Review

AI Predictive Models for Discharge Disposition in Patients Following Primary Total Knee Arthroplasty: Systematic Review and Meta-Analysis

Dongdong Xie[1,2], BSc; Rui Chen[3,4], MM; Chen Yue[2], PhD; Xiaoya Liu[1], MNS; Guorui Cao[2*], PhD; Ruina Jiao[3,5*], RN

[1]School of Nursing, Fujian University of Traditional Chinese Medicine, Fuzhou, China
[2]Knee Joint Diagnosis and Treatment Center, Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province, Luoyang, China
[3]Henan Provincial Artificial Intelligence Engineering Research Center for Orthopedic Trauma Rehabilitation, Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province, Luoyang, China
[4]School of Nursing, Shanghai Jiao Tong University, Shanghai, China
[5]Graduate School, Fujian University of Traditional Chinese Medicine, Fuzhou, China
*these authors contributed equally

Corresponding Author:

Ruina Jiao, RN
Henan Provincial Artificial Intelligence Engineering Research Center for Orthopedic Trauma Rehabilitation
Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province
82 Qiming South Road
Luoyang 471000
China
Phone: 86 13703889660
Email: jrnyx@126.com

Abstract

Background: Although a growing number of AI-based predictive models for discharge disposition in patients following primary total knee arthroplasty (TKA) have been constructed, their performance, quality, and applicability remain unclear.

Objective: This study aimed to evaluate the performance, quality, and applicability of published AI-based predictive models for discharge disposition in patients following primary TKA.

Methods: PubMed, Web of Science, Embase, the Cochrane Library, Scopus, China National Knowledge Infrastructure, Wanfang Database, China Science and Technology Journal Database, and SinoMed Database were searched from inception to May 26, 2025. We included original studies that developed AI-based predictive models for discharge disposition in adult patients (age ≥18 years) undergoing primary TKA. Studies involving simultaneous bilateral or revision TKA or those that did not develop a predictive model were excluded. Risk of bias and reporting quality were assessed using the PROBAST+AI (Prediction Model Risk of Bias Assessment Tool+AI) and TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI), respectively. Pooled area under the curve values and 95% CIs were estimated using a DerSimonian-Laird random-effects model with the Hartung-Knapp-Sidik-Jonkman adjustment. Heterogeneity was assessed using the Cochran Q test and I^2 statistic. Subgroup analyses were conducted according to algorithm type, categorized as machine learning (ML)–based and deep learning (DL)–based models.

Results: A total of 5 retrospective studies from the United States involving 4 ML-based models and 1 DL-based model were included, with sample sizes ranging from 175,042 to 434,550 patients. Age and sex were the most frequently used predictors. All included studies were rated as having a high risk of bias according to the PROBAST+AI checklist. The pooled area under the curve was 0.75 (95% CI 0.67-0.82; I^2=99.9%; P<.001) for internal validation and 0.77 (95% CI 0.51-1.00; I^2=99.8%; P<.001) for external validation, indicating a high degree of heterogeneity. Subgroup analysis showed no statistically significant difference between ML-based and DL-based models (P>.05).

Conclusions: AI-based predictive models for discharge disposition following primary TKA demonstrate a certain level of discriminative performance. However, the current evidence is limited, as all included studies were at high risk of bias and showed a high degree of heterogeneity. In addition, all included studies were conducted in the United States, which may

limit the generalizability of these findings to other health care settings. Future studies should follow PROBAST+AI and TRIPOD+AI checklists while developing and validating models across different countries and health care systems.

Trial Registration: PROSPERO CRD420251059218; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251059218

J Med Internet Res 2026;28:e94936; doi: 10.2196/94936

Keywords: AI; machine learning; deep learning; total knee arthroplasty; discharge disposition; predictive model

Introduction

The incidence of degenerative knee diseases is rising due to an aging population and increasing obesity rates [1-4]. Total knee arthroplasty (TKA) is the definitive intervention for end-stage knee diseases and is widely used for patients with knee osteoarthritis, rheumatoid arthritis, and posttraumatic arthritis [5,6]. TKA involves resurfacing the damaged joint with prosthetic implants, typically using metal alloys and high-density polyethylene. This intervention effectively alleviates pain, corrects deformities, and restores joint function, thereby enhancing patients' quality of life [7,8]. Over the past 2 decades, the annual volume of TKA in the US has grown steadily, from approximately 750,000 procedures currently to a projected 1.26 million by 2030 [9,10].

As the volume of TKA procedures continues to rise, the associated health care economic burden is expected to increase substantially [2]. The bundled payment model, implemented by the Centers for Medicare and Medicaid Services, requires health care institutions to manage total costs within target ranges [11]. Postacute care costs account for over 30% of total TKA bundled payments, thus representing a critical component of cost control [12]. Patient discharge dispositions are associated with the demand for postacute care services [13]. Therefore, accurately predicting TKA patients' discharge dispositions is crucial for controlling health care costs. At the patient level, individuals discharged to nonhome settings incur greater nursing costs and experience higher rates of complications, readmissions, and reoperations compared with those discharged to home [13-16]. Thus, accurately predicting discharge disposition assists health care teams in optimizing resource allocation, implementing tailored perioperative pathways, improving health care services, and potentially mitigating the risk of adverse outcomes [17].

AI is a branch of computer science that aims to perform tasks through algorithms that typically require human intelligence [18]. In clinical prediction modeling for discharge disposition following primary TKA, the most commonly applied AI algorithms are machine learning (ML) and deep learning (DL). ML algorithms learn patterns and relationships between input and output variables directly from data, iteratively improving their performance without being explicitly programmed for each task [19]. A more advanced subset of ML is DL, which is characterized by the use of multiple processing layers to learn more complex patterns and relationships in data [18]. The application of AI in health care has grown rapidly, particularly in the field of clinical prediction modeling. Discharge disposition following primary TKA is influenced by a multitude of interacting factors. Compared to traditional clinical risk predictive models, which often rely on predefined assumptions and linear relationships, AI-based predictive models can automatically learn high-dimensional, nonlinear patterns directly from complex data. This represents a powerful approach to enhancing both the accuracy of risk stratification and the efficiency of clinical decision-making [20]. Consequently, a systematic evaluation of existing AI-based predictive models for this purpose is warranted.

Despite the growing number of AI-based models for predicting discharge disposition following primary TKA, their performance, quality, and applicability remain unevaluated. Therefore, this systematic review and meta-analysis aimed to evaluate these aspects of published AI-based predictive models for discharge disposition in patients following primary TKA. To assess the potential clinical usefulness of AI-based models, this study primarily evaluated predictive performance, rather than focusing on the selection of individual predictors across studies.

Methods

Our protocol was registered on PROSPERO (CRD420251059218). This systematic review and meta-analysis was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist (Checklist 1).

Search Strategy

To reduce language bias and ensure comprehensive coverage, this study systematically searched both Chinese and English databases. Search terms included TKA, AI, ML, DL, prediction, and prognosis. Databases including PubMed, Web of Science, Embase, the Cochrane Library, Scopus, China National Knowledge Infrastructure, Wanfang Database, China Science and Technology Journal Database, and SinoMed were searched from their inception to May 26, 2025. The complete search strategy for each database is provided in Table S1 in Multimedia Appendix 1. Additionally, the reference lists of included studies and relevant systematic reviews were manually screened to identify additional eligible studies.

The PICOT strategy for this study is as follows: P (population)—adult patients (age ≥18 years) following primary unilateral TKA; I (intervention model)—published AI-based predictive models for discharge disposition in patients following primary TKA; C (comparator)—not applicable; O (outcome)—the outcome is discharge disposition, dichotomized as home discharge vs nonhome discharge

(the latter considered an adverse outcome); T (timing)—the outcome is predicted after evaluating patient demographics and baseline health, clinical comorbidities and clinical scores, and perioperative factors; and S (setting)—in the hospital setting.

Inclusion and Exclusion Criteria

This systematic review included original studies that developed AI-based predictive models for discharge disposition following primary TKA. The outcome, discharge disposition, was defined as a dichotomous variable: home discharge vs nonhome discharge. Home discharge was defined as returning home with or without home health services, while nonhome discharge referred to facilities such as rehabilitation or skilled nursing facilities. The study population was limited to adult patients (age ≥18 years) following primary TKA. Eligible study designs included cohort, case-control, and cross-sectional studies. We applied the following exclusion criteria: studies involving patients following simultaneous bilateral or revision TKA; studies that did not develop a predictive model; studies with unavailable full texts; and non-English or non-Chinese publications. AI algorithms from the included studies were categorized as either ML or DL. ML was defined as algorithms that were more complicated than regression models and that learned to make predictions from patterns within data, such as histogram-based gradient boosting and support vector machines [21]. DL algorithms were defined as neural networks containing 3 or more hidden layers.

Study Selection and Screening

Duplicate studies were removed using EndNote (Clarivate). Titles and abstracts of all retrieved studies were independently screened by DX and GC. Following this initial screening, the full texts of potentially eligible studies were independently evaluated by the same 2 reviewers. Furthermore, we manually searched the reference lists of all included studies to identify any other relevant publications. Any disagreements between the reviewers were resolved through discussion. If a consensus could not be reached, a third independent reviewer (RC) made the final decision.

Data Extraction

Based on the CHARMS (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies) checklist, we developed a structured data extraction form [22]. The extracted data were categorized into three domains: (1) basic study characteristics: author (year), country, study design, participants, data source, and sample size; (2) model development information: model, outcomes, predictors, predictor preprocessing, predictor selection criteria, handling missing values, and validation method; and (3) model performance: outcome, sample size, area under the curve (AUC) with 95% CI, and other relevant performance indicators.

For studies presenting multiple AI predictive models, the best-performing model that met all inclusion criteria was selected. Data were extracted by DX and independently verified by GC and RC. Any disagreements were resolved through consultation with Jiao to ensure the accuracy and consistency of the data extraction process.

Quality Assessment

DX, GC, and RC independently assessed the quality of the included studies, and any disagreements were resolved by RJ. The risk of bias and the applicability of the included predictive models were assessed using the PROBAST+AI (Prediction Model Risk of Bias Assessment Tool+AI) checklist, which is divided into 2 main sections: model development and model evaluation [23]. The model development section consists of 16 questions, categorized into 4 domains: participants and data sources, predictors, outcomes, and analyses. Similarly, the model evaluation section comprises 18 questions, which are also organized into the same 4 domains. During the assessment, each item was rated as yes (Y), probably yes (PY), probably no (PN), no (N), no information (NI), or, in some cases, not applicable (NA). A study was classified as having a high risk of bias if one or more items were rated as N or PN; otherwise, it was considered to be at low risk. The reporting quality of the studies was evaluated using the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI) checklist, which comprises 27 main items across the sections of title, background, objectives, methods, results, discussion, and registration [24]. Several of the main items contain multiple subitems.

Data Synthesis and Statistical Analysis

All meta-analyses were conducted using Stata software (version 18.0; Stata Corporation). Model performance was quantified by the AUC values and their 95% CIs. For studies that did not report SEs or 95% CIs, we estimated them using the Hanley and McNeil method [25]. Heterogeneity across studies was assessed using the Cochran Q test [26] (with significant heterogeneity set at $P \leq .10$) and the I^2 statistic, where values of 25%, 50%, and 75% represented low, moderate, and high heterogeneity, respectively. We anticipated a high degree of heterogeneity among the included studies, primarily due to differences in modeling algorithms, predictor selection, and cohort characteristics. A random-effects model with Hartung-Knapp-Sidik-Jonkman adjustment was applied to pool AUC values, and the results were visualized using forest plots [27]. Subgroup analyses were performed based on the algorithm types, comparing the predictive performance between ML-based models and DL-based models.

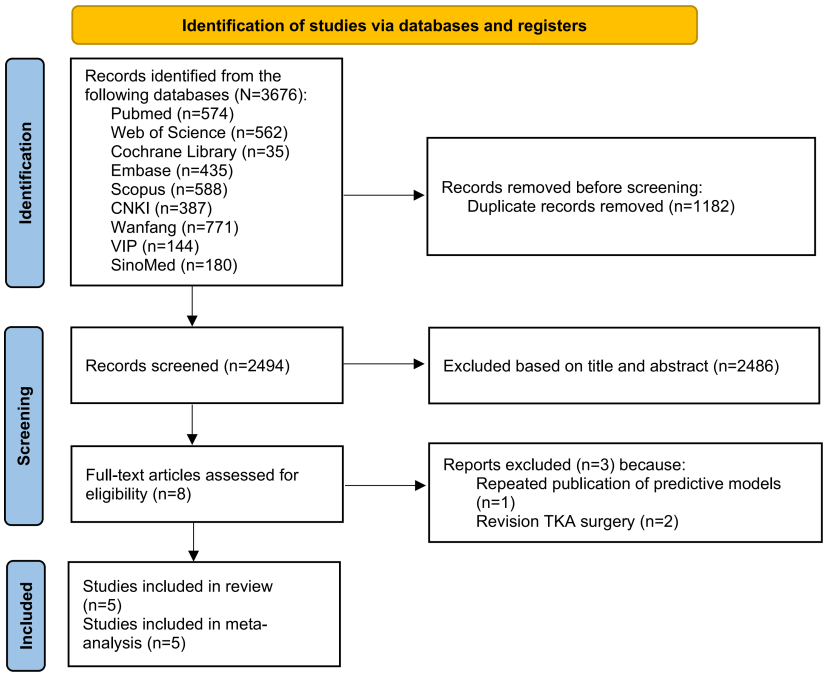
Results

Study Selection

The initial search identified 3676 records. After removing 1182 duplicates, the titles and abstracts of 2494 records were screened, which resulted in the exclusion of 2486 irrelevant studies. Full-text assessment of the remaining 8 studies led to the exclusion of 3 studies due to duplicate reporting of predictive models or revised surgical criteria. Ultimately, 5 studies [13,17,28-30] met the inclusion criteria

and were included in the systematic review. All selected studies were from English-language databases, as no eligible studies were identified from Chinese databases. Among the included studies, 3 focused on predicting discharge dispositions [28-30], while 2 aimed at predicting nonhome discharge [13,17] (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of the study selection process. CNKI: China National Knowledge Infrastructure; TKA: total knee arthroplasty; VIP: China Science and Technology Journal Database.



Study Characteristics

We summarized the study designs and participant characteristics of the 5 included studies [13,17,28-30]. All studies were retrospective analyses of national databases conducted in the United States and published between 2019 and 2024. The 5 studies all shared a primary focus on discharge disposition following primary TKA. Beyond this common outcome, their scope differed. One study also encompassed total hip arthroplasty patients [29], while another examined length of stay and inpatient mortality [28], and a third investigated length of stay and inpatient charges [30]. Sample sizes across the studies ranged from 175,042 to 434,550 patients (Table S2 in Multimedia Appendix 1).

All included studies applied AI algorithms to develop predictive models; 4 studies [13,17,28,29] used ML algorithms and 1 used a DL algorithm [30]. Among all predictive models, the most frequently used predictors were age and sex, which were present in all models. Other frequently used predictors, each appearing in 3 models, were body mass index, smoking status, hypertension, diabetes, cancer, dialysis, and race (Table S3 in Multimedia Appendix 1).

Reported AUC values ranged from 0.69 to 0.89 across the studies. Only one study provided a 95% CI for its AUC. For the remaining studies, the 95% CIs were estimated using the Hanley and McNeil [25] method in Stata software. Given the large sample sizes in all studies, the estimated 95% CIs were relatively narrow (Table 1).

Table 1. Model performance.[a]

Study	Outcome	Sample size	AUC[b] (95% CI)	Other performance indicators
Bacevich et al [13], 2025	Discharge disposition (home or nonhome discharge)	• Total: 365,240 • Training: 362,405 • Testing: 2835	ANN[c]: internal validation, 0.69 (estimated 95% CI 0.66-0.72)	Calibration metrics (testing set): ANN, slope 0.85; intercept 0.04
Chen et al [17], 2023	Discharge disposition (home or nonhome discharge)	• Total: 434,550 • Internal: 424,354 • External: 10,196	HGB[d]: internal validation, 0.84 (estimated 95% CI 0.84-0.84); external validation, 0.89 (estimated 95% CI 0.88-0.90)	Calibration metrics (external): HGB, slope 1.41; intercept –0.12, Brier 0.14
Zalikha et al [28], 2023	Discharge disposition (home or nonhome discharge)	• Total: 305,577 • Training: 64% (195,569)	LSVM[e]: internal validation: 0.723 (estimated 95% CI 0.72-0.73)	Reliability (accuracy): 79.83%

Study	Outcome	Sample size	AUC[b] (95% CI)	Other performance indicators
		• Testing: 16% (48,892) • Validation: 20% (61,115)		
Ramkumar et al [30], 2019	Discharge disposition (home or nonhome discharge)	• Total: 175,042 • Internal: 171,025 • External: 4017	ANN: internal validation, 0.761 (estimated 95% CI 0.76-0.76); external validation, 0.692 (estimated 95% CI 0.68-0.71)	Accuracy: internal: 69.4%; external: 64.4%
Booth et al [29], 2023	Discharge disposition (home or nonhome discharge)	• Total: 278,971 • Training cohort: 213,384 • Temporal validation cohort: 32,794	XGBoost[f]: internal validation, AUC 0.711 (95% CI 0.707–0.715); external validation, AUC 0.729 (95% CI 0.719–0.738)	Brier score: calibrated TKA[g] model: 0.077 (95% CI 0.075-0.079)

[a]For studies that did not report CIs, 95% CIs were estimated using the Hanley and McNeil method.
[b]AUC: area under the curve.
[c]ANN: artificial neural network.
[d]HGB: histogram-based gradient boosting.
[e]LSVM: linear support vector machine.
[f]XGBoost: extreme gradient boosting.
[g]TKA: total knee arthroplasty.

Model Validation

All included studies performed internal validation using established methods, including k-fold cross-validation, hold-out validation, or bootstrapping. External validation was conducted in 3 studies: one ML-based model [17] and one DL-based model [30] were each evaluated on an independent external cohort, while another ML-based model [29] applied temporal validation. In line with established classifications [31,32], temporal validation is considered a form of external validation for the purpose of this analysis (Table S3 in Multimedia Appendix 1).

The main characteristics, strengths, and limitations of the included models were summarized to facilitate comparison across models (Table 2).

Table 2. Summary of the main characteristics, strengths, and limitations of the included models.[a]

Study	Model characteristics	Strengths	Limitations
Bacevich et al [13], 2025	• Outcome: discharge disposition • Data source: ACS-NSQIP[b] database; 365,240 patients, nonhome discharge rates 18.4% • Algorithm: ANN[c], ML[d] • Predictors: demographic characteristics, comorbidities, functional status, ASA[e] classification, and recent preoperative conditions • Validation: internal validation using 5-fold cross-validation; AUC[f] 0.69 (estimated 95% CI 0.66-0.72)	Large national database; predictors based on established clinical risk variables; calibration metrics reported	Only internal validation; poor performance; limited reporting of model development details
Chen et al [17], 2023	• Outcome: discharge disposition • Data source: ACS-NSQIP database+institutional cohort; 434,550 patients (internal: 424,354; external: 10,196), nonhome discharge rates of 19.4% and 36.4% • Algorithm: HGB[g], ML • Predictors: demographic characteristics, comorbidities, and perioperative factors • Validation: internal validation using 5-fold cross-validation and external validation using an institutional cohort; AUC 0.84 (estimated 95% CI 0.84-0.84) and 0.89 (estimated 95% CI 0.88-0.90)	Included external validation; large sample size; calibration performance reported; high performance	External validation data from a single center; exclusion of cases with missing data; simplistic variable preprocessing (eg, race dichotomized)
Zalikha et al [28], 2023	• Outcome: discharge disposition	Used a large national database; incorporated socioeconomic	No external validation; limited predictor set; incomplete reporting of

Study	Model characteristics	Strengths	Limitations
	• Data source: NIS[h] database; 305,577 patients, nonhome discharge rates 20.3% • Algorithm: LSVM[i], ML • Predictors: demographic and socioeconomic characteristics, disease burden, and APRDRG[j] risk classification • Validation: hold-out internal validation; AUC 0.723 (estimated 95% CI 0.72-0.73)	and health care utilization-related predictors; reports accuracy	predictor preprocessing and selection criteria; exclusion of cases with missing data
Ramkumar et al [30], 2019	• Outcome: discharge disposition • Data source: NIS database + OME[k] dataset; 175,042 patients (internal: 171,025; external: 4,017), nonhome discharge rates of 45.6% and 47.1% • Algorithm: ANN, DL[l] • Predictors: demographic characteristics, admission characteristics, disease burden, comorbidities, and socioeconomic factors • Validation: internal validation using 10-fold cross-validation and external validation using an independent dataset; AUC 0.761 (estimated 95% CI 0.76-0.76) and 0.692 (estimated 95% CI 0.68-0.71)	External validation performed; evaluated performance across different datasets; includes admission process variables	Reduced performance in external validation; older Medicare population may limit applicability; incomplete reporting of predictor preprocessing and selection criteria; exclusion of cases with missing data
Booth et al [29], 2023	• Outcome: discharge disposition • Data source: ACS-NSQIP; training cohort 213,384 (nonhome 16.3%), temporal validation 32,794 (9.1%) • Algorithm: XGBoost[m], ML • Predictors: demographic characteristics, comorbidities, ASA classification, and baseline functional status • Validation: bootstrap internal validation and temporal external validation; AUC 0.711 (95% CI 0.707-0.715) and 0.729 (95% CI 0.719-0.738)	Temporal validation performed; Brier score reported; predictors based on established clinical risk variables	Temporal validation from the same database may not represent independent external validation; low nonhome discharge rate in the temporal validation cohort

[a]For studies that did not report CIs, 95% CIs were estimated using the Hanley and McNeil method.
[b]ACS-NSQIP: American College of Surgeons National Surgical Quality Improvement Program.
[c]ANN: artificial neural network.
[d]ML: machine learning.
[e]ASA: American Society of Anesthesiologists.
[f]AUC: area under the curve.
[g]HGB: histogram-based gradient boosting.
[h]NIS: national inpatient sample.
[i]LSVM: linear support vector machine.
[j]APRDRG: all patient refined diagnosis-related groups.
[k]OME: OrthoMiDaS (orthopedic minimal data set) episode of care.
[l]DL: deep learning.
[m]XGBoost: extreme gradient boosting.

Results of Quality Assessment

Based on the TRIPOD+AI checklist, the included studies showed good adherence to reporting standards, with the exception of one study [28], which reported on fewer than 70% of the items. The detailed assessment results are provided in Table S4 in Multimedia Appendix 1.

According to the PROBAST+AI checklist, all included studies were judged to be at high risk of bias across both the model development and model evaluation phases, indicating methodological concerns in these processes. In the participants and data sources domain, all studies [13,17,28-30] were rated at high risk of bias, largely due to their retrospective cohort designs. For predictors, 4 studies [13,28-30] were deemed at high risk of bias, primarily because of potential inconsistencies in defining and assessing predictive variables; one study [17] was considered at unclear risk of bias due to unreported blinding of predictor assessment to outcomes. The outcome domain was rated at unclear risk of bias for all 5 studies [13,17,28-30], as none reported whether outcome assessment was performed blinded to predictor information. Within the model development phase, the analysis domain was rated at high risk of bias in all studies [13,17,28-30], primarily due to inadequate handling of continuous and categorical variables, as well as inadequate handling of missing data. Similarly, in the model evaluation phase, the

analysis domain was judged at high risk of bias in 4 studies [17,28-30], largely due to a failure to appropriately address missing data (Table S5 in Multimedia Appendix 1).

In the applicability assessment, 2 studies were rated as high risk [28,30]. This high-risk rating was primarily attributed to the fact that these models predicted additional outcomes, such as length of stay, inpatient mortality, and hospital charges (Table S5 in Multimedia Appendix 1).

Meta-Analysis of Validation Models Included in the Review

The pooled AUC for internal validation across the 5 studies [13,17,28-30] was 0.75 (95% CI 0.67-0.82), with an I^2 value of 99.9% (P<.001), indicating a high degree of heterogeneity. Given that only one study [30] applied a DL algorithm and the remaining 4 used ML algorithms [13,17,28,29], the results of the subgroup analysis should thus be interpreted with caution. However, it offered preliminary evidence regarding performance differences between ML-based models and DL-based models in predicting discharge disposition following TKA. In the internal validation subgroup analysis, 4 ML-based models [13,17,28,29] showed a pooled AUC of 0.74 (95% CI 0.63-0.85; I^2=99.9%; P<.001), whereas the single DL-based model [30] achieved an AUC of 0.761 (95% CI 0.759-0.763; reported to 3 decimal places to present the precise range). Subgroup analysis showed no statistically significant difference between ML-based and DL-based models (P>.05; Figure 2; Figure 3).

Figure 2. Forest plot of pooled area under the curve (AUC) for internal validation of AI-based predictive models. Weights are from a random-effects model. Standard errors were estimated using the Hanley and McNeil method [13,17,28-30]. DL: deep learning; HKSJ: Hartung-Knapp-Sidik-Jonkman.

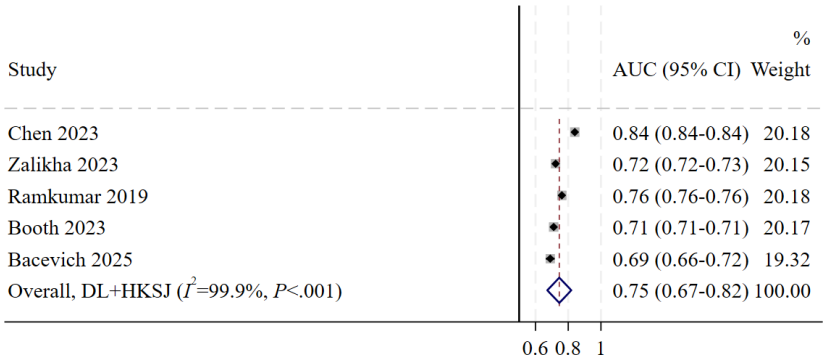
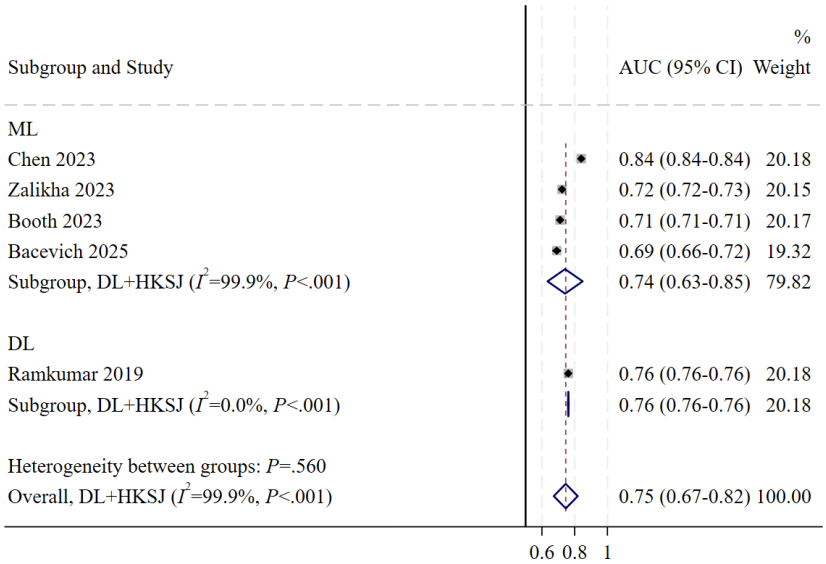


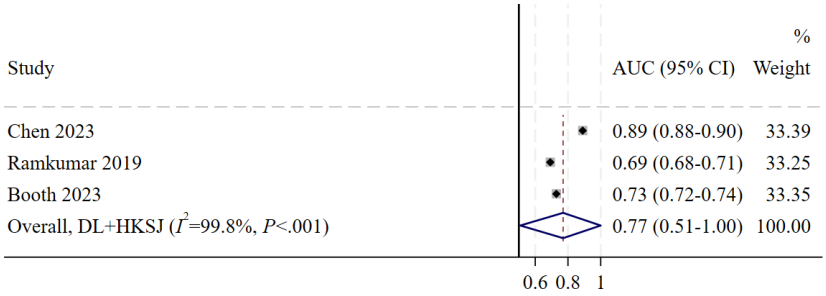
Figure 3. Subgroup analysis of internal validation performance by algorithm type. Weights and the between-subgroup heterogeneity test are from the random-effects model. SEs were calculated using the Hanley and McNeil method [13,17,28-30]. AUC: area under the curve; DL: deep learning subgroup; HKSJ: Hartung-Knapp-Sidik-Jonkman; ML: machine learning subgroup.



Three studies conducted external validation. The meta-analysis yielded a pooled AUC of 0.77 (95% CI 0.51-1.00; I^2=99.8%; P<0.001), indicating considerable heterogeneity [17,29,30]. The limited number of externally validated models resulted in wider CIs, and their upper bounds were truncated at 1.00 because of the theoretical maximum of the AUC [33,34]. An additional DerSimonian-Laird random-effects meta-analysis was performed for the external validation cohort and was considered a sensitivity analysis. This analysis yielded a pooled AUC of 0.77 (95%

CI, 0.64-0.90), which was consistent with the AUC point estimate obtained from the Hartung-Knapp-Sidik-Jonkman analysis. Subgroup analyzes were not performed for the external validation cohort because only 3 externally validated models were available, which was considered insufficient to provide reliable subgroup estimates [35] (Figure 4).

Figure 4. Forest plot of pooled area under the curve (AUC) for the external validation of AI-based predictive models. Weights are from the random-effects model. SEs were calculated using the Hanley and McNeil method. The upper bounds of CIs were truncated at 1.00 because of the theoretical maximum of the AUC [17,29,30]. DL: deep learning; HKSJ: Hartung-Knapp-Sidik-Jonkman.



Discussion

Principal Findings

Five predictive models from 5 studies [13,17,28-30], all conducted in the United States, were included in this review. The pooled AUC for internal validation across all studies was 0.75 (95% CI 0.67-0.82), and for external validation across 3 studies [17,29,30], it was 0.77 (95% CI 0.51-1.00), suggesting a certain level of discriminatory ability. However, all included studies were assessed as having a high risk of bias according to the PROBAST+AI checklist, and a high degree of heterogeneity was observed across the included models. Therefore, the existing evidence remains methodologically limited, and the high risk of bias and high degree of heterogeneity may limit the interpretation of the pooled performance estimates.

Several methodological limitations contributed to the high risk of bias identified in the included studies. A principal methodological limitation was the inadequate handling of missing data, with most studies opting to simply exclude cases with missing data rather than apply robust imputation methods. This approach likely introduced selection bias and may restrict the generalizability of the models to new populations. Additionally, all included predictive models were developed using retrospective cohort data. While this design allows for large sample sizes, the collection was not originally designed for model development. This may result in limitations in data completeness, predictor preprocessing, and consistency in variable definitions, potentially introducing selection and measurement bias. Moreover, most studies did not report CIs for model AUC values, which undermines the evaluation of estimate precision and increases the risk of bias. Owing to these limitations, the current AI-based predictive models for TKA discharge disposition demonstrate limited methodological quality and clinical utility.

Given the high risk of bias identified in the existing studies, future research should focus on improving the methodological quality and clinical utility of AI-based predictive models. Studies should follow the PROBAST+AI checklist throughout the model development and validation process. Prospective model development and validation studies are needed to minimize potential biases associated with retrospective data collection. In addition, standardized methods for handling missing data, such as multiple imputation, should be applied rather than simply excluding cases with missing data. Future studies should also provide comprehensive reporting of model performance metrics. For discrimination performance, studies should report the AUC together with its corresponding 95% CI to provide a complete description of model performance. In addition, calibration assessment should be strengthened by systematically evaluating the agreement between predicted probabilities and observed outcomes using appropriate measures, including calibration plots, calibration intercepts, and calibration slopes. Transparent reporting according to the TRIPOD+AI checklist should be encouraged to improve reproducibility and enable meaningful comparison between models. Finally, multicenter external validation using diverse patient populations and health care settings remains essential to evaluate model generalizability before these AI-based models can be considered for clinical implementation [23,24].

Our review also identified a high degree of heterogeneity across the included studies, which limits the clinical interpretation of the pooled AUC estimates. Although the pooled AUC values suggested a certain level of discriminative ability, the high heterogeneity indicates substantial variability among the included studies [21]. Therefore, the pooled AUC should not be interpreted as a universal estimate of the expected performance of AI-based predictive models in clinical practice. From a clinical implementation perspective, these AI-based predictive models require external validation in the target clinical environment before their clinical implementation [23].

Although subgroup analysis by algorithm type was performed, substantial heterogeneity remained in the ML subgroup (I^2=99.9%). One possible explanation is the large sample sizes of the included studies (ranging from 175,042

to 434,550 patients). Despite the relatively similar point estimates of AUC across studies, the large sample sizes resulted in extremely narrow CIs that were close to the point estimates. Consequently, the overlap between CIs across studies was limited, which may have contributed to the high observed heterogeneity.

Differences in patient populations and data sources may have contributed to the observed heterogeneity. The included AI-based models were developed using different datasets, with variations in baseline characteristics and cohort inclusion criteria. For example, the proportion of patients experiencing nonhome discharge varied substantially across cohorts, ranging from 9.1% to 47.1%. Furthermore, one study included only patients aged 65 years or older, whereas others enrolled broader adult populations [30]. These variations may have resulted in differences in case-mix across studies and contributed to heterogeneity across models. In addition, different datasets may include patients from distinct healthcare institutions, where clinical practices and discharge planning processes may vary [36]. Moreover, differences in data collection procedures, coding systems, and variable definitions across data sources may affect the consistency and standardization of predictor variables, further contributing to model heterogeneity [36].

Differences in predictor selection may have contributed to the observed heterogeneity. Although age and sex were included in all models, the remaining predictors varied across studies. Some models focused on demographic characteristics, comorbidities, and baseline clinical status [13,29], whereas others additionally incorporated socioeconomic factors, admission characteristics, or perioperative variables [17,28,30]. These variations may have resulted in models capturing different patient risk profiles and contributed to heterogeneity across models.

Differences in outcome definitions may represent another source of heterogeneity. Although all studies defined discharge disposition as home vs nonhome discharge, the composition of nonhome discharge categories varied. Most studies included rehabilitation and nursing facilities [13,17,28], whereas others additionally included acute rehabilitation centers, hospice care, or multilevel senior communities [29,30]. These variations may have resulted in different prediction targets, as different nonhome discharge categories represent distinct levels of postacute care needs and clinical complexity [37]. Standardized definitions and reporting frameworks for discharge disposition outcomes are needed in future studies.

The AI-based predictive models varied in algorithm types, modeling strategies, and validation methods, contributing to commonly observed heterogeneity across different models [21,38]. The included studies applied different algorithms, including artificial neural networks (ANNs), histogram-based gradient boosting, linear support vector machines, and extreme gradient boosting, which identify patterns from clinical data using distinct learning mechanisms [39-42]. The strategies used to define candidate predictors during model development also varied across studies, with some relying on established clinical risk calculators and others using predictors identified from previous literature [13,17,29]. Predictor preprocessing procedures also differed, such as restricting BMI values to a predefined range before model development [29]. Furthermore, variations in internal and external validation approaches may have contributed to increased heterogeneity across models.

Substantial inconsistencies were observed in the validation strategies adopted across the included studies. Chen et al [17] and Ramkumar et al [30] used data from public databases, applied k-fold cross-validation for internal validation, and further tested their models on an independent external cohort. This approach effectively assessed the external validity of the models while minimizing the risk of overfitting, representing a methodologically rigorous validation strategy. Temporal validation, a weaker form of external validation, was used by Booth et al [29]. They applied bootstrap resampling for internal validation on their training set and subsequently performed temporal validation using a cohort from the same database but from a different time period. Although this approach is superior to internal validation alone, the fact that the validation data originated from the same database as the development data limits its value in assessing the model's external applicability [43-45]. In comparison, Zalikha et al [13] and Bacevich et al [28] used only internal validation, such as k-fold cross-validation or a hold-out method, without an external validation component. While this approach provides a preliminary estimate of model performance, it often leads to overly optimistic performance estimates, carries a risk of overfitting, and fails to assess the model's external applicability [46]. Collectively, most existing AI-based models for predicting discharge disposition after TKA have not undergone extensive independent external validation. As a result, robust evidence supporting their clinical translation remains insufficient. Future studies should therefore perform rigorous, multicenter external validation to strengthen the generalizability and clinical applicability of these models.

Contrary to the typical expectation that model performance tends to decline in external validation, the study by Chen et al [17] reported a higher AUC in external validation than in internal validation (0.89 vs 0.84) [17,45]. The authors attributed this finding to the greater heterogeneity of patients in the external cohort, coupled with the finer granularity and more selective case inclusion in their institutional dataset. This increased data distribution variability and led to more extreme model predictions, thereby increasing the AUC. This suggests that differences between internal and external validation datasets do not necessarily lead to a decline in model performance on the external set; rather, the focus should be on whether such differences are conducive to model discrimination. Therefore, when assessing a model's external generalizability, it is essential to examine not only the presence of dataset discrepancies but also the nature of these differences. In another study, Booth et al [29] reported that their ML-based model achieved a higher AUC in external validation than in internal validation (0.729 vs 0.711) [29]. This study used temporal validation, a generally weaker form of external validation. Because the development and

validation sets differed only in time period, this approach increases the risk of overfitting and may yield overly optimistic performance estimates [43-45]. Therefore, to better evaluate a model's external validity, validation using an independent external cohort should be preferred. Additionally, as the class distribution in the validation set changed, the authors recalibrated the model prior to external validation to accommodate the declining trend in nonhome discharge rates. This step also enhances the model's applicability to the new dataset.

Among the included studies, the research by Ramkumar et al [30] used a DL algorithm, which is considered a relatively complex and advanced approach within the field of AI algorithms [30]. However, the authors noted that the predictive model they developed demonstrated poor performance in predicting discharge disposition, with an external validation AUC of only 0.69 (95% CI 0.68-0.71). We speculate that this outcome may be related to the algorithm's adaptability to different data structures. DL algorithms typically excel in handling unstructured, high-dimensional data such as images and audio [47]. In this study, however, the input data consisted largely of structured variables, including demographic and disease-related information. In such a data context, DL algorithms may not fully leverage their strengths, which could explain the model's relatively poor predictive performance in the external cohort [48]. Additionally, DL algorithms require high-quality data, but this requirement is often challenged by the prevalence of missing values and noise in public databases [49,50]. Developing models using such databases may further compromise their external generalizability. While regularization techniques such as dropout, early stopping, and weight decay were applied in the DL model to mitigate overfitting and enhance external validity, they do not eliminate the adverse effects of data missingness and noise on model performance [51-53]. Therefore, the selection of an AI algorithm should be guided by its suitability for the specific data characteristics rather than its complexity alone. Given that only one included study utilized a DL algorithm, our findings regarding DL remain preliminary and require validation in future high-quality research.

This study has several limitations. First, a high degree of heterogeneity was observed among the included studies, which may weaken the interpretability of the pooled results and warrants cautious interpretation. This heterogeneity also reflects the diversity in data sources and methodologies within this research field, highlighting the need for methodological standardization in future studies to enhance comparability and reproducibility. Second, this study has a geographical limitation because all included studies were conducted in the United States. Health care systems and discharge practices differ across countries due to variations in healthcare insurance systems, medical expenditure levels, post-acute care systems, and tiered health care delivery structures [54-56]. For example, the United States relies on postacute care services after hospital discharge, whereas other countries may adopt different models of rehabilitation and transitional care. Furthermore, only one DL-based model was included; our analysis of DL algorithms should be regarded as exploratory. Finally, most studies did not report SEs and CIs for AUC values. To perform the meta-analysis, we estimated the SEs and CIs using the Hanley and McNeil method. The Hanley and McNeil method for estimating AUC SEs and corresponding CIs relies on several statistical assumptions. Specifically, it assumes that the variance of AUC estimates can be appropriately approximated based on the observed AUC values and sample sizes, that observations from case and non-case groups are independent, and that the sampling distribution of AUC estimates can be approximated by a normal distribution for CI construction. However, these assumptions may not always be fully satisfied in real-world prediction model studies [25]. Although the Hanley and McNeil method has been widely documented, for example, in PASS software, it is an approximate estimation method based on limited summary-level data and may have issues with estimation inaccuracy. In the one study that did report a CI, the author-reported CI was concordant with our estimate, supporting the appropriateness of the estimation approach applied in this review [29].

Conclusions

To our knowledge, this is the first systematic review and meta-analysis of AI-based predictive models for discharge disposition following primary TKA. This systematic review included 5 predictive models from 5 studies [13,17,28-30]. The pooled AUC for internal validation was 0.75 (95% CI 0.67-0.82), while external validation across three studies yielded a pooled AUC of 0.77 (95% CI 0.51-1.00), indicating a certain level of discriminative ability. However, all included studies were rated as having a high risk of bias according to the PROBAST+AI checklist, with 2 studies [28,30] also showing high concerns regarding applicability. In addition, a high degree of heterogeneity was observed among the included studies. Therefore, the interpretation of the pooled AUC estimates should consider these limitations. Importantly, all included studies were conducted in the United States, and differences in health care systems and discharge practices across countries may limit the generalizability of these findings to other health care settings. Future studies should improve the methodological quality of AI-based predictive models by following the PROBAST+AI checklist throughout model development and validation and adhering to the TRIPOD+AI checklist for transparent reporting. Moreover, models should be developed and externally validated across different countries, health care systems, and patient populations.

Acknowledgments

The authors would like to express our sincere gratitude to Qiao Huang (Assistant Research Fellow) from Zhongnan Hospital of Wuhan University for his valuable methodological assistance during the preparation of this manuscript.

Generative AI Use

During the preparation of this manuscript, ChatGPT (OpenAI) was used to assist with translating portions of the manuscript from Chinese into English and language polishing. DeepSeek was used to check the grammar and verb tense of parts of the manuscript.

Funding

This research was supported by funding from Co-construction Project of the Henan Provincial Scientific Research Special Program on Traditional Chinese Medicine (2025LHZX5051), Project of scientific research for traditional Chinese medicine of the Henan province (2025ZY3094), Science and Technology Program of Luoyang City (2401115b), Special subject of scientific research of traditional Chinese medicine in Henan Province (2024ZY3067), Start-up project of postdoctoral research in Henan Province (365154), and the Special project of scientific research funding for overseas students in Henan Province (2024-04).

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. DX drafted the work and revisited it critically for important intellectual content. GC made a great contribution in the revision process. RC participated in the analysis or interpretation of data for the work. CY and XL prepared tables. RJ made substantial contributions to the conception and design of the work. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Database search strategies, study and model details, and assessment results.
[DOCX File (Microsoft Word File), 89 KB-Multimedia Appendix 1]

Checklist 1

PRISMA checklist.
[PDF File (Adobe File), 135 KB-Checklist 1]

References

1. Cram P, Lu X, Kates SL, Singh JA, Li Y, Wolf BR. Total knee arthroplasty volume, utilization, and outcomes among Medicare beneficiaries, 1991-2010. JAMA. Sep 26, 2012;308(12):1227-1236. [doi: 10.1001/2012.jama.11153] [Medline: 23011713]
2. Singh JA, Yu S, Chen L, Cleveland JD. Rates of total joint replacement in the United States: future projections to 2020-2040 using the National Inpatient Sample. J Rheumatol. Sep 2019;46(9):1134-1140. [doi: 10.3899/jrheum.170990] [Medline: 30988126]
3. Liu P, Li J, Liang W, et al. Evaluation of enhanced pain management programs for patients undergoing surgery for severe knee osteoarthritis: a retrospective study. Ann Med. Dec 2025;57(1):2517818. [doi: 10.1080/07853890.2025.2517818] [Medline: 40503668]
4. Jin W, Wang Q, Jin C, et al. Spatiotemporal distributions and regional disparities of rheumatoid arthritis in 953 global to local locations, 1980-2040, with deep learning-empowered forecasts and evaluation of interventional policies' benefits. Ann Rheum Dis. Jul 2025;84(7):1104-1116. [doi: 10.1016/j.ard.2025.04.009] [Medline: 40527715]
5. Vervullens S, Meert L, Baert I, et al. Prehabilitation before total knee arthroplasty: a systematic review on the use and efficacy of stratified care. Ann Phys Rehabil Med. May 2023;66(4):101705. [doi: 10.1016/j.rehab.2022.101705] [Medline: 36115573]
6. Borazjani R, Khorram R, Ghorbani M, et al. Aspirin versus other anticoagulants for the initial prevention of venous thromboembolism following elective total hip and knee arthroplasty: an umbrella review and meta-analysis. J Arthroplasty. Jan 2026;41(1):225-237. [doi: 10.1016/j.arth.2025.06.032] [Medline: 40513910]
7. AbuMoussa S, White CC 4th, Eichinger JK, Friedman RJ. All-polyethylene versus metal-backed tibial components in total knee arthroplasty. J Knee Surg. Aug 2019;32(8):714-718. [doi: 10.1055/s-0039-1683979] [Medline: 30959547]
8. Canovas F, Dagneaux L. Quality of life after total knee arthroplasty. Orthop Traumatol Surg Res. Feb 2018;104(1S):S41-S46. [doi: 10.1016/j.otsr.2017.04.017] [Medline: 29183821]
9. Sloan M, Premkumar A, Sheth NP. Projected volume of primary total joint arthroplasty in the U.S., 2014 to 2030. J Bone Joint Surg Am. Sep 5, 2018;100(17):1455-1460. [doi: 10.2106/JBJS.17.01617] [Medline: 30180053]

10. Shichman I, Roof M, Askew N, et al. Projections and epidemiology of primary hip and knee arthroplasty in Medicare patients to 2040-2060. JB JS Open Access. 2023;8(1):1. [doi: 10.2106/JBJS.OA.22.00112] [Medline: 36864906]
11. Ryan SP, Goltz DE, Howell CB, Attarian DE, Bolognesi MP, Seyler TM. Skilled nursing facilities after total knee arthroplasty: the time for selective partnerships is now! J Arthroplasty. Dec 2018;33(12):3612-3616. [doi: 10.1016/j.arth.2018.08.012] [Medline: 30197217]
12. Heligman JL. The effect of a discharge disposition algorithm on patient outcomes and satisfaction. Orthop Nurs. 2021;40(3):125-133. [doi: 10.1097/NOR.0000000000000753] [Medline: 34004610]
13. Bacevich BM, Chen TLW, Buddhiraju A, Shimizu MR, Seo HH, Kwon YM. Machine learning model outperforms the ACS Risk Calculator in predicting non-home discharge following primary total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc. Mar 2025;33(3):977-986. [doi: 10.1002/ksa.12492] [Medline: 39344759]
14. Bozic KJ, Ward L, Vail TP, Maze M. Bundled payments in total joint arthroplasty: targeting opportunities for quality improvement and cost reduction. Clin Orthop Relat Res. Jan 2014;472(1):188-193. [doi: 10.1007/s11999-013-3034-3] [Medline: 23649225]
15. Sabeh KG, Rosas S, Buller LT, Roche MW, Hernandez VH. The impact of discharge disposition on episode-of-care reimbursement after primary total hip arthroplasty. J Arthroplasty. Oct 2017;32(10):2969-2973. [doi: 10.1016/j.arth.2017.04.062] [Medline: 28601245]
16. Sattler L, Hing W, Rathbone E, Vertullo C. Intrinsic patient factors predictive of inpatient rehabilitation facility discharge following primary total knee arthroplasty: a systematic review and meta-analysis. BMC Musculoskelet Disord. Jul 22, 2020;21(1):481. [doi: 10.1186/s12891-020-03499-5] [Medline: 32698823]
17. Chen TLW, Buddhiraju A, Seo HH, Subih MA, Tuchinda P, Kwon YM. Internal and external validation of the generalizability of machine learning algorithms in predicting non-home discharge disposition following primary total knee joint arthroplasty. J Arthroplasty. Oct 2023;38(10):1973-1981. [doi: 10.1016/j.arth.2023.01.065] [Medline: 36764409]
18. Kuo RYL, Harrison C, Curran TA, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. Jul 2022;304(1):50-62. [doi: 10.1148/radiol.211785] [Medline: 35348381]
19. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. Jul 17, 2015;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
20. He Y, Liu N, Yang J, Hong Y, Ni H, Zhang Z. Comparison of artificial intelligence and logistic regression models for mortality prediction in acute respiratory distress syndrome: a systematic review and meta-analysis. Intensive Care Med Exp. Feb 21, 2025;13(1):23. [doi: 10.1186/s40635-024-00706-8] [Medline: 39982531]
21. Smith LA, Oakden-Rayner L, Bird A, et al. Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. Lancet Digit Health. Dec 2023;5(12):e872-e881. [doi: 10.1016/S2589-7500(23)00177-2] [Medline: 38000872]
22. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. Oct 2014;11(10):e1001744. [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
23. Moons KGM, Damen JAA, 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. Mar 24, 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
24. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. Apr 1982;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
26. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. Sep 6, 2003;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
27. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. Feb 18, 2014;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
28. Zalikha L, Court T, Nham F, El-Othmani MM, Shah RP. Improved performance of machine learning models in predicting length of stay, discharge disposition, and inpatient mortality after total knee arthroplasty using patient-specific variables. Arthroplasty. Jul 2, 2023;5(1):31. [doi: 10.1186/s42836-023-00187-2] [Medline: 37393281]
29. Booth GJ, Cole J, Geiger P, et al. Machine learning to predict discharge destination after total knee arthroplasty and total hip arthroplasty. J Surg Orthop Adv. 2023;32(4):252-258. [Medline: 38551234]

30. Ramkumar PN, Karnuta JM, Navarro SM, et al. Deep learning preoperatively predicts value metrics for primary total knee arthroplasty: development and validation of an artificial neural network model. J Arthroplasty. Oct 2019;34(10):2220-2227. [doi: 10.1016/j.arth.2019.05.034] [Medline: 31285089]
31. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development,Validation, and Updating. 1st ed. Springer; 2009. ISBN: 9780387772431
32. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. Dec 12, 2022;22(1):316. [doi: 10.1186/s12874-022-01801-8] [Medline: 36510134]
33. Chen Y, Ou Z, Deng Y, et al. Predictive value of machine learning for poststroke mortality risk: systematic review and meta-analysis. J Med Internet Res. Apr 2, 2026;28:e83821. [doi: 10.2196/83821] [Medline: 41926763]
34. Kusche H, Gundler C, Johanns O, et al. Leveraging machine-learning techniques to detect recurrences in cancer registry data: a multi-registry validation study using German lung cancer data. Eur J Cancer. Sep 9, 2025;227:115604. [doi: 10.1016/j.ejca.2025.115604] [Medline: 40674828]
35. Tatas Z, Kyriakou E, Koutsiouroumpa O, Seehra J, Mavridis D, Pandis N. Most meta-analyses in oral health do not have conclusive and robust results. J Dent. Oct 2024;149:105309. [doi: 10.1016/j.jdent.2024.105309] [Medline: 39142375]
36. Weiss A, Anderson JE, Chang DC. Comparing the National Surgical Quality Improvement Program with the Nationwide Inpatient Sample database. JAMA Surg. Aug 2015;150(8):815-816. [doi: 10.1001/jamasurg.2015.0962] [Medline: 26061977]
37. Keswani A, Tasi MC, Fields A, Lovy AJ, Moucha CS, Bozic KJ. Discharge destination after total joint arthroplasty: an analysis of postdischarge outcomes, placement risk factors, and recent trends. J Arthroplasty. Jun 2016;31(6):1155-1162. [doi: 10.1016/j.arth.2015.11.044] [Medline: 26860962]
38. Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. Feb 2023;154:8-22. [doi: 10.1016/j.jclinepi.2022.11.015] [Medline: 36436815]
39. Sapoval N, Aghazadeh A, Nute MG, et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun. Apr 1, 2022;13(1):1728. [doi: 10.1038/s41467-022-29268-7] [Medline: 35365602]
40. Kolasa K, Admassu B, Hołownia-Voloskova M, Kędzior KJ, Poirrier JE, Perni S. Systematic reviews of machine learning in healthcare: a literature review. Expert Rev Pharmacoecon Outcomes Res. Jan 2024;24(1):63-115. [doi: 10.1080/14737167.2023.2279107] [Medline: 37955147]
41. Nwanosike EM, Conway BR, Merchant HA, Hasan SS. Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review. Int J Med Inform. Mar 2022;159:104679. [doi: 10.1016/j.ijmedinf.2021.104679] [Medline: 34990939]
42. Zhou L, Song J, Li Z, Hu Y, Guo W. THGB: predicting ligand-receptor interactions by combining tree boosting and histogram-based gradient boosting. Sci Rep. Nov 28, 2024;14(1):29604. [doi: 10.1038/s41598-024-78954-7] [Medline: 39609487]
43. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. May 28, 2009;338:b605. [doi: 10.1136/bmj.b605] [Medline: 19477892]
44. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn Progn Res. 2019;3:16. [doi: 10.1186/s41512-019-0060-y] [Medline: 31463368]
45. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. Jan 2021;14(1):49-58. [doi: 10.1093/ckj/sfaa188] [Medline: 33564405]
46. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. Jan 2016;69:245-247. [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
47. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. Dec 2017;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
48. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Adv Neural Inf Process Syst. 2022:507-520. URL: https://dl.acm.org/doi/10.5555/3600270.3600307 [Accessed 2026-10-23]
49. Bhatt N, Bhatt N, Prajapati P, Sorathiya V, Alshathri S, El-Shafai W. A Data-Centric Approach to improve performance of deep learning models. Sci Rep. Sep 27, 2024;14(1):22329. [doi: 10.1038/s41598-024-73643-x] [Medline: 39333381]
50. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst. Nov 2023;34(11):8135-8153. [doi: 10.1109/TNNLS.2022.3152527] [Medline: 35254993]

51. Demir-Kavuk O, Kamada M, Akutsu T, Knapp EW. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics. Oct 25, 2011;12:412. [doi: 10.1186/1471-2105-12-412] [Medline: 22026913]
52. Xu C, Coen-Pirani P, Jiang X. Empirical study of overfitting in deep learning for predicting breast cancer metastasis. Cancers (Basel). Mar 25, 2023;15(7):7. [doi: 10.3390/cancers15071969] [Medline: 37046630]
53. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929-1958. [doi: 10.5555/2627435.2670313]
54. Fuchs VR. How and why US health care differs from that in other OECD countries. JAMA. Jan 2, 2013;309(1):33-34. [doi: 10.1001/jama.2012.125458] [Medline: 23280219]
55. Konetzka RT, Stuart EA, Werner RM. The effect of integration of hospitals and post-acute care providers on Medicare payment and patient outcomes. J Health Econ. Sep 2018;61:244-258. [doi: 10.1016/j.jhealeco.2018.01.005] [Medline: 29428772]
56. Lyndrup O, Kehlet H, Jørgensen CC, et al. Discharge destination of patients in fast-track primary hip and knee arthroplasty: results from a prospective Danish cohort. J Arthroplasty. Jul 2025;40(7):1690-1695. [doi: 10.1016/j.arth.2024.12.005] [Medline: 39653131]

Abbreviations

ANN: artificial neural network
AUC: area under the curve
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
DL: deep learning
HKSJ: Hartung-Knapp-Sidik-Jonkman
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool+AI
TKA: total knee arthroplasty
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI

Edited by Ivan Steenstra; peer-reviewed by Xiaolong Liang, Xiaoni Chai; submitted 09.Mar.2026; final revised version received 02.Oct.2026; accepted 06.Oct.2026; published 16.Nov.2026

Please cite as:
Xie D, Chen R, Yue C, Liu X, Cao G, Jiao R
AI Predictive Models for Discharge Disposition in Patients Following Primary Total Knee Arthroplasty: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e94936
URL: https://www.jmir.org/2026/1/e94936
doi: 10.2196/94936

© Dongdong Xie, Rui Chen, Chen Yue, Xiaoya Liu, Guorui Cao, Ruina Jiao. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.