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.

When high AUROC does not travel: internal-to-external performance differences in machine-learning models for antimicrobial resistance

Dripto Roy

connectwithdripto@gmail.com

University of Kalyani https://orcid.org/0009-0000-6443-2978

Research Article

Keywords: antimicrobial resistance, machine learning, artificial intelligence, external validation, transportability, AUROC, meta-research, reproducibility

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11124025/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Background: Machine-learning studies addressing antimicrobial resistance (AMR) often report high discrimination, but performance estimates may change when models are evaluated in new hospitals, countries, time periods, or data-generating environments. Recent AMR reviews have highlighted limited external validation and sparse calibration reporting, but the magnitude of internal-to-external performance differences has not been consistently synthesized.

Methods: We conducted a targeted meta-research evidence synthesis through 20 September 2026. Studies were eligible for the primary analysis when the same model and prediction target had a directly extractable AUROC/AUC from an internal or development evaluation and an independent external evaluation. The primary quantity was ΔAUROC = internal AUROC − external AUROC. Because multiple models and antibiotic targets can be reported within one article, we used the study as the primary analytical cluster and gave each study equal weight. Pair-level summaries and leave-one-study-out analyses were prespecified sensitivity analyses.

Results: Eight independent studies contributed 47 paired observations. Across individual pairs, the mean ΔAUROC was 0.0885 (median 0.0520; SD 0.1515), with lower external discrimination in 36/47 pairs (76.6%). When each study contributed one equal-weighted mean, the mean ΔAUROC was 0.0344 (SD 0.0776; exploratory 95% t-based CI −0.0305 to 0.0993). A 20,000-resample bootstrap over study means gave a similar interval (−0.0137 to 0.0859). Removing the largest contributing study reduced the study-cluster mean to 0.0180.

Conclusions: Internal-to-external performance differences are common but heterogeneous in the current AMR machine-learning evidence base. The apparent magnitude of the gap is highly sensitive to the unit of analysis, and the available evidence does not support a universal correction factor for internally reported AUROCs. Future AMR-AI studies should report independent validation, calibration, data-partitioning safeguards, sample counts, and paired internal/external performance explicitly.

Author Summary

I wanted to know whether the high predictive performance reported in studies using machine learning for antimicrobial resistance remains similar when the models are tested on genuinely new data. I assembled a targeted evidence set of published and public manuscript studies that reported directly comparable internal and external area-under-the-curve results. I then treated each paper as the main unit of evidence, because one paper can contribute many model and antibiotic comparisons that share the same patients and data-processing pipeline. Across 47 paired observations from eight studies, external performance was lower in most individual comparisons, but the average difference became much smaller when each study received equal weight. This shows why simply averaging every model result can exaggerate the apparent problem. I also found that calibration, temporal validation, and prospective evaluation remain much less commonly reported than discrimination. My findings support more transparent reporting and

stronger independent validation, but they do not justify a universal numerical correction to published antimicrobial-resistance models.

1. Introduction

Antimicrobial resistance (AMR) is a major global health problem, and surveillance systems increasingly combine microbiological, clinical, genomic, and laboratory data to characterize resistant infections and support treatment and infection-control decisions [1, 2]. Machine-learning (ML) methods have become common in this space because they can represent nonlinear associations and high-dimensional feature sets. Published studies have reported promising discrimination across electronic health records, whole-genome sequencing, matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry, and other data modalities [3].

However, predictive discrimination is conditional on the data-generating environment. A model can perform well within the population used for development and degrade when transported to another hospital, country, laboratory workflow, time period, or genomic background. In AMR prediction, such shifts may arise from changes in resistance prevalence, pathogen population structure, patient selection, antibiotic exposure, laboratory procedures, coding practices, specimen mix, or acquisition technology. These differences are not equivalent to statistical overfitting, and a fall in external AUROC should not automatically be interpreted as evidence that the development estimate was biased. In this article, “AMR-AI” denotes studies applying artificial intelligence or machine-learning methods to antimicrobial-resistance-related prediction.

Contemporary reporting standards for clinical prediction models emphasize transparent data partitioning, independent validation, calibration, and reproducibility [4]. Risk-of-bias frameworks similarly emphasize the consequences of inappropriate participant selection, predictors measured after the target is known, leakage across development and validation partitions, and inadequate validation design [5]. AMR-specific reviews have documented substantial heterogeneity and limited external validation. A 2026 scoping review of ML for AMR surveillance reported external, cross-site, or temporal validation in 15 of 41 core empirical studies, calibration or reliability metrics in only 5 of 41, and prospective evaluation in 1 of 41 [17]. A concurrent 2026 preprint independently appraised AMR-AI studies using reporting and risk-of-bias frameworks [18]. These studies establish the importance of validation quality, but they do not replace a direct synthesis of paired internal-versus-external performance differences.

The present study therefore addresses a narrower meta-research question: when an AMR-AI study reports both an internal or development performance estimate and an independent external estimate for the same model and target, how large is the difference, how heterogeneous is it, and how much does the estimate change when multiple results from the same article are treated as dependent? We use the descriptive quantity ΔAUROC = internal AUROC − external AUROC. We use the phrase “performance optimism” only as an informal description of the direction of this gap; ΔAUROC is not the classical bootstrap or resampling-based optimism statistic.

2. Results

Evidence base

The targeted search and verification workflow produced a working inventory of 59 candidate study records. Twenty-three studies received structured study-level extraction, and eight independent studies met the strict criteria for the locked primary paired analysis, yielding 47 directly comparable internal/external AUROC observations (Fig. 1). The primary set included electronic health-record, clinical/laboratory, and whole-genome sequencing modalities and spanned MRSA, multidrug-resistant organisms, Escherichia coli, Pseudomonas aeruginosa, and broader multidrug-resistance outcomes. The study-level composition is summarized in Table 1.

Table 1
Characteristics of the eight study clusters contributing to the primary paired analysis. Each study contributes multiple model/target pairs in some cases; the mean ΔAUROC is calculated within study before equal-weight study-level synthesis.

Study / citation	Modality	Target	External validation contrast	Pairs | mean ΔAUROC
S005 | Nigo et al. [6]	EHR time-series	MRSA-positive culture	Development/internal evaluation -> MIMIC-IV	3 | 0.049
S006 | Li et al. [7]	EHR	MDRO colonization/infection	PLAGH-ICU <-> MIMIC-IV	2 | 0.139
S007 | Nsubuga et al. [8]	Whole-genome sequencing	E. coli resistance to 3 antibiotics	England -> African cohorts	24 | 0.149
S008 | Yu et al. [9]	Clinical/laboratory	Carbapenem-resistant P. aeruginosa	Primary cohort -> independent cohort	1 | -0.061
S009 | Multi-task AMR [10]	EHR/clinical	Early AMR prediction	Korean institutions -> Ansan	9 | -0.000
S010 | CRO AutoML [11]	Clinical/EHR	Carbapenem-resistant organism infection	Development/testing -> independent cohort	6 | 0.044
S023 | MDR AECOPD [12]	EHR	MDR infection in AECOPD	MIMIC-IV -> external center	1 | 0.003
S061 | MDR pneumonia [14]	Clinical/EHR + laboratory	MDR pneumonia	MIMIC-IV -> China	1 | -0.047

Primary quantitative findings

Across all 47 paired observations, the pair-level mean ΔAUROC was 0.0885, with a median of 0.0520 and a sample standard deviation of 0.1515. The mean internal and external AUROCs were 0.7533 and

0.6647, respectively. External discrimination was lower in 36 of 47 pairs (76.6%) and higher in 11 pairs (23.4%). The full pair-level synthesis is summarized in Table 2.

Table 2
Primary quantitative synthesis. The study-cluster estimate is the primary descriptive quantity because individual model/antibiotic comparisons within a paper are not independent.

Analysis	Estimate	Method/definition	Interpretation
Pair-level observations	47	Direct matched internal/external comparisons	Sensitivity analysis; rows are clustered within studies
Independent study clusters	8	Number of parent studies	Primary analytical unit
Pair-level mean ΔAUROC	0.0885	Mean across 47 pairs	Sensitivity analysis
Pair-level median ΔAUROC	0.0520	Median across 47 pairs	Sensitivity analysis
Pair-level SD	0.1515	Sample SD across pairs	Sensitivity analysis
Positive ΔAUROC	36/47 (76.6%)	Internal > external	Direction of majority of pair-level comparisons
Negative ΔAUROC	11/47 (23.4%)	External > internal	Counterexamples retained
Study-cluster mean ΔAUROC	0.0344	Equal-weight mean of study-specific means	Primary descriptive estimate
Study-cluster median ΔAUROC	0.0234	Median of eight study-specific means	Descriptive
Study-cluster SD	0.0776	Sample SD across studies	Between-study heterogeneity
Exploratory 95% t-based CI	-0.0305 to 0.0993	df = 7	Crosses zero
Bootstrap 95% percentile CI	-0.0137 to 0.0859	20,000 resamples of eight study means	Sensitivity analysis; small-cluster caution
Excluding S007	0.0180	Equal-weight study mean after removing the 24-row WGS cluster	Sensitivity analysis
Leave-one-study-out range	0.0180 to 0.0481	Range of equal-weight means after omitting one study	No single exclusion reverses the sign

Dependence and sensitivity analyses

The difference between pair-level and study-level summaries was substantial. The pair-level mean of 0.0885 gives greater influence to studies reporting many model/antibiotic combinations. Nsubuga et al. contributed 24 of 47 pairs and had a study-level mean ΔAUROC of 0.1492, whereas several other studies

contributed only one or two pairs. Equal weighting across the eight studies reduced the mean to 0.0344. Removing the Nsubuga cluster reduced it further to 0.0180. A leave-one-study-out sensitivity analysis produced cluster means between 0.0180 and 0.0481. These analyses show that the numerical estimate is sensitive to the unit of evidence rather than being a stable property of “AMR-AI” as a whole.

The study-level pattern was also heterogeneous in direction. Nigo et al. reported smaller external AUROC values for all three extracted MRSA models, while Yu et al. and the US/China multidrug-resistant-pneumonia study reported higher external than internal AUROC for the extracted model. The existence of negative ΔAUROC values was retained in the primary data rather than interpreted as extraction error.

Reporting and validation characteristics

Among the 23 studies receiving structured study-level extraction in the targeted candidate set, 20 (87.0%) reported some form of external validation, six (26.1%) explicitly reported temporal validation, four (17.4%) reported a calibration or reliability metric, and one (4.3%) reported prospective evaluation (Table 3). These counts describe an enriched candidate set deliberately selected for transportability or validation relevance and therefore should not be interpreted as prevalence estimates for the full AMR-AI literature.

Table 3
Selected reporting characteristics in the 23-study structured candidate set. The candidate set was intentionally enriched for studies discussing external, temporal, cross-site, or related generalization questions.

Feature	Count	Interpretation
External validation reported	20/23 (87.0%)	Targeted candidate set; denominator = 23
Explicit temporal validation	6/23 (26.1%)	Targeted candidate set; denominator = 23
Calibration/reliability metric	4/23 (17.4%)	Targeted candidate set; denominator = 23
Prospective evaluation	1/23 (4.3%)	Targeted candidate set; denominator = 23

3. Discussion

Principal findings

This meta-research study identified a recurring but heterogeneous internal-to-external performance difference in AMR-AI. Across 47 paired observations, the simple row-level average was 0.089 AUROC. However, the more defensible study-cluster estimate was only 0.034, with an exploratory 95% t-based interval from − 0.030 to 0.099 and a bootstrap percentile interval from − 0.014 to 0.086. The fact that both intervals span zero, together with the modest number of independent studies, means that this analysis does not establish a universal field-wide performance penalty.

The most robust finding is methodological rather than numerical: the apparent magnitude of the internal-to-external gap changes materially when dependent model and antibiotic comparisons are aggregated within study. This is especially relevant to AMR because a single paper may report many outcomes, resistance phenotypes, model families, or antibiotics using the same patient cohort and preprocessing pipeline. Treating each row as independent can therefore make the evidence base appear much larger than it is.

Relationship to the current literature

The present analysis sits alongside, rather than replaces, recent AMR-AI reviews. The 2025 systematic review by Ardila et al. synthesized clinical AMR prediction studies and highlighted heterogeneity and limitations in validation and data processing [3]. The 2026 One Health scoping review likewise documented limited external or temporal validation, calibration, prospective evaluation, and operational deployment [17]. A September 2026 preprint provides an additional contemporaneous appraisal using reporting and risk-of-bias frameworks [18]. Our contribution is narrower: we focus on studies that make a direct internal-versus-external AUROC comparison and quantify how the estimate changes when the article, rather than every model/antibiotic result, is treated as the main unit of evidence.

The individual studies also illustrate why a universal adjustment is unlikely to work. Nigo et al. showed modest reductions when an electronic-health-record model was transported to MIMIC-IV [6], whereas Li et al. reported larger bidirectional cross-dataset reductions for multidrug-resistant-organism prediction [7]. Nsubuga et al. demonstrated much larger and outcome-specific differences in cross-country genomic prediction [8]. Conversely, other recent multicenter studies have found small differences or higher external than internal discrimination [9, 14, 15]. Related MALDI-TOF and genomic work further demonstrate that site, time, laboratory workflow, genomic population structure, and resistance prevalence can alter transportability [19–21].

Implications for AMR-AI study design

The findings support several practical reporting priorities. First, investigators should define the unit of splitting before model development and prevent preprocessing, feature selection, hyperparameter tuning, and other information from crossing the development-validation boundary. For genomic models, the relevant partition may need to account for lineage or closely related isolates rather than relying on an isolate-level random split. Second, independent external validation should be presented as a distinct evaluation environment, with sample counts and resistance prevalence reported for both development and external cohorts. Third, discrimination should be accompanied by calibration or another reliability assessment when the model output is intended to support clinical decisions [4, 5]. Fourth, when multiple antibiotics or model variants are reported, the dependence among comparisons should be made explicit.

The evidence also argues against using a single internally reported AUROC as a deployment-readiness signal. Recent AMR surveillance research increasingly emphasizes external validation, calibration, drift monitoring, and operational evaluation [17, 18]. These elements are complementary: discrimination

describes ranking ability, calibration describes the relationship between predicted and observed risk, and external validation tests transportability. None is a substitute for the others.

Implications for evidence synthesis

The study illustrates why quantitative synthesis of AMR-AI performance requires careful reconstruction of the evidence-generating hierarchy. AUROC is a common metric, but a pair of AUROCs from the same article may not be statistically independent from other pairs in that article. Moreover, many papers do not report the covariance or sufficient uncertainty information needed for a conventional inverse-variance meta-analysis of paired AUROCs. A robust future synthesis will need prespecified hierarchical units, explicit treatment of internal validation design, and standardized extraction of calibration, prevalence, sample size, and leakage safeguards.

A larger future database could support formal multilevel meta-regression examining whether external performance differences vary by validation design, data modality, pathogen, geographic separation, temporal separation, or methodological safeguards. The present dataset is intentionally too small to support that level of modeling responsibly: eight independent study clusters do not provide enough information for a stable multivariable model.

4. Materials and Methods

Study design and scope

We conducted a targeted evidence synthesis and meta-research analysis of AMR-AI prediction studies. The work was not prospectively registered and was not designed as a formal systematic review. The purpose was to identify primary studies that explicitly evaluated transportability, external validation, temporal validation, cross-site performance, cross-dataset performance, or closely related generalization questions and to extract directly comparable internal/external discrimination estimates.

The study was reported with attention to principles from TRIPOD + AI for transparent prediction-model reporting and PROBAST concepts for bias and applicability [4, 5]. Because the included modalities, study architectures, and validation designs were heterogeneous, we did not assign an aggregate risk-of-bias score and did not convert individual methodological features into an arbitrary quality scale.

Targeted evidence search

Searches were conducted through PubMed-indexed literature, PubMed Central, publisher repositories, and web-indexed scholarly sources, with the final targeted search cutoff set to 20 September 2026. Core searches combined terms for antimicrobial resistance with machine learning or artificial intelligence and validation, generalizability, external validation, temporal validation, cross-site, genomic, MALDI-TOF, and related concepts. Additional targeted searches were used when a candidate study appeared to contain a directly comparable internal and external performance table. The working search log and source URLs are included in the research workbench accompanying this manuscript.

Eligibility for the primary paired analysis

A study was eligible for the primary paired analysis when it met all of the following criteria: (1) the outcome represented a clinically or microbiologically meaningful AMR or resistance-prediction task; (2) a machine-learning or artificial-intelligence model was used; (3) the same model and target had an independently evaluated internal, held-out, or development-validation discrimination estimate and an external estimate; (4) AUROC/AUC was reported in both environments so that an exact internal-minus-external difference could be calculated; and (5) the external environment was genuinely independent in cohort, institution, country, dataset, or time, rather than merely being a random split from the same acquisition environment.

We excluded apparent training-set performance when no independently evaluated internal estimate was available, aggregate ranges that could not be paired to a specific model-target comparison, mismatched outcome definitions, and resistance-gene detection tasks without a comparable phenotype or clinically meaningful AMR outcome. Studies that failed the strict paired criteria but contained useful information on generalization, temporal drift, calibration, or validation were retained in a secondary evidence inventory rather than forced into the primary estimator.

Data extraction

Study-level extraction recorded publication year, data modality, pathogen or resistance target, development cohort, validation design, external setting, temporal separation, calibration/reliability reporting, prospective evaluation, sample scale, and potential leakage signals. For each primary pair, we recorded study identifier, model, target, internal evaluation type, internal setting, internal AUROC/AUC, external evaluation type, external setting, external AUROC/AUC, and source URL. All 47 primary pairs were retained as individual rows and assigned to a parent study identifier for clustering.

Primary estimand and dependence structure:

For paired observation i, the difference in AUROC was defined as:

$$\Delta\ AUROC_i = AUROC_{\text{internal},i} - AUROC_{\text{external},i}$$

where $AUROC_{\text{internal},i}$ is the AUROC reported for the model during internal or development validation for observation i, and $AUROC_{\text{external},i}$is the corresponding AUROC obtained during external validation. Positive values indicate lower external discrimination; negative values indicate higher external discrimination. No negative value was removed, transformed, or winsorized. We use “performance optimism” as informal shorthand for a positive internal-to-external gap, not as a claim that the internal estimate equals classical resampling optimism.

Because multiple rows can arise from the same article and share patients, preprocessing, hyperparameter optimization, or external datasets, the primary unit of evidence was the study. Each

study contributed its arithmetic mean of eligible paired ΔAUROC values, and these study-level means were then given equal weight. Pair-level summaries were treated as sensitivity analyses only.

Statistical analysis

We summarized the 47 paired observations using the mean, median, sample standard deviation, and counts of positive and negative ΔAUROC. For the primary study-cluster synthesis, we calculated the arithmetic mean and median of the eight study-level means and the sample standard deviation across studies. An exploratory two-sided 95% t-based interval used seven degrees of freedom. We also calculated a nonparametric percentile bootstrap interval using 20,000 resamples of the eight study-level means. Because the cluster count was small, both intervals are interpreted as descriptive uncertainty summaries rather than as evidence of a confirmatory meta-analytic effect.

Sensitivity analyses included exclusion of the Nsubuga et al. cluster, which supplied 24 of 47 observations, and leave-one-study-out recalculation. We did not perform a multivariable meta-regression or an inverse-variance random-effects meta-analysis because most publications did not supply the covariance or standard-error information required for a paired synthesis, and the number of independent studies was too small for a reliable multivariable model.

Reporting-characteristics analysis

For the 23 studies receiving structured study-level extraction, we tabulated whether external validation, explicit temporal validation, calibration/reliability metrics, and prospective evaluation were reported. These counts were descriptive and were not treated as population prevalence estimates because the targeted candidate set was enriched for validation and generalization questions.

Reproducibility and data provenance

All primary paired observations are provided as a machine-readable CSV and as the S1 supporting-information table. The complete extraction workbook records study identifiers, source URLs, validation characteristics, and secondary evidence. The analysis script reproduces the descriptive statistics, study-cluster sensitivity analyses, and manuscript figures from the locked paired dataset. No patient-level dataset was accessed or reconstructed for this study.

Declarations

5. Data availability

The study-level extracted data underlying the primary quantitative results are provided in the accompanying supporting-information workbook and CSV. Source publications are identified by URL in the extraction table. The analysis code is provided with the submission package. The study did not use patient-level data, and no patient-level information is included in the released dataset.

6. Ethics statement

This study used published aggregate study-level information and did not involve recruitment, intervention, or access to identifiable human-subject data. Institutional ethics approval was therefore not sought for this literature-based analysis.

7. Author contributions

D.R.: Conceptualization; methodology; investigation; data curation; formal analysis; software; visualization; writing—original draft; writing—review and editing.

8. Acknowledgments

The author thanks the investigators who made AMR-AI studies and supporting data available through open or publicly accessible sources. Generative artificial intelligence (AI)-assisted tools were used for drafting and document-preparation support; all scientific decisions, data extraction, calculations, source checking, and final manuscript content were reviewed by the author.

9. Funding

No dedicated external funding was received for this work.

10. Competing interests

The author declares no competing interests.

11. Supporting information captions

S1 Table. Directly paired internal/development and external AUROC extraction dataset. Each row represents a matched model-target comparison used in the primary analysis; source URLs are provided.

S2 Table. Study-level characteristics and reporting framework for the eight primary study clusters and selected references used to contextualize the analysis.

S3 Code. Reproducible Python analysis pipeline used to calculate the reported summary statistics, sensitivity analyses, and figures.

References

1. World Health Organization. Global antibiotic resistance surveillance report 2025. Geneva: World Health Organization (2025) Available from: https://www.who.int/publications/i/item/9789240116337
2. Antimicrobial Resistance Collaborators (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 399(10325):629–655. 10.1016/S0140-6736(21)02724-0

3. Ardila CM et al (2025) Machine-learning-based prediction of antimicrobial resistance in critical and high-priority pathogens: a systematic review. PLoS ONE 20(2):e0319460. 10.1371/journal.pone.0319460
4. Collins GS, Moons KGM, Dhiman P et al (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use artificial intelligence. BMJ 385:e078378. 10.1136/bmj-2023-078378
5. Moons KGM, Wolff RF, Riley RD et al (2019) PROBAST: a tool to assess risk of bias and applicability of prediction model studies. Ann Intern Med 170(1):51–58. 10.7326/M18-1376
6. Nigo K et al (2024) Deep learning model for personalized prediction of positive MRSA culture using time-series electronic health records. Nat Commun. 10.1038/s41467-024-46211-0
7. Li Y et al (2024) Development and validation of machine learning models to predict MDRO colonization or infection on ICU admission by using electronic health record data. Antimicrob Resist Infect Control. 10.1186/s13756-024-01428-y
8. Nsubuga M et al (2024) Generalizability of machine learning in predicting antimicrobial resistance in Escherichia coli: a multi-country case study in Africa. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11273897/
9. Yu et al (2026) Prediction of carbapenem-resistant Pseudomonas aeruginosa using machine learning with independent external validation. Front Cell Infect Microbiol. 10.3389/fcimb.2026.1663528
10. Multi task learning (2026) based early prediction model for antibiotic resistance using multi institutional cohort data. Sci Rep. 10.1038/s41598-026-41185-z
11. Carbapenem-resistant organism (2026) prediction using automated machine learning with independent external validation. Front Microbiol. 10.3389/fmicb.2026.1775344
12. Gu Y, Xu W, Xu J et al (2026) Construction and internal-external validation of a machine learning-based risk prediction model for multidrug resistance in ICU patients with acute exacerbation of chronic obstructive pulmonary disease. Front Med 13:1806672. 10.3389/fmed.2026.1806672
13. Development (2026) internal-external validation of a machine learning-based risk prediction model for multidrug resistance in mechanically ventilated ICU patients with chronic obstructive pulmonary disease. Front Med 13:1899252. 10.3389/fmed.2026.1899252
14. Predicting multidrug-resistant pneumonia: an interpretable machine learning model validated in US and Chinese patient cohorts (2026) Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC13033201/
15. Wang H, Yang C, Zhao M et al (2026) Comparison and validation of multiple machine learning algorithms for predicting MDRO infection in catheter-related bloodstream patients: a multicenter cohort study. Microbiol Spectr 14(3):e0371325. 10.1128/spectrum.03713-25
16. Xu J, Li Y, Wang Y et al (2026) Machine-learning and gene-synergy networks reveal interpretable drivers of antibiotic resistance in Staphylococcus aureus. Int J Antimicrob Agents 67(2):107690. 10.1016/j.ijantimicag.2025.107690

17. From surveillance to (2026) intelligence: a scoping review of machine learning for antimicrobial resistance surveillance intelligence across One Health. Front Public Health. 10.3389/fpubh.2026.1922265
18. Gulyás-Oldal V et al (2026) Validation, calibration and transparency, not accuracy: an appraisal of artificial intelligence against antimicrobial resistance and a sustainable, testable diagnostic architecture. Preprints.org. Available from: https://www.preprints.org/manuscript/202609.0097
19. Nguyen HA et al (2024) Predicting Pseudomonas aeruginosa drug resistance using artificial intelligence and clinical MALDI-TOF mass spectra. mSystems 9(9):e00789–e00724. 10.1128/msystems.00789-24
20. Ren et al (2024) Repurposing MALDI-TOF mass spectrometry for antimicrobial-resistance prediction in Staphylococcus epidermidis. JAC Antimicrob Resist. 10.1093/jacamr/dlae176
21. MSDeepAMR: deep-learning antimicrobial-resistance prediction across DRIAMS sites (2024) Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11062410/
22. VAMPr (2019) VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. PLoS Comput Biol 15(12):e1007511. 10.1371/journal.pcbi.1007511

Figures

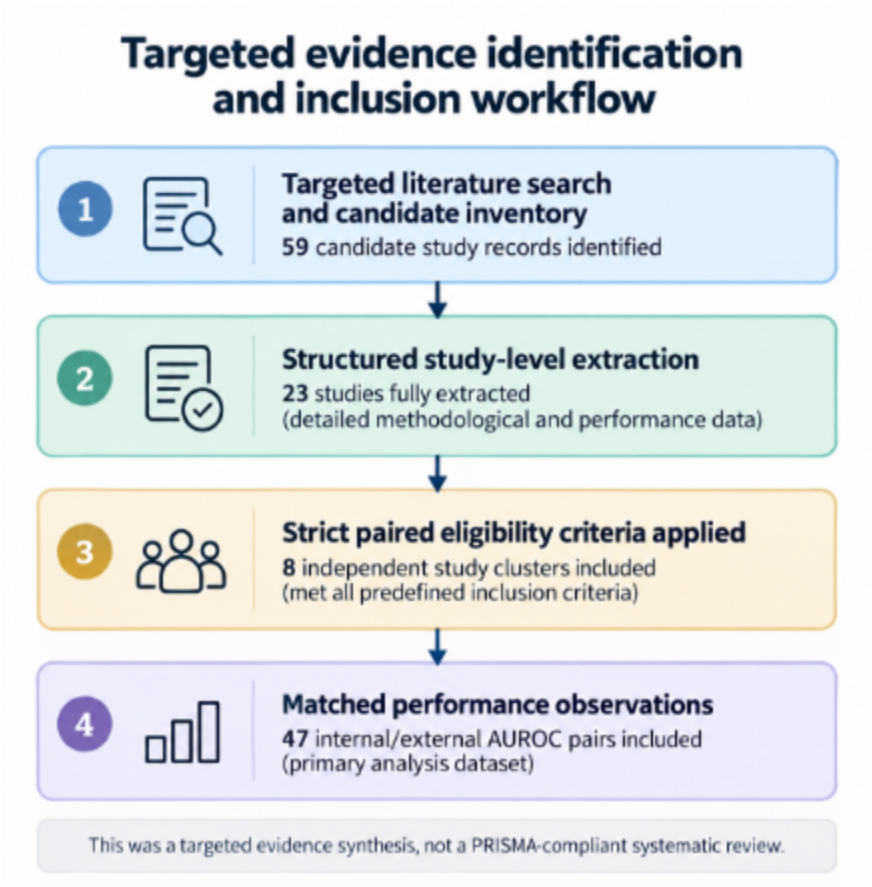


Figure 1

Targeted evidence identification and inclusion workflow. The 59-record inventory and 23 structured extractions describe the targeted evidence process used for this meta-research study; they are not intended to represent an exhaustive PRISMA-compliant systematic-review census.

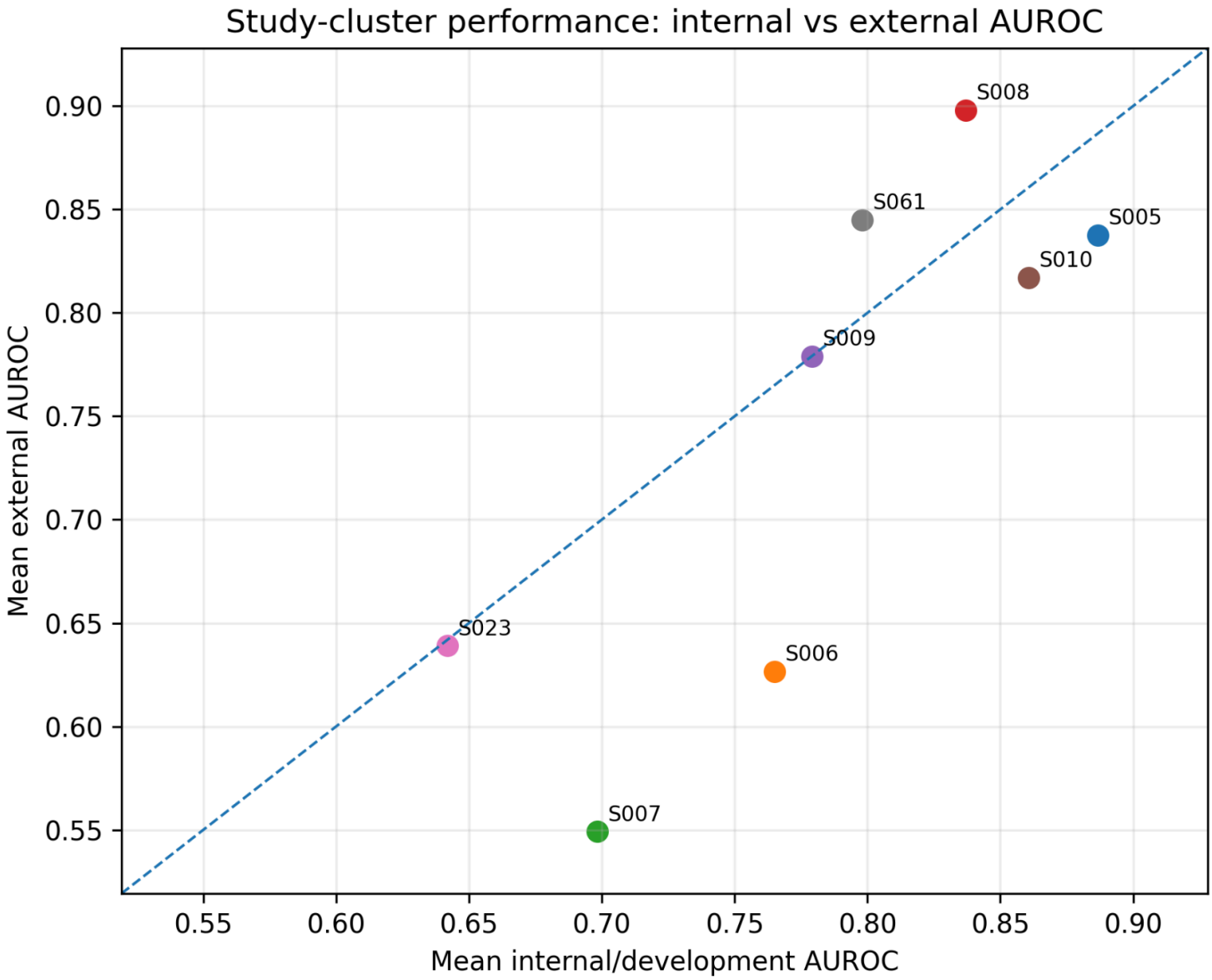


Figure 2

Mean internal/development versus mean external AUROC by study cluster. The diagonal indicates equal performance; points below the diagonal indicate lower external discrimination.

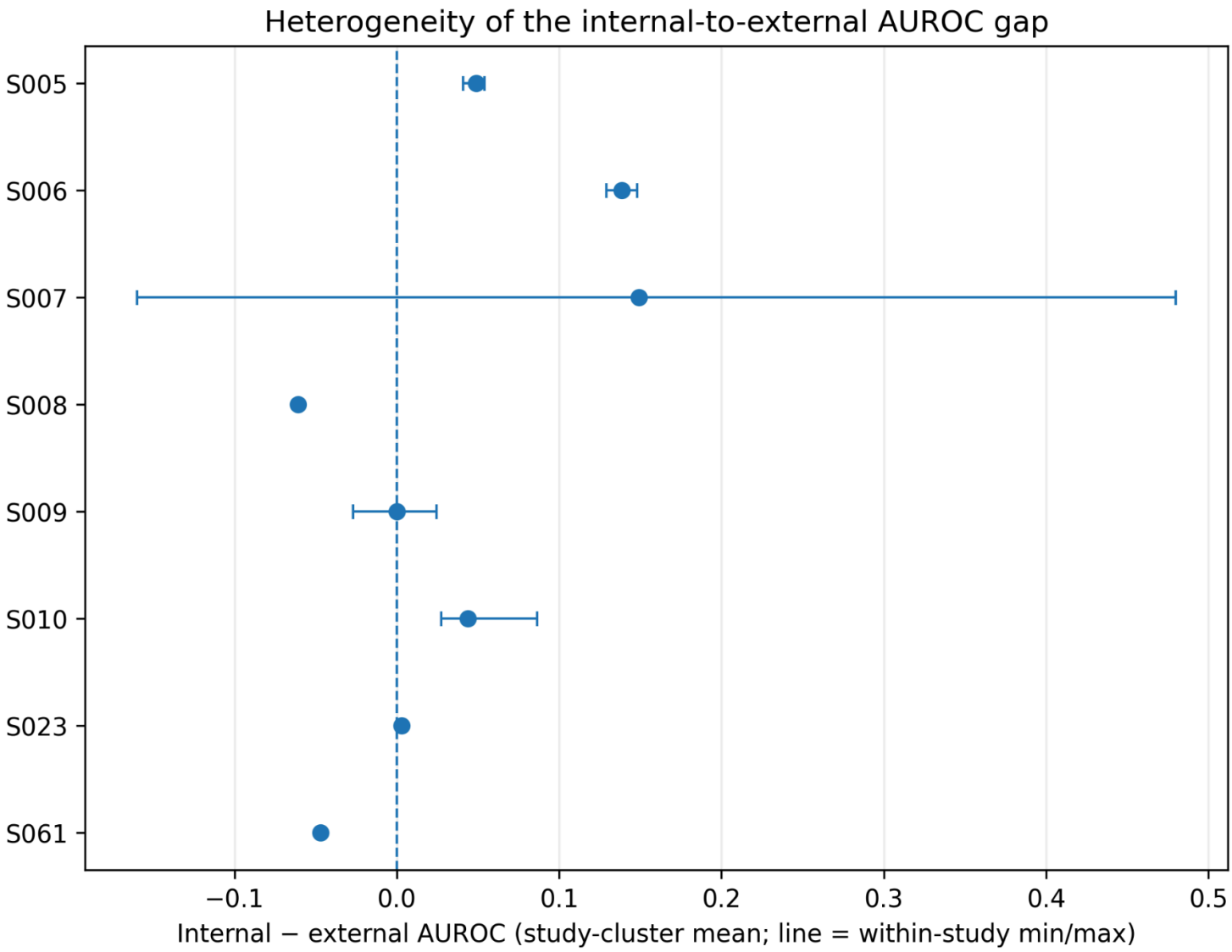


Figure 3

Heterogeneity of the internal-to-external AUROC gap. Points are study-cluster means; whiskers span the minimum and maximum paired ΔAUROC within each study.

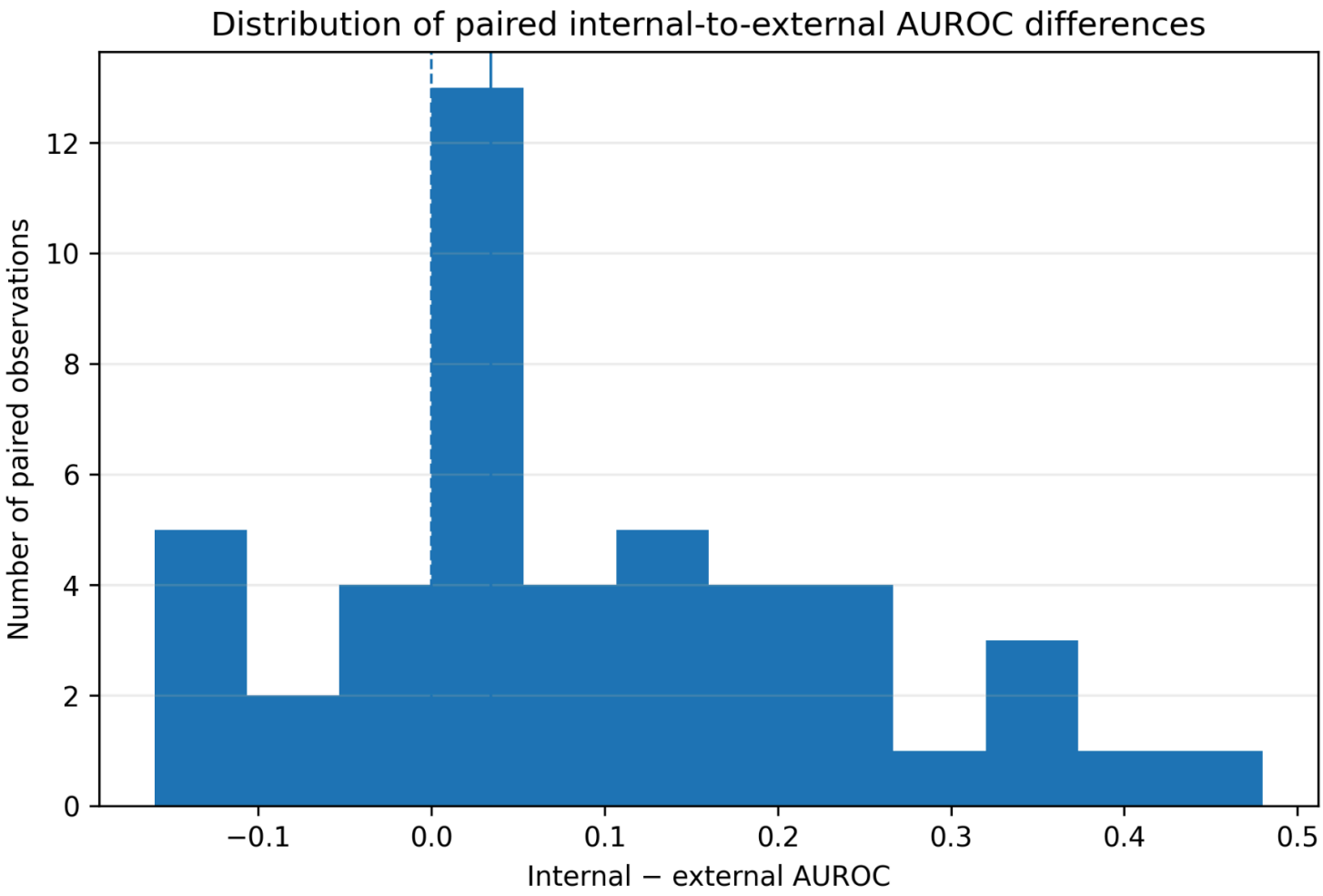


Figure 4

Distribution of paired internal-to-external AUROC differences. Positive values indicate lower external discrimination; the dashed horizontal reference marks zero.