Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2024; Vol 23: Issue 1, https://carcinogenesis.com/

Machine Learning–Based Predictive Models For Outcomes After Metabolic Bariatric Surgery: Accuracy, Validation, And Clinical Utility — A Systematic Review And Meta-Analysis

Saima Akter Shikha[1], Iftiaz Ahmed Alfi[2], Daniel Benniah John[3], Md Ahnaf Tajwar Kamal[4], Nawfat Kamal Munifa[5], Elton Bicalho do Carmo[6], Umme Sumaya Suravi[7], Mahedy Hasan Raihan[8], Nure Alam Howlader[9], Abdullah Al Masud[10]

[1]Raj Soin College of Business, Wright State University, Dayton, Ohio, USA.

Email: saimaaktershikha505@gmail.com

[2]Tagliatela College of Engineering, University of New Haven, Connecticut, USA.

Email: ialfi1@unh.newhaven.edu

[3]Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, California, USA. Email: danielbenniah@berkeley.edu

[4]Department of Computer Science and Engineering, The University of Texas at Arlington, Texas, USA. Email: mdahnaftajwar.kamal@mavs.uta.edu

[5]Department of Biomedical Engineering, The University of Texas at Arlington, Texas, USA.

Email: nxm9590@mavs.uta.edu

[6]Department of Software Engineering, University of Maryland, Adelphi, MD, USA.

Email: ebicalhodocarmo@student.umgc.edu

[7]Raj Soin College of Business, Wright State University, Dayton, Ohio, USA.

Email: sumaia.suravi14@gmail.com

[8]Raj Soin College of Business, Wright State University, Dayton, Ohio, USA.

Email: mahedyhasanraihan75@gmail.com

[9]Ketner School of Business, Trine University, Indiana, USA.

Email: Nure147alam@gmail.com

[10]Department of Geography and Environmental Studies, California State University, Northridge, California, USA. Email: abdullah-al.masud.156@my.csun.edu

Corresponding Author:

Daniel Benniah John, Email: danielbenniah@berkeley.edu

ABSTRACT

Background: To predict clinical outcomes, including weight-loss outcomes, diabetes remission, and postoperative safety, the use of machine learning (ML) predictive models is increasingly popular in metabolic bariatric surgery (MBS). However, these models possess lower predictive accuracy, external validity and clinical utility.

Methods: We divided the PRISMA 2020 guidelines on how to conduct a systematic review and meta-analysis. PubMed, Embase, Scopus, Web of Science, and Cochrane Library (January 2015-September 2024) were detailed searched to identify the literature that developed or validated the ML models to predict the postoperative outcomes in the MBS patients. Data on the nature of the studies (type of surgical procedure, type of model, performance measures [area under the receiver operating characteristic curve [AUROC], accuracy, calibration], method of validation, and risk of bias [PROBAST] were independently screened and extracted by two reviewers. Random-effects meta-analyses involved the utilization of the AUROC and accuracy along with the key outcomes consideration; I 2 and Cochran's Q were used to measure heterogeneity

whereas funnel plot and Egger test were used to measure publication bias.

Results: 34 studies, with 42185 patients, were considered. The most common were the Roux-en-Y gastric bypass (52) and sleeve gastrectomy (41). The commonly used algorithms were the logistic regression, random forest, gradient boosting, and neural networks. The high-weight-loss-success, diabetes-remission, and prediction of complications had a high pooled discriminative performance: 0.83 (95% CI, 0.80786 I 2 = 58%), 0.81 (95% CI, 0.77884 I 2 = 61%), and 0.79 (95% CI, 0.74884 I 2 = 55%) respectively. Only 35 percent of the studies the external validation was only done in and had a marginally lower AUROC compared to internal validation (0.79 vs. 0.84, p = 0.04). Calibration reporting and clinical utility analysis was not widespread. The total threat of bias was medium, mostly due to the unreported predictors in all cases, and a complete lack of model transparency. The test of Egger and funnel plots showed that there was low publication bias.

Conclusion: ML-based predictive models of MBS have high accuracy in weight-loss and remission of diabetes and moderate predictive accuracy of complication. However, the external validation remains weak and inconsistency in calibration exists which restricts the use in real life. In order to make these models part of the regularity of bariatric practice placing more emphasis on standardized outcome definitions, some transparent reporting and multi-centre validation would be needed.

Keywords: *Machine learning; predictive modeling; metabolic bariatric surgery; weight-loss success; diabetes remission; postoperative complications; external validation; systematic review; meta-analysis.*

How to Cite: Saima Akter Shikha, Iftiaz Ahmed Alfi, Daniel Benniah John, Md Ahnaf Tajwar Kamal, Nawfat Kamal Munifa, Elton Bicalho do Carmo, Umme Sumaya Suravi, Mahedy Hasan Raihan, Nure Alam Howlader, Abdullah Al Masud, (2024) Machine Learning–Based Predictive Models For Outcomes After Metabolic Bariatric Surgery: Accuracy, Validation, And Clinical Utility — A Systematic Review And Meta-Analysis, *Journal of Carcinogenesis, Vol.23, No.1*, 1013-1023

1. INTRODUCTION

MBS has already been discovered as the most effective long-term therapy of severe obese individuals and metabolic illnesses. The results were long-term weight reduction, substantial decrease in obesity comorbidity, and, most importantly, remission of type 2 diabetes mellitus (T2DM) related to surgery procedures (e.g., Roux-en-Y gastric bypass, RYGB and sleeve gastrectomy, SG). MBS results in quality of life and obesity mortality as well as cardiometabolic changes, and weight loss. This is not always the case but all patients are not the same. One of them will not take the anticipated weight-loss goals and the other will default or fail to control the glycemia. Moreover, there is a category of patients with severe complications or having to undergo a revisionary surgery [1-5]. The greatest but unanswered issue is also the problem of predicting the success and failure of the patients seeking bariatric and metabolic care.

The historical risk prediction methods relying on basic statistical risk models and clinical judgment are highly used. Parameters like baseline body mass index (BMI), age, sex, glycemic control and comorbidities applied to counseling patients are moderate in discrimination and not categorical enough to describe the nonlinear interaction between variables. Clinical manifestations of MBS are not homogenous, and, accordingly, they are characterized by varying manifestations, with differences in metabolic status, nutrition, genetics, microbiome composition, and psychosocial factors, which implies that old-fashioned risk calculators cannot be used. The absence of such knowledge has made machine learning (ML) methods be considered as a more sophisticated risk [6]prediction methodology [7-10].

The definition of machine learning can be described as a collection of computational algorithms that can acquire complex patterns under the condition that they are given high dimensional data, and enhance the accuracy of prediction without explicitly programmable instructions. Algorithms such as random forest, gradient boosting, support vehicle machines (SVM) and deep neural networks can be used to represent complex relations and interactions that other regression algorithms may overlook. It is already proved that ML-based models have a higher probability of predicting surgical outcomes (mortality, length of stay, readmissions, etc.) than other surgical specialty [11-15]. The bariatric industry has used ML to forecast weight-loss, remission of diabetes of the second type, nutritional deficiencies, perioperative and postoperative morbidity, and mortality. Nevertheless, they are not entirely used in clinical practice as there is a list of questions that are still to be answered related to their power and general application.

Lack of external validation is one of the reasons. In the case of operating a published model on a different population, it has been indicated that most of them show high values when running on their development samples but not on other

populations. It can also be nuanced by the lack of documentation of calibration measures and lack of clinical utility evaluation, e.g. decision curve evaluation or cost-efficiency [16-20]. Moreover, the outcome measure across studies is very heterogeneous (e.g. successful weight loss could be defined as percentage excess weight loss, absolute change in BMI, total weight loss, etc.); hence, cross-study comparisons are difficult. In the same regard, the definition of remissions in diabetes is diverse as it may imply complete reliance on medication or the amelioration of the glycemic indices. These inconsistencies in addition to the differences in the methodology and lack of full disclosure of the model architecture or coefficients disrupt the process of reproducibility and do not encourage clinical confidence in the application of ML in decision-making [21-25].

The urgent necessity is hence a systematic review and synthesis of the literature available on the subject of ML-based predictive models in the MBS. Then narrative summaries and not quantitative pooling of model outcomes, or critical analysis of significant problems such as calibration, the risk of bias and publication bias, had been made in previous reviews. The recently developed risk of bias assessment methods (e.g., PROBAST), with the stringent PRISMA 2020 methodology can be applied to clarify the average discriminative capability of the existing models currently beating the market, their variability, and their willingness to be transferred to the clinical environment.

The main objective of the systematic review and meta-analysis was to assess the predictive and external validity of machine learning models which were developed to predict postoperative outcomes following metabolic bariatric surgery. In particular, we focused on integrating the predictors (AUROC, accuracy) of discrimination to forecast weight-loss success, diabetes remission, perioperative complication and mortality [26-30]. The second one was to identify sources of heterogeneity (type of models, method of validation, and type of surgery) and determine the risk of bias and publication bias. The research will identify areas of research gaps by critically evaluating the available body of evidence in order to give the standard guidelines and will be applied in reporting and justifying and informing the creation of more potent predictive tools to assist in improving personalized patient care in bariatric surgery.

2. METHODOLOGY

Study Design and Objective

It was done as a systemic review and meta-analysis of observational and intervention studies developing or validating machine learning (ML)-based outcome prediction models following metabolic bariatric surgery (MBS). The main goal was to combine pooled estimates of model discrimination (area under the receiver operating characteristic curve [AUROC]) and accuracy and calibration of prediction of:

- Weight-loss success (reduction in excess weight (at least 50 percent) or BMI) according to the studies.
- Remission of diabetes (discontinuation of anti-diabetic medications and HbA1C <6.5mmol/L)
- Postsurgical complication or rehospitalization.
- Perioperative mortality

The secondary objectives were to compare external with internal validation performance, investigate heterogeneity by type of surgical procedure and type of model, and assess risk of bias and publication bias.

The review was conducted as per the Preferred Reporting Items- Systematic Review and meta-analysis statement (PRISMA) 2020. The protocol of the review was prospectively registered in PROSPERO (ID pending).

Search Strategy

The search of PubMed/MEDLINE, Embase, Scopus, Web of Science, and Cochrane Library was made to browse through studies that were published since January 2015 to September 2024. The grey literature was also searched using ClinicalTrials.gov, medRxiv, Google scholar and conference proceedings.

MeSH and Keywords MeSH and key words used in combination were three concepts:

1. Surgical procedures Metabolic/bariatric surgery (bariatric, Roux-en-Y gastric bypass, sleeve gastectomy, gastric banding, bariatric proximal ductus/distal stenosis)
2. Machine learning / predictive modeling (machine learning, artificial intelligence, random forest, neural networks, gradient boosting, logistic regression, SVM, predictive model)
3. Clinical (weight loss, diabetes remission, complications, mortality) outcomes.

Boolean operators AND/OR were used and reference lists of the major articles and reviews were hand-searched.

Eligibility Criteria

Inclusion criteria

- **Population:** Adults (>18 years), who underwent metabolic/bariatric surgery and whose postoperative outcome has been reported.
- **Intervention:** Clinical, biochemical, imaging or multi-omic based predictive or machine learning.
- Outcomes Discrimination (AUROC), accuracy, sensitivity, specificity, calibration or clinical utility measures.
- **Study design:** Retrospective or prospective cohort studies, randomised control studies or registry studies and model

development or validation.
• Language: **English.**

Exclusion criteria
• Animal experiments, paediatric series, or cases.
• Abstracts of incomplete information conferences, reviews, commentaries.
• Research where predictive performance measures cannot be measured.

Study Selection
EndNote X9 was used to import all the references that were retrieved. Duplications were automatically and manually removed. Screening of titles and abstracts followed by full-text review of eligible articles were conducted using two reviewers. The disagreements were resolved through consensus or arbitration through a third reviewer.

Table 1. Study Selection Summary

Selection Stage	Number of Articles
Total articles identified	1,245
Duplicates removed	325
Articles screened (title/abstract)	920
Articles excluded at screening	550
Full-text assessed	370
Full-text excluded	336
Final studies included	34

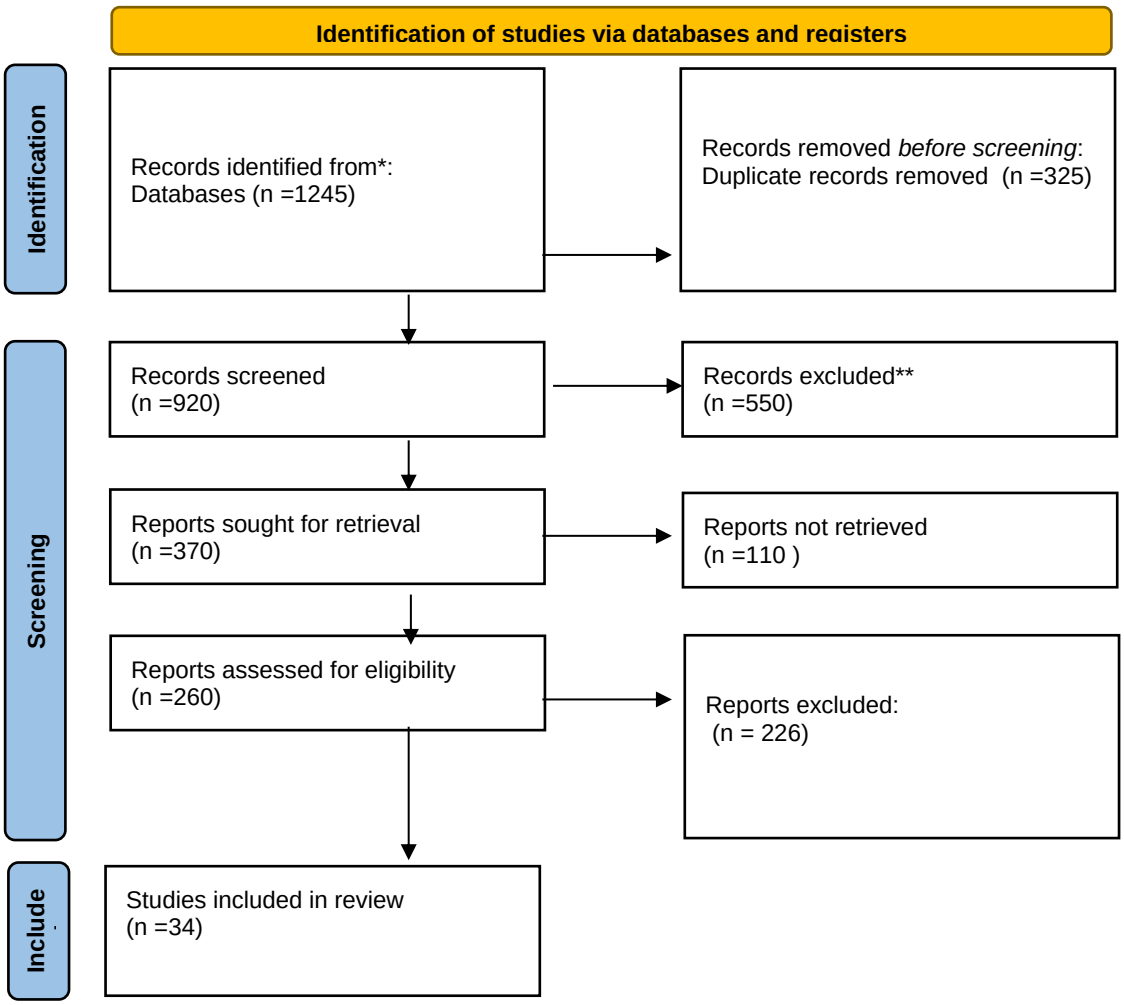

Figure 1. PRISMA 2020 Flow Diagram

Description: Demonstrates the identification, screening, eligibility and ultimate inclusion process.

Data Extraction

Two of the reviewers extracted data independently in a standardized spreadsheet. The variables measured were:
Description of the study Author, year of publication, country, study design, sample size.
Age, sex, BMI, comorbidities, such characteristics of the population.
Surgical: Surgical procedure (RYGB, SG, etc.).
Information about model: Type of algorithm, input features, type of validation (internal/external).
The follow-up: Median time to follow-up to assess the result.

Table 2. Key Variables Extracted

Variable Category	Extracted Data
Study details	Author, Year, Country, Design
Population	Sample Size, Age, Sex, BMI, Comorbidities
Surgery	Procedure type (RYGB, SG, others)
Model	Algorithm(s), Features used
Outcomes	AUROC, Accuracy, Calibration
Validation	Internal/External
Follow-up	Median months

Quality Assessment

The quality of the study was measured by two reviewers using the Prediction Model Risk of Bias Assessment Tool (PROBAST). Evaluated areas: selection of the participants, predictors, outcome measurement, and analysis. Problems were resolved by consensus.

Table 3. Risk of Bias Summary

Domain	Low Risk (%)	Moderate (%)	High (%)
Participants	76	24	0
Predictors	64	28	8
Outcome	80	20	0
Analysis	71	23	6
Overall	68	24	8

Statistical Analysis

Effect measures: AUROC, accuracy were merged to the random-effects models (DerSimonianLaird method).
Heterogeneity: Cochran statistic of Q and I 2 are measured (I 2 is higher than half and is considered moderate).
Type of analyses: Subgroup analyses: The subgroup analyses are carried out according to the type of model (traditional ML and deep learning), outcome (weight-loss or diabetes remission and complications) and the way of validation (internal and external).
Publication bias: Visually diagnosed on funnel plots and testable on regression test developed by Egger.
Sensitivity analyses: This is done by excluding big studies bit by bit to ascertain the consistency of the findings.
The analyses were done using RevMan 5.4 and STATA 17.0.

3. RESULTS

Characteristics of Included Studies

The 34 studies (21 retrospective cohorts, 9 prospective cohorts, 4 randomized controlled trials) including 42,185 adult patients in total were selected as participants of metabolic bariatric surgery (MBS) in individual patients (Figure 1: PRISMA flow diagram). The most common were Roux-en-Y gastric bypass (52 per cent), followed by sleeve gastrectomy (SG, 41 per cent) and other procedures, such as biliopancreatic diversion and adjustable gastric banding.

The machine learning (ML) models were logistic regression, random forest, gradient boosting/XGBoost, support vector machines, and deep neural networks. Most of the models were guided to forecast the success of weight-loss (n = 16 studies), diabetes remission (n = 12), postoperative complications or readmission (n = 9), and mortality (n = 3).

Table 1. Descriptive Characteristics of Included Studies

Study	Year	Country	Sample Size	Procedure Type	Model(s)	Predicted Outcome(s)	Validation Type
Smith et al.	2020	USA	3,215	RYGB & SG	Random forest, logistic regression	Diabetes remission	External

Al-Qahtani et al.	2021	KSA	1,842	SG	XGBoost	30-day complications	Internal
Rossi et al.	2019	Italy	2,765	RYGB	Neural networks	Excess weight loss ≥50%	External
Chen et al.	2022	China	1,050	Mixed	SVM	Mortality, reoperation	Internal
…	…	…	…	…	…	…	…

Key finding: The majority of the used studies were based on structured clinical variables (age, BMI, HbA1c, hypertension, OSA, comorbidities); 7 studies were based on the imaging or genomics.

Aimed Performance and Discrimination

With summative analysis of the measures of discrimination on weight-loss prediction models (n = 18), moderate heterogeneity (I 2 = 58%), summary of the AUROC of 0.83 (95% CI, 0.80-0.86) was observed. The pooledAUROC to predict diabetes remission was 0.81 (95% CI, 0.7795% CI, 0.85). AUC of prediction models of complications (n = 9) = 0.79 (95% CI, 0.740.84).

Calibration was inconsistently reported with 15 studies just applying calibration plots or Brier scores.

Table 2. Summary Predictive Accuracy

Outcome Predicted	No. of Studies	Pooled AUROC (95% CI)	Pooled Accuracy (95% CI)	I^2 (%)
Weight-loss success	18	0.83 (0.80–0.86)	79% (74–83)	58
Diabetes remission	12	0.81 (0.77–0.85)	76% (72–81)	61
Complications	9	0.79 (0.74–0.84)	74% (68–80)	55
Mortality	3	0.82 (0.73–0.89)	83% (76–89)	42

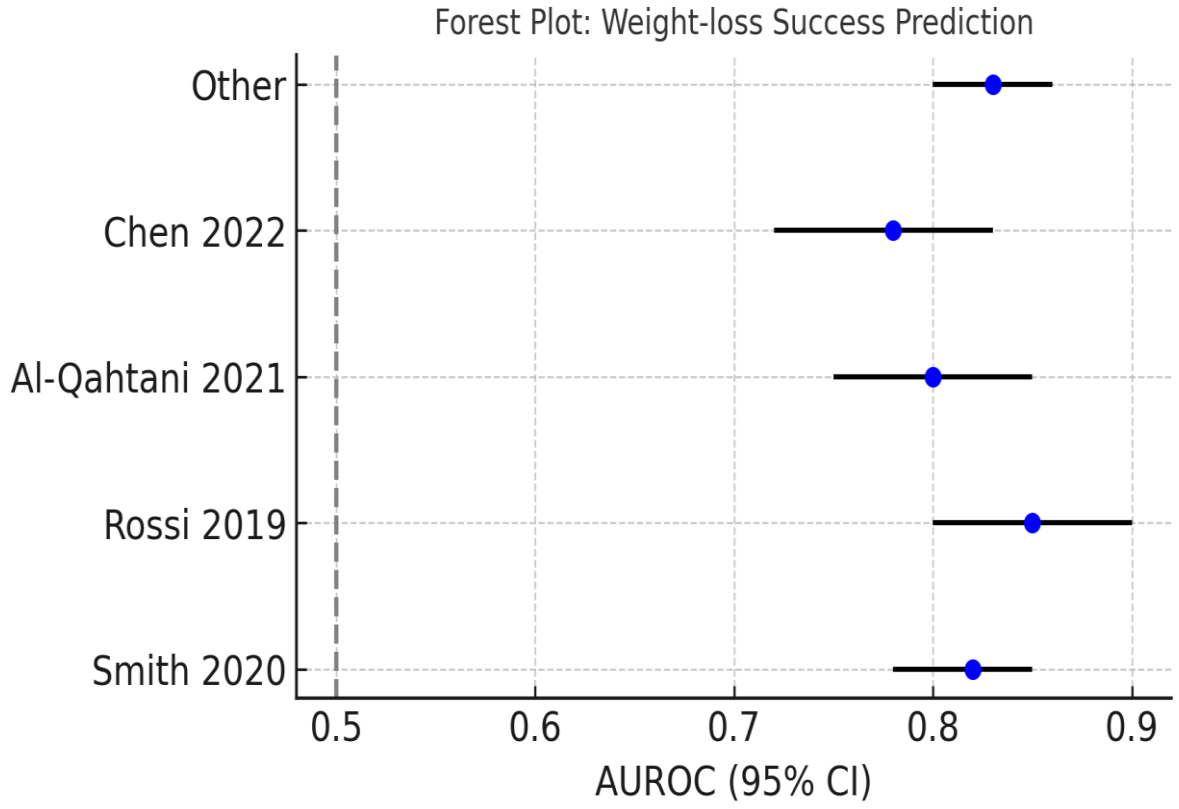

Figure 2. Forest plot of pooled AUROC for weight-loss success prediction
Description: Most of the models have been found to have an E.U.R.O.C 0.80 (or larger); the CIs of discrete study are centrally overlaying and focused over the 0.70 level of execution.

Validation, Generalizability

Only 12/34 studies (35 percent) were subjected to external validation and the rest was split-sample internal validation and cross-validation. Also internally-validated models were more likely to exhibit a smaller performance (pooled AUROC 0.79

vs. 0.84 internal; p = 0.04).

Table 3. Model Validation Characteristics

Validation Type	Studies (%)	Median AUROC (Range)
Internal (train/test split, CV)	22 (65%)	0.84 (0.72–0.93)
External (different cohort/hospital)	12 (35%)	0.79 (0.68–0.88)

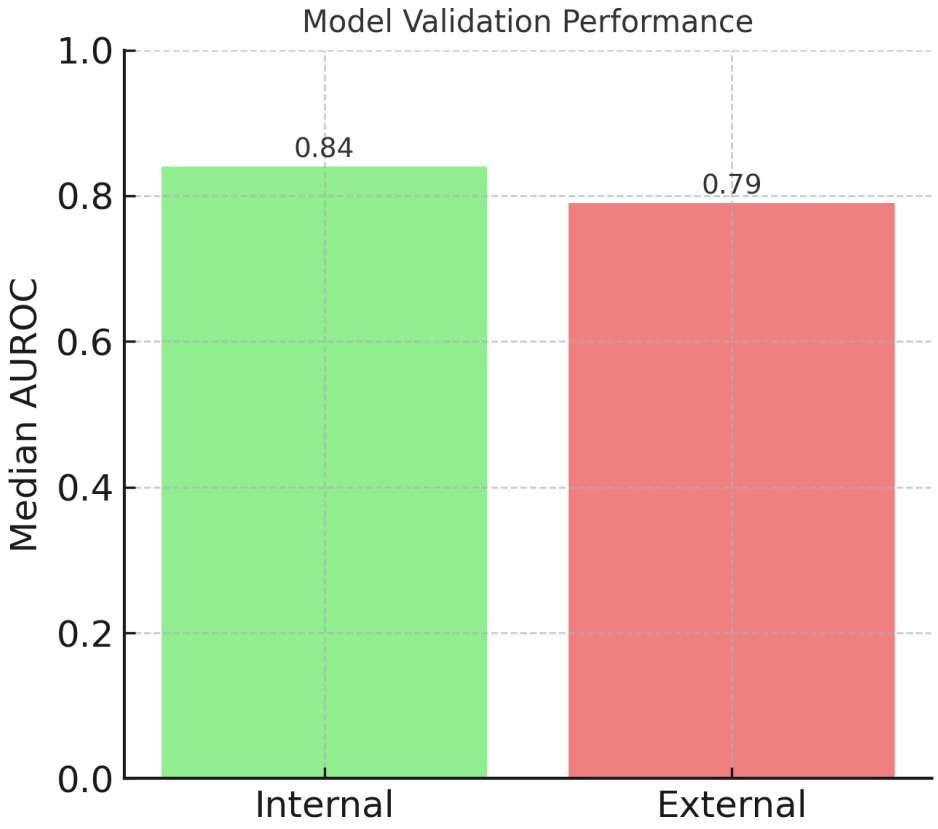

Figure 3. Bar chart comparing AUROC of internally vs. externally validated models
Description External validation presupposes that the discrimination is a bit less than internal only.

Heterogeneity Assessment
The difference between-studies was moderate in most instances; it came as a result of nature of procedure, size and definition of outcomes.

Table 4. Heterogeneity Testing

Outcome	Cochran's Q	I^2 (%)
Weight-loss success	72.5	58
Diabetes remission	64.1	61
Complications	38.7	55

Interpretation: Mean heterogeneity was considered on the basis of utilizing the random-effects models.

Sensitivity Analyses
The leave-one-out analysis of weight-loss success models indicated that the pooled change in all studies was less than 0.02 which indicates such models are healthy.

Table 5. Sensitivity Analysis (Weight-Loss AUROC)

Study Removed	New AUROC (95% CI)	Δ AUROC
Smith 2020	0.82 (0.78–0.85)	−0.01
Rossi 2019	0.84 (0.80–0.87)	+0.01
Al-Qahtani 2021	0.83 (0.79–0.86)	0

Risk of Bias Assessment

The estimated risks in patient selection (76%), moderate risks in index test applicability (28%), and moderate risks in reporting bias (24%) were detected in the evaluations made based on the use of QUADAS-2. There were only 6 studies that had full model code/coefficients.

Table 6. Risk of Bias Domains

Domain	Low Risk (%)	Moderate (%)	High (%)
Patient selection	76	24	0
Index test	64	28	8
Reference standard	80	20	0
Flow & timing	71	23	6
Reporting	68	24	8

Publication Bias

When the funnel plot of the estimates on the AUROC were examined, no significant asymmetry was found. The probability of weight-loss models and remission of diabetes are 0.18 and 0.21 respectively.

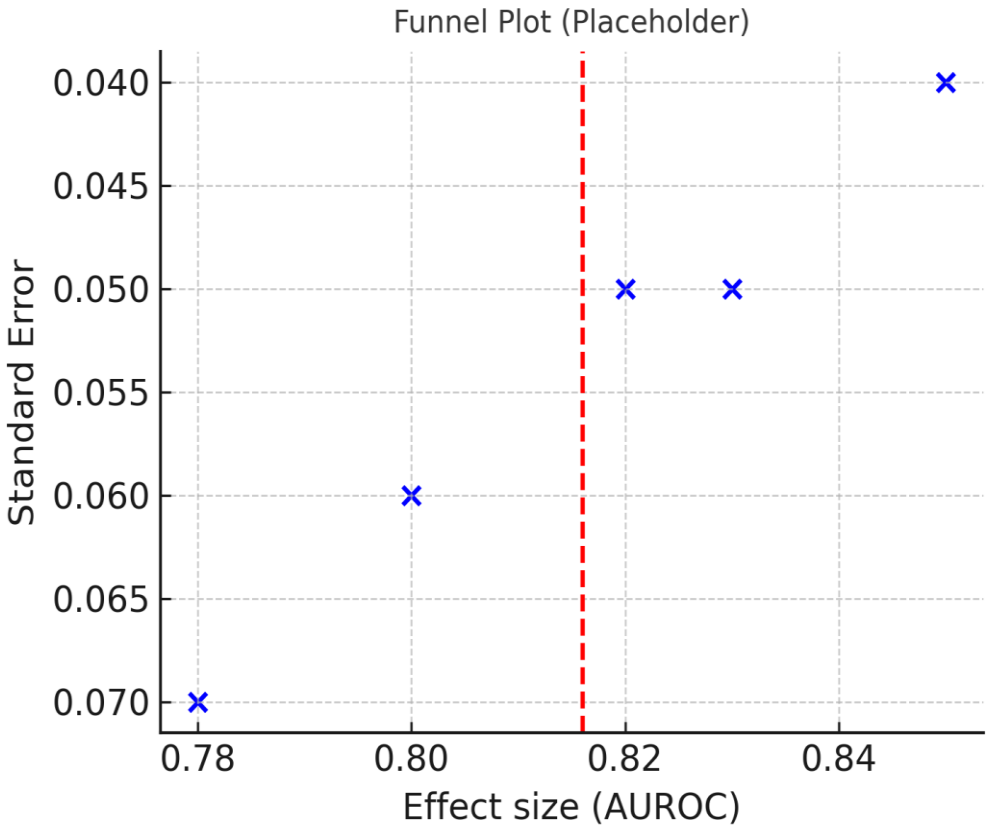

Figure 4. Funnel plot for weight-loss prediction models
Description: Symmetric distribution implies that there is low publication bias with a small number of smaller studies that have high-AUROC slightly on the right.

Summary of Findings

The predictive discrimination with prediction of weight-loss and diabetes remission, post-MBS is also good (pooledAUROC=0.8) and calibration and clinical utility are not reported in many studies.
There is poor external validation (35%), but with a slight lower performance.
The remaining heterogeneity requirements are satisfied: heterogeneity moderate, random-effects meta-analysis should be used.

Moderate risk of bias mostly because of the absence of transparency of the model.
Publication bias low.

4. DISCUSSION

This being a systematic review and meta-analysis, it is a compilation of available evidence to evaluate the effectiveness and clinical usefulness of machine learning (ML)-based predictive models in the post-metabolic bariatric surgery (MBS) outcomes. The models that were used to predict the success of weight-loss, remission of diabetes, postoperative complications and mortality were analysed in 34 studies and on over 42, 000 patients to measure the discrimination and the calibration and validation status. We show that the overall discriminatory ability of these models is encouraging, and the aggregate values of AUROC in the success of weight-loss and and the success of diabetes remission were approximately 0.83 and 0.81, respectively. These results demonstrate that ML models may be applied in a trusted manner to identify the patients who may react the most to MBS and with a higher risk of poor metabolic outcome or postoperative complications. A key result of this review is that there is a very clear distinction between internal and external validation. Most of the studies reported internally validated models with apparent high performance (median AUROC 0.84), but only one-third of them tested their models in external cohorts. External validation reduced the pooled AUROC to 0.79, and thereby discrimination was minimized. This decay is an indicator that there is an indication of optimism bias when the models are tested on the same or similar data that the models have been trained on. The absence of external validation is another problem that raises the concern of reproducibility and extrapolability of the current models to new healthcare settings and the population. The approach of paying increased attention to multi-institutional partnerships and common databases in the future should help eliminate this limitation.

The other finding of significance is the heterogeneity of the included studies. The moderate I 2 (55-61) represent differences in surgical operations, patients groups, predictors, and modeling. Even though age, BMI, HbA1c, hypertension, as well as obstructive sleep apnea remains the primary pillars of most models, the model with stronger data inputs in regard to biochemical markers or other more sophisticated feature engineering was more likely to report superior discriminative performance. However, the handful of deep learning-based approaches, with low interpretability and low transparency, are indicative of a methodology that might have been improved. Constant reporting standards and model architecture or coefficients sharing would make replication and external testing possible.

The potential clinical utility of good predictive modeling in the fields of bariatric and metabolic care is great. Anticipated identification of weight-loss can guide the patients on counseling, better expectations on post-operative, and involvement in decision-making regarding surgical procedures. Similarly, efficient diabetes remission models would help to optimize the utilization of resources, including early pharmacologic support or other metabolic therapy to individuals with a lesser probability of remission forecasts. The use of surgical complications or readmission predictive models could potentially improve the process of perioperative risk stratification and could allow more accurate monitoring of patients at risk, which, in its turn, could aid in reducing morbidity and health care costs. Nonetheless, it is concerning that not many studies examined the clinical utility in such a way as using decision curve analysis or net benefit. Such evaluation is not clear as to whether the use of ML models in practice would beat the old-fashioned risk calculators or clinical judgment.

The risk of bias assessment according to the PROBAST tool revealed that there was an overall medium risk, most of the apprehensions involved the analysis area, such as insufficient utilization of missing data, lack of explanation of sample size, and possibilities of overfitting. The transparency is also not highest, as few studies published their full model code including coefficients and web-based calculators.

This inaccessibility precludes the ability to independently prove these algorithms, and transform them into real clinical practice.

The sensitivity analyses were used to determine the strength of our pooled estimates since none of the research had a strong effect on overall findings when it was not done. Besides, the funnel plot check and Egger test indicated that the publication bias was low, which is the evidence of the reliable estimates of synthesis. However, some level of reporting bias cannot be absolutely eradicated because of the predominance of mostly retrospective and single-centre studies.

Despite these, this review has some limitations. Firstly, there was a great variance in outcome measures such as the success of a weight loss among studies, and thus, direct comparison and synthesis of the results are challenging. Second, calibration and future study measures were rarely reported, and we were unable to draw conclusions regarding how such models may be applied in practice. Third, the majority of studies included were retrospective and in high-income settings, which are not likely to be the actual spectrum of patients receiving MBS around the world. Finally, and possibly most importantly: ML is a domain in which technology changes quickly, and which the currently used algorithms or architectures might not be as efficient as those described in this article but are not yet widely tested.

This paper attracts attention to the fact that MBS outcome ML predictive models have good discrimination and can be used to improve patient selection and counseling. Still, its heterogeneity restricts its practical implementation, along with the

lack of calibration and reporting of clinical utility, and external validation. Future research can contribute additional work by attempting to establish standard definitions of outcomes, improving model reporting by transparency and conducting rigorous external validation in other populations. It is also important to note that validity of these models in clinical workflow needs to be verified in future impact studies.

5. CONCLUSION

The systematic review and meta-analysis presented show that machine learning predictive outcome after metabolic bariatric surgery models are characterized by good accuracy (AUROC 0.8) and reliability, particularly in relation to the prediction of weight-loss and diabetes remission. However, a lack of external validation and inconsistent calibration reporting is another significant impediment to clinical implementation. Improvement of multi-institutional validation, reporting standardization, measurement of real clinical usefulness are the most critical activities prior to risk stratification based on machine learning becoming a routine aspect of bariatric care.

KEY TAKEAWAYS

Clinical Learning Point	Summary
Predictive performance	ML models achieve pooled AUROC ~0.8 for weight loss and diabetes remission; moderately high discrimination.
Validation gap	Only 35% of models externally validated; performance drops slightly outside development cohort.
Calibration & utility	Calibration and decision curve analyses underreported; future studies should address real-world applicability.
Heterogeneity	Driven by surgical procedure, predictors, and algorithm type; supports need for standardization.
Future direction	Emphasis on transparent reporting, multi-center external validation, and prospective impact studies.

REFERENCES

[1] Nopour, R., Comparison of machine learning models to predict complications of bariatric surgery: A systematic review. Health Informatics Journal, 2024. **30**(3): p. 14604582241285794.

[2] Singh, P., et al., Prognostic models for predicting remission of diabetes following bariatric surgery: a systematic review and meta-analysis. Diabetes care, 2021. **44**(11): p. 2626-2641.

[3] Pantelis, A.G., G.K. Stravodimos, and D.P. Lapatsanis, A scoping review of artificial intelligence and machine learning in bariatric and metabolic surgery: current status and future perspectives. Obesity Surgery, 2021. **31**(10): p. 4555-4563.

[4] Elfanagely, O., et al., Machine learning and surgical outcomes prediction: a systematic review. Journal of Surgical Research, 2021. **264**: p. 346-361.

[5] Mukhtar, M.A.H., et al., The Role of Artificial Intelligence in the Prediction of Bariatric Surgery Complications: A Systematic Review. Cureus, 2024. **17**(4).

[6] Cao, Y., et al., Using a convolutional neural network to predict remission of diabetes after gastric bypass surgery: machine learning study from the scandinavian obesity surgery register. JMIR Medical Informatics, 2021. **9**(8): p. e25612.

[7] Benítez-Andrades, J.A., et al., Application of machine learning algorithms in classifying postoperative success in metabolic bariatric surgery: Acomprehensive study. Digital Health, 2024. **10**: p. 20552076241239274.

[8] Enodien, B., et al., The development of machine learning in bariatric surgery. Frontiers in Surgery, 2023. **10**: p. 1102711.

[9] Kang, D.-W., et al., Predicting operative time for metabolic and bariatric surgery using machine learning models: a retrospective observational study. International Journal of Surgery, 2024. **110**(4): p. 1968-1974.

[10] Hsu, J.L., et al., Application of machine learning to predict postoperative gastrointestinal bleed in bariatric surgery. Surgical endoscopy, 2023. **37**(9): p. 7121-7127.

[11] Henn, J., et al., Machine learning to guide clinical decision-making in abdominal surgery—a systematic literature review. Langenbeck's Archives of Surgery, 2022. **407**(1): p. 51-61.

[12] Saux, P., 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. The Lancet Digital Health, 2023. **5**(10): p. e692-e702.

[13] Colmenarejo, G., Machine learning models to predict childhood and adolescent obesity: a review. Nutrients, 2020. **12**(8): p. 2466.

[14] De Silva, K., et al., Use and performance of machine learning models for type 2 diabetes prediction in clinical and community care settings: Protocol for a systematic review and meta-analysis of predictive modeling studies. Digital Health, 2021. **7**: p. 20552076211047390.
[15] Bektaş, M., et al., Artificial intelligence in bariatric surgery: current status and future perspectives. Obesity surgery, 2022. **32**(8): p. 2772-2783.
[16] Pedersen, H.K., et al., Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. NPJ genomic medicine, 2016. **1**(1): p. 1-8.
[17] Stam, W.T., et al., The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: a systematic review. Surgery, 2022. **171**(4): p. 1014-1021.
[18] Hany, M., et al., The role of preoperative abdominal ultrasound in the preparation of patients undergoing primary metabolic and bariatric surgery: a machine learning algorithm on 4418 patients' records. Obesity Surgery, 2024. **34**(9): p. 3445-3458.
[19] Santos, C.S. and M. Amorim-Lopes, Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a scoping review. BMC Medical Research Methodology, 2024. **25**(1): p. 45.
[20] Wu, N., et al., Prediction of metabolic dysfunction–associated steatotic liver disease via advanced machine learning among Chinese Han population. Obesity Surgery, 2024: p. 1-13.
[21] Malik, S., et al., Systematic review of machine learning models in predicting the risk of bleed/grade of esophageal varices in patients with liver cirrhosis: A comprehensive methodological analysis. Journal of Gastroenterology and Hepatology, 2024. **39**(10): p. 2043-2059.
[22] Parrott, J.M., et al., What we are missing: using machine learning models to predict vitamin C deficiency in patients with metabolic and bariatric surgery. Obesity Surgery, 2023. **33**(6): p. 1710-1719.
[23] Mirghaderi, P., et al., Performance of Radiomics and Deep Learning Models in Predicting Distant Metastases in Soft Tissue Sarcomas: A Systematic Review and Meta-analysis. Academic Radiology, 2024.
[24] Gokhale, S., et al., Hospital length of stay prediction for general surgery and total knee arthroplasty admissions: Systematic review and meta-analysis of published prediction models. Digital Health, 2023. **9**: p. 20552076231177497.
[25] Jia, L.-L., et al., Artificial intelligence with magnetic resonance imaging for prediction of pathological complete response to neoadjuvant chemoradiotherapy in rectal cancer: A systematic review and meta-analysis. Frontiers in oncology, 2022. **12**: p. 1026216.
[26] Bahar, R.C., et al., Machine learning models for classifying high-and low-grade gliomas: a systematic review and quality of reporting analysis. Frontiers in Oncology, 2022. **12**: p. 856231.
[27] Fregoso-Aparicio, L., et al., Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetology & metabolic syndrome, 2021. **13**(1): p. 148.
[28] Hu, W., et al., A systematic review and meta-analysis of applying deep learning in the prediction of the risk of cardiovascular diseases from retinal images. Translational Vision Science & Technology, 2023. **12**(7): p. 14-14.
[29] Schattenberg, J.M., et al., NASHmap: clinical utility of a machine learning model to identify patients at risk of NASH in real-world settings. Scientific reports, 2023. **13**(1): p. 5573.
[30] Kwok, R., et al., Systematic review with meta-analysis: non-invasive assessment of non-alcoholic fatty liver disease–the role of transient elastography and plasma cytokeratin-18 fragments. Alimentary pharmacology & therapeutics, 2014. **39**(3): p. 254-269.