

Systematic Review

Open Access



Machine learning and artificial intelligence for predicting short and long-term complications following metabolic bariatric surgery - a systematic review

Athanasios G. Pantelis , Panagiota Epiphaniou, Dimitris P. Lapatsanis

Surgical Department of Obesity and Metabolic Disorders, Athens Medical Group, Psychiko Clinic, Athens 115 25, Greece.

Correspondence to: Dr. Athanasios G. Pantelis, Surgical Department of Obesity and Metabolic Disorders, Athens Medical Group, Psychiko Clinic, 1, Andersen str., Athens 115 25, Greece. E-mail: ath.pantelis@gmail.com

How to cite this article: 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. *Art Int Surg*. 2025;5:322-44. <https://dx.doi.org/10.20517/ais.2024.104>

Received: 9 Dec 2024 **First Decision:** 30 Apr 2025 **Revised:** 29 May 2025 **Accepted:** 11 Jun 2025 **Published:** 2 Jul 2025

Academic Editor: Andrew Gumbs **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Background: Machine learning (ML) and other applications of artificial intelligence (AI) are revolutionizing medicine, particularly in the field of surgery. These models have the potential to outperform traditional predictive tools, aiding clinicians in decision making and enhancing operative safety through improved patient selection.

Methods: A systematic search was conducted across PubMed/MEDLINE and Google Scholar, guided by the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement, to identify studies employing ML and AI algorithms to predict postoperative complications following metabolic bariatric surgery (MBS). The search included primary studies published in English up to November 2024. The area under the receiver operating characteristic curve (AUROC) was used as a surrogate metric for algorithm performance, with values exceeding 0.8 considered clinically significant; however, studies were not excluded based on AUROC thresholds.

Results: The search identified 23 studies meeting the inclusion criteria. These were categorized into seven domains: general complications (8 studies, 34.8%), readmissions after MBS (4 studies, 17.4%), hemorrhage (1 study, 4.3%), leaks (1 study, 4.3%), venous thromboembolism (3 studies, 13.0%), nutritional deficiencies (4 studies, 17.4%), and miscellaneous complications such as gastroesophageal reflux disease, gallbladder disease,



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



and major adverse cardiovascular events (MACE) (3 studies, 13.0%). The studies spanned from 2007 to 2024, with 87.0% (20/23) published in or after 2019. In total, 87 AI/ML algorithms were analyzed. While several studies reported AUROC values exceeding 0.7, the highest achieved was 0.94. However, most studies exhibited methodological limitations, including a lack of external validation and inadequate handling of imbalanced datasets, where complication events were markedly fewer than non-events.

Conclusions: While AI and ML approaches generally outperform traditional predictive models in forecasting postoperative complications following MBS, few algorithms demonstrated clinically significant performance with AUROC values above 0.8. Future research should adopt more rigorous methodologies and implement strategies to address imbalanced datasets, ensuring broader clinical applicability of AI/ML tools.

Keywords: Metabolic bariatric surgery, artificial intelligence, machine learning, postoperative complications, leak, hemorrhage, bleeding, thromboembolism

INTRODUCTION

The integration of artificial intelligence (AI) and machine learning (ML) into healthcare has been transformative, with their applications in surgery growing exponentially. ML is widely regarded as a subset of AI, encompassing algorithms that improve performance through data exposure. In this review, we use AI as an umbrella term that includes ML, deep learning (DL), and natural language processing (NLP). While these terms are sometimes used interchangeably in the literature, we refer specifically to ML when discussing algorithmic models for complication prediction, and reserve “AI” for broader decision-support or data-driven systems. These technologies (primarily ML algorithms as the most widely implemented subset of AI) are now leveraged for preoperative planning, intraoperative guidance, and postoperative monitoring, significantly enhancing surgical precision and decision making. In the field of metabolic bariatric surgery (MBS), ML models and other AI-based tools (e.g., DL or NLP applications) have shown promise in areas such as patient selection, outcome prediction, and complication identification, paving the way for a new era of personalized and data-driven surgical care^[1-3].

Over the years, the safety profile of MBS has markedly improved, thanks to advancements in surgical techniques, optimization of perioperative protocols, and the evolution of surgical equipment. Accumulated experience among bariatric surgeons has also contributed to reducing complications, leading to better patient outcomes. Additionally, the widespread accreditation of dedicated metabolic and bariatric surgical centers has further contributed to safety improvements by promoting standardization of care, implementation of best practices, and multidisciplinary team-based approaches^[4]. As a result, procedures such as sleeve gastrectomy and Roux-en-Y gastric bypass (RYGB) have become safer and more widely accepted as effective interventions for managing obesity and its associated comorbidities. For instance, according to a review published almost a decade ago, bariatric surgery has a safety profile comparable to that of many operations deemed as “routine”, including laparoscopic cholecystectomy, appendectomy, and colectomy, provided it is performed by specialized surgeons in accredited, high-volume centers^[5]. The ASMBS/IFSO guidelines also report a perioperative mortality rate between 0.03% and 0.2%, confirming the procedure’s overall safety profile^[6].

Despite these advancements, the potential for complications persists, making postoperative monitoring and risk stratification critical. Here, AI and ML present an exciting frontier. By analyzing vast datasets of patient characteristics, surgical details, and outcomes, these technologies can identify patterns and predict complications with remarkable accuracy. This predictive capability could enable early intervention and better allocation of resources, ultimately improving patient safety and long-term outcomes.

In this context, exploring the application of AI in predicting complications following MBS offers significant potential. By systematically reviewing the current literature on this topic, we aim to synthesize evidence on the utility of AI and ML in enhancing postoperative care for bariatric patients. This approach will help identify gaps in existing knowledge, highlight promising applications, and provide a roadmap for future research to optimize outcomes and further improve the safety profile of MBS in an era of rapidly advancing medical technology.

METHODS

This systematic review and meta-analysis were conducted according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines to investigate the utility of ML and other applications of AI in predicting complications after MBS^[7]. The study protocol was registered online with [Protocolsof.io](https://www.protocolsof.io).

The PICOS framework was employed to define the study eligibility criteria as follows:

Population (P): Individuals living with obesity, aged 18-65 years, who had undergone MBS of any type. This included, but was not limited to, laparoscopic sleeve gastrectomy (LSG), RYGB, one anastomosis gastric bypass (OAGB), single anastomosis duodenal switch with sleeve gastrectomy (SADI-S), and adjustable gastric banding (AGB), except when AGB was the sole bariatric intervention studied. Both index or revisional procedures were eligible.

Intervention (I): Application of ML (supervised and unsupervised) [Supervised learning refers to algorithms that are trained on labeled datasets with known outcomes (e.g., presence or absence of complications), while unsupervised learning involves identifying patterns or clusters in data without predefined outcome labels], DL, or other AI algorithms to predict and analyze postoperative complications, either within 90 days (early complications) or during a later phase (late complications). This classification is primarily pathophysiological rather than strictly chronological, given that complications such as bleeding, leakage, and venous thromboembolism (VTE) typically occur during the immediate postoperative period, whereas nutritional deficiencies and gastroesophageal reflux disease (GERD) tend to arise in the later phase, and typically beyond the first 90 postoperative days.

Comparison (C): Performance of the AI algorithm(s) compared with conventional or established predictive tests was desirable but not mandatory.

Outcomes (O): Algorithm performance metrics, such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC), as reported in each study. Comparisons with conventional predictive algorithms were desirable but not required. AUROC values were not used as an inclusion or exclusion criterion. Nonetheless, values > 0.8 were considered clinically relevant during the interpretation of model performance.

Study design (S): Retrospective and prospective studies involving any number of participants were included.

Literature search strategy

A systematic literature search of the electronic databases PubMed (MEDLINE) and Google Scholar was conducted by two independent reviewers (AGP, PE). The search terms were organized into three groups:

Group A: Terms related to MBS, including “bariatric”, “gastric band*”, “sleeve”, “gastric bypass”, “duodenal switch”, and “SADI”.

Group B: Terms associated with postoperative complications, such as “complications”, “adverse events”, “morbidity”, “leak*”, “erosion”, “hemorrhage”, “bleeding”, “chole*”, “fever”, “infect*”, “thrombosis”, “embolism”, “pneumonia”, “respiratory”, “cardiovascular”, “infarction”, “kidney”, “renal”, “acute”, “nutritional deficien*”, “anemia”, “calcium”, “vitamin”, “reflux”, “GERD”, “failure”, “hernia”, “weight recurrence”, and “readmission”.

Group C: Terms relevant to AI algorithms, including “artificial intelligence”, “machine learning”, “deep learning”, “natural language processing”, and “neural network”.

Search terms from Groups A, B, and C were combined using Boolean operators (AND and OR) to ensure a comprehensive search strategy. Additionally, the reference lists of included studies were reviewed to identify potentially eligible studies. The search was restricted to English-language publications and included studies available up to November 30, 2024.

Study selection

We considered all studies published in English up to November 30, 2024, conducted on human populations. Duplicate search results were removed before screening abstracts for eligibility. Only primary studies were included; reviews (narrative, scoping, systematic, or meta-analyses), case reports, editorials, letters to the editor, and commentaries were excluded. Additionally, only studies with accessible full texts were considered, leading to the exclusion of conference abstracts.

Studies were excluded if they focused on topics unrelated to our scope, such as education and the learning curve of bariatric procedures, comorbidities [e.g., major adverse cardiovascular events (MACE) as comorbidities of obesity rather than postoperative complications], quality of life, bariatric outcomes (e.g., weight loss or resolution of obesity-related comorbidities), obesity in general (without bariatric surgery), complications of endoscopic interventions (e.g., endoscopic sleeve gastropasty), or computer vision analysis. Studies focused exclusively on robotic surgery or the surgical learning curve were also excluded, as their primary aim did not align with the prediction of postoperative complications.

Full texts of the remaining studies were retrieved for further evaluation by two independent reviewers (AGP, PE). Any selection discrepancies were resolved through discussion, and if consensus could not be reached, a third researcher (DPL) provided the final decision.

Risk of bias assessment

The PROBAST risk of bias (RoB) tool was independently applied by two reviewers (AGP, PE) to assess the methodological quality of each study included in the analysis of ML and other AI models^[8,9]. This tool evaluates RoB across four domains: participant selection, predictors, outcomes, and analysis, providing an overall RoB assessment based on these categories.

Data extraction

The included studies were referenced using Zotero (Corporation for Digital Scholarship), and Microsoft Excel was utilized during the screening and data extraction process. Data were extracted by two independent reviewers (AGP, PE) into an Excel spreadsheet for the following parameters: first author; year of publication; country or countries of the institution(s) involved; DOI (or PMID if the DOI was

unavailable); type of complication; type of surgery; study design (retrospective or prospective); purpose of the prediction (prognostic or diagnostic); total cohort size; number of complications; sizes of the training, test, and validation datasets; top-ranked variables (features); AI algorithms studied; methods used to address data imbalance; and performance metrics, including accuracy, sensitivity, specificity, F1-score, AUROC, area under the precision-recall curve (AUPRC), positive predictive value (PPV), and negative predictive value (NPV).

Data synthesis

A descriptive summary was used to categorize the types of ML/AI models based on the type of complication post-MBS. The discriminative ability of each algorithm was evaluated using metrics such as sensitivity, specificity, accuracy, F1-score, PPV, NPV, AUROC, and AUPRC, as reported for each outcome. Among these metrics, AUROC was considered the most reliable, as it accounts for the model's true positive rate and false positive rate across various cutoff thresholds. AUROC values range from 0.5 (indicating random guessing) to 1.0 (indicating perfect classification), with values > 0.80 generally regarded as clinically useful^[10].

A comparative meta-analysis of the ML/AI models was not feasible due to heterogeneity in study methodologies, outcome reporting, and the lack of comparisons between AI/ML algorithms and conventional predictive models in most studies.

RESULTS

The search strategy identified a total of 1,398 articles after duplicates were removed. These studies were screened for eligibility based on their titles and abstracts. Following this screening, 1,368 studies were excluded, leaving 30 articles for full-text assessment. Ultimately, 23 studies met the criteria for inclusion in the final review. The selection process is summarized in the flowchart presented in [Figure 1](#). [Supplementary Table 1](#) provides details of the studies excluded during the eligibility phase ($n = 7$), along with the reasons for their exclusion.

Of the 23 primary studies included, 20 (87.0%) were published from 2019 onward, with 10 (43.5%) appearing between 2023 and 2024. [Figure 2](#) illustrates the temporal evolution of the included studies. The most common country of origin was the USA, contributing 10 studies (43.5%), followed by Sweden (4 studies, 17.4%), and Iran and China (2 studies each, 8.7%). [Figure 3](#) shows the geographical distribution of these studies.

In terms of AI/ML algorithms, 87 analyses were conducted across the included studies. The most frequently used algorithm was logistic regression (LR), appearing in 16 analyses (18.4%), followed by neural networks (NNs) (not otherwise specified) in 10 analyses (11.5%), random forest (RF) in 9 analyses (10.3%), multilayer perceptron (MLP) in 8 analyses (9.2%), and support vector machine (SVM) and eXtreme gradient boosting (XGB) in 7 analyses each (8.0%). [Figure 4](#) summarizes the frequency of each algorithm's implementation.

The included studies were further categorized based on the type of complication investigated, resulting in seven groups: (1) complications in general (8 studies, 34.8%)^[11-18]; (2) readmissions after MBS (4 studies, 17.4%)^[19-22]; (3) hemorrhage (1 study, 4.3%)^[23]; (4) leak (1 study, 4.3%)^[24]; (5) VTE [including deep venous thrombosis (DVT), pulmonary embolism (PE), and portomesenteric and splenic vein thrombosis (PMSVT), 3 studies, 13.0%]^[24-26]; (6) nutritional deficiencies (including anemia, vitamin deficiencies, micronutrient deficiencies, hypocalcemia, *etc.*, 4 studies, 17.4%)^[27-30]; and (7) miscellaneous, including GERD^[31], gallbladder disease^[32], and MACE (myocardial infarction, cerebrovascular accident, cardiac

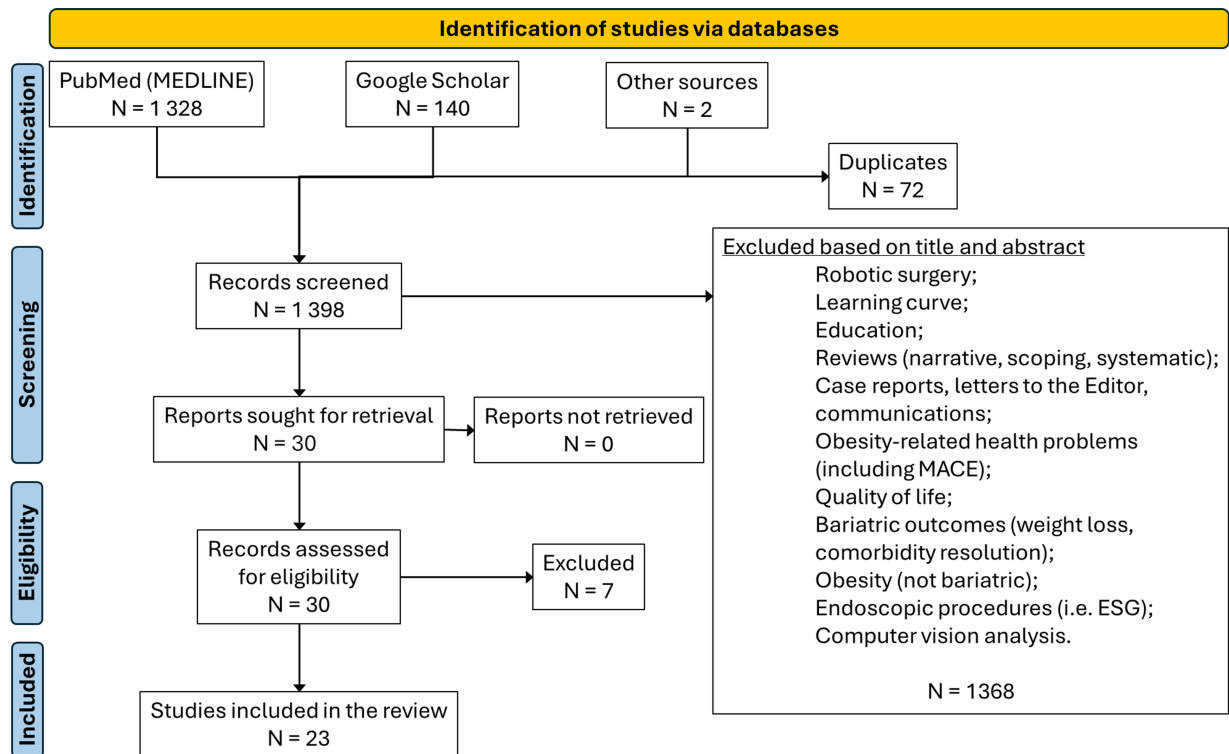


Figure 1. PRISMA flowchart, illustrating the process of selecting eligible publications for inclusion in the systematic review. MACE: Major adverse cardiovascular events; ESG: endoscopic sleeve gastropasty; PRISMA: preferred reporting items for systematic reviews and meta-analyses.

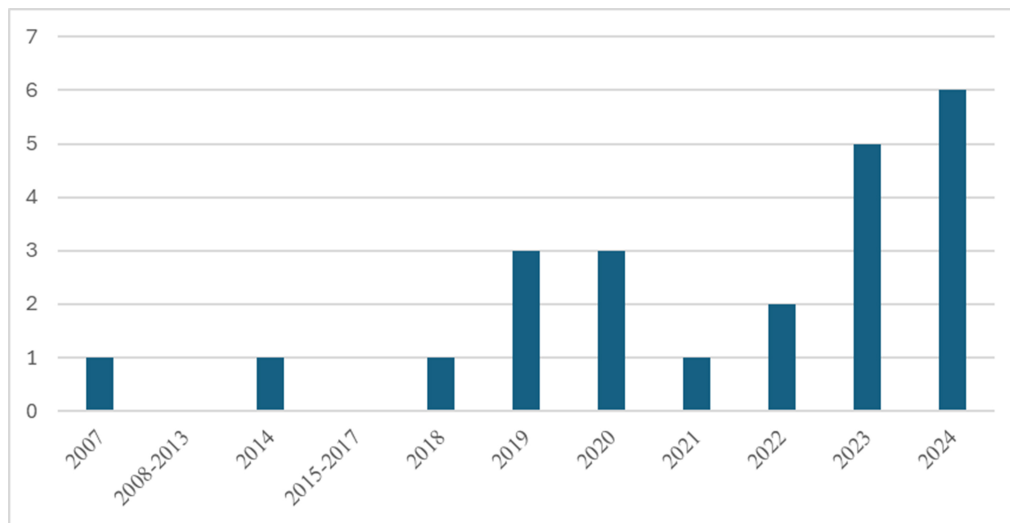


Figure 2. Temporal evolution of the included studies.

arrhythmias, congestive heart failure, cardiac arrest)^[33] (3 studies, 13.0%). Notably, one study belonged to 2 categories (leak and VTE)^[24]. [Figure 5](#) summarizes the distribution of the included studies across categories.

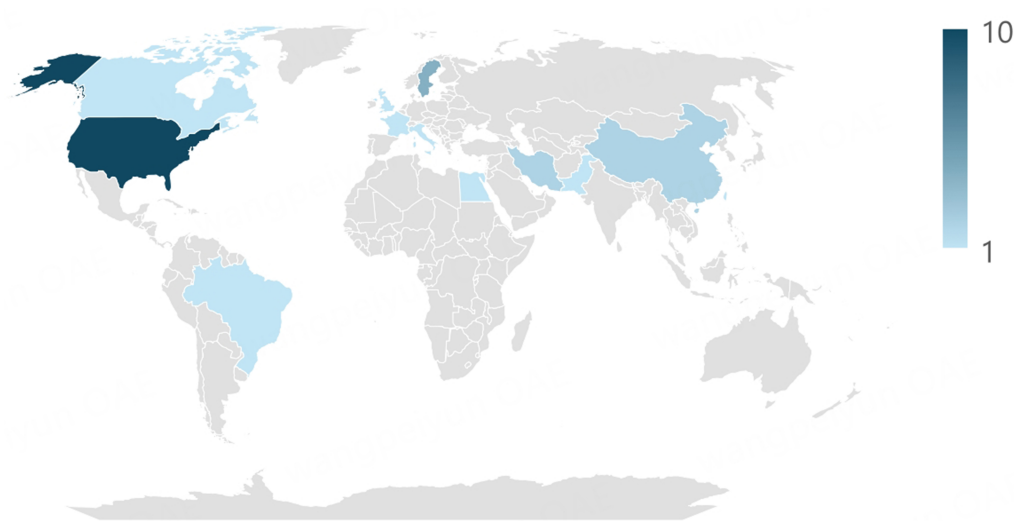


Figure 3. Geographical distribution of the included studies.

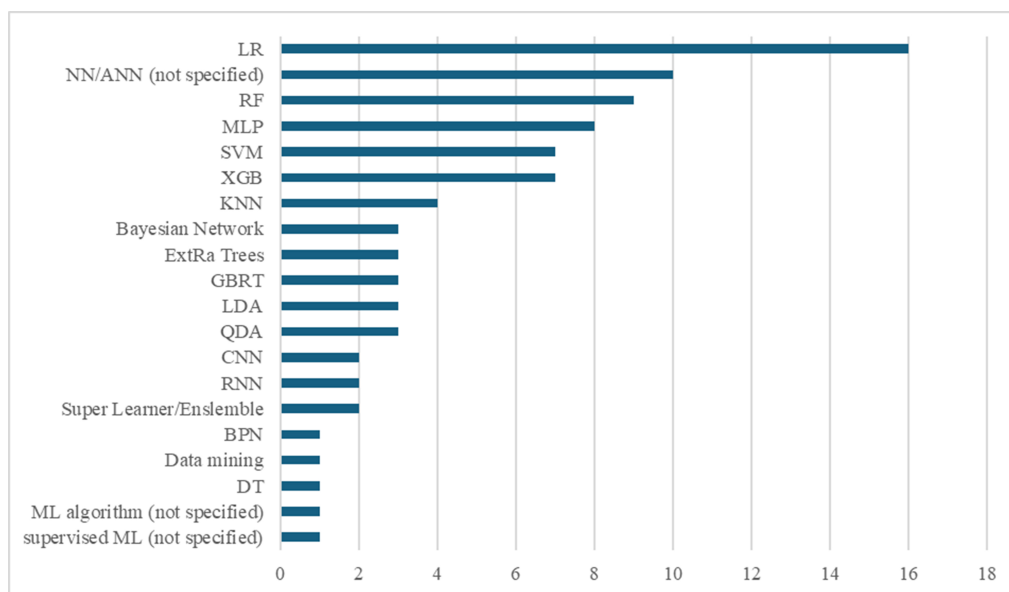


Figure 4. Types and frequency of algorithms implemented in the included studies. LR: Logistic regression; NN: neural network; ANN: artificial neural network; MLP: multilayer perceptron; RF: random forest; SVM: support vector machine; XGB: eXtreme gradient boosting; KNN: k-nearest neighbor; GBRT: gradient boosted regression tree; LDA: linear discriminant analysis; QDA: quadratic discriminant analysis; CNN: convoluted neural network; RNN: recurrent neural network; BPN: backpropagation neural network; DT: decision tree; ML: machine learning.

Complications of MBS in general

As highlighted in [Figure 5](#), this category of complications, in general, included eight primary studies^[11-18] that explored the utility of ML algorithms in predicting postoperative complications cumulatively, without focusing on specific complications.

Cao *et al.* published two studies on this topic. The first evaluated the performance of 29 supervised ML algorithms (both base and ensemble models) in predicting post-MBS complications in a cohort of 37,811 patients from the Scandinavian Obesity Surgery Registry (SOREg) database^[11]. To address the significant

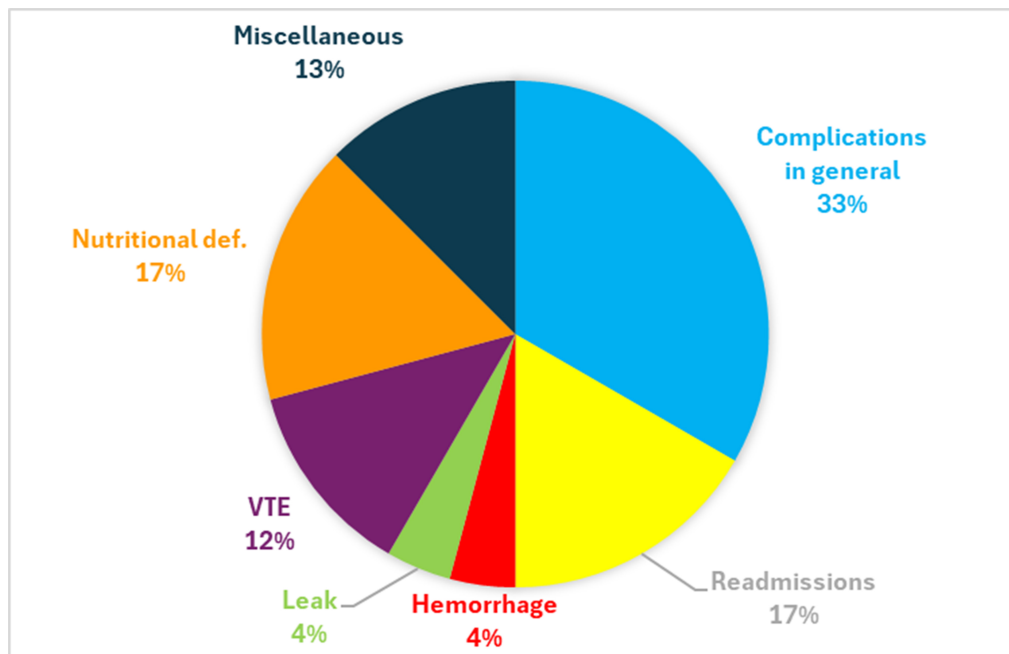


Figure 5. Distribution of the included studies by category. VTE: Venous thromboembolism; def.: deficiency; miscellaneous included gastroesophageal reflux disease, gallbladder disease, and major adverse cardiovascular events.

data imbalance between bariatric surgery outcomes and the occurrence of complications, they employed the synthetic minority oversampling technique (SMOTE). The best-performing algorithms based on AUROC were oversampling RF (0.99), oversampling AdaExtra Trees (0.98), oversampling AdaGradient Tree (0.97), and oversampling bagging k-nearest neighbor (KNN, 0.94) for the training set. However, performance dropped substantially in the test set, with the best algorithm (gradient regression tree and bagging MLP) achieving an AUROC of only 0.58.

In their second study, Cao *et al.* assessed the performance of three supervised DL models [MLP, convolutional neural network (CNN), and recurrent neural network (RNN)], individually and with oversampling techniques (SMOTE), using the same SOReg population^[12]. The best-performing algorithm was oversampling MLP for the training set (AUROC: 0.84, 95%CI: 0.83-0.85). For the test set, MLP and oversampling CNN performed best, each with an AUROC of 0.57 (95%CI: 0.55-0.59 for MLP and 0.55-0.61 for CNN). These studies highlighted that, despite good predictive performance in training datasets after appropriate tuning, there remains significant room for improvement in the clinically relevant test datasets. Notably, the same group of investigators published a relevant study in 2018 on severe complications^[15]. However, that study primarily focused on features, and the analysis was conducted exclusively using multivariate LR, without reporting the AUROC value.

Wise *et al.* published two studies examining the predictive value of LR and artificial neural networks (ANN) for 30-day morbidity and mortality following LSG^[16] and duodenal switch^[17]. Using data from the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) database, the studies analyzed populations of 101,721 and 2,907 patients, respectively. ANN outperformed LR in both cases, but AUROC values ranged from 0.581 to 0.685 - below the clinically meaningful threshold of 0.80. The same group also evaluated LR and ANN for predicting outcomes after revisional RYGB following LSG, with AUROC values ranging from 0.587 to 0.604^[13].

Sheikhtaheri *et al.* utilized MLP models to predict complications following OAGB at 10-day, 1-month, and 3-month intervals postoperatively^[14]. In contrast to earlier findings, this study achieved an AUROC of 0.996 in the validation dataset at the 10-day mark. A notable limitation, however, was the use of data from only five hospitals over five years, yielding a significantly smaller population compared to nationwide registries such as SOReg and MBSAQIP.

More recently, Zucchini *et al.* conducted a retrospective analysis of several ML models (LR, SVM, RF, KNN, MLP, and XGB) for predicting 30-day complications at their bariatric center^[18]. RF achieved an AUROC of 0.94 in the training set and 0.88 in the test set, both outperforming the MBSAQIP predictive tool (AUROC 0.64). Three models (KNN, XGB, and RF) had AUROC values above 0.80, but only XGB achieved 100% sensitivity with a cutoff of 0.05, allowing it to accurately identify both high- and low-risk patients. The authors noted that the study's primary limitation was its small sample size ($N = 424$), which limits the generalizability of the findings.

Table 1 provides a summary of the key characteristics of studies in this category.

Readmissions following MBS

Hospital readmissions within the immediate postoperative period (usually within 30 days after surgery) are a recognized indicator of care quality and have been widely used to evaluate surgical and non-surgical complications across various procedures^[34,35]. This review identified four studies^[19-22] that explored the application of ML algorithms in predicting readmissions following MBS.

Two studies utilized the MBSAQIP database. Torquati *et al.* analyzed a cohort of 393,833 patients with a 3.9% readmission rate^[21]. The study compared the super learner (SL) algorithm with LR, finding that SL demonstrated superior predictive performance (AUROC: 0.674, 95%CI: 0.670-0.679 *vs.* LR: 0.252, 95%CI: 0.249-0.255). However, the predictive value of SL did not reach clinical utility.

Butler *et al.* examined 863,348 patients with a readmission rate of 4.52%^[19], comparing four algorithms: XGB, RF, NNs, and LR. XGB and RF performed equally well (AUROC: 0.785, 95%CI 0.784-0.786 for XGB; 0.784-0.785 for RF), followed by NN (AUROC: 0.754, 95%CI: 0.753-0.754) and LR (AUROC: 0.620, 95%CI: 0.620-0.621; $P < 0.001$). Performance improved when algorithms incorporated all predischarge variables rather than just preoperative and intraoperative factors. Nonetheless, none of the models achieved clinical significance (AUROC > 0.8).

Zhang *et al.* explored ML algorithms for readmission prediction using lab test data from 1,262 patients with a 7.69% readmission rate^[22]. The study applied five algorithms - SVM, LR, MLP, RF, and XGB. AUROC values ranged from 0.743 (95%CI: 0.641-0.845) for XGB to 0.784 (95%CI: 0.691-0.866) for GLM. Although the models performed better than in larger datasets, the small sample size limited the study's generalizability.

Charles-Nelson *et al.* employed formal concept analysis (FCA), a data-mining method rather than an ML algorithm, to analyze one-year readmissions post-MBS^[20]. Using data from the Programme de Mé dicalisation des Systèmes d'Information (PMSI) database, the study included 198,389 procedures performed on 196,323 patients. FCA identified 12 primary reasons for readmissions, including dysphagia (0.4%), vomiting (0.4%), acute gastric hemorrhage (0.6%), ventral hernia (0.7%), medical abortion (0.7%), fistula and acute peritonitis (1.3%), abdominal/pelvic pain (1.9%), cholelithiasis (2.2%), regular follow-up (14.9%), and combinations of the above. While FCA enhances interpretability and clinical pattern recognition, it

Table 1. Study characteristics for complications in general after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|--------------------------------|---------------------|---------|------------|----------------------|--|---------------------|------------------|--------------|---|
| Cao Y ^[11] | 2019 | Sweden | 44,061 | SOREg | SMOTE | LR | 0.560 | 0.540 | Training: age, BMI, HbA1c, HTN, DM, dyspepsia, previous VTE, RBS Test: WC, HbA1c, DLP, RBS |
| | | | | | | LR ^{AB} | 0.480 | 0.470 | |
| | | | | | | LR ^O | 0.430 | 0.510 | |
| | | | | | | LDA | 0.560 | 0.540 | |
| | | | | | | LDA ^B | 0.560 | 0.540 | |
| | | | | | | LDA ^O | 0.460 | 0.520 | |
| | | | | | | QDA | 0.550 | 0.570 | |
| | | | | | | QDA ^B | 0.540 | 0.560 | |
| | | | | | | QDA ^O | 0.550 | 0.480 | |
| | | | | | | DT | 0.500 | 0.500 | |
| | | | | | | RF | 0.520 | 0.520 | |
| | | | | | | RF ^O | 0.990 | 0.510 | |
| | | | | | | XTR | 0.510 | 0.510 | |
| | | | | | | XTR ^{AB} | 0.510 | 0.510 | |
| | | | | | | XTR ^O | 0.980 | 0.480 | |
| | | | | | | GRT | 0.540 | 0.580 | |
| | | | | | | GRT ^{AB} | 0.510 | 0.520 | |
| | | | | | | GRT ^O | 0.970 | 0.510 | |
| | | | | | | KNN | 0.520 | 0.540 | |
| | | | | | | KNN ^B | 0.520 | 0.530 | |
| | | | | | | KNN ^O | 0.940 | 0.540 | |
| | | | | | | SVM | 0.460 | 0.500 | |
| | | | | | | SVM ^{AB} | 0.520 | 0.490 | |
| | | | | | | SVM ^O | 0.440 | 0.490 | |
| | | | | | | MLP | 0.530 | 0.500 | |
| | | | | | | MLP ^B | 0.550 | 0.580 | |
| | | | | | | MLP ^O | 0.370 | 0.540 | |
| | | | | | | DL-NN | 0.550 | 0.540 | |
| | | | | | | DL-NN ^O | 0.670 | 0.560 | |
| Cao Y ^[12] | 2020 | Sweden | 44,061 | SOREg | SMOTE | MLP | 0.600 | 0.570 | Not specified (in total, 5 continuous, and 11 dichotomous) |
| | | | | | | MLP ^O | 0.840 | 0.540 | |
| | | | | | | CNN | 0.580 | 0.550 | |
| | | | | | | CNN ^O | 0.790 | 0.570 | |
| | | | | | | RNN | 0.580 | 0.560 | |
| Scott AW ^[13] | 2024 | USA | 8,895 | MBSAQIP | N/A | RNN ^O | 0.650 | 0.550 | Non-white race, initial BMI, therapeutic anticoagulation |
| | | | | | | Multiv. LR | - | 0.587 | |
| Sheikhtaheri A ^[14] | 2019 | Iran | 1,493 | 5 regional hospitals | SMOTE | ANN | 0.601 | 0.600 | Not specified (in total 32) |
| | | | | | | 10-dy MLP | - | 1.000 | |
| | | | | | | 1-mo MLP | - | 1.000 | |
| Stenberg E ^[15] | 2018 | Sweden | 37,811 | SOREg | N/A | 3-mo MLP | - | 0.930 | RBS, age, BMI, WC, operation year, GERD, dyspepsia |
| | | | | | | LR | - | N/A | |

| | | | | | | | | | |
|----------------------------|------|------------|---------|------------------------|-----|----------------------|--------------|--------------|--|
| Wise ES ^[16] | 2020 | USA | 101,721 | MBSAQIP | N/A | Multiv. LR | - | 0.572 | Age, non-white race, initial BMI, severe HTN, DM, RBS, functional status |
| | | | | | | ANN | 0.581 | 0.590 | |
| Wise E ^[17] | 2023 | USA | 2,907 | MBSAQIP | N/A | Multiv. LR | - | 0.619 | Age, non-white race, cardiac Hx, HTN on ≥ 3 medications, RBS, OSA, Cr |
| | | | | | | ANN | 0.656 | 0.690 | |
| Zucchini N ^[18] | 2024 | USA, Italy | 424 | Local bariatric center | N/A | MLP | 0.670 | 0.650 | ALP, PLT, TG, HbA1c, albumin |
| | | | | | | LR | 0.700 | 0.650 | |
| | | | | | | SVM | 0.790 | 0.780 | |
| | | | | | | KNN | 0.840 | 0.820 | |
| | | | | | | XGB | 0.850 | 0.830 | |
| | | | | | | RF | 0.920 | 0.910 | |
| | | | | | | MBSAQIP [*] | 0.610 | 0.630 | |

^{AB} AdaBoost applied to previous algorithm. ^B Bagging applied to previous algorithm. ^O Oversampling applied to previous algorithm. ^{*} MBSAQIP perioperative risk calculator. Numbers in bold signify clinically meaningful values of AUROC (> 0.80). MBS: Metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; SOReg: Scandinavian Obesity Surgery Registry; MBSAQIP: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program registry; SMOTE: synthetic minority oversampling technique; N/A: not available; LR: logistic regression; LDA: linear discriminant analysis; QDA: quadratic discriminant analysis; DT: decision tree; RF: random forest; XTR: ExtRa trees; GRT: gradient regression trees; KNN: k-nearest neighbor; SVM: support vector machine; MLP: multilayer perceptron; DL-NN: deep-learning neural network; ANN: artificial neural network; multiv.: multivariate; XGB: eXtreme gradient boosting; BMI: body mass index; HbA1c: glycated hemoglobin; HTN: hypertension; DM: diabetes mellitus; VTE: venous thromboembolism; RBS: revisional bariatric surgery; WC: waist circumference; DLP: dyslipidemia; Hx: history; OSA: obstructive sleep apnea; Cr: creatinine; ALP: alkaline phosphatase; PLT: platelets; TG: triglycerides.

does not generate AUROC values, precluding direct comparison with ML-based studies.

Notably, none of these studies addressed the issue of unbalanced data. Table 2 summarizes the key characteristics of studies investigating AI/ML for predicting readmissions after MBS.

Hemorrhage after MBS

Our search identified only one relevant study^[23], which is noteworthy given that postoperative bleeding is the most commonly reported complication following MBS, with an incidence ranging from 0.4%-4.4% after RYGB and 0.4%-3.4% after LSG^[36]. Post-MBS hemorrhage may be intraluminal, intraabdominal, or a combination of both.

In this MBSAQIP-based study, Hsu *et al.* assessed the predictive performance of four ML algorithms - RF, XGB, deep neural networks (NN), and LR - for postoperative gastrointestinal bleeding^[23]. Among these, RF demonstrated the highest predictive accuracy, with an AUROC of 0.764 (±0.019), while LR had the lowest performance, with an AUROC of 0.709 (±0.018).

Table 2. Study characteristics for readmissions after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|----------------------------------|---------------------|---------|------------|------------------------|--|--|--------------------------------------|--|--|
| Butler LR ^[19] | 2024 | USA | 863,348 | MBSAQIP | N/A | LR ^P LR ^{PI} RF ^P RF ^{PI} XGB ^P XGB ^{PI} NN ^P NN ^{PI} | - - - - - - - - | 0.620 0.615 0.785 0.617 0.785 0.640 0.754 0.558 | Intervention or reoperation prior to discharge, unplanned ICU admission, initial procedure, intraoperative transfusion |
| Charles-Nelson A ^[20] | 2020 | France | 196,323 | PMSI | N/A | FCA | - | - | N/A |
| Torquati M ^[21] | 2023 | USA | 393,833 | MBSAQIP | N/A | SL LR | - - | 0.674 0.650 | Bypass, change in BMI, sleeve, HTN on ≥ 3 medications |
| Zhang M ^[22] | 2024 | China | 1,262 | Local bariatric center | N/A | SVM LR MLP RF XGB | - - - - - | 0.784 0.779 0.778 0.751 0.743 | RBC, CRP, UA |

^PAll predischarge variables. ^{PI}Only preoperative and intraoperative variables. MBS: Metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; MBSAQIP: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program registry; PMSI: Programme de Médicalisation des Systèmes d'Information; N/A: not available; LR: logistic regression; RF: random forest; XGB: eXtreme gradient boosting; NN: neural network; FCA: formal concept analysis; SL: super learner; SVM: support vector machine; MLP: multilayer perceptron; ICU: intensive care unit; BMI: body mass index; HTN: hypertension; RBC: red blood cell count; CRP: c-reactive protein; UA: uric acid.

Table 3 summarizes the characteristics and findings of this study.

Leak after MBS

Leaks and fistulas are among the most feared complications of bariatric surgery, with an incidence ranging from 0.5% to 2% in high-volume centers, depending on the type of surgery and whether it is an index or revisional procedure^[37]. Despite the significance of these complications, our search identified only one relevant study.

Specifically, Nudel *et al.* developed and validated three models (ANN, XGB, and LR) to predict two different types of complications post-MBS: leaks and VTE (the latter analyzed in the next section), using the MBSAQIP database^[24]. In this series, the incidence of leaks was 0.7%. ANN demonstrated the highest predictive performance, with an AUROC of 0.75 (95%CI: 0.73-0.78), followed by XGB (0.70, 95%CI: 0.68-0.72) and LR (0.63, 95%CI: 0.61-0.65), with significant differences among all models ($P < 0.001$).

Table 3. Study characteristics for hemorrhage after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|-----------------------------|---------------------|---------|------------|-------------|--|-----------------------|------------------|----------------------------------|--|
| Hsu JL ^[23] | 2023 | USA | 159,950 | MBSAQIP | N/A | LR RF XGB NN | - - - - | 0.709 0.764 0.746 0.741 | Procedure type, Hct, age, operation length, Cr |

MBS: Metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; MBSAQIP: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program registry; N/A: not available; LR: logistic regression; RF: random forest; XGB: eXtreme gradient boosting; NN: neural network; Hct: hematocrit; Cr: creatinine.

Table 4 summarizes the key points of this study.

Venous thromboembolic events after MBS

Bariatric patients are particularly vulnerable to VTE due to a combination of factors, including chronic inflammation that disrupts the venous intimal lining and induces hypercoagulability, surgical stress, anatomical factors, limited mobility leading to venous stasis, the mechanics of laparoscopic surgery, and prolonged operation times. According to a comprehensive meta-analysis of 87 studies with over 2.5 million patients, the cumulative in-hospital incidence of VTE in the laparoscopic era is 0.15%, while the incidence within the first 30 postoperative days rises to 0.50%^[38]. A rarer, yet potentially catastrophic form of VTE unique to MBS is PMSVT, with an incidence of approximately 0.1%^[39].

In our analysis, we identified three relevant studies^[24-26]. As mentioned earlier, Nudel *et al.* developed and validated a series of MBSAQIP-data-driven algorithms to predict both leaks and VTE post-MBS^[24]. For VTE prediction, they evaluated the performance of three ML algorithms (ANN, XGB, LR) and Bariclot, a linear, forward regression statistical model that is less complex than typical ML algorithms. In their study, all three ML algorithms showed similar performances (AUROC: 0.64-0.67, 95%CI: 0.61-0.70), which are clearly below the clinically useful threshold of 0.8, although they performed better than Bariclot. Dang *et al.* compared Bariclot with clinical risk scores such as Caprini and Finks and found a slightly better performance for Bariclot (AUROC 0.602 vs. 0.553-0.582), still below clinical significance^[26].

In a more recent study, Ali *et al.* developed an MBSAQIP-data-driven supervised ML algorithm that incorporated the regression coefficients of six predictors from a pool of 26 features^[25]. This risk model achieved an AUROC of 0.79 (95%CI: 0.63-0.81), which is considered borderline clinically significant but notably better than the previous models.

Table 5 summarizes the key findings from the studies on VTE prediction.

Table 4. Study characteristics for hemorrhage after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|-----------------------------|---------------------|---------|------------|-------------|---|---------------------|------------------|-------------------------|---|
| Nudel J ^[24] | 2021 | USA | 436,807 | MBSAQIP | Over-sampling w/imbalanced-learn Python library | ANN XGB LR | - - - | 0.750 0.700 0.630 | Age, preop BMI, change in BMI, weight, Hct, height, 1st assistant training (attending), albumin |

MBS: Metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; MBSAQIP: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program registry; ANN: artificial neural network; XGB: eXtreme gradient boosting; LR: logistic regression; preope: preoperative; BMI: body mass index; Hct: hematocrit.

Table 5. Study characteristics for VTE after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|-----------------------------|---------------------|---------------|------------|-------------|---|---|------------------|----------------------------------|--|
| Ali H ^[25] | 2024 | USA, Pakistan | 6,526 | MBSAQIP | N/A | Supervised ML - RYGB - LSG - ESG | - - - | 0.790 0.630 0.760 | COPD, length of stay, prior DVT, HbA1c, venous stasis, preop anti-coagulants |
| Dang JT ^[26] | 2019 | Canada | 274,221 | MBSAQIP | N/A | Bariclot (multiv. LR) Finks Caprini | - - - | 0.602 0.582 0.553 | Hx of VTE, operative time, functional status |
| Nudel J ^[24] | 2023 | USA | 436,807 | MBSAQIP | Over-sampling w/imbalanced-learn Python library | ANN XGB LR Bariclot | - - - - | 0.650 0.670 0.640 0.600 | Bypass, change in BMI, sleeve, HTN on ≥ 3 medications |

VTE: Venous thromboembolism; MBS: metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; MBSAQIP: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program registry; PMSI: Programme de Médicalisation des Systèmes d’Information; N/A: not available; LR: logistic regression; RF: random forest; XGB: eXtreme gradient boosting; NN: neural network; FCA: formal concept analysis; SL: super learner; SVM: support vector machine; MLP: multilayer perceptron; ICU: intensive care unit; BMI: body mass index; HTN: hypertension; RBC: red blood cell count; CRP: c-reactive protein; UA: uric acid.

Nutritional deficiencies after MBS

All metabolic bariatric procedures, to varying extents, alter the anatomy and physiology of the gastrointestinal tract. These changes increase patients’ susceptibility to deficiencies in both macro- and micronutrients, potentially leading to serious conditions such as anemia, osteoporosis, and protein malnutrition^[40].

Our search identified four relevant studies that met the inclusion criteria. One study published in 2014 explored the use of a Bayesian network decision-making system for predicting iron deficiency anemia (IDA), folate deficiency, vitamin B12 deficiency, thiamine deficiency, and malnutrition^[27]. While the performance metrics of this study were strong, the small study population limits its generalizability. Nonetheless, it represents pioneering work in the application of AI in healthcare.

In 2023, three studies were published, indicating increased scientific interest in this area. Lenér *et al.* applied ML (RF) to assess the effectiveness of iron supplementation post-RYGB for preventing IDA^[28]. However, neither AUROC values nor details on handling imbalanced data were provided in this study. Pan *et al.* examined the predictive performance of ML for IDA after LSG in premenopausal women^[29]. Their algorithm achieved an AUROC of 0.858 on the training dataset and 0.799 on the test dataset, showing promising clinical potential. Finally, Parrott *et al.* used three ML models to analyze the incidence of vitamin C deficiency in post-MBS patients, revealing a higher prevalence than previously reported in the literature^[30]. Of the three models (Bayesian network with 18 laboratory variables, Bayesian network with 47 demographic variables, and RF with 81 variables), the RF model demonstrated the best predictive performance (AUROC 0.708).

Table 6 summarizes the key characteristics and findings of all four studies.

Miscellaneous complications after MBS

This category encompasses the remaining studies that utilized ML methods to predict complications following MBS but do not fit into any of the previously described categories.

GERD is a well-documented long-term complication that may develop in a subset of patients who have undergone LSG^[41]. Emile *et al.* developed an ensemble model to predict GERD after LSG in a cohort of 441 patients^[31]. Their algorithm achieved an AUROC of 0.93 (95%CI: 0.88-0.99), providing robust evidence of the clinical applicability of AI. One critique of this study is that it did not account for endoscopic findings in a standardized manner, nor did it incorporate Hill's classification. Nevertheless, it paves the way for meaningful future research in the field.

It is well known that cholelithiasis develops in 30%-50% of patients after bariatric surgery, and these individuals face an increased risk of complications, including biliary colic, acute cholecystitis, acute pancreatitis, and bile duct stones^[42,43]. Liew *et al.* were the first to report an ANN-based model for predicting gallbladder disease after MBS^[32]. Their model outperformed traditional LR, achieving an average correct classification rate of 97.14% compared to 88.2%, with a lower type II error rate.

Bariatric patients are at increased risk for cardiovascular complications in the postoperative period due to obesity and its sequelae, with a documented incidence of 1 in 1,000 procedures^[44]. Romero-Velez *et al.* compared three ML models (LR, a single-layer NN, and XGB) to predict MACE within the first 30 postoperative days after MBS, using the MBSAQIP database^[33]. The NN outperformed the other models, with an AUROC close to the threshold of clinical significance (0.798).

Table 7 provides an overview of the key characteristics and findings from all three studies.

RoB assessment

As mentioned earlier, the included studies were evaluated across four domains (Participants, Predictors, Outcome, Analysis). Each study was rated as having a low, unclear, or high probability of bias based on the

Table 6. Study characteristics for nutritional deficiencies after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|-----------------------------|---------------------|---------|------------|------------------------|--|--|-----------------------|--|--|
| Cruz MR ^[27] | 2014 | Brazil | 60 | Shell Netica | N/A | Bayesian network - IDA - THFA - B12 - B1 Malnutrition | - - - - - | 0.839 1.000 1.000 0.982 1.000 | Gender, age, surgery time, Hgb, Hct, MCV, albumin, ferritin, vit-B12, THFA, food intake, physical signs and symptoms of nutrient deficiency |
| Lenér F ^[28] | 2023 | Sweden | 971 | Local bariatric center | N/A | RF LR | - - | N/A N/A | Hgb, TIBC, ferritin, vit-B12, THFA, ESR |
| Pan Y ^[29] | 2023 | China | 407 | Local bariatric center | SMOTE | Linear SMV | 0.858 | 0.799 | Preop ferritin, age, Hgb, Cr, FCP |
| Parrott JM ^[30] | 2023 | UK, USA | 187 | Local bariatric center | Random under-sampling | ML models - BN (18) - BN (47) - RF (81) | - - - | 0.700 0.693 0.708 | Fid 30-100, RDW, GFR, FID > 100, ALT, WBC, RBC, AST, MCHC, CRP, Hct Ethnicity, race, domestic partner, BMI, primary procedure, no. of surgeries |

Numbers in bold signify clinically meaningful values of AUROC (> 0.80). MBS: Metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; SMOTE: synthetic minority oversampling technique; IDA: iron-deficiency anemia; THFA: folate; B12: vitamin B12; B1: vitamin B1; RF: random forest; LR: logistic regression; SMV: support vector machine; BN (18): Bayesian network with 18 lab variables; BN (47): Bayesian network with 47 demographic variables; RF 81: random forest with 81 variables; Hgb: hemoglobin; Hct: hematocrit; MCV: mean corpuscular volume; TIBC: total iron-binding capacity; ESR: erythrocyte sedimentation rate; preop.: preoperative; Cr: creatinine; FCP: fasting C-peptide; FID 30-100: functional iron deficiency with ferritin levels 30-100; FID > 100: functional iron deficiency with ferritin levels > 100; RDW: red cell distribution width; GFR: glomerular filtration rate; ALT: alanine aminotransferase; WBC: white blood cell; RBC: red blood cell; AST: aspartate aminotransferase; MCHC: mean corpuscular hemoglobin concentration; CRP: C-reactive protein; BMI: body mass index; no.: number.

PROBAST criteria, as outlined in the relevant paper by Moons *et al.*^[9]. The overall risk assessment was derived from these four domains, using the following criteria:

- Low risk: All domains rated as low risk, or one domain rated as unclear and the rest low risk.
- Unclear risk:
 - 2-4 domains rated as unclear, rest low risk;
 - or one domain rated as high risk and the rest low risk;
 - or one high risk, one unclear, and the rest low risk.
- High risk:
 - One high risk + 2-3 unclear;
 - or 2-4 domains rated as high risk.

Table 7. Study characteristics for miscellaneous complications after MBS

| First author (citation no.) | Year of publication | Country | Population | Database(s) | Method of dealing with imbalanced data | Algorithms examined | AUROC (training) | AUROC (test) | Top-ranked features/variables |
|--------------------------------|---------------------|---------|------------|------------------------|--|---------------------|------------------|-------------------------|--|
| Emile SH ^[31] | 2022 | Egypt | 441 | Local bariatric center | N/A | Ensemble model | - | 0.93 | Age, weight, preop GERD, bougie size, distance of 1st stapler from pylorus |
| Liew PL ^[32] | 2007 | Taiwan | 117 | Local bariatric center | BPN | ANN | - | - | Chronic inflammation, HbA1c, DBP |
| Romero-Velez G ^[33] | 2024 | USA | 755,506 | MBSAQIP | N/A | LR ANN XGB | - - - | 0.790 0.798 0.787 | Sex, ethnicity, HTN, GERD, COPD, DLP, chronic steroid use, renal insufficiency, dialysis, Hx of DVT/PE, venous stasis, therapeutic anticoag., O ₂ -dependent, OSA, need for mobility device, Hx of MI/PCI, previous cardiac surgery, IDDM, type of surgery, age, BMI, albumin, operative time |

Numbers in bold signify clinically meaningful values of AUROC (> 0.80). MBS: Metabolic bariatric surgery; AUROC: area under the receiver operating characteristic curve; MBSAQIP: Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program registry; BPN: backpropagation; ANN: artificial neural network; LR: logistic regression; XGB: eXtreme gradient boosting; preop: preoperative; GERD: gastroesophageal reflux disease; HbA1c: glycated hemoglobin; DBP: diastolic blood pressure; HTN: hypertension; COPD: chronic obstructive pulmonary disease; DLP: dyslipidemia; Hx: history; DVT: deep venous thrombosis; PE: pulmonary embolism; anticoag.: anticoagulation; O₂: oxygen; OSA: obstructive sleep apnea; MI: myocardial infarction; PCI: percutaneous coronary intervention; IDDM: insulin-dependent diabetes mellitus; BMI: body mass index.

The majority of studies ($N = 18$, 78.2%) were rated as having an unclear risk, primarily due to a high number of studies ($N = 17$, 73.9%) being assessed as having a high probability of bias in the “Analysis” domain. This was largely attributable to the fact that many studies did not employ methods to address imbalanced data, an inherent limitation of studying rare phenomena such as postoperative complications. Notably, three studies were assessed as having an overall “low risk” of bias^[24,29,30]. The detailed results of the RoB assessment for the included studies are presented in [Figure 6](#).

DISCUSSION

This systematic review highlights the evolving role of AI and ML in predicting postoperative complications following MBS. By synthesizing data from diverse studies, our findings reveal key trends, challenges, and opportunities for the integration of these technologies into clinical practice.

AI/ML models have shown promising predictive performance across various postoperative complications, with AUROC values frequently exceeding 0.7 and peaking at 0.94 in certain models. These results emphasize the potential utility of AI/ML in the early identification of high-risk patients, enabling personalized perioperative management strategies. Specific findings, such as robust predictions for VTE and infections, underscore the feasibility of targeted interventions to mitigate complications. Importantly, while we focused on AUROC for consistency and comparability, several of the included studies also reported sensitivity, specificity, or accuracy, while only one reported the F1-score^[18]. In line with current recommendations for reporting ML models in biomedical research, such as those outlined by Luo *et al.*, future studies should aim for more consistent inclusion of complementary metrics to enhance clinical interpretability and support meaningful comparisons^[45].

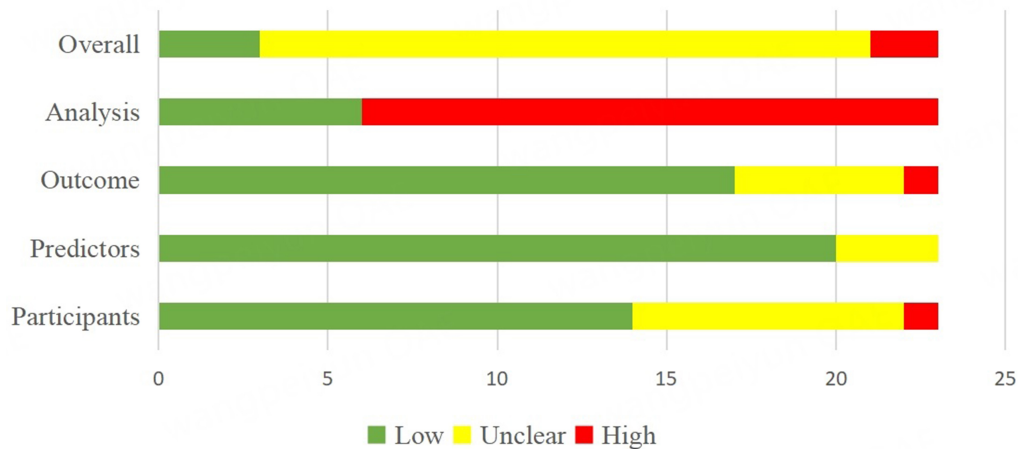


Figure 6. RoB in the included studies, according to the PROBAST tool. The X-axis represents the number of included studies ($N = 23$). RoB: Risk of bias.

Notably, cardiovascular complications, often multifactorial and influenced by patient comorbidities, were associated with high predictive accuracy in studies using large datasets. This aligns with the critical need for enhanced cardiovascular risk stratification in MBS patients, a population inherently predisposed to cardiovascular events due to obesity-related pathophysiology.

Despite these promising results, significant limitations exist in the current body of research. Several studies suffered from small sample sizes or imbalanced datasets, particularly in predicting rare complications. This may lead to biased models and reduced generalizability. Some studies also demonstrated signs of overfitting^[11,12], with models performing well in training but significantly worse in test datasets, suggesting that they captured noise rather than generalizable patterns^[46]. Future studies should mitigate this by employing techniques such as cross-validation, regularization, or dropout in DL models^[47]. This issue underscores the importance of applying mitigation techniques such as cross-validation, regularization, and dropout, particularly in DL models. Importantly, we did not exclude studies based on AUROC thresholds; models with values below 0.8 were retained when they addressed clinically relevant outcomes or introduced novel methodological approaches. The difficulty of achieving generalizable predictions was particularly evident in algorithms predicting miscellaneous complications, which demonstrated inconsistent performance, likely due to sparse data. Moreover, the issue of imbalanced data has been addressed by several authors^[11,12,48]. The rarity of postoperative complications underscores the need for future research to address this challenge by employing appropriate methods, such as SMOTE. This issue was a key concern regarding methodological quality and the potential for bias in the included studies. Several authors employed mitigation strategies such as SMOTE, which is widely regarded as an appropriate method to address class imbalance by synthetically generating minority class examples, thus enhancing algorithm learning and reducing bias toward the majority class.

Additionally, a notable gap was the lack of external validation in most studies, raising concerns about the applicability of these models across different clinical settings. Only a few studies employed multi-center data or tested models on independent cohorts. Without external validation, it is difficult to determine whether a model trained on a specific patient population will perform equally well in another clinical context. As a result, most of the included models remain investigational and cannot yet be implemented in routine decision making. Notably, the majority of the included studies drew from large, well-established registries such as MBSAQIP and SOReg, which enhances the relevance and generalizability of the findings; however,

even registry-based studies often lacked external validation. To bridge this gap, future research should emphasize prospective validation in independent and diverse cohorts, ideally across multiple institutions. Finally, while advanced algorithms like DL were associated with higher accuracy, their “black-box” nature presents challenges in clinical adoption. Interpretability remains a critical hurdle, as clinicians require clear explanations of predictions to trust and act on model outputs.

Our findings align with prior reviews that underscore the promise of AI/ML in surgical outcome prediction. AI/ML is a promising tool that outperforms traditional predictive tools and may assist decision making^[49-53]. In a recent meta-analysis in the field of gastrointestinal surgery, Wang *et al.* compared 62 LR models with 143 ML models, reporting that the ML models demonstrated superior mean performance (difference in AUROC: 0.07; 95%CI: 0.004-0.009; $P < 0.001$)^[53].

Our review uniquely focuses on MBS, highlighting its distinct challenges, such as heterogeneity in patient profiles and surgical techniques. Compared to other surgical fields, MBS research on AI/ML appears to lag in terms of external validation and real-world implementation, reflecting a need for concerted efforts to standardize methodologies. A recent review in the field of MBS focused on the application of ML in predicting postoperative complications^[54]. The author identified seven studies, four of which were also included in our analysis. However, this review did not distinguish between complications arising from MBS and those related to obesity itself. Additionally, it excluded studies addressing readmissions and those focused on specific complications such as leaks, hemorrhage, VTE, and GERD. Along the same lines, another recent review investigated the role of AI in predicting bariatric surgery complications^[55]. This review also included seven studies, with significant overlap with the previous one. Notably, at least one of the included studies did not focus on postoperative complications but rather on long-term outcomes, such as weight loss and remission of obesity-related health problems. We believe our study offers a broader yet more focused perspective by incorporating a larger number of studies specifically dedicated to both short- and long-term complications.

Another important methodological consideration is the heterogeneity of surgical procedures across the included studies. However, we found that the overwhelming majority focused on LSG and RYGB, either explicitly or as part of registry-based cohorts where these procedures dominate (e.g., MBSAQIP, SOReg). Only two studies exclusively examined other operations: Sheikhtaheri *et al.* focused on OAGB^[14], and Wise *et al.* analyzed outcomes after duodenal switch^[17]. Although some studies included a small proportion of other procedures, such as gastric banding, their findings are still highly relevant to LSG/RYGB populations^[20]. Conversely, other studies that included procedures such as single-anastomosis sleeve-jejunal bypass and transit bipartition were considered outliers and interpreted accordingly^[22]. While this variability is a limitation, it does not substantially compromise the generalizability of our conclusions for the two dominant procedures globally. [Supplementary Table 2](#) has also been provided to list the subset of studies that focused exclusively on LSG and/or RYGB.

A related but distinct source of heterogeneity is the inclusion of revisional bariatric procedures. Revisional surgeries are known to carry higher perioperative risk and may affect model performance if not explicitly accounted for^[56,57]. In our review, several studies excluded revisional cases altogether^[22-26,28-33], while others included them as covariates but did not report stratified performance metrics^[11,12,15]. Only one study (Scott *et al.*) focused exclusively on revisional procedures and reported separate predictive values^[13]. The lack of stratification in the remainder of the studies limited our ability to evaluate the impact of revisional status on model accuracy. This introduces an additional layer of complexity in interpreting pooled findings and highlights the need for future models to either stratify or develop dedicated predictive tools for revisional

bariatric surgery.

The clinical relevance of ML in bariatric surgery is growing, with early successes pointing toward both preoperative and postoperative applications. A notable example is the SOPHIA study, which developed and externally validated a ML-based calculator using data from 10 prospective cohorts and 2 randomized trials to predict 5-year postoperative weight trajectories^[58]. This tool exemplifies how interpretable models can inform patient selection, support shared decision making, and guide long-term follow-up. In parallel, embedded systems enhanced by AI and connected through the Internet of Things are emerging as real-time supports during the perioperative course. These include tools capable of recognizing surgical steps through computer vision, predicting remaining surgery duration, or analyzing postoperative physiological data (e.g., heart rate patterns) via wearable devices to detect complications such as leaks. Our recent work outlines how such technologies may enhance both the safety and personalization of MBS, while also highlighting the ethical and implementation challenges involved. For these tools to reach clinical maturity, future efforts must focus on prospective validation, seamless EHR integration, and multidisciplinary collaboration that ensures clinical interpretability and accountability.

AI/ML holds the potential to revolutionize perioperative care in MBS by enabling risk stratification, optimizing resource allocation, and guiding tailored interventions. However, clinicians and researchers must address the limitations highlighted to realize this potential fully. Future efforts should prioritize the use of multi-center datasets to enhance model robustness and generalizability, integration of explainable AI frameworks to improve transparency and clinical acceptance, and rigorous validation studies that assess the impact of these models on clinical outcomes and cost-effectiveness. The field is poised for significant advancements with the adoption of newer techniques such as federated learning^[59,60], which allows collaborative model development without data sharing, and transformer-based models^[61], known for their superior contextual understanding. Additionally, research must expand to explore the ethical implications of AI/ML in MBS, particularly concerning patient consent and data privacy.

In conclusion, in this review, we provide a comprehensive overview of AI/ML applications in predicting complications after MBS. While the findings are encouraging, substantial work remains to translate these advancements into clinical practice. Future research should focus on addressing the identified limitations and leveraging emerging technologies to enhance predictive accuracy, interpretability, and real-world utility. By doing so, AI/ML can become a cornerstone in improving the safety and outcomes of MBS.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study and performed data analysis and interpretation: Pantelis AG, Epiphaniou P

Performed data acquisition, as well as providing administrative, technical, and material support: Pantelis AG, Lapatsanis DP

Availability of data and materials

Source data are available from the corresponding author upon reasonable request.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Pantelis AG, Stravodimos GK, Lapatsanis DP. A scoping review of artificial intelligence and machine learning in bariatric and metabolic surgery: current status and future perspectives. *Obes Surg*. 2021;31:4555-63. DOI PubMed
2. Bellini V, Valente M, Turetti M, et al. Current applications of artificial intelligence in bariatric surgery. *Obes Surg*. 2022;32:2717-33. DOI PubMed PMC
3. Bektaş M, Reiber BMM, Pereira JC, Burchell GL, van der Peet DL. Artificial intelligence in bariatric surgery: current status and future perspectives. *Obes Surg*. 2022;32:2772-83. DOI PubMed PMC
4. Al-Mazrou AM, Bellorin O, Dakin G, Pomp A, Unruh MA, Afaneh C. Implementation of the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program and outcomes of bariatric surgery. *Am J Surg*. 2023;225:362-6. DOI PubMed
5. Aminian A, Brethauer SA, Kirwan JP, Kashyap SR, Burguera B, Schauer PR. How safe is metabolic/diabetes surgery? *Diabetes Obes Metab*. 2015;17:198-201. DOI PubMed
6. Eisenberg D, Shikora SA, Aarts E, et al. 2022 American Society of Metabolic and Bariatric Surgery (ASMBS) and International Federation for the Surgery of Obesity and Metabolic Disorders (IFSO) Indications for Metabolic and Bariatric Surgery. *Obes Surg*. 2023;33:3-14. DOI PubMed PMC
7. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. DOI PubMed PMC
8. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51-8. DOI PubMed
9. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170:W1-33. DOI PubMed
10. Çorbacioğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value. *Turk J Emerg Med*. 2023;23:195-8. DOI PubMed PMC
11. Cao Y, Fang X, Ottosson J, Näslund E, Stenberg E. A comparative study of machine learning algorithms in predicting severe complications after bariatric surgery. *J Clin Med*. 2019;8:668. DOI PubMed PMC
12. Cao Y, Montgomery S, Ottosson J, Näslund E, Stenberg E. Deep learning neural networks to predict serious complications after bariatric surgery: analysis of scandinavian obesity surgery registry data. *JMIR Med Inform*. 2020;8:e15992. DOI PubMed PMC
13. Scott AW, Amateau SK, Leslie DB, Ikramuddin S, Wise ES. Prediction of 30-day morbidity and mortality after conversion of sleeve gastrectomy to Roux-en-Y gastric bypass: use of an artificial neural network. *Am Surg*. 2024;90:1202-10. DOI PubMed
14. Sheikhtaheri A, Orooji A, Pazouki A, Beitollahi M. A clinical decision support system for predicting the early complications of one-anastomosis gastric bypass surgery. *Obes Surg*. 2019;29:2276-86. DOI PubMed
15. Stenberg E, Cao Y, Szabo E, Näslund E, Näslund I, Ottosson J. Risk prediction model for severe postoperative complication in bariatric surgery. *Obes Surg*. 2018;28:1869-75. DOI PubMed PMC
16. Wise ES, Amateau SK, Ikramuddin S, Leslie DB. Prediction of thirty-day morbidity and mortality after laparoscopic sleeve gastrectomy: data from an artificial neural network. *Surg Endosc*. 2020;34:3590-6. DOI PubMed
17. Wise E, Leslie D, Amateau S, et al. Prediction of thirty-day morbidity and mortality after duodenal switch using an artificial neural network. *Surg Endosc*. 2023;37:1440-8. DOI PubMed
18. Zucchini N, Capozzella E, Giuffrè M, et al. Advanced non-linear modeling and explainable artificial intelligence techniques for predicting 30-day complications in bariatric surgery: a single-center study. *Obes Surg*. 2024;34:3627-38. DOI PubMed
19. 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;19:1236-44. DOI PubMed PMC
20. Charles-Nelson A, Lazzati A, Katsahian S. Analysis of trajectories of care after bariatric surgery using data mining method and health administrative information systems. *Obes Surg*. 2020;30:2206-16. DOI PubMed
21. Torquati M, Mendis M, Xu H, et al. Using the super learner algorithm to predict risk of 30-day readmission after bariatric surgery in the United States. *Surgery*. 2022;171:621-7. DOI PubMed

22. Zhang M, Chen R, Yang Y, Sun X, Shan X. Machine learning analysis of lab tests to predict bariatric readmissions. *Sci Rep*. 2024;14:16845. DOI PubMed PMC
23. Hsu JL, Chen KA, Butler LR, et al. Application of machine learning to predict postoperative gastrointestinal bleed in bariatric surgery. *Surg Endosc*. 2023;37:7121-7. DOI PubMed PMC
24. Nudel J, Bishara AM, de Geus SWL, et al. Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database. *Surg Endosc*. 2021;35:182-91. DOI PubMed PMC
25. Ali H, Inayat F, Moond V, et al. Predicting short-term thromboembolic risk following Roux-en-Y gastric bypass using supervised machine learning. *World J Gastrointest Surg*. 2024;16:1097-108. DOI PubMed PMC
26. Dang JT, Switzer N, Delisle M, et al. Predicting venous thromboembolism following laparoscopic bariatric surgery: development of the BariClot tool using the MBSAQIP database. *Surg Endosc*. 2019;33:821-31. DOI PubMed
27. Cruz MR, Martins C, Dias J, Pinto JS. A validation of an intelligent decision-making support system for the nutrition diagnosis of bariatric surgery patients. *JMIR Med Inform*. 2014;2:e8. DOI PubMed PMC
28. Lenér F, Höskuldsdóttir G, Landin-Wilhelmsen K, et al. Anaemia in patients with self-reported use of iron supplements in the Bariatric surgery SUBstitution and nutrition study: a prospective cohort study. *Nutr Metab Cardiovasc Dis*. 2023;33:998-1006. DOI PubMed
29. Pan Y, Du R, Han X, et al. machine learning prediction of iron deficiency anemia in Chinese premenopausal women 12 months after sleeve gastrectomy. *Nutrients*. 2023;15:3385. DOI PubMed PMC
30. Parrott JM, Parrott AJ, Rouhi AD, Parrott JS, Dumon KR. What we are missing: using machine learning models to predict vitamin C deficiency in patients with metabolic and bariatric surgery. *Obes Surg*. 2023;33:1710-9. DOI PubMed
31. Emile SH, Ghareeb W, Elfeki H, El Sorogy M, Fouad A, Elrefai M. Development and validation of an artificial intelligence-based model to predict gastroesophageal reflux disease after sleeve gastrectomy. *Obes Surg*. 2022;32:2537-47. DOI PubMed PMC
32. Liew PL, Lee YC, Lin YC, et al. Comparison of artificial neural networks with logistic regression in prediction of gallbladder disease among obese patients. *Dig Liver Dis*. 2007;39:356-62. DOI PubMed
33. Romero-Velez G, Dang J, Barajas-Gamboa JS, et al. Machine learning prediction of major adverse cardiac events after elective bariatric surgery. *Surg Endosc*. 2024;38:319-26. DOI PubMed
34. McCormack R, Michels R, Ramos N, Hutzler L, Slover JD, Bosco JA. Thirty-day readmission rates as a measure of quality: causes of readmission after orthopedic surgeries and accuracy of administrative data. *J Healthc Manag*. 2013;58:64-76; discussion 76. PubMed
35. Wilson MP, Jack AS, Nataraj A, Chow M. Thirty-day readmission rate as a surrogate marker for quality of care in neurosurgical patients: a single-center Canadian experience. *J Neurosurg*. 2019;130:1692-8. DOI PubMed
36. Kollmann L, Gruber M, Lock JF, Germer CT, Seyfried F. Clinical management of major postoperative bleeding after bariatric surgery. *Obes Surg*. 2024;34:751-9. DOI PubMed PMC
37. Firkins SA, Simons-Linares R. Management of leakage and fistulas after bariatric surgery. *Best Pract Res Clin Gastroenterol*. 2024;70:101926. DOI PubMed
38. El Ansari W, El-Menyar A, El-Ansari K, Al-Ansari A, Lock M. Cumulative incidence of venous thromboembolic events in-hospital, and at 1, 3, 6, and 12 months after metabolic and bariatric surgery: systematic review of 87 studies and meta-analysis of 2,731,797 patients. *Obes Surg*. 2024;34:2154-76. DOI PubMed PMC
39. Gomes R, Costa-Pinho A, Ramalho-Vasconcelos F, et al. Portomesenteric venous thrombosis after bariatric surgery: a case series and systematic review comparing LSG and LRYGB. *J Pers Med*. 2024;14:722. DOI PubMed PMC
40. Lupoli R, Lembo E, Saldalamacchia G, Avola CK, Angrisani L, Capaldo B. Bariatric surgery and long-term nutritional issues. *World J Diabetes*. 2017;8:464-74. DOI PubMed PMC
41. Serra FE, Cohen RV. Gastroesophageal reflux disease after sleeve gastrectomy. *Dig Med Res*. 2024;7. DOI
42. Tsirlina VB, Keilani ZM, El Djouzi S, et al. How frequently and when do patients undergo cholecystectomy after bariatric surgery? *Surg Obes Relat Dis*. 2014;10:313-21. DOI PubMed
43. Tustumi F, Bernardo WM, Santo MA, Cecconello I. Cholecystectomy in patients submitted to bariatric procedure: a systematic review and meta-analysis. *Obes Surg*. 2018;28:3312-20. DOI PubMed
44. Khorgami Z, Jackson TN, Aminian A, Sahawneh JM, Sclabas GM, Chow GS. Early cardiac complications after bariatric surgery: does the type of procedure matter? *Surg Obes Relat Dis*. 2019;15:1132-7. DOI PubMed
45. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323. DOI PubMed PMC
46. Aliferis C, Simon G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In: Simon GJ, Aliferis C, editors. Artificial intelligence and machine learning in health care and medical sciences. Cham: Springer International Publishing; 2024. pp. 477-524. DOI
47. Kiremitci S, Seven G, Silahtaroglu G, Kochan K, Degirmencioglu Tosun S, Senturk H. The role of artificial intelligence and deep learning in determining the histopathological grade of pancreatic neuroendocrine tumors by using EUS images. *Endosc Ultrasound*. 2025;14:48-56. DOI PubMed PMC
48. Razzaghi T, Safo I, Ewing J, Sadrifaridpour E, Scott JD. Predictive models for bariatric surgery risks with imbalanced medical datasets. *Ann Oper Res*. 2019;280:1-18. DOI
49. Xue B, Li D, Lu C, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to

- identify risks of postoperative complications. *JAMA Netw Open*. 2021;4:e212240. DOI PubMed PMC
50. Stam WT, Goedknecht LK, Ingwersen EW, Schoonmade LJ, Bruns ERJ, Daams F. The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: a systematic review. *Surgery*. 2022;171:1014-21. DOI PubMed
51. Henn J, Bunes A, Schmid M, Kalff JC, Matthaei H. Machine learning to guide clinical decision-making in abdominal surgery-a systematic literature review. *Langenbecks Arch Surg*. 2022;407:51-61. DOI PubMed PMC
52. Ravenel M, Joliat GR, Demartines N, Uldry E, Melloul E, Labgaa I. Machine learning to predict postoperative complications after digestive surgery: a scoping review. *Br J Surg*. 2023;110:1646-9. DOI PubMed PMC
53. Wang J, Tozzi F, Ashraf Ganjouei A, et al. Machine learning improves prediction of postoperative outcomes after gastrointestinal surgery: a systematic review and meta-analysis. *J Gastrointest Surg*. 2024;28:956-65. DOI PubMed
54. Nopour R. Comparison of machine learning models to predict complications of bariatric surgery: a systematic review. *Health Informatics J*. 2024;30:14604582241285794. DOI PubMed
55. Hassan Mukhtar MA, Babiker Ahmed AU, Siddig Mohammed MA, Ibrahim Omer NO, Altom DS, Elnour MAA. The role of artificial intelligence in the prediction of bariatric surgery complications: a systematic review. *Cureus*. 2025;17:e82461. DOI PubMed PMC
56. Abu-Abeid A, Dvir N, Lessing Y, et al. Primary versus revisional bariatric and metabolic surgery in patients with a body mass index ≥ 50 kg/m²-90-day outcomes and risk of perioperative mortality. *Obes Surg*. 2024;34:2872-9. DOI PubMed PMC
57. Giannopoulos S, Li WS, Kalantar Motamedi SM, Embry M, Stefanidis D. Outcome comparison between primary and revisional bariatric surgery: a propensity-matched analysis. *Surgery*. 2024;175:592-8. DOI PubMed
58. Saux P, Bauvin P, Raverdy V, et al. Development and validation of an interpretable machine learning-based calculator for predicting 5-year weight trajectories after bariatric surgery: a multinational retrospective cohort SOPHIA study. *Lancet Digit Health*. 2023;5:e692-702. DOI PubMed
59. Lee GH, Shin SY. Federated learning on clinical benchmark data: performance assessment. *J Med Internet Res*. 2020;22:e20891. DOI PubMed PMC
60. Fathima AS, Basha SM, Ahmed ST, et al. Federated learning based futuristic biomedical big-data analysis and standardization. *PLoS One*. 2023;18:e0291631. DOI PubMed PMC
61. Cho HN, Jun TJ, Kim YH, et al. Task-specific transformer-based language models in health care: scoping review. *JMIR Med Inform*. 2024;12:e49724. DOI PubMed PMC