Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transportability of artificial intelligence models for predicting hip and knee osteoarthritis progression: a systematic review and meta-analysis of within-model changes in discrimination

Qianyu Niu
The Third Clinical Medical College of Guangzhou University of Chinese Medicine
Jiaqi Yang
The Third Clinical Medical College of Guangzhou University of Chinese Medicine
Xinwei Liu
The Third Clinical Medical College of Guangzhou University of Chinese Medicine
Fan Yang
The Third Clinical Medical College of Guangzhou University of Chinese Medicine
Wei He
The Third Clinical Medical College of Guangzhou University of Chinese Medicine
Xiaoqiang Yang
401728942@qq.com
The Third Clinical Medical College of Guangzhou University of Chinese Medicine

Research Article

Keywords: osteoarthritis, artificial intelligence, machine learning, deep learning, external validation, transportability, prognosis, meta-analysis

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11403368/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Artificial intelligence (AI) models can achieve strong discrimination for osteoarthritis (OA) progression in development cohorts, but clinical translation depends on whether performance is maintained in new populations and settings. Previous OA reviews have mainly summarized absolute predictive performance rather than within-model changes after transport.

Methods

We searched PubMed, Embase, Web of Science Core Collection, and the Cochrane Library/CENTRAL from inception through 10 September 2026. We included reports evaluating AI or machine-learning models for future hip or knee OA outcomes and retained prespecified contextual evidence relevant to transportability. Validation was classified as independent external evaluation, internal-external or leave-one-cohort-out evaluation, quasi-external evaluation, or target-domain adaptation. The primary effect was the matched change in discrimination (ΔAUC/AUROC = AUC_external − AUC_internal). Risk of bias and applicability were assessed with PROBAST + AI. Random-effects synthesis used REML with publication-cluster robust inference; Hartung-Knapp, publication-level, leave-one-publication-out, and preprint sensitivity analyses were also performed.

Results

Of 14,171 records, 28 reports were included; 24 concerned knee OA and four hip OA. OAI was used in 24 reports and MOST in 15. The primary meta-analysis included 18 matched AUC/AUROC effects from five peer-reviewed publication clusters; 16 of 18 effects were negative. The pooled change was − 0.044 (publication-cluster robust 95% CI, − 0.064 to − 0.024; I^2 = 31.3%), with a similar effect-level estimate (− 0.044; 95% CI, − 0.062 to − 0.027). Adding one externally evaluated preprint yielded − 0.050 (95% CI, − 0.074 to − 0.026). All five publication-level mean changes were negative. Overall model-evaluation risk of bias was low in 7 reports, unclear in 10, and high in 11. External-domain evidence was heavily concentrated on OAI-to-MOST transport.

Conclusions

Across available matched comparisons, external evaluation was associated with an average reduction of approximately 0.044 AUC/AUROC units. Because the evidence base was small and repeatedly reused OAI and MOST, the pooled estimate should be interpreted as an empirical summary of available transport contrasts rather than a general transportability penalty. Future studies should evaluate frozen models in genuinely independent settings and report calibration and clinical utility alongside discrimination.

Introduction

Osteoarthritis (OA) is a heterogeneous chronic disease in which the timing and extent of structural deterioration, symptomatic worsening, and progression to joint replacement vary markedly among individuals. Prediction models that identify patients at high risk of progression could support trial enrichment, risk-adapted monitoring, and potentially earlier intervention. Over the past decade, machine-learning and deep-learning methods have increasingly been applied to radiographs, magnetic resonance imaging, clinical variables, biomarkers, and multimodal data to forecast knee and hip OA outcomes.

Clinical translation depends on whether predictive performance is maintained outside the model's development setting. In medical AI, such transport can be affected by differences in patient case mix, disease prevalence and severity, imaging hardware and processing, predictor availability, and outcome ascertainment. In a broad review of externally validated radiologic deep-learning algorithms, Yu and colleagues found that 81% of algorithms performed worse on external data, with a median external-minus-development performance difference of approximately − 0.046 (1).

Recent OA-specific reviews have shown that progression models can achieve apparently useful discrimination, while also highlighting substantial heterogeneity and limited independent external evaluation. Liu et al. pooled absolute performance across model classes, input modalities, progression definitions, and validation strategies in 32 knee OA studies (2). Gillani et al. subsequently reviewed 33 deep-learning imaging studies and reported that 97% used the Osteoarthritis Initiative (OAI) for training, whereas only 27% performed external validation; the descriptive median AUC was 0.81 internally and 0.75 externally (3). These reviews provide important benchmarks, but pooling absolute AUC or C-index values from different models, outcomes, and populations does not directly quantify the change in discrimination experienced by the same model after transport to a new domain.

OA prediction studies repeatedly reuse OAI, the Multicenter Osteoarthritis Study (MOST), and related benchmark resources. Multiple publications, model variants, prediction horizons, knees, or outcome definitions can therefore generate many effect estimates without contributing an equivalent amount of independent transportability evidence. Recent AI reviews increasingly distinguish reports from unique cohorts and explicitly assess cohort overlap (4, 5). Treating every published model or effect estimate as independent can overstate precision and obscure the number of genuinely distinct external domains that have been evaluated.

We conducted a systematic review and meta-analysis focused on transportability rather than absolute predictive performance. The primary objective was to quantify the matched within-model change in discrimination between internal and external evaluation. We also mapped the validation architecture of the literature across independent external evaluation, internal-external or leave-one-cohort-out evaluation, quasi-external evaluation, and target-domain adaptation. Publication/report, model family, effect estimate, and dataset-domain comparison were treated as distinct units when interpreting the amount of independent evidence.

Methods

Registration and reporting standards

This systematic review and meta-analysis was conducted and reported in accordance with the PRISMA 2020 statement (6). The protocol was prospectively registered in PROSPERO (CRD420261526312). Terminology for prediction-model development and evaluation followed TRIPOD + AI, including the distinction between development data and data used to evaluate model performance (7). Risk of bias and applicability were assessed using PROBAST + AI (8). The review question, data items, and quantitative synthesis were additionally informed by established guidance for systematic reviews and meta-analyses of prediction-model performance (9).

Eligibility criteria

We included reports involving adults with, or at risk of, hip or knee OA that developed or evaluated an AI or machine-learning model for a future OA-related outcome. Eligible prognostic outcomes included incident radiographic OA, structural progression, symptomatic or pain progression, combined structural-symptomatic progression, and progression to total knee or hip replacement. Deep-learning, conventional machine-learning, and multimodal AI approaches were eligible. Conventional regression models were retained when they provided a directly relevant external-validation comparator. The formal prognostic evidence set comprised reports predicting future OA-related outcomes. A separately labelled contextual stratum was retained to map validation architecture, including internal-only progression studies, quasi-external evaluations, target-domain adaptation studies, and one closely related multicentre external AI evaluation with a non-prognostic OA endpoint. Contextual evidence did not contribute to the primary prognostic meta-analysis.

We did not require the terms external validation, transportability, or generalizability to appear in the title or abstract. Reports were not excluded solely because they were preprints or conference proceedings, because prediction horizons differed modestly between development and external cohorts, because discrimination was reported as a C-index rather than an AUC, or because target-domain fine-tuning was performed. Instead, these features were captured and addressed through stratification or sensitivity analysis. Reviews, editorials, protocols, non-OA populations, non-predictive tasks without relevant transportability context, and reports without usable model-performance information were excluded.

Information sources and search strategy

PubMed, Embase, Web of Science Core Collection, and the Cochrane Library/CENTRAL were searched from database inception through 29 September 2026. The search combined controlled vocabulary and free-text terms for knee or hip OA, artificial intelligence, machine learning, prediction modelling, prognosis, progression, incident disease, worsening, and arthroplasty. No restrictions on date, language, or publication type were applied at the search stage. Complete database-specific search strategies and yields are provided in Supplementary Table S5.

Study selection and data extraction

Study selection and data extraction
All retrieved records were managed in EndNote 21. Two reviewers (QN and JY) independently screened titles and abstracts and then assessed full-text reports against the prespecified eligibility criteria. Disagreements were resolved by discussion or, when necessary, consultation with a third reviewer (WH). The same two reviewers independently extracted data using a standardized spreadsheet. Extracted items included first author and publication year; joint; development/internal and external/validation data sources; sample size, when reported; target outcome and prediction horizon; predictor modality; model or algorithm family; validation design; discrimination metric; internal and external performance estimates with measures of uncertainty; the method used to estimate source/internal performance (for example, repeated cross-validation or a

held-out source-domain test set); and whether target-domain adaptation or fine-tuning occurred before evaluation. Publication/report, model family, and dataset-domain comparison were treated as distinct analytic units because repeated OAI-to-MOST evaluations across publications do not constitute independent external domains.

Validation taxonomy

For this review, validation designs were grouped into four operational categories; these labels were used descriptively and were not intended as a universal or ordinal hierarchy. L1 denoted independent external evaluation in a cohort, institution, or data resource distinct from that used for model development. L2 denoted internal-external evaluation, including leave-one-cohort-out and temporal designs. L3 denoted quasi-external evaluation using a held-out center or subset within the same overarching data resource. Target-domain adaptation, including transfer learning or fine-tuning with target-cohort data, was retained as a separate category because post-adaptation performance does not estimate frozen-model transportability. Surgical endpoints and conventional statistical comparators were also labelled separately to avoid combining clinically or statistically non-equivalent evidence. Report-level classifications are summarized in Supplementary Table S1.

Risk of bias and applicability

Two reviewers independently assessed the included reports using PROBAST + AI and finalized all judgments by agreement. Because the central review question concerned model performance outside the development setting, the primary quality assessment focused on model-evaluation risk of bias across participants/data sources, predictors, outcome, and analysis, together with applicability. Detailed report-level judgments and rationales are provided in Supplementary Table S3.

We did not apply a formal certainty-of-evidence grading framework such as GRADE. Instead, we evaluated the robustness and credibility of the evidence across complementary dimensions relevant to model transportability: PROBAST + AI risk of bias and applicability, consistency and heterogeneity of effect estimates, precision, independence of datasets and publication clusters, and prespecified sensitivity analyses. We therefore did not assign formal high, moderate, low, or very low certainty ratings.

Effect measure and quantitative synthesis

The primary transportability effect was the matched within-model change in discrimination: ΔAUC/AUROC = AUC_external - AUC_internal. Negative values indicate lower discrimination in the external evaluation dataset. AUC and AUROC were treated as equivalent measures of discrimination. C-index results were summarized separately because their interpretation in time-to-event settings is not necessarily interchangeable with AUC estimates from binary outcomes. Surgical endpoints were likewise analyzed separately from non-surgical OA progression.

When 95% confidence intervals were reported, standard errors were derived under an approximate normality assumption. If internal and external performance were evaluated in independent cohorts and their covariance was unavailable, the variance of ΔAUC was estimated as the sum of the component variances. For Wang et al., uncertainty was reconstructed from the available event and non-event counts using the Hanley-McNeil approximation for AUC sampling variability (10). Studies reporting only run-level or fold-level dispersion, rather than patient-sampling uncertainty, were retained descriptively but were not assigned conventional sampling standard errors for the primary meta-analysis.

The primary quantitative synthesis was restricted to peer-reviewed reports with a matched internal-external AUC/AUROC contrast and defensible sampling uncertainty. A random-effects model with restricted maximum likelihood (REML) was used to estimate between-effect heterogeneity. Because several publications contributed multiple effects, the REML point estimate was paired with a publication-cluster robust sandwich standard error and a t reference distribution with G-1 degrees of freedom to obtain a dependence-aware confidence interval (11). This clustering addresses multiplicity within a publication but cannot fully account for dependence created when the same source or target cohorts are reused across different publications. Only five publication clusters contributed to the primary synthesis, and robust-variance methods are asymptotic in the number of clusters; the resulting interval was therefore treated as exploratory rather than confirmatory. An effect-level Hartung-Knapp random-effects analysis served as a complementary sensitivity analysis (12). Formal domain-level meta-analysis was not attempted because too few independent source-to-target transport domains were available. The effect-level dataset is reported in Supplementary Table S2.

Additional sensitivity analyses included equal weighting of publication-level means, leave-one-publication-out analysis, and addition of the 2026 Blanco et al. preprint. Effects were also described according to how source/internal performance had been estimated, and repeated OAI-to-MOST or MOST-to-OAI evaluations were collapsed descriptively at the transport-domain level. These comparisons were not subjected to subgroup inference because the number of publication clusters and distinct domains was too small. The funnel plot was treated as an exploratory visualization only; no formal Egger-type test was interpreted because the 18 effects arose from just five publication clusters. Formal meta-regression was likewise not undertaken. A descriptive publication-level bubble plot explored the relationship between internal

discrimination and transportability change without inferential testing. Sensitivity and secondary analyses are summarized in Supplementary Table S4.

Results

Study selection

The database search identified 14,171 records: 3,942 from PubMed, 5,907 from Embase, 4,013 from Web of Science Core Collection, and 309 from the Cochrane Library/CENTRAL. After 3,235 duplicates were removed, 10,936 records underwent title and abstract screening, of which 8,872 were excluded. We sought 2,064 full-text reports; 13 could not be retrieved, leaving 2,051 reports for eligibility assessment. A further 2,023 reports were excluded because of the wrong population or disease (n = 369), no eligible prediction model (n = 522), the wrong target outcome or a non-prognostic task (n = 621), no relevant model-performance or validation data (n = 295), or an insufficient report (n = 216). Twenty-eight reports were included in the systematic review (Fig. 1).

Characteristics of the evidence base

The evidence map contained 28 reports published between 2019 and 2026; 24 concerned knee OA and four concerned hip OA. Twenty-seven reports evaluated future OA-related outcomes. One multicentre cross-sectional severity-grading study was retained only as contextual evidence on external-validation architecture and did not contribute to the prognostic quantitative synthesis. Most evidence came from a small number of public longitudinal cohorts: OAI appeared in 24 reports and MOST in 15. Independent external evaluations frequently used the OAI-to-MOST transport axis, whereas fewer studies evaluated models in substantially different clinical settings, including a Taiwanese hospital, PROCOAC in Spain, JoCoOA, and the multi-cohort World COACH resource. Outcomes included structural, symptomatic, combined, and surgical progression, and predictor sets included radiographs, MRI, clinical variables, biomarkers, and multimodal combinations. Five peer-reviewed publication clusters contributed to the primary paired AUC/AUROC meta-analysis. The individual reports are summarized in Table 1 (13–40), with their validation roles mapped in Supplementary Table S1.

Table 1
Characteristics of the 28 included reports and their validation designs.

Study	Joint	Development / internal data	External / validation data	Target outcome	Prediction horizon	Predictor modality	Model / algorithm	Validation category
Han et al., 2022	Knee	OAI	MOST	Radiographic/KL structural progression	NA	Radiograph + clinical	AI/ML prediction model	L1 independent external
Almhdie-Imjabbar et al., 2022	Knee	OAI and MOST	OAI ↔ MOST	mJSN structural progression	NA	Radiographic bone texture	ML / Siamese-network-derived radiologic descriptors	L1 independent external
Salis et al., 2024	Knee	OAI	MOST	End-stage symptomatic + radiographic OA	~ 2 to 2.5 y; ~4 to 5 y	Clinical + radiographic predictors	ML models (full and reduced predictor sets)	L1 independent external
Wang et al., 2025	Knee	OAI test set (n = 578)	Independent Taiwan hospital (n = 274)	KL progression	48 months	Radiograph + clinical	Deep learning / multimodal prediction	L1 independent external
Nair et al., 2024	Knee	OAI	MOST	Clinically important WOMAC change	2 years	Clinical; clinical + imaging	Gradient boosting machine	L1 independent external
Dunn et al., 2023	Knee	OABC/FNIH	JoCoOA and OAI	Radiographic structural progression	NA	Epigenetic biomarkers	Machine-learning biomarker model	L1 independent external
van den Berg et al., 2026 (morphology)	Hip	World COACH pooled cohorts (35,984 hips)	Leave-one-cohort-out across 9 cohorts	Incident radiographic hip OA	4–8 years	Radiographic morphology + demographics	GLMM and random forest	L2 internal-external / LOCO
van den Berg et al., 2026 (deep learning)	Hip	World COACH pooled cohorts (25,090 hips)	Leave-one-cohort-out across 7 cohorts	Incident radiographic hip OA	4–8 years	DL radiographic features + demographics	Deep-learning feature models / hybrid models	L2 internal-external / LOCO
Blanco et al., 2026	Knee	OAI	PROCOAC	Pain / clinical progression	NA	Clinical + imaging	Machine learning	L1 independent external
Rajamohan et al., 2024	Knee	OAI	MOST	Incident and progressive OA	NA	Imaging + clinical	Machine learning	External / contextual
Rajamohan et al., 2026	Knee	OAI	MOST	Incident and progressive OA	NA	Imaging + clinical	Sequential / continual-learning framework	External / contextual
Rajamohan et al., 2023	Knee	OAI	MOST	Future total knee replacement	NA	MRI / radiographic imaging	Deep learning	L1 external; surgical endpoint
Rajamohan et al., 2025 (RiskFORM)	Knee	OAI	MOST	Future total knee replacement	1, 2, and 4 years	Radiograph + risk factors	RiskFORM prediction model	L1 external; surgical endpoint

Abbreviations: AI, artificial intelligence; DL, deep learning; GLMM, generalized linear mixed model; KL, Kellgren–Lawrence; KOA, knee osteoarthritis; LOCO, leave-one-cohort-out; ML, machine learning; MOST, Multicenter Osteoarthritis Study; OAI, Osteoarthritis Initiative; OA, osteoarthritis; PROCOAC, Prospective Cohort of Osteoarthritis from A Coruña; TKR, total knee replacement; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index. L1, independent external validation in a distinct cohort or clinical setting; L2, internal-external validation including LOCO designs; L3, quasi-external validation within the same overarching data resource. NA, not reported or not applicable.

Study	Joint	Development / internal data	External / validation data	Target outcome	Prediction horizon	Predictor modality	Model / algorithm	Validation category
Cigdem et al., 2024	Knee	OAI	MOST and independent hospital	Time to total knee replacement	NA	X-ray / MRI features	Survival / machine-learning models	L1 external; surgical endpoint
Cigdem et al., 2025	Knee	OAI	MOST and/or independent hospital	Time to total knee replacement	NA	Imaging + clinical	Machine-learning survival model	L1 external; surgical endpoint
Zhang et al., 2025 (MR-Transformer)	Knee	OAI	MOST	Future total knee replacement	NA	MRI	Transformer-based model	L1 external; surgical endpoint
McCabe et al., 2022	Knee	OAI	MOST	Future OA occurrence / progression	NA	Clinical / radiographic	Cox prediction model	Independent external comparator
Paz-González et al., 2024	Knee	OAI	PROCOAC	Incident radiographic KOA	96 months	Clinical / radiographic	Stepwise multivariable logistic regression	Independent external comparator
Tiulpin et al., 2019	Knee	OAI	MOST	OA progression / TKR composite	NA	Radiograph + clinical	Deep learning + clinical prediction	External / contextual
Lee et al., 2025	Knee	OAI	MOST	OA progression	NA	NA	AI/ML prediction model	External / contextual
Castagno et al., 2025	Knee	FNIH/OAI	POMA	Rapid radiographic and/or pain progression	NA	Imaging + clinical	Machine learning	L3 quasi-external
Yin et al., 2024	Knee	OAI	OAI center/subset holdout	Incident radiographic OA	NA	Imaging	Deep learning	L3 quasi-external
Panfilov et al., 2022	Knee	OAI	OAI held-out data	OA progression	NA	Imaging	Deep learning	L3 quasi-external
Attur et al., 2020	Knee	NYU / OAI	Replication cohort	OA progression	NA	Biomarkers + clinical	Machine learning	External / contextual
IMI-APPROACH validation	Knee	European source cohorts	IMI-APPROACH prospective cohort	Pain and structural progression	NA	Multimodal clinical / imaging	Machine learning	External / contextual
Herrera et al., 2026	Knee	OAI / MOST	Target cohort after adaptation	Structural progression	NA	Imaging	Transfer-learning model	Adapted external validation
Hidaka et al., 2024	Hip	Single-center hip OA cohort (361 patients; 1,697 radiographs)	Five-fold cross-validation	Progression from KL 0–2 to KL ≥ 3	3, 4, and 5 years	Radiograph + demographics	EfficientNet-B4 features + gradient boosting	Internal validation only
Xu et al., 2025	Hip	Two Chinese hospitals (20,745 radiographs)	Third hospital (1,928 radiographs) + OAI (1,249 hips)	Current KL severity grading	Cross-sectional	Radiograph	CBAM-enhanced ResNet-50	External; cross-sectional context

Abbreviations: AI, artificial intelligence; DL, deep learning; GLMM, generalized linear mixed model; KL, Kellgren–Lawrence; KOA, knee osteoarthritis; LOCO, leave-one-cohort-out; ML, machine learning; MOST, Multicenter Osteoarthritis Study; OAI, Osteoarthritis Initiative; OA, osteoarthritis; PROCOAC, Prospective Cohort of Osteoarthritis from A Coruña; TKR, total knee replacement; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index. L1, independent external validation in a distinct cohort or clinical setting; L2, internal-external validation including LOCO designs; L3, quasi-external validation within the same overarching data resource. NA, not reported or not applicable.

Risk of bias and applicability

Overall model-evaluation risk of bias was rated low in 7 of 28 reports, unclear in 10, and high in 11. The analysis domain accounted for most concerns: 8 reports were rated low risk, 11 unclear, and 9 high risk. Outcome assessment, by contrast, was rated low risk in all 28 reports under the final review judgments. Overall applicability was of low concern in 19 reports, unclear in five, and high concern in four. Thus, the principal limitations arose from model-development and evaluation procedures, sample size, performance estimation, target-domain adaptation, and dependence between development and evaluation data rather than from outcome ascertainment (Fig. 2). Study-level judgments are shown in Supplementary Figure S1 and detailed in Supplementary Table S3.

Primary within-model transportability meta-analysis

The primary synthesis included 18 matched AUC/AUROC effects from five peer-reviewed publication clusters (Han 2022, Almhdie-Imjabbar 2022, Salis 2024, Wang 2025, and Nair 2024) (13–17). Sixteen of the 18 effects were negative, one was zero, and one was positive; all five publication-level mean changes were negative. Individual matched changes ranged from − 0.110 to + 0.010 (Fig. 3; Supplementary Table S2).

The pooled external-minus-internal change in discrimination was − 0.044, with a publication-cluster robust 95% CI of -0.064 to -0.024 (Fig. 4). Effect-level heterogeneity was modest to moderate (I^2 = 31.3%). The effect-level REML/Hartung-Knapp sensitivity analysis produced the same point estimate (-0.044; 95% CI, -0.062 to -0.027). Point estimates were therefore similar across the two analytic approaches, but uncertainty remains substantial because only five publication clusters contributed to the synthesis and several publications reused the same benchmark cohorts.

Sensitivity and secondary analyses

Equal weighting of the five peer-reviewed publication clusters yielded a mean ΔAUC of approximately − 0.052. In leave-one-publication-out analyses, the mean remained negative regardless of which publication was omitted, ranging from approximately − 0.062 to -0.041 (Supplementary Figure S3). At the publication level, four clusters used a held-out source-domain evaluation for the internal comparator (mean ΔAUC approximately − 0.050), whereas one cluster used repeated cross-validation (mean ΔAUC approximately − 0.062); this comparison was descriptive only. Adding the independently evaluated Blanco et al. preprint (21) increased the analysis to 19 effects from six publication clusters and yielded a cluster-robust estimate of -0.050 (95% CI, -0.074 to -0.026). These sensitivity analyses are summarized in Supplementary Table S4.

Domain-level patterns were treated descriptively because external datasets were repeatedly reused. Across the four peer-reviewed primary publication clusters that transported models between OAI and MOST, the mean change was approximately − 0.040. The independent OAI-to-Taiwan hospital evaluation showed a change of -0.099, whereas the OAI-to-PROCOAC preprint showed − 0.117. Dunn et al. provided additional independent replication evidence, with an average decrease of approximately 0.055 but nonstandard uncertainty reporting (18). World COACH leave-one-cohort-out studies showed substantial between-cohort variability (19, 20) (Fig. 5). No formal pooled confidence interval was calculated at the transport-domain level because the number of genuinely independent domains was insufficient.

Surgical endpoints were analyzed separately. In the RiskFORM analysis (32), the OAI-to-MOST change in AUROC was approximately − 0.100 at 1 year, -0.060 at 2 years, and − 0.050 at 4 years (Supplementary Figure S4; Supplementary Table S4). C-index-based time-to-event studies also generally indicated lower external discrimination (27, 33) but were not pooled with AUC/AUROC effects. The exploratory funnel plot (Supplementary Figure S2) was presented without formal small-study-effect testing because of the dependence among effects and the small number of publication clusters. The publication-level bubble plot (Supplementary Figure S5) likewise remained descriptive and showed little relationship between internal discrimination and transportability change (R^2 approximately 0.02).

Discussion

Principal findings

Across 18 matched AUC/AUROC contrasts from five peer-reviewed publication clusters, external evaluation was associated with a mean reduction of approximately 0.044 AUC units. Similar point estimates were obtained in the effect-level and publication-level sensitivity analyses. The direction of the summary remained negative when each primary publication was omitted in turn and when a recent preprint was added, although the small number of publication clusters limits the precision of these comparisons.

The evidence base was highly concentrated: OAI appeared in 24 reports and MOST in 15, and four of the five peer-reviewed publication clusters contributing to the primary synthesis involved transport between OAI and MOST. Repeated use of these cohorts allows reproducibility to be examined across modelling approaches, but it does not provide the same evidence as replication across independent health systems, geographic regions, acquisition protocols, or clinical populations. Publication-level clustering accounts for multiple estimates reported within the same article, but it cannot fully account for participant and domain overlap when OAI or MOST is reused across different publications. The

effective amount of independent transportability evidence is therefore smaller than the number of effects or publications suggests, and the pooled estimate should be read as a summary of the available published contrasts rather than as a precise estimate of performance loss across independent clinical settings.

Relation to previous evidence

Recent OA reviews have addressed a related but different question. Liu et al. pooled absolute C-index and diagnostic performance across 32 knee OA studies and reported substantial uncertainty across future settings (2). Gillani et al. focused on deep-learning imaging models and described a decline from a median internal AUC of 0.81 to an external median of 0.75; only 27% of the included studies performed external validation, and 97% used OAI for training (3). Those reviews describe the broader performance landscape and the concentration of datasets. Our analysis instead estimates the matched external-minus-internal change for the same model, which reduces heterogeneity attributable to comparisons across unrelated model architectures, predictor modalities, and outcome definitions.

Yu et al. reported a median external-minus-development performance difference of -0.046 across radiologic deep-learning algorithms, with 81% showing some reduction on external data (1). This numerical similarity should be regarded as descriptive rather than confirmatory because the clinical tasks, validation designs, model classes, performance measures, and underlying populations differ substantially between the two evidence bases.

Recent AI reviews have also handled study structure more explicitly. A 2026 npj Digital Medicine review of AI for glaucoma progression grouped 46 reports into 43 unique study cohorts and identified generalizability as a major translational limitation (4). A 2026 npj Digital Medicine meta-analysis of AI for imaging-based prediction of lung-cancer metastasis screened explicitly for cohort overlap and reported detailed study and model characteristics in supplementary tables (5). Accordingly, report count, model count, and independent cohort or domain count were treated as distinct descriptors of the evidence base.

Interpretation of the within-model change estimand

The matched external-minus-internal contrast was selected to reduce between-model heterogeneity in architecture, predictor modality, and outcome definition. Because each contrast refers to the same fitted modelling approach, it provides a more direct description of observed performance change after transport than comparisons of unrelated internal and external studies.

ΔAUC should not be interpreted as a causal or model-intrinsic measure of degradation. It may reflect differences in case mix, disease and outcome spectrum, prevalence, predictor measurement, imaging protocol, and outcome ascertainment. It can also reflect how internal performance was estimated, because apparent, cross-validated, and held-out source-domain estimates carry different degrees of optimism. A lower external AUC may therefore arise even when a model is well calibrated, whereas a higher external AUC can occur when the target cohort has a wider risk spectrum. Future external evaluations should report calibration, calibration-in-the-large, calibration slope, decision-curve net benefit, and subgroup performance alongside discrimination.

Validation designs and degree of domain separation

The included studies differed substantially in the degree of separation between development and evaluation data. Independent external evaluation in a distinct cohort (L1), leave-one-cohort-out or other internal-external evaluation (L2), a held-out site within the same overarching resource (L3), and evaluation after target-domain fine-tuning represent different transport challenges. The L1/L2/L3 labels are operational categories used in this review rather than an established universal hierarchy. Adapted models address whether performance can be recovered after exposure to target-domain data; they do not estimate the performance of a frozen model before adaptation.

Herrera et al. showed that cross-cohort transfer learning between OAI and MOST can improve performance in the target cohort (24). Such approaches may be useful for deployment, but post-adaptation AUC is not an unbiased estimate of pre-adaptation generalizability. External-evaluation reports should therefore present frozen-model performance before adaptation, describe the target-domain information used for adaptation, and report the incremental change after adaptation.

Implications for model evaluation

The limited number of independent transport domains indicates that further external evaluation in new clinical populations would add more information about generalizability than additional analyses confined to the same benchmark cohorts. OAI-to-MOST validation remains useful as a common reference, but evidence from additional health systems, countries, community cohorts, and imaging protocols is needed. Multi-cohort resources such as World COACH show how leave-one-cohort-out evaluation can reveal heterogeneity that is not visible in a single pooled test set (19, 20).

Prediction endpoints also need to match intended clinical use. Structural progression, pain worsening, combined end-stage OA, and arthroplasty are not interchangeable. Arthroplasty is clinically important but is influenced by access to care, patient preference, surgical thresholds, and competing morbidity; surgical endpoints were therefore analyzed separately. Uncertainty should reflect patient sampling rather than only

variability across training seeds or cross-validation folds. Reports should also identify development, tuning, internal evaluation, external evaluation, and adaptation datasets explicitly, consistent with TRIPOD + AI terminology (7).

Calibration was too inconsistently reported for meaningful synthesis in the current literature, despite being necessary when a model is intended to provide individual risk estimates. Future studies should report Brier scores, calibration plots and slopes, precision-recall performance for imbalanced outcomes, prespecified decision thresholds, decision-curve analyses, and clinically relevant subgroup performance where appropriate. PROBAST + AI provides a framework for identifying design and analysis weaknesses before model evaluation is undertaken (8).

Strengths and limitations

This review has several strengths. The search covered four major databases through early September 2026 and deliberately prioritized sensitivity. We separated inclusion in the systematic review from eligibility for quantitative synthesis, allowing the evidence map to capture independent external evaluation, quasi-external designs, internal-only evidence, and adaptation while restricting the primary pooled estimate to comparable paired AUC/AUROC effects with defensible uncertainty. We used a matched within-model estimand rather than combining absolute performance across unrelated models. We also distinguished publication, model family, and dataset domain, and used publication-cluster robust inference to acknowledge multiple effects contributed by the same report. Risk of bias and applicability were assessed with the updated PROBAST + AI framework and finalized after independent dual-reviewer assessment.

Several limitations warrant emphasis. First, only five peer-reviewed publication clusters contributed to the primary synthesis. Robust sandwich variance estimators are asymptotic in the number of independent clusters, and simulation work suggests that substantially larger numbers are preferable (11). We therefore regard the publication-cluster robust interval as dependence-aware but exploratory and place similar weight on the publication-level and Hartung-Knapp sensitivity analyses. Publication-level clustering also cannot fully account for participant-level dependence created when OAI or MOST is reused across different publications; the effective amount of independent information is likely smaller than the number of publication clusters suggests. Second, the internal component of the matched contrast was not estimated using an identical strategy across studies. Differences between repeated cross-validation and held-out source-domain evaluation may contribute to variation in ΔAUC independently of dataset transport. Third, covariance between internal and external AUC estimates was generally unavailable. For independent cohorts, we used the sum of the component variances; where uncertainty had to be reconstructed, additional assumptions were required. Fourth, AUC was the only metric reported consistently enough for the primary matched synthesis. Calibration and clinical utility could not be pooled, and C-index results were kept separate.

The broad evidence map also included a small, explicitly labelled contextual stratum, including one non-prognostic multicentre external AI evaluation. Contextual reports were excluded from the primary prognosis meta-analysis but increase heterogeneity in the qualitative evidence base. Selective reporting remains possible: models that fail external evaluation may be less likely to be published, and studies may preferentially report the best-performing model or prediction horizon. The exploratory funnel plot cannot resolve this concern because only five publication clusters contributed to the primary analysis. Finally, ΔAUC reflects both model transport and differences between source and target evaluation settings. The pooled estimate is therefore an empirical summary of observed performance changes across the available domains, not a universal penalty to be applied to future models.

Conclusion

Across the available matched comparisons, discrimination was lower in external than in internal evaluation by an average of approximately 0.044 AUC/AUROC units. This estimate was derived from a small and highly concentrated evidence base with substantial reuse of OAI and MOST and should be interpreted as a summary of observed performance changes in the available transport domains rather than as a general transportability penalty. Future external evaluations should test frozen models in genuinely independent populations and report calibration and clinical utility in addition to discrimination. Pre-adaptation and post-adaptation performance should also be reported separately as transfer-learning approaches become more common.

Declarations

Ethics approval and consent to participate

Clinical trial number: not applicable. This systematic review and meta-analysis used previously published aggregate data and involved no new recruitment of human participants or animals.

Consent for publication

Not applicable.

Competing interests

The authors declare no commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

Guangdong Basic and Applied Basic Research Fund Guangzhou United Fund (2021A1515110698); Guangzhou Science and Technology Planning Project (2025A04J3965); National Natural Cultivation Project of the Third Affiliated Hospital of Guangzhou University of Chinese Medicine (Sy2026026); State Key Laboratory of Traditional Chinese Medicine Syndrome Postgraduate Program (SKLKY2024A0003).

Author Contribution

Q.N. contributed to data curation, formal analysis, and writing the original draft. J.Y. contributed to methodology, data curation, and reviewing and editing the manuscript. X.L. contributed to investigation and writing the original draft. F.Y. contributed to supervision, validation, and reviewing and editing the manuscript. W.H. contributed to conceptualization, formal analysis, and reviewing and editing the manuscript. X.Y. contributed to conceptualization and reviewing and editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

All aggregate data underlying the quantitative analyses are provided in the main table and accompanying supplementary material. Additional analytic details are available from the corresponding authors upon reasonable request.

References

1. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiol Artif Intell. 2022;4(3):e210064.
2. Liu Y, Xiao G, Zhang Y. Predictive Value of Machine Learning in Knee Osteoarthritis Progression: Systematic Review and Meta-Analysis. J Med Internet Res. 2025;27:e80430.
3. Gillani A, von Eisenhart-Rothe R, Woertler K, Hirschmann MT, Rueckert D, Hinterwimmer F. Promises and limitations of deep learning for predicting knee osteoarthritis progression from medical imaging: a systematic review. Knee Surg Sports Traumatol Arthrosc; 2026.
4. Liang YG, Fan L, Teixeira-Pinto A. A systematic review of AI for predicting glaucoma progression: challenges and recommendations towards clinical implementation. npj Digit Med. 2026;9:140.
5. Sun Y, Jin L, Wang B. Systematic review and meta-analysis of AI in lung cancer metastasis imaging for diagnosis and prognosis. npj Digit Med. 2026;9:688.
6. Page MJ, McKenzie JE, Bossuyt PM. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
7. Collins GS, Moons KGM, Dhiman P. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
8. Moons KGM, Damen JAA, Kaul T. 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.
9. Debray TPA, Damen J, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.
10. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
11. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods. 2010;1(1):39–65.
12. Rover C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol. 2015;15:99.

13. Han T, Kather JN, Pedersoli F. Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. Nat Mach Intell. 2022;4:1029–39.
14. Almhdie-Imjabbar A, Nguyen KL, Toumi H, Jennane R, Lespessailles E. Prediction of knee osteoarthritis progression using radiological descriptors obtained from bone texture analysis and Siamese neural networks: data from OAI and MOST cohorts. Arthritis Res Ther. 2022;24:66.
15. Salis Z, Driban JB, McAlindon TE. Predicting the onset of end-stage knee osteoarthritis over two- and five-years using machine learning. Semin Arthritis Rheum. 2024;66:152433.
16. Wang CT, Chang KT, Lai F, Pao JL, Lin SM, Chang CH. Simplifying Knee OA Prognosis: A Deep Learning Approach Using Radiographs and Minimal Clinical Inputs. Diagnostics (Basel). 2025;15(19):2543.
17. Nair A, Alagha MA, Cobb J, Jones G. Assessing the Value of Imaging Data in Machine Learning Models to Predict Patient-Reported Outcome Measures in Knee Osteoarthritis Patients. Bioeng (Basel). 2024;11(8):824.
18. Dunn CM, Sturdy C, Velasco C. Peripheral Blood DNA Methylation-Based Machine Learning Models for Prediction of Knee Osteoarthritis Progression: Biologic Specimens and Data From the Osteoarthritis Initiative and Johnston County Osteoarthritis Project. Arthritis Rheumatol. 2023;75:28–40.
19. van den Berg MA, Boel F, van Buuren MMA. Arthritis Care Res (Hoboken). 2026;78:489–500. Hip Morphology-Based Osteoarthritis Risk Prediction Models: Development and External Validation Using Individual Participant Data From the World COACH Consortium.
20. van den Berg MA, van Tulder G, Ahedi H. Integrating deep learning-based radiographic features with demographic variables for hip osteoarthritis prediction: development and validation using individual participant data from the World COACH Consortium. Osteoarthr Imaging. 2026.
21. Blanco FJ, Martinez-Sotodosos L, Oreiro N. Risk stratification for the rapid pain progression phenotype in knee osteoarthritis using interpretable multimodal machine learning: development in the Osteoarthritis Initiative and external evaluation in the Prospective Cohort of Osteoarthritis from A Coruna. 2026.
22. McCabe PG, Lisboa P, Baltzopoulos B, Olier I. Externally validated models for first diagnosis and risk of progression of knee osteoarthritis. PLoS ONE. 2022;17(7):e0270652.
23. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. Sci Rep. 2019;9:20038.
24. Herrera D, Almhdie-Imjabbar A, Ibrahim N. Deep transfer learning-based approaches with multimodal fusion to improve the prediction of knee osteoarthritis progression: data from the OAI and MOST cohorts. Sci Rep. 2026;16:29715.
25. Hidaka R, Matsuda K, Igari T. Development and accuracy of an artificial intelligence model for predicting the progression of hip osteoarthritis using plain radiographs and clinical data: a retrospective study. BMC Musculoskelet Disord. 2024;25:893.
26. Rajamohan HR, Wang T, Leung K. Prediction of total knee replacement using deep learning analysis of knee MRI. Sci Rep. 2023;13:6922.
27. Cigdem O, Chen S, Zhang C, Cho K, Kijowski R, Deniz CM. Estimating time-to-total knee replacement on radiographs and MRI: a multimodal approach using self-supervised deep learning. Radiol Adv. 2024;1(4):umae030.
28. Zhang C, Chen S, Cigdem O, Rajamohan HR, Cho K, Kijowski R, et al. MR-Transformer: A Vision Transformer-based Deep Learning Model for Total Knee Replacement Prediction Using MRI. Radiol Artif Intell. 2025;7(5):e240373.
29. Castagno S, Birch M, van der Schaar M, McCaskie A. Predicting rapid progression in knee osteoarthritis: a novel and interpretable automated machine learning approach, with specific focus on young patients and early disease. Ann Rheum Dis. 2025;84(1):124–35.
30. Rajamohan HR, Cho K, Kijowski R, Deniz CM, editors. Deep learning for prediction of knee osteoarthritis structural progression and incidence2024.
31. Rajamohan HR, Xu Y, Zhu W, Kijowski R, Cho K, Geras KJ, et al. Robust disease prognosis via diagnostic knowledge preservation: a sequential learning approach. PLoS ONE. 2026;21(5):e0344600.
32. Rajamohan HR, Kijowski R, Cho K, Deniz CM. A Progressive Risk Formulation for Enhanced Deep Learning based Total Knee Replacement Prediction in Knee Osteoarthritis. 2025.
33. Cigdem O, Hedayati E, Rajamohan HR, Cho K, Chang G, Kijowski R, et al. Estimation of time-to-total knee replacement surgery with multimodal modeling and artificial intelligence. Comput Biol Med. 2025;193:110364.
34. Paz-Gonzalez R, Balboa-Barreiro V, Lourido L, Calamia V, Fernandez-Puente P, Oreiro N, et al. Prognostic model to predict the incidence of radiographic knee osteoarthritis. Ann Rheum Dis. 2024;83(5):661–8.
35. Lee DW, Han HS, Ro DH, Lee YS. Development of the machine learning model that is highly validated and easily applicable to predict radiographic knee osteoarthritis progression. J Orthop Res. 2025;43(1):128–38.
36. Yin R, Chen H, Tao T, Zhang K, Yang G, Shi F, et al. Expanding from unilateral to bilateral: A robust deep learning-based approach for predicting radiographic osteoarthritis progression. Osteoarthritis Cartilage. 2024;32(3):338–47.

37. Panfilov E, Saarakkala S, Nieminen MT, Tiulpin A. Predicting Knee Osteoarthritis Progression from Structural MRI using Deep Learning. 2022.
38. Attur M, Krasnokutsky S, Zhou H, Samuels J, Chang G, Bencardino J, et al. The combination of an inflammatory peripheral blood gene expression and imaging biomarkers enhance prediction of radiographic progression in knee osteoarthritis. Arthritis Res Ther. 2020;22:208.
39. Widera P, Welsing PMJ, Danso SO. Development and validation of a machine learning-supported strategy of patient selection for osteoarthritis clinical trials: the IMI-APPROACH study. Osteoarthr Cartil Open. 2023;5(4):100406.
40. Xu S, Guo C, Zuo J. Development and validation of a deep learning model for automatic severity grading of hip osteoarthritis: a multi-center study. Ann Med. 2025;57(1):2584361.

Figures

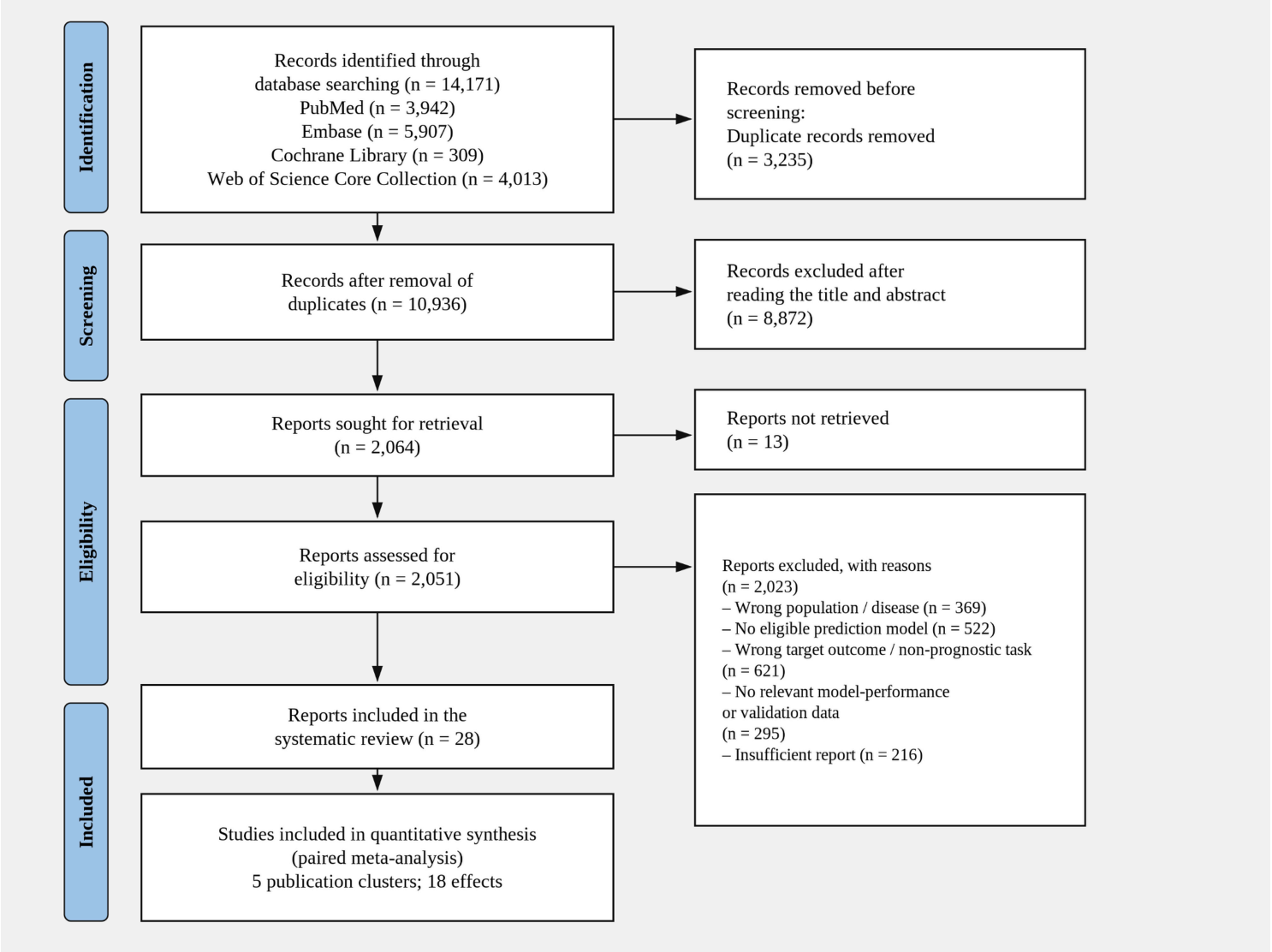


Figure 1

PRISMA 2020 flow diagram of study selection. Searches of PubMed, Embase, Web of Science Core Collection, and the Cochrane Library/CENTRAL identified 14,171 records. After duplicate removal and screening, 2,051 full-text reports were assessed for eligibility, 2,023 were excluded, and 28 reports were included in the systematic review. Of these, 27 contributed prognostic evidence and one cross-sectional report was retained solely as contextual non-prognostic external-validation evidence. The primary quantitative synthesis comprised 18 matched effects from five publication clusters.

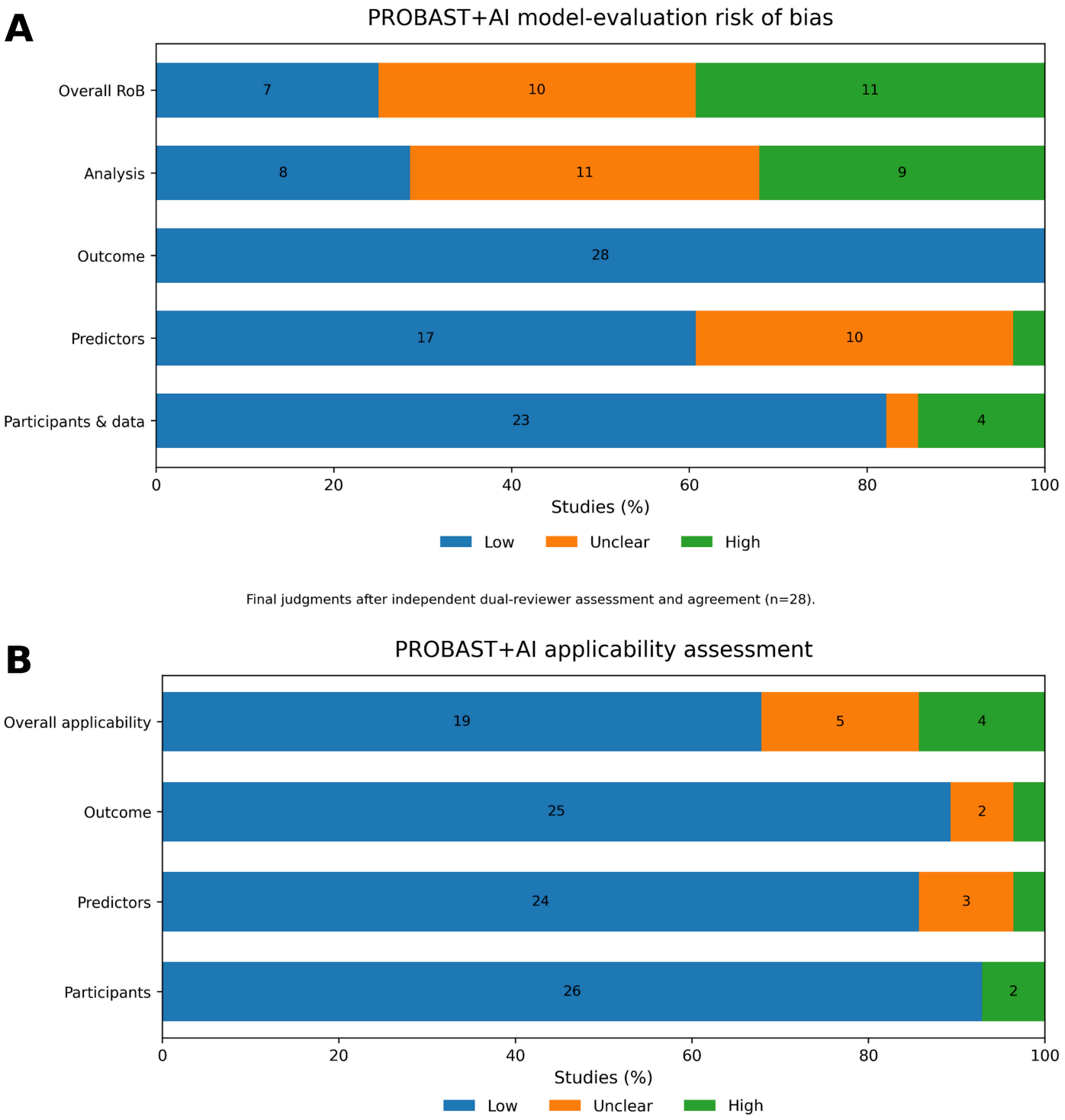


Figure 2

PROBAST+AI assessment of model-evaluation risk of bias and applicability across the 28 included reports. (A) Distribution of risk-of-bias judgments for participants/data sources, predictors, outcome, analysis, and the overall model-evaluation judgment. (B) Distribution of applicability concerns for participants, predictors, outcome, and the overall applicability judgment. Bar segments represent the percentage of reports in each category, with counts shown within segments. All judgments shown are the final consensus ratings after independent dual-reviewer assessment.

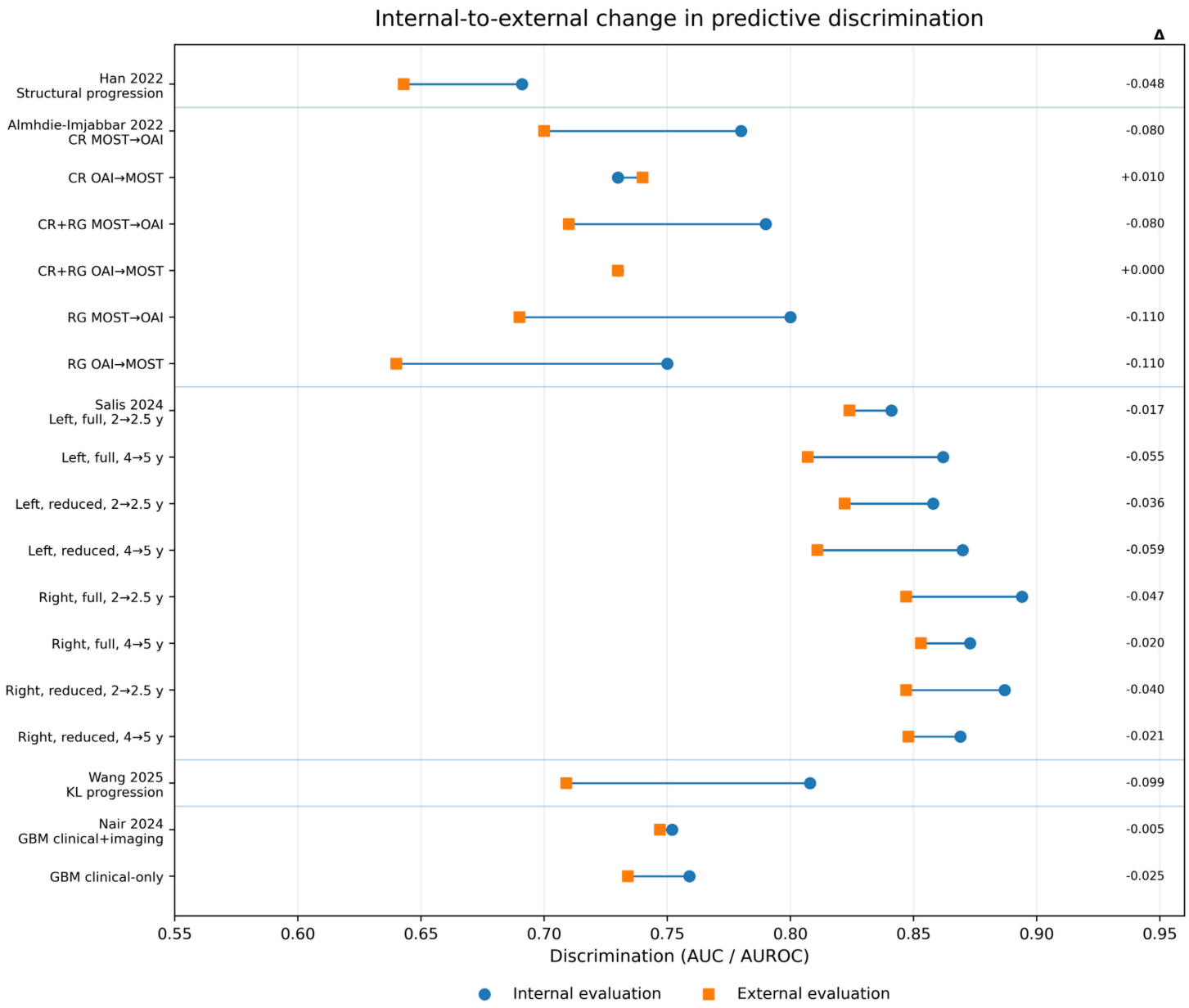


Figure 3

Paired internal-to-external discrimination for the 18 primary AUC/AUROC effects. For each effect, the internal and external discrimination estimates from the same model comparison are connected by a line; multiple eligible effects from the same publication are displayed separately. ΔAUC/AUROC is defined as external minus internal discrimination, so negative values indicate lower discrimination after transport.

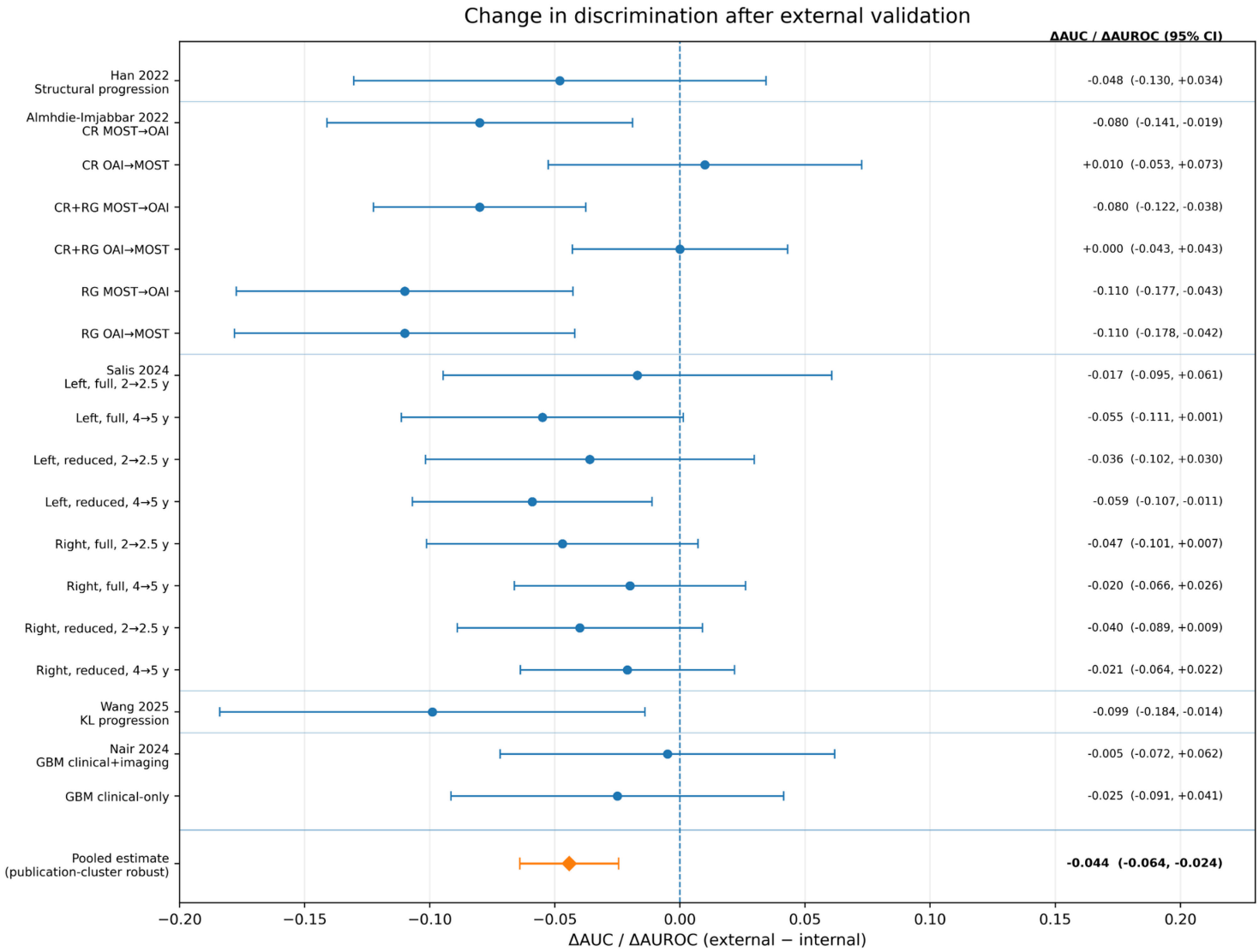


Figure 4

Forest plot of matched within-model changes in AUC/AUROC. Points and horizontal lines show effect-specific ΔAUC/AUROC estimates and 95% confidence intervals. The vertical reference line denotes no change (Δ = 0). Negative values indicate lower external discrimination. The pooled point estimate was obtained from a REML random-effects model; the primary pooled 95% confidence interval uses publication-cluster robust inference to account for multiple effects contributed by the same publication cluster.

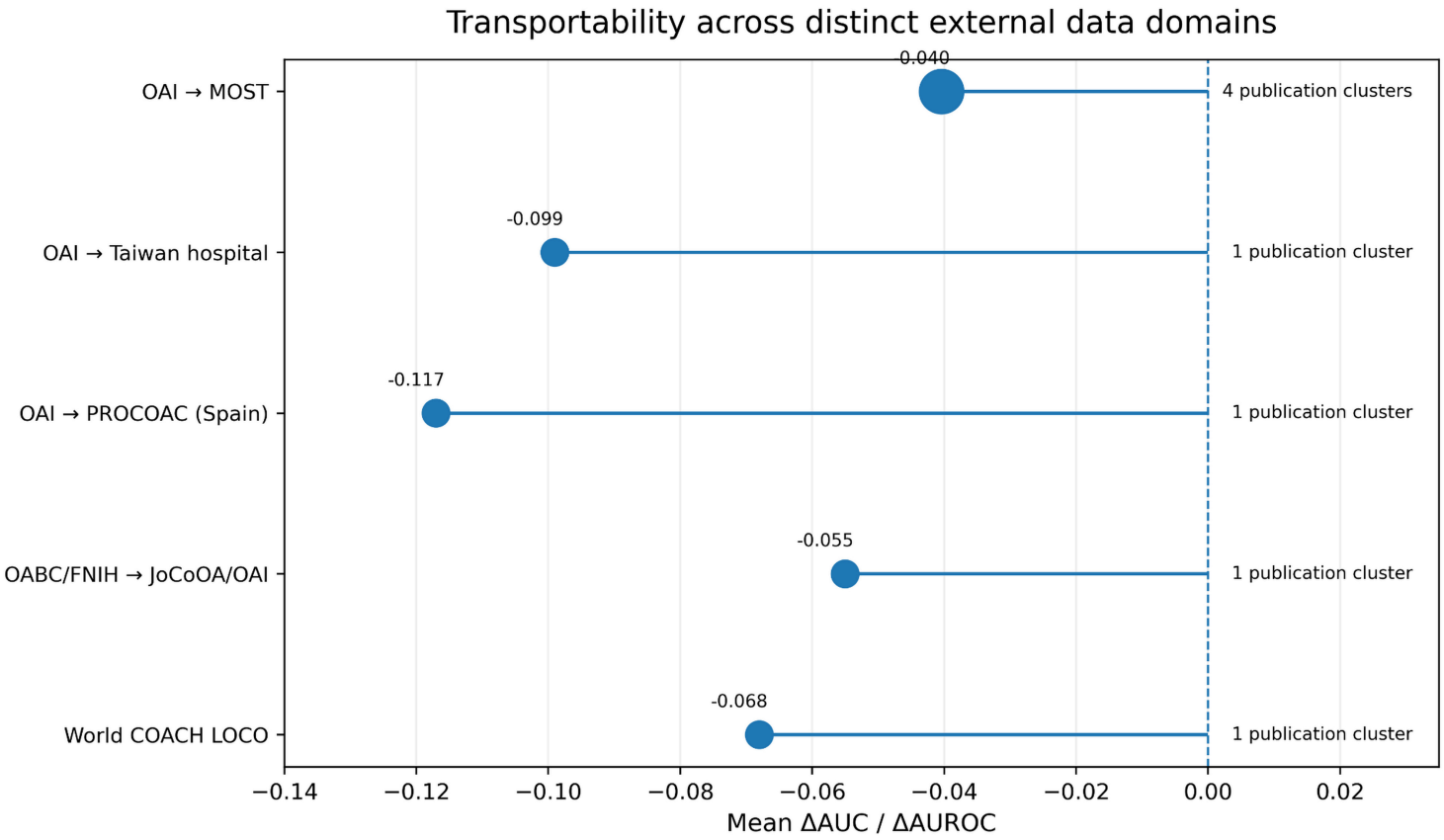


Figure 5

Descriptive landscape of transportability across distinct external dataset domains. Points represent mean external-minus-internal changes in AUC/AUROC for the indicated transport pathways; marker size reflects the number of contributing publication clusters. The OAI-to-MOST point summarizes four peer-reviewed primary publication clusters, whereas the other pathways are based on single publication clusters or descriptive cohort-level summaries. Because datasets are repeatedly reused across publications, these domain-level values should not be interpreted as independent formal meta-analytic estimates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- SupplementaryFiguresS1S5.docx
- SupplementaryTablesS1S4.xlsx
- SupplementaryTableS5.docx