Cureus
Part of SPRINGER NATURE

Open Access Review Article

Machine Learning-Based Prediction of Hypocalcaemia Following Total Thyroidectomy: A Systematic Review

Review began 11/05/2026
Review ended 11/05/2026
Published 11/08/2026

© **Copyright** 2026
Alhazmi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119319

Mohammed A. Alhazmi [1], Raum Y. Musawi [2], Rahaf A. Omer [2], Reham T. Alruwaili [3], Rayim A. Oraybi [2], Yazan Y. Nahari [2], Amani M. Ibrahim [4], Saud K. Alahmari [5], Aryam A. Althobaiti [4], Tahani M. Ahmed [6], May H. Alqaed [3], Amal A. Badri [2]

1. Endocrine Unit and Diabetic Center, King Fahad Central Hospital, Jazan, SAU 2. Faculty of Medicine, Jazan University, Jazan, SAU 3. College of Medicine, Jouf University, Sakakah, SAU 4. College of Medicine, Taif University, Taif, SAU 5. College of Medicine, King Khalid University, Abha, SAU 6. Department of Otolaryngology-Head and Neck Surgery, Al Salam Health Medical Hospital, Riyadh, SAU

Corresponding author: Raum Y. Musawi, raoom596@gmail.com

Abstract

Machine learning (ML) has increasingly been applied to predict hypocalcaemia following total thyroidectomy; however, its predictive performance and potential advantages over conventional statistical approaches remain unclear. This review evaluated studies developing or validating ML models for predicting postoperative hypocalcaemia or hypoparathyroidism after total or completion thyroidectomy. Six studies involving 20,845 patients from multiple countries were included. All studies focused on the prediction of early or transient hypocalcaemia, while none addressed permanent hypoparathyroidism. The performance of the best-performing ML models varied considerably, with area under the receiver operating characteristic curve (AUC) values ranging from 0.65 to 0.98. In studies directly comparing ML models with logistic regression using the same datasets, differences in discrimination were minimal, indicating no consistent advantage of ML over conventional methods. Models incorporating perioperative parathyroid hormone (PTH) measurements generally achieved the highest predictive performance, whereas those relying solely on clinical variables demonstrated more modest discrimination. Despite promising results in some studies, all models exhibited important methodological limitations, lacked external validation, and provided insufficient assessment of calibration. Overall, current evidence suggests that ML models can predict postoperative hypocalcaemia with variable accuracy but have not consistently outperformed conventional logistic regression. Predictive performance appears to depend largely on the inclusion of perioperative PTH measurements. Given the limitations of the available evidence, these models cannot yet be recommended for routine clinical use. Future research should focus on rigorous external validation, comprehensive calibration assessment, and evaluation in independent patient populations before clinical implementation.

Categories: Endocrinology/Diabetes/Metabolism
Keywords: hypocalcaemia, logistic regression, machine learning, parathyroid hormone, postoperative complications, prediction models, systematic review, total thyroidectomy

Introduction And Background

Hypocalcaemia is the most common complication of total thyroidectomy, usually resulting from transient hypoparathyroidism caused by parathyroid gland injury, devascularisation, or inadvertent removal [1,2]. Although most cases are temporary, postoperative hypocalcaemia can cause paraesthesia, muscle cramps, tetany, and cardiac arrhythmias. Severe or persistent cases may require lifelong calcium and vitamin D supplementation and are associated with increased morbidity [3,4]. Postoperative hypocalcaemia can also prolong hospitalisation and increase emergency visits and readmissions, highlighting the importance of early risk identification and appropriate perioperative management [3,5].

Postoperative parathyroid hormone (PTH) measurement is widely used to predict hypocalcaemia and guide calcium supplementation and discharge decisions. However, its predictive thresholds and optimal measurement timing vary across institutions [2,4]. Other factors, including age, sex, Graves' disease, thyroid malignancy, central neck dissection, inadvertent parathyroidectomy, parathyroid autotransplantation, and preoperative vitamin D deficiency, also contribute to postoperative risk. Because no single predictor adequately captures individual risk, multivariable prediction models have been developed to integrate patient characteristics, clinical variables, surgical factors, and biochemical measurements [1,3,6].

Machine learning (ML) methods, including random forests, gradient boosting, and neural networks, offer the potential to identify complex nonlinear relationships and interactions among predictors [2,4]. However, their clinical usefulness depends on adequate sample sizes, appropriate model development, reliable calibration, and external validation [1,5]. Despite their increasing application in clinical prediction, ML models have not consistently demonstrated superior performance to conventional logistic regression. Furthermore, many published models are limited by small development datasets, inadequate handling of

missing data, overfitting, and insufficient validation [3,6].

Several ML models have recently been developed to predict hypocalcaemia following thyroidectomy using different algorithms, predictors, outcome definitions, and validation strategies [3-6]. However, their predictive performance and methodological quality remain unclear, and it is uncertain whether they offer meaningful advantages over conventional statistical approaches. The Prediction Model Risk of Bias Assessment Tool with Artificial Intelligence (PROBAST+AI) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis with Artificial Intelligence (TRIPOD+AI) provide contemporary frameworks for evaluating the methodological quality and reporting of prediction models developed using statistical or ML approaches [7,8].

This systematic review aimed to identify and describe ML models developed or validated to predict hypocalcaemia following total thyroidectomy, critically evaluate their predictive performance, compare their performance with conventional logistic regression where direct comparisons were available, and assess their risk of bias and reporting quality using PROBAST+AI and TRIPOD+AI.

Review

Methods

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [9].

Eligibility Criteria

Eligibility criteria were based on the Population, Intervention, Comparator, Outcome, Timing, and Setting framework [10]. We included prospective and retrospective cohort studies involving adults undergoing total or completion thyroidectomy, with or without neck dissection, that developed or validated ML models to predict postoperative hypocalcaemia or hypoparathyroidism. Eligible algorithms included random forests, gradient boosting, artificial neural networks, support vector machines, decision trees, and ensemble methods. Studies were required to report discrimination or classification performance. Logistic regression models evaluated in the same cohorts were extracted for comparison. We excluded studies evaluating conventional regression alone, individual predictors without a prediction model, intraoperative parathyroid identification without patient-level outcome prediction, animal studies, and non-original research publications.

Information Sources and Search Strategy

PubMed/MEDLINE, Scopus, Web of Science Core Collection, and the Cochrane Central Register of Controlled Trials (CENTRAL) were searched from inception to 20 September 2026. The search combined terms relating to thyroid surgery, hypocalcaemia or hypoparathyroidism, and ML or artificial intelligence (AI). Controlled vocabulary and free-text terms were adapted for each database. No date or language restrictions were applied during the search.

Study Selection

Duplicates were removed using Mendeley Reference Manager (Elsevier, Amsterdam, Netherlands). Two reviewers independently screened titles and abstracts, followed by full-text assessment. Disagreements were resolved by a third investigator. Reasons for exclusion were documented, and the selection process was presented in a PRISMA 2020 flow diagram.

Data Extraction

Two reviewers independently extracted data using a standardised form based on the CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) checklist, with disagreements resolved by a third investigator. Extracted information included study characteristics, participant demographics, surgical procedures, outcome definitions, predictors, ML algorithms, model development methods, missing-data handling, validation strategies, and reported performance measures, including discrimination, calibration, and classification performance.

Risk of Bias and Applicability

Two reviewers independently assessed risk of bias and applicability using PROBAST+AI. The four assessment domains were participants and data sources, predictors, outcome, and analysis. Each domain was rated as low, high, or unclear risk of bias. Overall risk was considered high if any domain was rated as high. Applicability was assessed separately for the participants, predictors, and outcome domains. Disagreements were resolved by a third investigator [8].

Reporting Quality

Reporting quality was assessed independently by two reviewers using 12 pre-specified items from the TRIPOD+AI statement. Items covered study identification, predictor and outcome definitions, sample size, missing-data handling, model development, validation, performance reporting, calibration, model availability, participant flow, and fairness. Each item was scored as fully reported (1), partially reported (0.5), or not reported (0). Overall adherence was expressed as a percentage of the maximum possible score. This focused assessment did not represent a complete evaluation of the TRIPOD+AI statement [9].

Data Synthesis

Findings were synthesised descriptively using tables and narrative summaries. Model characteristics, predictors, validation strategies, and reported performance were compared across studies. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC) or C-statistic. When multiple algorithms were evaluated, their reported performance was extracted, with emphasis on the best-performing model using the most rigorous validation approach available. Comparisons between ML and logistic regression were conducted descriptively within studies using the same participants and validation datasets. Model performance was also examined according to predictor timing, particularly the inclusion of intraoperative or postoperative PTH measurements. Calibration, external validation, and other reported performance measures were considered when evaluating clinical applicability.

Results

Study Selection

The database searches identified 644 records. After removing 129 duplicates, 515 records underwent title and abstract screening, of which 495 were excluded. Twenty full-text articles were assessed, and 14 were excluded because of ineligible study design (n=6), population (n=3), or outcome (n=5). Six studies were included in the systematic review (Figure *1*).

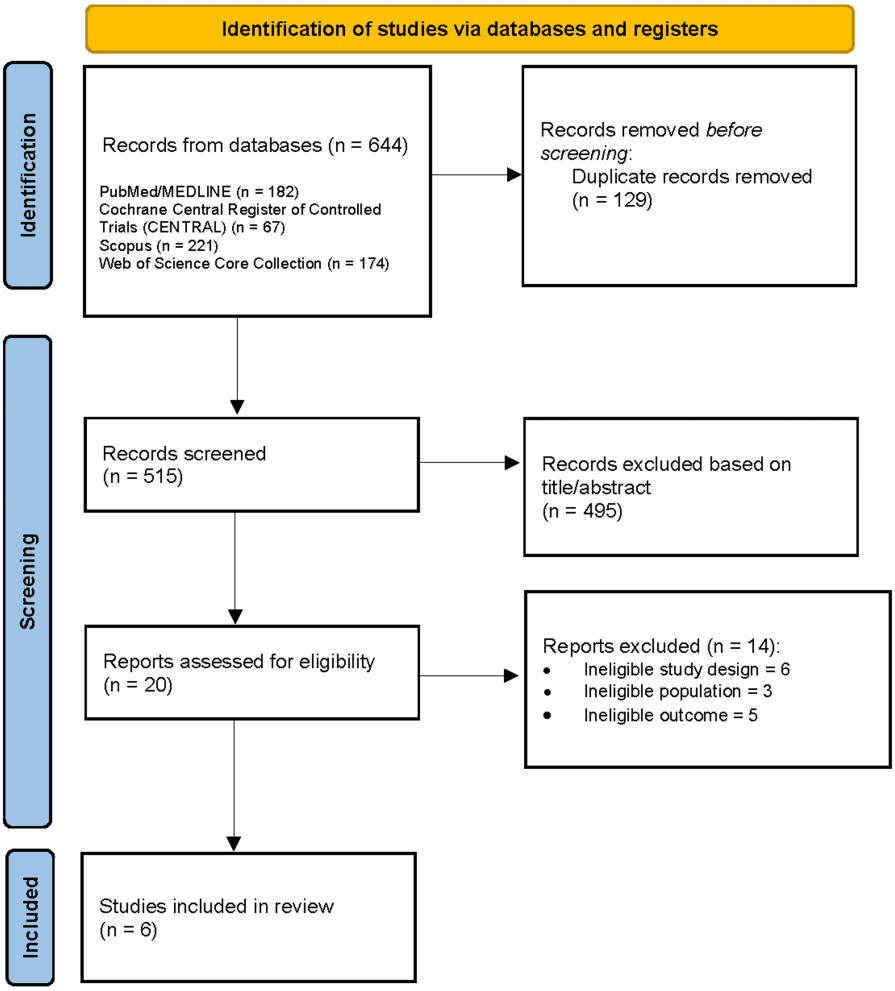

FIGURE 1: PRISMA flow diagram depicting the study selection process for the systematic review

PRISMA flow diagram detailing the study selection process for this systematic review. Following the identification of records through database searching and other sources, duplicates were removed, and the remaining records were screened. Full-text articles were assessed for eligibility, with exclusions documented along with reasons. Studies meeting the inclusion criteria were included in the final synthesis.

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Characteristics of the Included Studies

The six studies originated from France, Brazil, China, the United States, India, and Turkey and included 20,845 patients. Four were single-centre retrospective cohorts, one used a multicentre registry, and one combined retrospective model development with prospective validation. Individual study populations ranged from 196 to 11,259 patients. Four studies exclusively included patients undergoing total thyroidectomy, whereas the remaining two also included other thyroid procedures (Table *1*).

Study ID	Country/setting	Design, period	Population and surgery	n (events, %)	Outcome definition and timing
Muller et al. [1]	France; CHU Lille; single centre; 5 surgeons	Retrospective; derivation cohort Nov 2019-Mar 2022; temporal validation Apr-Sep 2022	Total thyroidectomy (100%); central ND 33.9%; Graves disease 18%	Derivation: 610 (100, 16.4%); validation: 118 (26, 22%)	Serum calcium <80 mg/L at any time after surgery (early/transient hypocalcaemia)
de Carvalho et al. [2]	Brazil; A.C. Camargo Cancer Center; single centre	Retrospective (2014-Jan 2024) plus prospective (Jan-Oct 2024)	Total thyroidectomy (100%); ND 14.2%; malignant disease 74.3%	Development: 7,066 (15.6%); test: 786 (62, 7.9%) [c]; prospective: 149 (31)	Transient clinical hypocalcaemia (symptomatic); calcium <8.0 mg/dL and/or PTH <12 pg/mL used as supportive criteria
Zhong et al. [3]	China; Sir Run Run Shaw Hospital; single centre	Retrospective; Feb 2023-Jan 2026	Differentiated thyroid cancer; bilateral surgery 68.6% [c], unilateral surgery 31.4%; lateral ND 29.8%	876 (335, 38.2%); training set: 701; test set: 175	Albumin-corrected calcium <2.0 mmol/L or typical symptoms within 24 h postoperatively
Seib et al. [4]	USA; ACS NSQIP (91-112 hospitals/year)	Registry-based study; 2016-2018	Non-lobectomy thyroidectomy (total/subtotal thyroidectomy ± ND)	11,259 (723, 6.4%)	NSQIP-recorded hypocalcaemia before discharge (30-day incidence also reported: 985/10,995)
Rao et al. [5]	India; AIIMS Raipur; single centre	Retrospective; 2019/2020-2022 (inconsistently reported)	Total thyroidectomy (100%); benign disease 47%; malignant disease 38%; CCC 55%	196 (25, 12.8%)	Hypocalcaemia graded as mild, moderate, or severe; diagnostic threshold not reported
Yavuz et al. [6]	Turkey; Çukurova University; single centre	Retrospective; 2015-2021	Total thyroidectomy (100%); malignant disease 41%; permanent hypocalcaemia excluded	571 (113, 19.8%); test cohort approximately 113 (approximately 23) [c]	Total calcium <8.5 mg/dL at the 2-week outpatient follow-up visit

TABLE 1: Characteristics of the included studies

This table summarises the key characteristics of the studies included in the systematic review, including study setting, design, patient population, surgical procedures, sample size, event rates, and definitions of postoperative hypocalcaemia. Outcome definitions and timing are reported as described by the original authors.

Ca: calcium; CCC: central compartment clearance; DTC: differentiated thyroid cancer; ND: neck dissection; PTH: parathyroid hormone; TT: total thyroidectomy

Percentages refer to the proportion of participants within each study cohort unless otherwise specified. Event rates represent the number and proportion of patients meeting the study-defined hypocalcaemia outcome. "[c]" indicates information derived or estimated from data reported in the original publication.

Reported hypocalcaemia incidence ranged from 6.4% to 38.2%, reflecting differences in outcome definitions and assessment timing. Three studies used biochemical definitions, one used clinical symptoms, one relied on registry coding, and one did not clearly define the outcome. All studies investigated early or transient hypocalcaemia; none evaluated permanent hypoparathyroidism [1-6].

Characteristics of the ML Models

All six studies developed at least one ML model, but none externally validated a previously developed model. Random forests, gradient boosting, and neural networks were the most frequently evaluated algorithms. Five studies compared multiple algorithms, including conventional logistic regression. The best-performing ML algorithms included extreme gradient boosting (XGBoost), random forests, ensemble learning, and multilayer perceptrons. However, two studies ultimately selected logistic regression as their preferred model [1-6] (Table *2*).

Study ID	Algorithms compared	Headline ML model	Final predictors (timing)	Missing data	EPV	Validation	Availability/explainability
Muller et al. [1]	LR, L1-regularised LR, decision tree, RF, gradient boosting, neural network	RF	%PTH², PTH at induction, %PTH × sex, BMI, age² (intraoperative/skin-closure PTH)	Not reported (BMI 6-18% missing)	≤6.7 [c]	Repeated CV plus temporal validation	Online Shiny application
de Carvalho et al. [2]	XGBoost, RF, KNN, LR	XGBoost (three-stage models)	Top SHAP predictors: postoperative PTH, operative time, surgeon workload, bilateral ND, BMI, sex, age (preoperative, intraoperative, and postoperative variables)	XGBoost native handling; PTH 8.1% missing	NR	Temporal holdout, sliding-window validation, and prospective validation	SHAP analysis; clinical decision-support tool in development
Zhong et al. [3]	LR, SVM, RF, XGBoost, LightGBM, CatBoost, ANN	XGBoost (authors selected LR nomogram)	Fifteen variables including preoperative calcium, magnesium, PTH, BMI, lateral ND, and parathyroid autotransplantation (preoperative and intraoperative variables only)	Median/mode imputation within CV	~12-16	Fivefold CV plus 20% holdout set	SHAP analysis; LR nomogram
Seib et al. [4]	Super Learner ensemble versus full and curated GLMs	Super Learner ensemble	79 preoperative NSQIP variables (preoperative variables only)	Imputation plus missingness indicators	~6 [c]	Nested CV	None
Rao et al. [5]	ANN only	MLP (one hidden layer)	Nine inputs including 24-h postoperative PTH, sternotomy, and RLN sacrifice (includes postoperative PTH)	Not reported	2.8 [c]	Ten random training/test splits	None
Yavuz et al. [6]	LR, RF, gradient boosting, XGBoost, LDA, MLP	XGBoost (LR performed best)	LR importance-selected predictor set; preoperative, intraoperative, and postoperative day-1 laboratory variables	Mean/mode imputation; variables with >85% missingness excluded	Low (113 events)	80/20 holdout set plus fivefold CV for tuning	None

TABLE 2: Characteristics of the ML models

This table summarises the ML and statistical models assessed in each study, including candidate algorithms, the selected model, predictor variables, handling of missing data, EPV, validation strategy, and model availability or explainability features.

ANN: artificial neural network; BMI: body mass index; CatBoost: categorical boosting algorithm; CV: cross-validation; EPV: events per candidate predictor variable; GLM: generalised linear model; KNN: k-nearest neighbours; LDA: linear discriminant analysis; LightGBM: Light Gradient Boosting Machine; LR: logistic regression; ML: machine learning; MLP: multilayer perceptron; ND: neck dissection; NSQIP: National Surgical Quality Improvement Program; PTH: parathyroid hormone; RF: random forest; RLN: recurrent laryngeal nerve; SHAP: Shapley additive explanations; SVM: support vector machine; XGBoost: extreme gradient boosting

Predictor timing is reported according to when data were available for model development (preoperative, intraoperative, or postoperative). EPV reflects the number of outcome events available per candidate predictor variable and is included as an indicator of model stability. "[c]" indicates information derived or estimated from published data. "NR" indicates not reported.

Three studies incorporated intraoperative or postoperative PTH measurements, which were among the most influential predictors. Other models relied primarily on preoperative and intraoperative clinical variables. Two studies provided accessible prediction tools, whereas the remaining models were not made available for practical clinical use.

Risk of Bias and Reporting Quality

All six studies were assessed as having a high overall risk of bias using PROBAST+AI. The main concerns involved inadequate sample sizes, limited validation, insufficient handling of missing data, and incomplete assessment of model performance. Applicability concerns were also identified in most studies.

Adherence to the 12 selected items of TRIPOD+AI ranged from 29% to 79%, with a median of 63%. None of the studies provided numerical calibration measures or adequate sample size justification. Three studies did

not report confidence intervals for discrimination, and only two made their prediction models accessible.

Predictive Performance

The AUC of the best-performing ML model in each study ranged from 0.564 to 0.982. Performance varied according to the predictors included and the validation approach. Models incorporating intraoperative or postoperative PTH generally achieved higher AUCs (0.882-0.982), whereas models relying exclusively on preoperative or intraoperative clinical information achieved lower values (0.645-0.777). In the prospective validation cohort of one study, the preoperative model achieved an AUC of 0.53, compared with 0.94 for the postoperative model [1-6] (Table *3*).

Study ID	Model/set	AUC (95% CI)	Calibration	Classification performance at the stated threshold	Decision curve analysis
Muller et al. [1]	RF, temporal validation cohort; RF, CV	0.928 (0.86-0.97); 0.902 (0.829-0.970)	NR	Risk threshold 7%: sensitivity 92%, specificity 73%; risk threshold 20%: sensitivity 79%, specificity 90% (derivation cohort)	No
de Carvalho et al. [2]	XGBoost postoperative model, test set; intraoperative model, test set; preoperative model, test set; postoperative model, prospective cohort	0.882 (0.828-0.936); 0.700 (0.618-0.783); 0.645 (0.563-0.727); 0.940 (NR)	Calibration plots only; Brier scores: 0.073, 0.140, and 0.152	Threshold ≤0.025: sensitivity 0.758, specificity 0.783; prospective cohort: sensitivity 0.70, specificity 0.97	Yes
Zhong et al. [3]	XGBoost, holdout set; LR, holdout set/CV	0.777 (NR); 0.760 (NR)/0.796 (NR)	NR	LR threshold 0.372: sensitivity 0.806, specificity 0.532 (test set)	Yes
Seib et al. [4]	Super Learner ensemble, nested CV; 30-day outcome	0.720 (0.702-0.739); 0.656 (0.638-0.673)	NR	Risk ≥15%: sensitivity 0.20, specificity 0.95, PPV 0.22, NPV 0.95	No
Rao et al. [5]	ANN; evaluation set unclear	0.982 (NR)	NR	96.3% correctly classified (test set)	No
Yavuz et al. [6]	XGBoost, holdout set; LR, holdout set	0.564 (NR); 0.610 (NR)	NR	LR: sensitivity 0.565, precision 0.295	No

TABLE 3: Predictive performance of the included models

This table summarises the discrimination, calibration, classification performance, and decision-curve analysis results of the prediction models evaluated in the included studies. Where multiple models or datasets were assessed, the principal reported results are shown.

ANN: artificial neural network; AUC: area under the receiver operating characteristic curve; CI: confidence interval; CV: cross-validation; LR: logistic regression; NPV: negative predictive value; PPV: positive predictive value; RF: random forest; XGBoost: extreme gradient boosting

AUC values quantify model discrimination, with higher values indicating better distinction between patients with and without postoperative hypocalcaemia. Sensitivity and specificity are reported at the thresholds specified by the original authors. Decision-curve analysis indicates whether clinical utility was formally evaluated. "NR" indicates not reported. Where no 95% CI was reported, CIs used in the meta-analysis were derived using the Hanley-McNeil method based on sample size and the number of outcome events.

Only two studies evaluated models using data collected after the development period, and neither demonstrated validation in an independent hospital. Overall, the findings indicate considerable variation in predictive performance and limited evidence of generalisability.

Comparison With Logistic Regression

Five studies compared ML with logistic regression using the same participants and validation datasets. Differences in AUC between the approaches ranged from -0.046 to +0.017. ML achieved marginally higher discrimination in three studies, performed equally in one study, and performed worse in one study. Logistic regression also outperformed individual ML algorithms in several comparisons [1-4,6] (Table *4*).

Study ID	Dataset	Best ML algorithm (AUC)	LR (AUC)	ML − LR	Authors' conclusion
Muller et al. [1]	CV	RF, 0.902	0.898	0.004	RF selected as the final model
de Carvalho et al. [2]	Holdout test sets (postoperative/intraoperative/preoperative models)	XGBoost, 0.88/0.70/0.65	0.88/0.72/0.68	0.00/−0.02/−0.03	XGBoost selected because of more balanced sensitivity and kappa performance
Zhong et al. [3]	Holdout test set (CV)	XGBoost, 0.777 (best ML in CV: 0.789)	0.760 (CV: 0.796, best overall)	+0.017 (CV: −0.007)	LR selected because of greater transparency and interpretability
Seib et al. [4]	Nested CV	Super Learner ensemble, 0.720	0.711 (full GLM); 0.704 (curated GLM)	+0.009/+0.016 (p=0.027/<0.001)	Improvement considered "marginal" and not clinically meaningful
Rao et al. [5]	Not reported	ANN, 0.982	Not evaluated	–	–
Yavuz et al. [6]	Holdout test set	XGBoost, 0.564	0.61	−0.046 (DeLong test, not significant)	LR performed best; ML models offered no improvement

TABLE 4: ML versus LR in the same patients and validation set

This table directly compares the best-performing ML model with LR when evaluated in the same patient population and validation dataset. Differences in discrimination are expressed as the absolute difference in AUC between ML and LR models.

ANN: artificial neural network; AUC: area under the receiver operating characteristic curve; CV: cross-validation; GLM: generalised linear model; LR: logistic regression; ML: machine learning; RF: random forest; XGBoost: extreme gradient boosting

Positive values in the "ML − LR" column indicate superior discrimination by the ML model, whereas negative values favour LR. Reported statistical comparisons are reproduced from the original studies. Where LR was not evaluated, comparison was not possible.

Only one study directly compared an ML model with PTH alone, reporting AUCs of 0.902 and 0.889, respectively. In another study, removing PTH from the postoperative model reduced sensitivity from 0.565 to 0.375, highlighting the importance of PTH measurements for prediction [1,2].

Calibration and Clinical Applicability

None of the studies reported numerical calibration measures. One study presented calibration plots and Brier scores, whereas the remaining studies did not adequately assess calibration. Classification performance was reported using different thresholds, limiting comparisons across studies. Only two studies performed decision-curve analysis. The absence of external validation, limited calibration assessment, and restricted model availability currently limit the clinical applicability of these prediction models.

Risk of Bias

Risk of bias was assessed using PROBAST+AI and is summarised in Table *5*. All included studies were judged to have a high overall risk of bias, primarily because of limitations in model development and analysis, including inadequate validation, insufficient handling of missing data, limited calibration assessment, and lack of external validation. Applicability concerns were identified in several studies due to variations in study populations, predictors, and outcome definitions.

Study ID	Participants	Predictors	Outcome	Analysis	Overall risk of bias	Applicability	Main reasons
Muller et al. [1]	Unclear	Low	Low	High	High	Low	EPV ≤6.7; calibration not reported; handling of missing data not described; exclusions based on calcium supplementation
de Carvalho et al. [2]	Low	Low	High	Unclear	High	High (outcome)	Symptom-based outcome definition with PTH used as a supportive criterion; calibration presented only as plots; categorised predictors
Zhong et al. [3]	Low	Unclear	Unclear	High	High	High (participants)	Mixed surgical extent (31.4% unilateral surgery); prophylactic calcium included as a predictor; composite outcome definition; calibration not reported; univariable predictor screening
Seib et al. [4]	Low	Low	Unclear	High	High	High (participants)	Registry-based outcome without a detailed definition; subtotal thyroidectomy included; clustering effects not accounted for; calibration not reported
Rao et al. [5]	Unclear	Unclear	High	High	High	High (outcome)	Outcome definition not reported; EPV 2.8; evaluation dataset for AUC unclear; calibration not reported
Yavuz et al. [6]	High	Low	Unclear	High	High	Unclear	Exclusion of permanent hypocalcaemia (outcome-related exclusion); mean imputation; confidence intervals not reported; calibration not reported

TABLE 5: Risk of bias and applicability assessment using PROBAST+AI

This table presents the risk of bias and applicability assessment of the included prediction model studies using the PROBAST+AI framework. Assessments are reported across the domains of participants, predictors, outcomes, and analysis, together with overall risk of bias and applicability concerns.

AUC: area under the receiver operating characteristic curve; EPV: events per candidate predictor variable; PTH: parathyroid hormone; PROBAST+AI: Prediction Model Risk of Bias Assessment Tool with Artificial Intelligence

Overall risk of bias was judged according to PROBAST+AI guidance and reflects the highest-risk domain. Applicability concerns indicate the extent to which study participants, predictors, and outcomes align with the review question. Examples of issues contributing to high risk of bias include inadequate handling of missing data, insufficient events per predictor variable, unreported calibration, poorly defined outcomes, and outcome-related exclusions.

Discussion

The findings of this review align with previous evidence demonstrating the central role of postoperative PTH measurement in risk stratification after thyroidectomy. A systematic review of 69 studies involving 9,163 patients reported that early postoperative PTH measurements achieved median sensitivity and specificity values of approximately 85-90% for predicting transient hypocalcaemia, supporting their continued use in clinical practice. Furthermore, the American Thyroid Association (ATA) statement on postoperative hypoparathyroidism recommends the use of early postoperative PTH assessment, noting that a postoperative PTH level below 15 pg/mL identifies patients at increased risk of acute hypoparathyroidism and may guide calcium and vitamin D supplementation strategies [11-13].

The lack of a consistent advantage of ML over logistic regression observed in the present review is also consistent with broader prediction-model literature, which has shown that increases in algorithmic complexity do not necessarily translate into clinically meaningful improvements in discrimination, particularly when key predictors such as PTH already provide substantial prognostic information [10,13].

Importantly, none of the included studies evaluated permanent hypoparathyroidism, despite its clinical significance. Recent evidence suggests that postoperative day 1 PTH levels may also help identify patients at risk for permanent hypoparathyroidism, with a threshold of ⩽7 pg/mL demonstrating high negative predictive value in long-term follow-up studies [11]. Future research should therefore expand beyond transient hypocalcaemia and evaluate whether ML-based approaches can improve prediction of permanent hypoparathyroidism.

Limitations

This review has several limitations. First, only six studies were included, limiting the precision of pooled estimates and preventing subgroup analyses. Second, substantial heterogeneity in study populations, outcome definitions, predictors, and validation methods may affect the comparability and generalisability of the findings. Third, some confidence intervals and sample sizes had to be estimated from reported data. Fourth, selecting the best-performing ML model from each study may have introduced optimism into the pooled performance estimate. Fifth, no included study evaluated permanent hypoparathyroidism; therefore, the findings are limited to early or transient hypocalcaemia. Finally, the exclusion of grey literature and non-English-language publications, together with the absence of an Embase search, may have resulted in relevant studies being missed.

Conclusions

ML models demonstrate good but variable performance in predicting hypocalcaemia after total thyroidectomy, with no clear advantage over logistic regression. Their predictive performance relies largely on postoperative PTH levels. However, the available models remain at high risk of bias, lack external validation, and have insufficient calibration reporting. Overall certainty of the evidence is low to very low. Future research should prioritise external validation and calibration of existing models rather than developing new ones. Until reliable models become available, established PTH-based clinical pathways should remain the standard of care.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Raum Y. Musawi, Amani M. Ibrahim, Amal A. Badri

Acquisition, analysis, or interpretation of data: Raum Y. Musawi, Mohammed A. Alhazmi, Rahaf A. Omer, Reham T. Alruwaili, Rayim A. Oraybi, Yazan Y. Nahari, Saud K. Alahmari, Aryam A. Althobaiti, Tahani M. Ahmed, May H. Alqaed

Drafting of the manuscript: Raum Y. Musawi, Reham T. Alruwaili, Rayim A. Oraybi, Yazan Y. Nahari, Amani M. Ibrahim, Saud K. Alahmari, Tahani M. Ahmed, May H. Alqaed

Critical review of the manuscript for important intellectual content: Raum Y. Musawi, Mohammed A. Alhazmi, Rahaf A. Omer, Aryam A. Althobaiti, Amal A. Badri

Disclosures

AI use statement: The authors used Copilot (Microsoft AI, Redmond, Washington, United States) as follows: language polishing in Methods. The authors confirm that they have reviewed and verified all AI-assisted content and take full responsibility for the work. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Muller O, Bauvin P, Bacoeur O, et al.: Machine learning-based algorithm for the early prediction of postoperative hypocalcemia risk after thyroidectomy. Ann Surg. 2024, 280:835-41. 10.1097/SLA.0000000000006480
2. de Carvalho GB, Martins L, Valieris R, et al.: Artificial intelligence for predicting transient hypocalcemia after total thyroidectomy. Sci Rep. 2026, 16:22145. 10.1038/s41598-026-49133-7
3. Zhong F, Zhou C, Zhang L, et al.: An interpretable nomogram for predicting early acute postoperative hypocalcemia in differentiated thyroid cancer: development and internal validation. Front Endocrinol (Lausanne). 2026, 17:1840443. 10.3389/fendo.2026.1840443
4. Seib CD, Roose JP, Hubbard AE, Suh I: Ensemble machine learning for the prediction of patient-level outcomes following thyroidectomy. Am J Surg. 2021, 222:347-53. 10.1016/j.amjsurg.2020.11.055
5. Rao KN, Arora R, Rajguru R, Nagarkar NM: Artificial neural network to predict post-operative hypocalcemia following total thyroidectomy. Indian J Otolaryngol Head Neck Surg. 2024, 76:3094-102. 10.1007/s12070-024-04608-9
6. Yavuz B, Atar C, Topal U, Cem Eray İ, Dalcı K, Parsak CK, Sakman G: Prediction of transient hypocalcemia following total thyroidectomy using machine learning methods. Curr Probl Surg. 2026, 80:102060.

10.1016/j.cpsurg.2026.102060

7. Moons KG, Damen JA, Kaul T, et al.: PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025, 388:e082505. 10.1136/bmj-2024-082505
8. Collins GS, Moons KG, Dhiman P, et al.: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods: a Korean translation. Ewha Med J. 2025, 48:e48. 10.12771/emj.2025.00668
9. Page MJ, McKenzie JE, Bossuyt PM, et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Rev Esp Cardiol (Engl Ed). 2021, 74:790-9. 10.1016/j.rec.2021.07.010
10. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P: Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007, 7:16. 10.1186/1472-6947-7-16
11. Mathur A, Nagarajan N, Kahan S, Schneider EB, Zeiger MA: Association of parathyroid hormone level with postthyroidectomy hypocalcemia: a systematic review. JAMA Surg. 2018, 153:69-76. 10.1001/jamasurg.2017.3398
12. Orloff LA, Wiseman SM, Bernet VJ, et al.: American Thyroid Association statement on postoperative hypoparathyroidism: diagnosis, prevention, and management in adults. Thyroid. 2018, 28:830-41. 10.1089/thy.2017.0309
13. Akmercan A, Batun KD, Atalay AH, Ergenç M, Güllüoğlu BM, Uğurlu MÜ: Predicting permanent hypoparathyroidism after thyroidectomy: role of early postoperative parathyroid hormone levels and surgical factors. BMC Endocr Disord. 2026, 26:267. 10.1186/s12902-026-02369-w