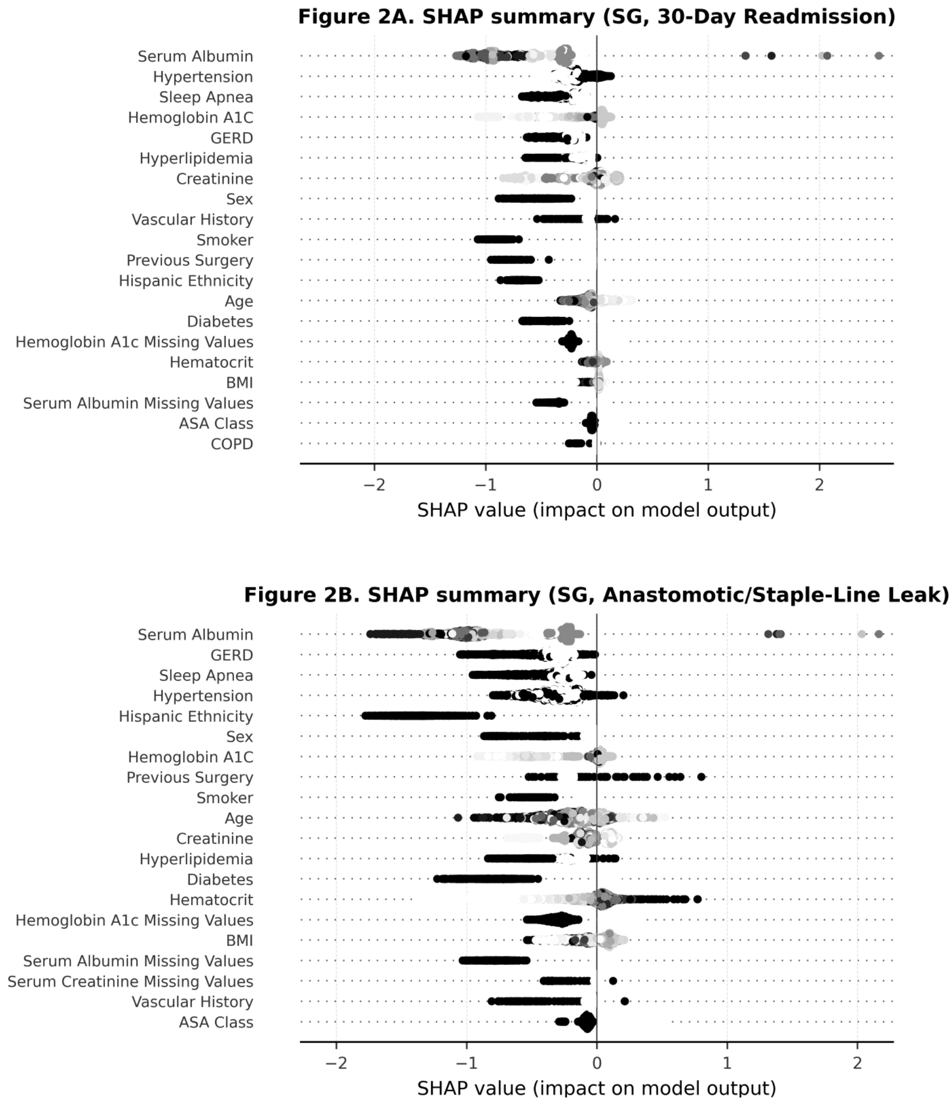

Figure 2. Feature attribution for sleeve gastrectomy (XGBoost, 2023 temporal validation cohort).
Panel A (upper): Key predictors of 30-day readmission; Panel B (lower): Key predictors of 30-day anastomotic/staple-line leak.
Both panels display SHAP summary plots. Each point represents an individual patient; horizontal position indicates the feature's impact on the predicted risk (right = increased risk; left = decreased risk). Features are ranked from top to bottom according to overall influence on model output. Darker shading reflects higher feature values.
Across both outcomes, serum albumin, sleep apnea, gastroesophageal reflux disease (GERD), and hypertension were among the most influential preoperative factors. Variables labeled as "Missing Values" (e.g., serum albumin or hemoglobin A1c missing) indicate that whether the test was obtained also contributed predictive signal and should be interpreted as a marker of preoperative testing patterns rather than physiologic effect. Because anastomotic/staple-line leak was rare in this cohort, predictive performance and calibration for Panel B should be interpreted with caution.

Table 3. Model calibration and clinical utility for key outcomes in the sleeve gastrectomy subgroup (2023 temporal validation cohort)

Outcome	Top SHAP predictors (XGBoost)	Calibration (2023 Validation Cohort)	Decision-Curve Utility (1–10% Risk Thresholds)
30-Day Readmission (SG)	Serum albumin; Hypertension; Sleep apnea; Hemoglobin A1c; GERD	XGBoost: Brier 0.0221, Intercept +0.088, Slope 0.976 Logistic regression: Brier 0.0221, Intercept +0.080, Slope 1.021	Limited net benefit was observed across evaluated thresholds (1–10%) for both models.
Anastomotic/Staple-Line Leak (SG)	Serum albumin; GERD; Sleep apnea	Poor calibration due to extremely low event rate. Slopes unstable for both XGBoost and logistic regression.	No meaningful net benefit for either model across any clinically plausible risk threshold.

- Models used preoperative variables only.
- Calibration interpreted on the 2023 temporal validation cohort.
- Net benefit evaluated using decision-curve analysis at thresholds 1–10% in 0.5% steps.
- Full SHAP summary plots provided in Figure 2.

SG, Sleeve gastrectomy; GERD, Gastroesophageal reflux disease; SHAP, Shapley additive explanations.

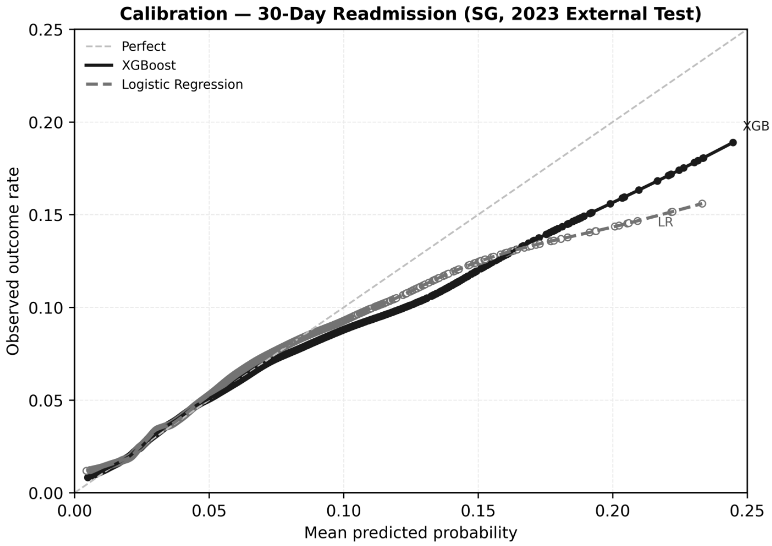

Figure 3. Calibration for 30-day readmission in sleeve gastrectomy (SG), 2023 temporal validation cohort.
A locally weighted smoothing method (LOESS) was used to display the overall relationship between predicted and observed risk. The diagonal represents perfect calibration. Both models showed generally good alignment between predicted and observed risk across the range of probabilities. Brier score, calibration intercept, and slope are reported for each model. Decision-curve analysis for this outcome is provided in Table 3 (threshold range 1–10%).

yield substantial performance gains.

Relation to the MBSAQIP Risk Calculator

The MBSAQIP risk/benefit calculator was developed to support patient counseling and shared decision-making by providing expected risks at the population level rather than precise individual classification.[2] Our findings do not contradict this purpose. Instead, by directly comparing a traditional regression approach with a flexible machine-learning approach within the same registry setting, we demonstrate that modern ML does not automatically enable substantially better individual prediction when model inputs are limited to standard preoperative registry fields. This helps explain why registry-based calculators can be useful for setting expectations and identifying higher-risk groups while still exhibiting modest discrimination for many specific outcomes. In practical terms, such tools may inform counseling but should not be used as binary gatekeepers for individual clinical decisions (for example, deferring surgery based solely on a predicted leak risk) when patient-level separation is limited.

Population-level Calibration Versus Individual Discrimination

For selected outcomes, (most notably readmission after sleeve gastrectomy), predicted risks tracked observed event rates reasonably well at the population level. However, this did not translate into strong separation between higher- and lower-risk individual patients, particularly for rare outcomes such as anastomotic or staple-line leak and mortality. This distinction is clinically important: a model may appear reasonable on average yet remain weak for individual decision-making when event prevalence is low and available predictors provide limited signal.

What the Models Learned

In the XGBoost models, the most influential predictors were a relatively small set of familiar preoperative factors, including serum albumin, sleep apnea, gastroesophageal reflux disease, and hypertension. Laboratory missingness indicators also contributed to predictions, suggesting that test ordering or documentation patterns may encode information beyond physiology alone. We used SHAP to interpret tree-based model behavior, a widely used and well-described approach for explaining feature contributions in these models.[16] The highlighted predictors align with prior bariatric and perioperative evidence: low albumin has been associated with postoperative complications and leak risk in MBSAQIP analyses, obstructive sleep apnea has been linked to postoperative complications, and GERD is relevant to outcomes and patient selection considerations in sleeve gastrectomy populations.[10,11,12] In contrast, hemoglobin A1c is less consistently associated with short-term surgical outcomes, and its apparent influence may partly reflect testing practices and missingness patterns rather than a direct causal effect, consistent with documented variability and trends in preoperative A1c testing.[13]

Clinical Utility

Decision-curve analysis demonstrated, at best, small net clinical benefit for selected common outcomes and little to no benefit for rare outcomes. This pattern is expected when event prevalence is low and discrimination is modest, conditions under which threshold-based decision support is inherently unstable and unlikely to meaningfully change management. In this context, statistically significant associations or highly ranked features do not necessarily translate into actionable clinical risk thresholds.

Context within the Literature

Machine learning has been increasingly applied to surgical risk prediction, often with the expectation that flexible algorithms will outperform conventional regression.[4] In bariatric surgery, some ML studies report improved discrimination for selected endpoints such as readmission, but many focus on single outcomes, narrower cohorts, or alternative feature sets and validation strategies, limiting generalizability.[5] For example, recent machine-learning models incorporating laboratory features have demonstrated moderate discrimination for bariatric readmission, although performance varies across cohorts and modeling approaches.[14] Systematic reviews of AI applications in bariatric surgery emphasize heterogeneity in methods and outcomes, and note that while ML may outperform traditional models in some settings, few algorithms achieve clinically strong discrimination.[6] Outside bariatrics, ML-based surgical calculators show that clinically useful performance can be achieved when prediction tasks have stronger signal and/or richer inputs.[8] Similarly, more complex approaches such as deep neural networks have demonstrated promise in selected perioperative mortality settings, typically relying on more detailed perioperative data than standard preoperative registry variables.[15] Taken together, the broader literature supports a straightforward interpretation of our findings: ML can be helpful in some contexts, but when restricted to structured preoperative registry variables, it often matches rather than surpasses logistic regression, and substantial gains usually require additional information beyond what registries routinely capture.[4]

Limitations

This observational registry analysis is subject to residual confounding and the inherent constraints of registry data. Rare outcomes limited model stability and calibration precision. Laboratory values required imputation, and missingness indicators may reflect practice patterns rather than physiology. Our analysis was restricted to preoperative predictors; intraoperative factors, physiologic time-series data, imaging, and longitudinal postoperative trajectories were not incorporated and may meaningfully improve prediction. This is also the most plausible pathway to clinically meaningful gains: when preoperative registry variables reach an intrinsic ceiling, improvement is more likely to come from richer perioperative inputs rather than algorithm changes alone. Finally, validation was temporal, using a held-out 2023 cohort within MBSAQIP-participating centers; performance may differ in other healthcare systems or data environments.

Conclusion

In a large national MBSAQIP cohort, preoperative models demonstrated modest and often limited patient-level prediction for 30-day outcomes after bariatric surgery, and XGBoost did not provide consistent improvement over traditional logistic regression. These findings support the use of registry-based prediction primarily for population-level counseling rather than high-accuracy individual risk classification, particularly for rare events. Future work aiming for clinically actionable prediction will likely require integration of richer perioperative inputs, including intraoperative variables and longitudinal physiologic data, rather than reliance on algorithmic complexity alone.

Acknowledgments

This study used the fully de-identified MBSAQIP Participant Use File. No external funding was received. We thank the American College of Surgeons and participating centers for data collection and stewardship.

Authors' Contribution

Conceptualization: Amir Monshizadeh, Masoud Rezvani, Khosrow Najjari
Data curation: Amir Hossein Latif, Seyed Hossein Hosseini Nourzad, Amir Monshizadeh
Formal Analysis: Amir Monshizadeh
Methodology: Amir Monshizadeh, Masoud Rezvani
Supervision: Masoud Rezvani, Khosrow Najjari
Writing – original draft: Amir Monshizadeh
Writing – review & editing: Seyed Hossein Hosseini Nourzad, Masoud Rezvani

Competing Interests

The authors declare that they have no competing interests.

Consent for Publication

Not applicable.

Data Availability Statement

The datasets supporting the conclusions of this article are derived from the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) Participant Use File (PUF) and are available from the American College of Surgeons (ACS) under a data-use agreement. Due to licensing and data-use restrictions, the underlying PUF datasets are not publicly available; access may be obtained from ACS through the MBSAQIP PUF application and approval process. Analytic code used for this study is available from the corresponding author upon reasonable request.

Ethical Approval
This study used the fully de-identified Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) Participant Use File (PUF) provided by the American College of Surgeons (ACS). The PUF is HIPAA-compliant and contains no direct or indirect patient identifiers. Under U.S. federal regulations (45 CFR 46.102), analyses of fully de-identified data do not constitute human-subjects research and were therefore deemed exempt from Institutional Review Board oversight; informed consent was not required.

Funding
No specific funding was received for this research.

Intelligence Disclosure
Artificial intelligence tools were used for language editing and manuscript refinement. The authors take full responsibility for the content, interpretation, and accuracy of the manuscript.

Supplementary Files
Supplementary File 1. Definitions of 30-day post-operative composite variables (panel A), list of secondary outcomes (panel B) and predictors (panel C).
Supplementary File 2. Average Precision (AP) for Logistic Regression (LR) and XGBOOST Across Outcomes and Surgical Subgroups
Supplementary File 3. Logistic Regression Model Outputs Across Surgical Subgroups and Outcomes
Supplementary File 4. Significant Predictors from Logistic Regression Models Across Selected Surgical Subgroups
Supplementary File 5. Subgroup-specific XGBoost models trained independently within procedure type and procedural intent strata (sensitivity analysis).

References
1. Arterburn DE, Courcoulas AP. Bariatric surgery for obesity and metabolic conditions in adults. BMJ 2014;349:g3961. doi:10.1136/bmj.g3961
2. Grieco A, Huffman KM, Cohen ME, Hall BL, Morton JM, Ko CY. The Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program bariatric surgical risk/benefit calculator: 1-year weight. Surg Obes Relat Dis 2023;19(7):690-6. doi:10.1016/j.soard.2022.12.028
3. Grieco A, Cohen ME, Hall BL, Morton JM, Ko CY. The Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program bariatric surgical risk/benefit calculator. Surg Obes Relat Dis 2021;17(5):912-20. doi:10.1016/j.soard.2021.01.020
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347-58. doi:10.1056/NEJMra1814259
5. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HM. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive optimal trees in emergency surgery risk (POTTER) calculator. Ann Surg 2018;268(4):574-83. doi:10.1097/sla.0000000000002956
6. Butler LR, Chen KA, Hsu J, Kapadia MR, Gomez SM, Farrell TM. Predicting readmission after bariatric surgery using machine learning. Surg Obes Relat Dis 2023;19(11):1236-44. doi:10.1016/j.soard.2023.05.025
7. Pantelis AG, Epiphaniou P, Lapatsanis DP. Machine learning and artificial intelligence for predicting short and long-term complications following metabolic bariatric surgery-a systematic review. Artif Intell Surg 2025;5(3):322-44. doi:10.20517/ais.2024.104
8. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
9. American College of Surgeons (ACS). User Guide for the 2023 Participant Use Data File (PUF). Chicago, IL: ACS; 2024.
10. Alizadeh RF, Li S, Inaba C, Penalosa P, Hinojosa MW, Smith BR, et al. Risk factors for gastrointestinal leak after bariatric surgery: MBASQIP analysis. J Am Coll Surg 2018;227(1):135-41. doi:10.1016/j.jamcollsurg.2018.03.030
11. Mutter TC, Chateau D, Moffatt M, Ramsey C, Roos LL, Kryger M. A matched cohort study of postoperative outcomes in obstructive sleep apnea: could preoperative diagnosis and treatment prevent complications? Anesthesiology 2014;121(4):707-18. doi:10.1097/aln.0000000000000407
12. DuPree CE, Blair K, Steele SR, Martin MJ. Laparoscopic sleeve gastrectomy in patients with preexisting gastroesophageal reflux disease: a national analysis. JAMA Surg 2014;149(4):328-34. doi:10.1001/jamasurg.2013.4323
13. Hamid SA, Graetz E, Zolfaghari EJ, Schultz KS, Schneider EB, Gibbs KE. Characterization of trends in preoperative hemoglobin A1c testing prior to metabolic and bariatric surgery: a retrospective, observational study. Perioper Med (Lond) 2025;14(1):3. doi:10.1186/s13741-024-00483-8
14. Zhang M, Chen R, Yang Y, Sun X, Shan X. Machine learning analysis of lab tests to predict bariatric readmissions. Sci Rep 2024;14(1):16845. doi:10.1038/s41598-024-67710-6
15. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. Anesthesiology 2018;129(4):649-62. doi:10.1097/aln.0000000000002186
16. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable ai for trees. Nat Mach Intell 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9