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.

Dataset-structured evidence synthesis for machine-learning prediction models: a methodological framework and worked example

Osama Abdelhay

o.abdelhay@psut.edu.jo

Princess Sumaya University for Technology

Da'ad Abdel-Hay

King Hussein Cancer Centre

Abdullah F. Qasem

Independent Researcher

Taghreed Altamimi

Alfaisal University

Research Article

Keywords: machine learning, prediction models, meta-analysis, evidence synthesis, external validation, calibration, clinical decision support, systematic reviews, dataset reuse, performance evaluation

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10424152/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

Systematic reviews of machine-learning prediction models often pool performance estimates by publication or algorithm label. This can misrepresent the evidence when algorithms share participants, benchmark datasets recur across reports, internal and external validation are combined, or discrimination is interpreted without calibration. A previous commentary proposed a dataset-structured perspective. We extend that perspective into an operational framework and a worked example.

Methods

We define a hierarchy of evidence units for prediction-model synthesis: data source or cohort, analytic dataset, validation exercise, modeling pipeline, performance estimate, and publication/report. We apply the framework to a published systematic review and meta-analysis of machine-learning models for early prediction of ventilator-associated pneumonia (VAP). The aim was not to update the clinical review, but to reconstruct its evidence base using dataset-structured rules. We coded data sources, analytic datasets, validation design, performance domains, data-source reuse, dependent estimates, and implications for pooling, using one row per validation exercise whenever possible.

Results

Under a conventional paper/model count, the VAP evidence base appeared to contain 10 included reports, 10 primary report-level model entries, and 10 extractable AUROC/c-statistic summaries. Dataset-structured reconstruction identified 7 unique data-source groups, 10 analytic datasets, and 11 validation exercises. MIMIC-III recurred in 4 reports; one abstract was treated as companion evidence to a preprint, and several reports contributed dependent algorithm or prediction-window estimates. No validation exercise met our conservative definition of fully external validation, and none provided calibration-ready or decision-analytic evidence for the VAP target. A single pooled AUROC was therefore not an appropriate primary estimand.

Conclusions

Dataset-structured synthesis changes both the descriptive count and the interpretation of machine-learning prediction evidence. In the VAP example, the conclusion shifted from high pooled discrimination to promising internal or temporal discrimination with unproven transportability, calibration, and clinical usefulness. Future reviews should define the validation exercise as an evidence unit, reconstruct dependence before pooling, distinguish validation designs, and report performance domains beyond discrimination.

Background

Machine-learning prediction models are now developed for diagnosis, prognosis, triage, screening, surveillance, monitoring, and treatment selection. Their performance is increasingly summarized in systematic reviews and meta-analyses. This is an important development because prediction models intended for health care should be judged against accumulated evidence rather than isolated claims of technical performance. Yet many syntheses still proceed by identifying papers, extracting reported performance metrics, and pooling them. A previous commentary proposed a dataset-structured response to this problem [1].

The difficulty is not that machine learning lies outside statistics. Clinical prediction models developed with logistic regression, penalized regression, tree-based methods, support vector machines, gradient boosting, deep learning, or other approaches share the same broad scientific problem: a model is fitted or specified, and its predictions are evaluated in a target context. The quantity being synthesized is therefore not an intervention effect but model performance in a validation setting. A c-statistic, calibration slope, Brier score, sensitivity at a threshold, or net benefit is not a property of an algorithm alone. It reflects a model or modeling pipeline evaluated on a particular dataset, population, outcome definition, prediction horizon, measurement process, and validation design [2–7].

That commentary argued that the validation exercise - a model or pipeline evaluated on a defined dataset for a defined clinical task - should often be treated as the central evidential unit [1]. The present article deliberately builds on that work. Its contribution is operational: an evidence-unit taxonomy, conservative coding rules, a synthesis-decision algorithm, a worked reconstruction of a VAP review, and reusable extraction and reporting materials. We examine how conclusions change when publications, data sources, analytic datasets, validation exercises, modeling pipelines, and performance domains are separated before pooling.

The paper is framed as a methodological framework and case study. It is relevant to systematic reviewers, clinical epidemiologists, informaticians, and developers of clinical decision-support models because it addresses a common practical problem: when the literature contains multiple algorithms, repeated benchmark datasets, heterogeneous validation designs, and incomplete performance reporting, what exactly should a meta-analysis count and pool?

Methods

Framework: evidence units and estimands

We defined a dataset-structured hierarchy for prediction-model evidence synthesis (Fig. 1). The data source or cohort is the population source, registry, public database, imaging archive, institution, or clinical network from which records arise. The analytic dataset is the subset of records used for a defined modeling or validation task after applying eligibility criteria, time windows, inclusion/exclusion criteria, outcome definitions, and prediction horizons. The validation exercise is the evaluation of a

defined model or pipeline on a defined analytic dataset for a defined clinical prediction target under a specified validation design. The modeling pipeline includes predictor measurement, preprocessing, missing-data handling, feature engineering, feature selection, algorithm family, tuning, calibration, and threshold selection. The performance estimate is the reported metric, uncertainty interval, or performance curve derived from the validation exercise. The publication is treated as a reporting container rather than the default evidential unit.

This hierarchy was used to separate four common review targets. A review may ask about the average performance of a fixed model in new settings, the performance of a retrained pipeline, the relative performance of algorithm families, or the readiness of a clinical implementation. These are different estimands. A fixed-model validation review should prioritize external validation of the locked model. A pipeline-performance review should examine whether preprocessing, tuning, and calibration protocols are sufficiently similar. An algorithm-comparison review should prefer within-dataset contrasts when evaluating algorithms on the same participants. A clinical-usefulness review should require calibration and decision-relevant performance, not discrimination alone.

Worked example selection

We purposively selected a published systematic review and meta-analysis of machine-learning models for early prediction of ventilator-associated pneumonia (VAP) as a methodological case study [8]. VAP was chosen because it is clinically interpretable; the review was recent and compact; the source evidence included multiple algorithms and prediction windows; several reports used MIMIC-III; and calibration- and deployment-relevant evidence was sparse. The case study was not intended to update the VAP review or make a new clinical recommendation. Its purpose was to demonstrate how a dataset-structured framework changes evidence counting and synthesis decisions.

The source review summarized the performance of prediction models among adults undergoing invasive mechanical ventilation [8]. We reconstructed the included evidence at the levels of publication, data source, analytic dataset, validation exercise, modeling pipeline, and performance domain. Information was taken from the source review and, where available, the cited primary reports. No participant-level data were obtained or analyzed.

Extraction and coding

For each report in the worked example, we recorded the publication identifier, clinical target, data source or cohort, country or region, analytic cohort, sample size where available, prediction horizon or window, validation design, algorithms or pipelines evaluated, AUROC/c-statistic or other performance measures, calibration reporting, decision-analytic reporting, and notes on dependence. We used one row per validation exercise whenever the source reports allowed. Estimates based on the same participants, task, or source data were coded as dependent rather than as independent validation evidence.

Validation design was coded conservatively. Apparent performance, cross-validation, bootstrap validation, and random split-sample evaluation were coded as internal validation unless the report clearly showed participant separation from development data and a clinically meaningful source difference. Temporal same-site validation was coded separately from fully external validation. Fully external validation requires participant separation from model development and a meaningful difference in institution, geography, measurement process, or clinical workflow. If the validation design was unclear, it was coded as internal/unclear rather than external.

Dataset identity and overlap were coded using dataset names, institutions, public database version, recruitment years, clinical setting, and companion-report evidence. Repeated use of the same public source was treated as reuse of a data-source family, not as repeated independent external validation. Where two reports appeared to describe the same MIMIC-III v1.3 analysis with the same model set and 24-hour performance pattern, the later abstract was treated as companion evidence for counting. This conservative classification does not imply duplicate publication.

Synthesis-decision rules

Before considering quantitative pooling, we applied decision rules based on the review question, validation design, dependence, comparability, and performance domains. A single headline pooled AUROC was considered inappropriate as the primary estimand when the evidence mixed internal and temporal validation, reused the same data source, contained multiple dependent algorithm or window estimates, lacked calibration, or combined non-comparable targets such as static diagnosis, early prediction, surveillance, and dynamic transition modeling. In those circumstances, the primary synthesis modes were an evidence map and a dataset-structured narrative synthesis. Any quantitative summaries should be stratified by validation design and data-source family and should not treat dependent estimates as independent.

Results

Evidence-unit taxonomy

The resulting framework is summarized in Table 1 and Fig. 1. The central change is that the validation exercise replaces the paper as the default unit for inspection before pooling. This does not mean that every validation exercise must be meta-analyzed. Rather, it means that the reviewer first reconstructs what each performance estimate represents, then decides whether the estimates are comparable and sufficiently independent for quantitative synthesis. Table 2 presents the minimum extraction schema used in the worked example and is recommended for future reviews of machine-learning prediction model performance.

Table 1
Evidence units in dataset-structured synthesis.

Unit	Definition	Use in synthesis
Publication/report	A paper, preprint, abstract, or report that contains one or more model evaluations.	Used for traceability and citation; should not automatically define independent evidence.
Data source/cohort	The underlying clinical source such as a hospital, registry, public database, trial cohort, or imaging archive.	Defines participant provenance, benchmark reuse, broad overlap, case mix, and applicability.
Analytic dataset	The subset of records produced by eligibility criteria, time period, outcome definition, prediction horizon, and preprocessing.	Defines the actual dataset on which a model or pipeline was evaluated.
Validation exercise	A defined model or pipeline evaluated on a defined analytic dataset for a defined clinical task and validation design.	Primary evidential unit for performance synthesis; determines whether pooling, stratification, or narrative synthesis is appropriate.
Modelling pipeline	Predictors, preprocessing, missing-data handling, feature engineering, algorithm, tuning, calibration, and thresholding.	Clarifies whether the review compares fixed models, retrained pipelines, or broad algorithm families.
Performance estimate	A reported metric, curve, threshold result, uncertainty interval, or decision-analytic measure.	Determines whether conclusions concern discrimination, calibration, overall accuracy, threshold performance, or clinical utility.

Table 2
Minimum extraction schema for dataset-structured reviews of machine-learning prediction models.

Level	Minimum extraction items	Synthesis role
Data source/cohort	Source name; institution or database; country or region; public/private status; recruitment dates; clinical setting; eligibility; sample size; events.	Reconstructs independence, overlap, benchmark reuse, case mix, and applicability.
Analytic dataset	Inclusion/exclusion criteria; time window; outcome definition; prediction horizon; predictor measurement process; event/non-event counts.	Defines the dataset-specific target for performance estimation and comparability.
Validation exercise	Apparent, resampling-based, split-sample, temporal, geographic, internal-external, fully external, or unclear validation; model development source; validation source.	Separates development-stage performance from transportability evidence.
Modelling pipeline	Algorithm family; predictors/features; preprocessing; missing data; feature selection; tuning; calibration; threshold selection; locked code/model availability.	Distinguishes algorithm labels from actual validated pipelines.
Performance domain	Discrimination; calibration; overall performance; threshold metrics; decision-analytic measures; uncertainty; events at horizon.	Prevents discrimination-only conclusions from being interpreted as clinical usefulness.
Dependence and reuse	Repeated dataset use; companion reports; multiple algorithms; multiple horizons/windows; shared thresholds; partial overlap.	Determines whether estimates are independent, clustered, paired, or unsuitable for pooling.
Risk of bias/applicability	Participants, predictors, outcome, analysis, leakage, overfitting, sample size, missing data, target-setting applicability.	Determines primary versus secondary analyses, sensitivity analyses, and strength of inference.

VAP evidence reconstruction

Under the conventional paper/model count used for the worked example, the VAP evidence base appeared to contain 10 included reports, 10 primary report-level model entries, and 10 extractable AUROC/c-statistic summaries. The dataset-structured reconstruction changed this picture substantially (Table 3; Fig. 2). We identified 7 unique data-source groups: UMCU, the National Trauma Data Bank, OUTCOMEREA, Hospital Tacchini, the Qatar/Hamad trauma registry, LifeCenter Hospital, and the MIMIC-III source family. The non-MIMIC sources were represented by Schurink et al., Pearl and Bar-Or, Liquet et al., dos Santos et al., Abujaber et al., and Amador et al. [9–14]. Together, the evidence base yielded 10 analytic datasets and 11 validation exercises after separating tasks, temporal evaluations, and companion reporting.

MIMIC-III was the principal reused data source [15]. Giang and colleagues evaluated several machine-learning methods on an internal hold-out test set [16]. Liang and colleagues developed a random-forest model with cross-validation [17]. Faucher and colleagues compared logistic regression, explainable boosting machines, and XGBoost across post-intubation windows [18]. A Calvert/Faucher abstract appeared to report companion evidence from the same MIMIC-III v1.3 setup [19]. These reports should therefore not be interpreted as four independent clinical settings.

No validation exercise met our conservative definition of fully external validation. Two reports used temporal same-site or future-cohort designs: dos Santos et al. at Hospital Tacchini [12] and Amador et al. at LifeCenter Hospital [14]. These designs provide stronger evidence than random split-sample evaluation or cross-validation, but they do not establish transportability across institutions or measurement systems. The remaining validation exercises were internal, random split, cross-validation, or unclear.

Calibration-ready and decision-analytic evidence were absent for the VAP target in the reconstructed dataset. Several reports focused on discrimination or classification accuracy, and some evidence could not be placed within a static AUROC framework. Pearl and Bar-Or [10] reported Gini, RMSE, and accuracy rather than conventional AUROC, whereas Liquet and colleagues [11] used a dynamic multistate frailty model rather than a static supervised classifier. These reports remain relevant to the evidence map but should not be forced into a single AUROC pool.

Table 3
Before/after evidence count in the VAP worked example.

Evidence descriptor	Conventional paper/model count	Dataset-structured count
Included reports	10	-
Primary report-level model entries	10 primary report-level model entries	-
Extractable AUROC/c-statistic summaries	10 extractable summary AUROC/c-statistic entries	-
Unique data-source/cohort groups	-	7
Unique analytic datasets	-	10
Unique validation exercises	-	11
Validation exercises meeting the fully external definition	-	0
Internal, temporal, or unclear validation exercises	-	11
Validation exercises with calibration-ready evidence	-	0
Validation exercises with decision-analytic evidence	-	0
Reused or overlapping datasets	-	1 data-source family: MIMIC-III
Reports contributing multiple dependent estimates	-	5 reports / 4 study groups after companion merge

Table note. The conventional count treats each included report as contributing its main report-level performance entry. It does not count every algorithm-by-time-window-by-task combination as independent evidence; those combinations are captured as dependent estimates. Temporal same-site validation was not counted as fully external validation.

Dataset-structured synthesis decision

The synthesis decision is summarized in Table 4 and Fig. 3. The central result is not a new pooled AUROC; it is a change in the primary interpretation. A conventional approach might summarize the literature as 10 studies with high reported discrimination. After reconstruction, the evidence base was dominated by internal or temporal validation, repeated MIMIC-III analyses, companion reporting, dependent estimates based on algorithms or windows, and the absence of calibration or decision-analytic evidence. The appropriate primary conclusion is therefore that several pipelines show promising

internal or temporal discrimination, but independent external transportability and clinical usefulness remain unproven.

This conclusion also changes how algorithm comparisons should be interpreted. Separately pooled AUROCs for algorithm families would be vulnerable to confounding by data-source difficulty, outcome definitions, prediction windows, and tuning effort. Models compared within the same MIMIC-III analytic cohort are paired by design, whereas models evaluated on unrelated cohorts are not directly comparable as algorithm contrasts. Algorithm ranking should therefore be limited to within-dataset contrasts under comparable prediction windows and outcome definitions or avoided when such contrasts are sparse. Prior comparative evidence also cautions against treating broad algorithm labels as intrinsically superior to regression-based approaches [20].

Table 4
Dataset-structured synthesis decision for the VAP worked example.

Review question	Evidence available after reconstruction	Dataset-structured synthesis decision
What is the average discrimination of ML models for VAP prediction?	AUROC/c-statistic was commonly reported, but estimates came from heterogeneous validation exercises, different prediction horizons, repeated MIMIC-III analyses, and multiple dependent algorithms/windows within the same datasets.	Do not use a single headline pooled AUROC as the primary result. A discrimination summary is possible only as a secondary, stratified analysis after separating validation design, source dataset, and dependent estimates. If reported, label it as internal/temporal discrimination, not transportable clinical performance.
Is model performance transportable to new clinical settings?	No validation exercise met the fully external definition. Two reports used temporal same-site or future-cohort validation; most evidence was internal, cross-validation, random split, or unclear.	Primary conclusion should focus on lack of independent external validation. Reported internal/temporal discrimination is promising, but transportability remains unproven.
Which algorithm is best?	Several reports compared multiple algorithms, often within the same dataset and sometimes across multiple prediction windows. MIMIC-III was reused across Giang, Liang, Faucher, and the Calvert/Faucher abstract.	Avoid ranking algorithms from separately pooled AUROCs. Restrict algorithm comparison to within-dataset contrasts where models were evaluated on the same participants under the same outcome definition and prediction horizon.
Are predictions clinically useful?	Discrimination was reported more often than calibration or clinical utility. Calibration-ready evidence and decision-analytic evidence were not available for synthesis.	Clinical usefulness cannot be inferred from AUROC alone. State explicitly that absolute risk reliability, threshold performance, net benefit, and impact evidence remain unproven.
Is the evidence mature enough for deployment claims?	Conventional counting suggests 10 reports/models, but reconstruction shows fewer independent data-source groups, repeated MIMIC-III use, companion reporting, 0 validation exercises meeting the fully external definition, and 0 calibration-ready validation exercises.	Evidence maturity is lower than the paper/model count suggests. The evidence supports methodological promise, not deployment readiness.
Should all reported AUROCs/c-statistics be pooled?	The evidence includes nested estimates from the same studies, repeated source cohorts, internal and temporal validation designs, and non-comparable clinical targets such as diagnostic support, early prediction, surveillance, and dynamic transition modelling.	No, not as a primary analysis. A pooled AUROC across all studies would mix estimands and overweight dependent evidence. Use an evidence map or exploratory stratified summaries.
How should MIMIC-III studies be handled?	MIMIC-III appears repeatedly with different analytic cohorts, prediction windows, and algorithms. One	Treat MIMIC-III as a reused source family, not four independent clinical settings. Merge companion evidence

Review question	Evidence available after reconstruction	Dataset-structured synthesis decision
	abstract appears to be companion evidence to a preprint.	for counting and cluster or sensitivity-analyze the MIMIC-III family.
How should temporal same-site validations be interpreted?	Hospital Tacchini and LifeCenter used future or validation cohorts from the same institution or institution group.	Report temporal validation separately from fully external validation. These designs are stronger than random split/CV, but do not establish geographic or institutional transportability.
How should non-static or non-AUROC evidence be handled?	Pearl and Bar-Or reported Gini/RMSE/accuracy rather than AUROC; Liquet used dynamic multistate frailty modelling rather than a static classifier.	Do not force non-comparable evidence into an AUROC pool. Retain these studies in the evidence map and narrative synthesis.
What is the final synthesis decision?	The reconstructed evidence base contains promising discrimination signals, but they are largely internal, dependent, heterogeneous, and calibration-poor.	Primary synthesis mode: dataset-structured narrative synthesis plus evidence map. A new pooled AUROC is not necessary and may be misleading.

Reusable methodological outputs

The framework produces five reusable outputs: (1) an evidence reconstruction table with one row per validation exercise; (2) a before/after evidence count comparing conventional paper/model counts with dataset-structured counts; (3) an evidence map showing publications, data-source families, validation exercises, and performance domains; (4) a dataset-structured synthesis decision table; and (5) an operational supplement containing an extraction template, coding dictionary, worked-example dataset, code for evidence counts, and a reporting checklist. Table 5 provides conservative defaults for common ambiguities.

Table 5
Operational defaults for ambiguity, dependence, and incomplete performance reporting.

Problem	Conservative coding	Primary synthesis implication
Dataset identity unclear	Code as unclear provenance; contact authors if the ambiguity affects independence or eligibility.	Do not count as an independent external validation unless independence is documented; test exclusion of unclear sources.
Same public source or registry reused	Code as reused data-source family; record database version, time period, and analytic cohort.	Cluster, merge companion reports for counting, or conduct leave-one-source-family-out sensitivity analysis.
Participant overlap probable but not quantified	Classify as definite, probable, possible, unclear, or unlikely using source names, dates, institutions, registries, and eligibility criteria.	Treat definite/probable overlap as dependent; test possible/unclear overlap in sensitivity analyses.
Multiple algorithms or windows on same participants	Code as dependent algorithm/window estimates.	Prefer within-dataset contrasts for comparison or select one prespecified estimate per validation exercise.
Validation design unclear	Code as internal/unclear rather than external.	Exclude from the primary external-validation synthesis or place in a separate stratum.
Calibration missing	Code as not assessable for absolute risk accuracy.	Do not infer clinical usefulness or deployment readiness from discrimination alone.
Decision-analytic evidence missing	Code net benefit/clinical utility as not assessable.	Do not claim clinical utility; state that decision thresholds and impact remain unproven.
Uncertainty missing	Code precision as unclear.	Avoid inverse-variance pooling based on guessed precision unless assumptions and sensitivity analyses are explicit.

Discussion

Principal findings

This article converts the previously published dataset-structured argument into an operational framework and a worked example [1]. The VAP case study shows that paper counts, model counts, and AUROC counts can overstate the maturity of a machine-learning prediction literature. A review that appears to contain 10 reports and 10 AUROC-type summaries may, after reconstruction, contain a smaller set of data-source families, no validation exercise meeting a fully external definition, repeated benchmark use, multiple dependent estimates, and no calibration-ready evidence. This does not render the literature uninformative; it changes what it can support.

The most important implication is that the primary synthesis question should be defined before extracting the most convenient performance metric. If the question is transportability, external validation should be the primary evidence. If the question is algorithm comparison, within-dataset contrasts should be prioritized over separately pooled raw performance. If the question is clinical usefulness, calibration, threshold performance, and decision-analytic evidence are essential. AUROC-only pooling can describe ranking performance, but it cannot establish absolute risk accuracy or deployment readiness.

Relation to reporting and appraisal guidance

The proposed framework is not a replacement for TRIPOD-SRMA, TRIPOD + AI, PROBAST + AI, the original PROBAST, CHARMS, or guidance on prediction-model meta-analysis [2–5, 7, 21]. Instead, it uses the variables highlighted by those frameworks as analytic variables. Source population, target population, outcome definition, prediction horizon, validation design, model details, performance measures, risk of bias, and applicability are not merely reporting fields. They define the estimand, the evidence unit, the dependence structure, and the interpretation of any pooled estimate.

The framework also aligns with current thinking on clinical prediction model evaluation. Internal validation, internal-external validation, temporal validation, and fully external validation answer different questions [22–26]. Calibration and decision-analytic evaluation answer questions that discrimination cannot answer [27, 28]. Dataset-structured synthesis makes these distinctions visible at the evidence-synthesis stage, where they are often blurred by paper-centered extraction.

Implications for quantitative synthesis

Dataset-structured synthesis does not imply that quantitative meta-analysis should be avoided. It implies that quantitative synthesis should follow evidence reconstruction. When independent external validations are sufficient, and performance measures are comparable, random-effects meta-analysis on appropriate scales may be useful [5, 29]. When multiple algorithms or horizons are evaluated on the same participants, multivariate or multilevel models, paired contrasts, robust variance estimation, or prespecified selection of one estimate per dataset may be appropriate [30, 31]. When algorithm comparisons are incomplete, network approaches may be considered, but only when nodes represent comparable models or pipelines and transitivity is defensible [32].

Conversely, a finding that pooling is not defensible is a valid synthesis result. In the VAP example, the evidence map itself is the primary result because it shows that transportability and clinical usefulness are not established. A pooled AUROC that ignores internal validation, repeated use of MIMIC-III, companion reporting, and missing calibration would be easier to report but would be weaker as clinical evidence.

Strengths and limitations

The main strength of this work is its operational focus. The article defines evidence units, provides conservative coding rules, applies them to a real, published machine-learning prediction review, and produces outputs that other reviewers can reuse. The VAP example is compact enough to audit

manually, yet complex enough to demonstrate repeated use of public data sources, dependent estimates, temporal validation, heterogeneous prediction targets, and missing calibration.

Several limitations should be acknowledged. First, the worked example was a methodological case study, not an updated systematic review of VAP prediction, and the extraction depended on information available in the source review and cited reports. Second, coding was performed by a single reviewer; explicit rules and an auditable reconstruction table improve transparency but do not replace independent duplicate extraction. Third, dataset overlap can be difficult to reconstruct from published reports alone, and the classification of companion reporting may require clarification from the authors in a formal review. Fourth, we did not re-estimate performance from participant-level data or shared predictions. Fifth, the conservative definition of fully external validation may classify temporal same-site validations as stronger than internal validation but not as fully external; this is intentional for transportability claims but should be reported transparently. Finally, the framework was illustrated in one clinical area and should be tested in other areas of machine-learning prediction, including imaging, time-to-event prognosis, and multimodal deep learning. These limitations are consistent with broader reviews documenting weaknesses in methodological conduct and reporting across clinical machine-learning prediction studies [33–35], as well as concerns about benchmark-driven evaluation in medical imaging [36].

Future research

Several methodological questions remain. Simulation studies are needed to quantify the bias and undercoverage introduced when repeated benchmark datasets and multiple dependent estimates from different algorithms are treated as independent. More work is needed on the joint synthesis of discrimination and calibration, especially given the inconsistent reporting of calibration. Comparative synthesis of algorithms requires further development of contrast-based, multilevel, and network approaches. Finally, individual participant data or shared prediction outputs would allow recalculation of performance under harmonized definitions, dataset-specific recalibration, and clinically meaningful thresholds. Until such data become common, aggregate-data reviews can still improve substantially by reconstructing evidence units before pooling.

Conclusions

Machine-learning prediction models do not make meta-analysis obsolete; they make its epidemiologic assumptions more visible. The central question is not whether a collection of algorithms achieves a high average AUROC in the published literature, but whether a defined model or pipeline has been validated in sufficiently independent, relevant, and clinically meaningful settings, with performance domains adequate to support the intended conclusion.

The VAP worked example shows that dataset-structured reconstruction can change a review from an apparently model-rich evidence base to a smaller, more dependent set of validation exercises with unproven external transportability and no calibration-ready evidence. Future reviews should count

datasets as well as papers, define validation exercises before pooling, separate internal from external validation, model or avoid dependence, and treat calibration and decision-relevant performance as necessary evidence for clinical usefulness. This is not a rejection of meta-analysis; it is a return to its purpose: summarizing evidence without changing the meaning of the quantities being combined.

Declarations

Ethics approval and consent to participate:

Not applicable. This methodological study used information from published aggregate reports and did not involve human participants, patient-level data, or animal data.

Consent for publication:

Not applicable.

Competing interests:

The author declares no competing interests.

Funding:

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

O.A. and D.A. conceptualized the study, conducted the initial screening for the VAP systematic review, and drafted the manuscript. A.Q. reviewed the methodology, served as the third reviewer in the systematic review, and edited the manuscript. T.A. developed the software and Excel templates, contributed to the systematic review, and proofread and critically reviewed the manuscript. All authors reviewed and approved the final version of the manuscript.

Acknowledgments:

None.

Declaration of generative AI and AI-assisted technologies: During the preparation of this manuscript, ChatGPT 5.5, Grammarly, Jenni, and EndNote were used to support drafting, organization, and

proofreading. The author reviewed, verified, and edited the content and takes full responsibility for the manuscript.

Availability of data and materials:

The evidence reconstruction template, coding dictionary, worked-example extraction table, and code used to reproduce the evidence counts will be supplied as additional files with the manuscript. No patient-level data were used.

References

1. Abdelhay O. Intelligence-Based Med. 2026;15:100422. 10.1016/j.ibmed.2026.100422. Meta-analysis of machine-learning prediction models: a dataset-structured approach to evidence synthesis.
2. Snell KIE, Levis B, Damen JAA, Dhiman P, Debray TPA, Hooft L, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). BMJ. 2023;381:e073538. 10.1136/bmj-2022-073538.
3. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
4. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro CL, Dhiman P, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. 10.1136/bmj-2024-082505.
5. Debray TPA, Damen JAAG, 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. 10.1136/bmj.i6460.
6. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019. 10.1007/978-3-030-16399-0.
7. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11:e1001744. 10.1371/journal.pmed.1001744.
8. Frondelius T, Atkova I, Miettunen J, Rello J, Vesty G, Chew HSJ, et al. Early prediction of ventilator-associated pneumonia with machine learning models: a systematic review and meta-analysis of prediction model performance. Eur J Intern Med. 2024;121:76–87. 10.1016/j.ejim.2023.11.009.
9. Schurink CAM, Visscher S, Lucas PJF, van Leeuwen HJ, Buskens E, Hoff RG, et al. A Bayesian decision-support system for diagnosing ventilator-associated pneumonia. Intensive Care Med. 2007;33:1379–86. 10.1007/s00134-007-0728-6.
10. Pearl A, Bar-Or D. Decision support in trauma management: predicting potential cases of ventilator associated pneumonia. Stud Health Technol Inf. 2012;180:305–9.

11. Liquet B, Timsit JF, Rondeau V. Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units. BMC Med Res Methodol. 2012;12:79. 10.1186/1471-2288-12-79.
12. Dos Santos RP, Silva D, Menezes A, Lukasewicz S, Dalmora CH, Carvalho O, et al. Automated healthcare-associated infection surveillance using an artificial intelligence algorithm. Infect Prev Pract. 2021;3:100167. 10.1016/j.infpip.2021.100167.
13. Abujaber A, Fadlalla A, Gammoh D, Al-Thani H, El-Menyar A. Machine learning model to predict ventilator associated pneumonia in patients with traumatic brain injury: the C.5 decision tree approach. Brain Inj. 2021;35:1095–102. 10.1080/02699052.2021.1959060.
14. Amador T, Saturnino S, Veloso A, Ziviani N. Early identification of ICU patients at risk of complications: regularization based on robustness and stability of explanations. Artif Intell Med. 2022;128:102283. 10.1016/j.artmed.2022.102283.
15. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. 10.1038/sdata.2016.35.
16. Giang C, Calvert J, Rahmani K, Barnes G, Siefkas A, Green-Saxena A, et al. Predicting ventilator-associated pneumonia with machine learning. Med (Baltim). 2021;100:e26246. 10.1097/MD.0000000000026246.
17. Liang Y, Zhu C, Tian C, Lin Q, Li Z, Li Z, et al. Early prediction of ventilator-associated pneumonia in critical care patients: a machine learning model. BMC Pulm Med. 2022;22:250. 10.1186/s12890-022-02031-w.
18. Faucher M, Casie Chetty S, Shokouhi S, Barnes G, Rahmani K, Calvert J, et al. Early prediction of ventilator-associated pneumonia in ICU patients using an interpretable machine learning algorithm. Preprints. 2022;2022060149. 10.20944/preprints202206.0149.v1.
19. Calvert J, Faucher M, Casie Chetty C, Shokouhi S, Evans D, Barnes G, Mao Q. Early prediction of ventilator-associated pneumonia in intensive care unit patients using an intelligible machine learning algorithm. Am J Respir Crit Care Med. 2022;205:A3451.
20. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
21. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170:51–8. 10.7326/M18-1376.
22. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. 10.1136/bmj-2023-074819.
23. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Collins GS, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. 10.1136/bmj-2023-074820.

24. Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM, Cochrane IPD, Meta-analysis Methods Group. Individual participant data meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015;12:e1001886. 10.1371/journal.pmed.1001886.
25. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. 10.1136/bmj.i3140.
26. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245–7. 10.1016/j.jclinepi.2015.04.005.
27. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
28. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26:565–74. 10.1177/0272989X06295361.
29. Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? Stat Methods Med Res. 2018;27:3505–22. 10.1177/0962280217705678.
30. Snell KIE, Hua H, Debray TPA, Ensor J, Look MP, Moons KGM, Riley RD. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. J Clin Epidemiol. 2016;69:40–50. 10.1016/j.jclinepi.2015.05.009.
31. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods. 2010;1:39–65. 10.1002/jrsm.5.
32. Yusufujiang M, Damen JAA, Idema DL, Schuit E, Moons KGM, de Jong VMT. Network meta-analysis of prediction models using aggregate or individual participant data: a scoping review and recommendations for reporting and conduct. J Clin Epidemiol. 2025;188:112006. 10.1016/j.jclinepi.2025.112006.
33. Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2023;154:8–22. 10.1016/j.jclinepi.2022.11.015.
34. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol. 2022;22:12. 10.1186/s12874-021-01469-6.
35. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. 2022;22:101. 10.1186/s12874-022-01577-x.
36. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5:48. 10.1038/s41746-022-00592-y.

Figures

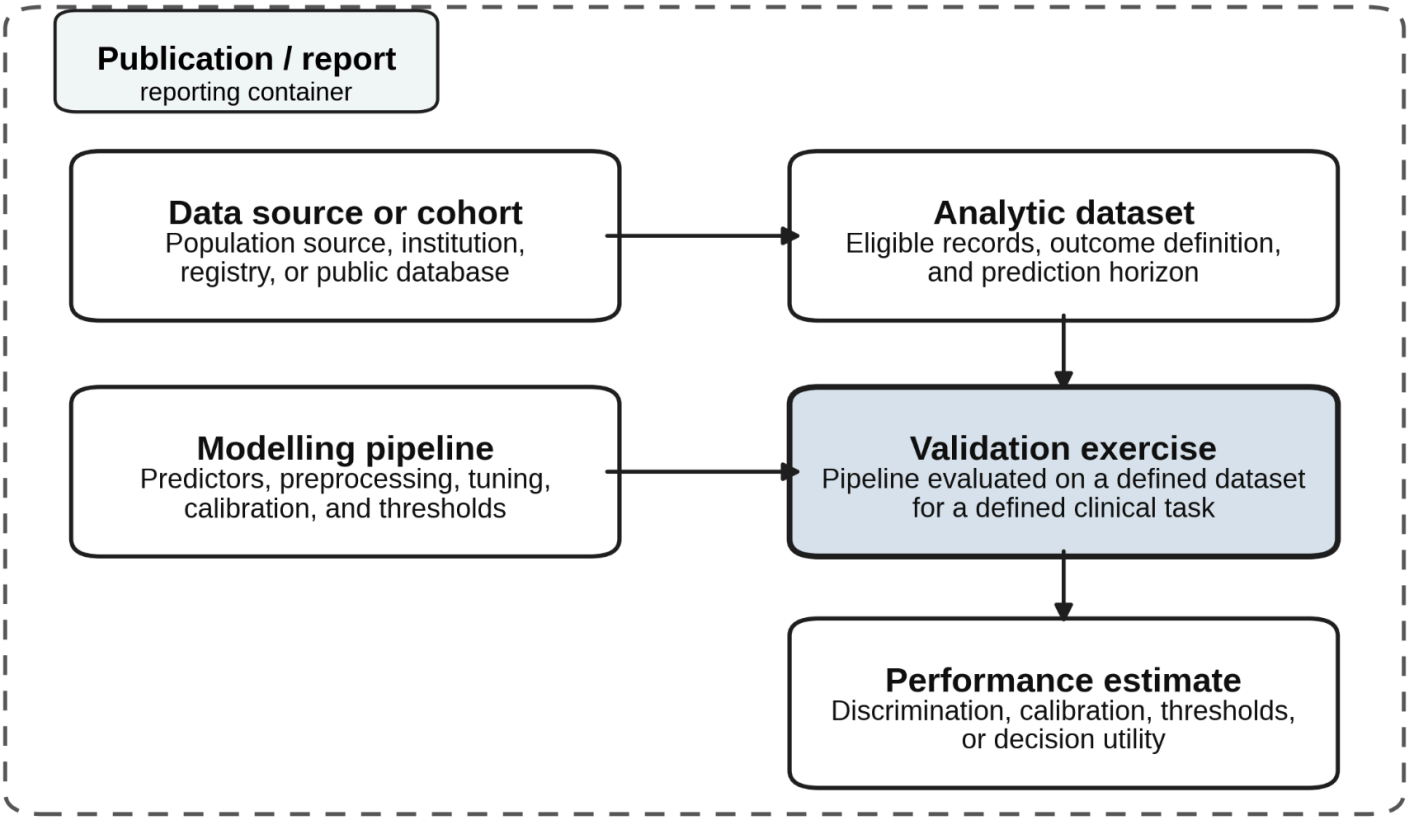


Figure 1

Dataset-structured evidence hierarchy for machine-learning prediction-model synthesis. The publication is the reporting unit; the validation exercise is often the evidential unit.

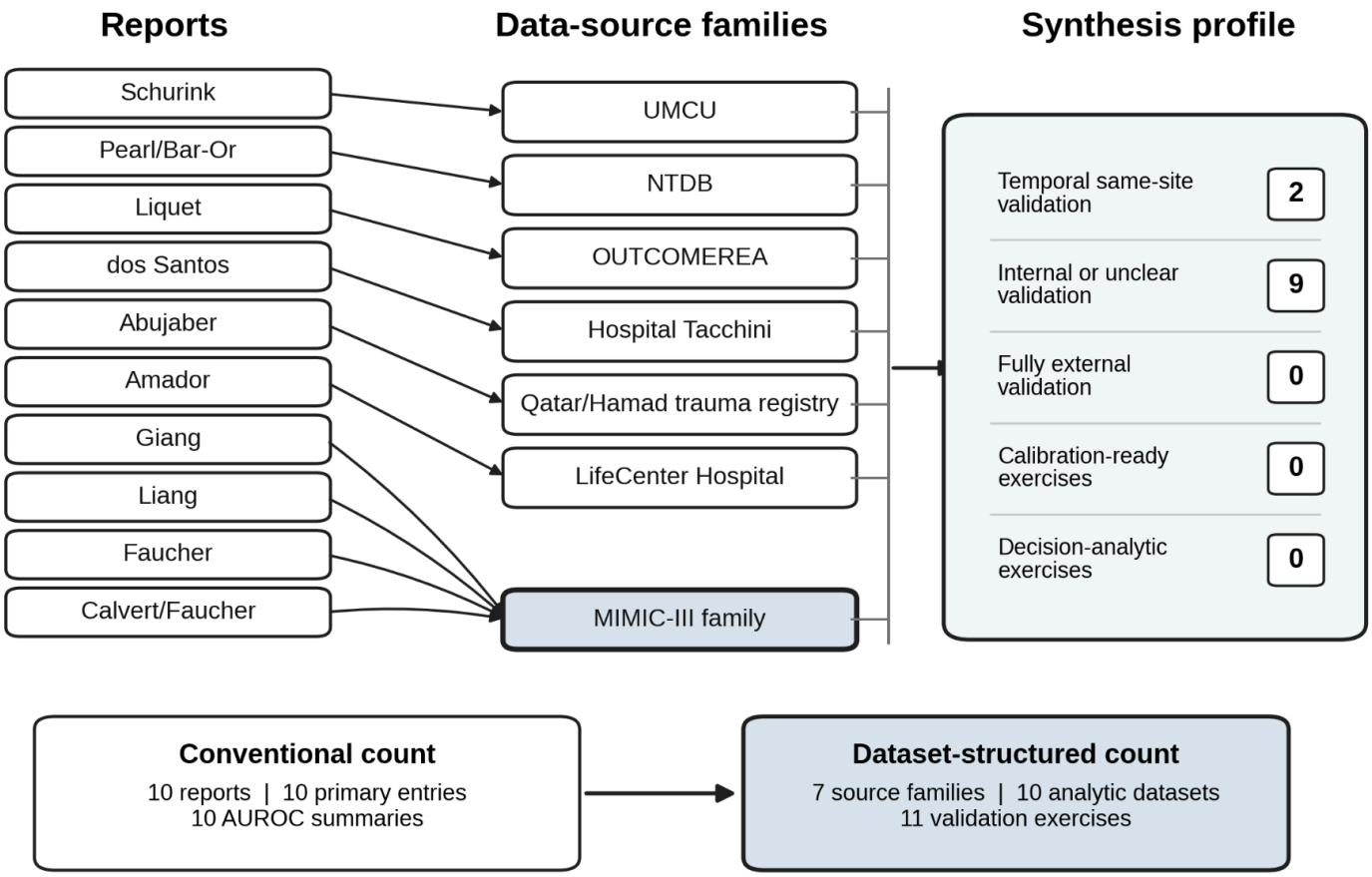


Figure 2

Evidence map for the VAP worked example. Repeated use of MIMIC-III and dependent estimates reduces the effective number of independent validation sources.

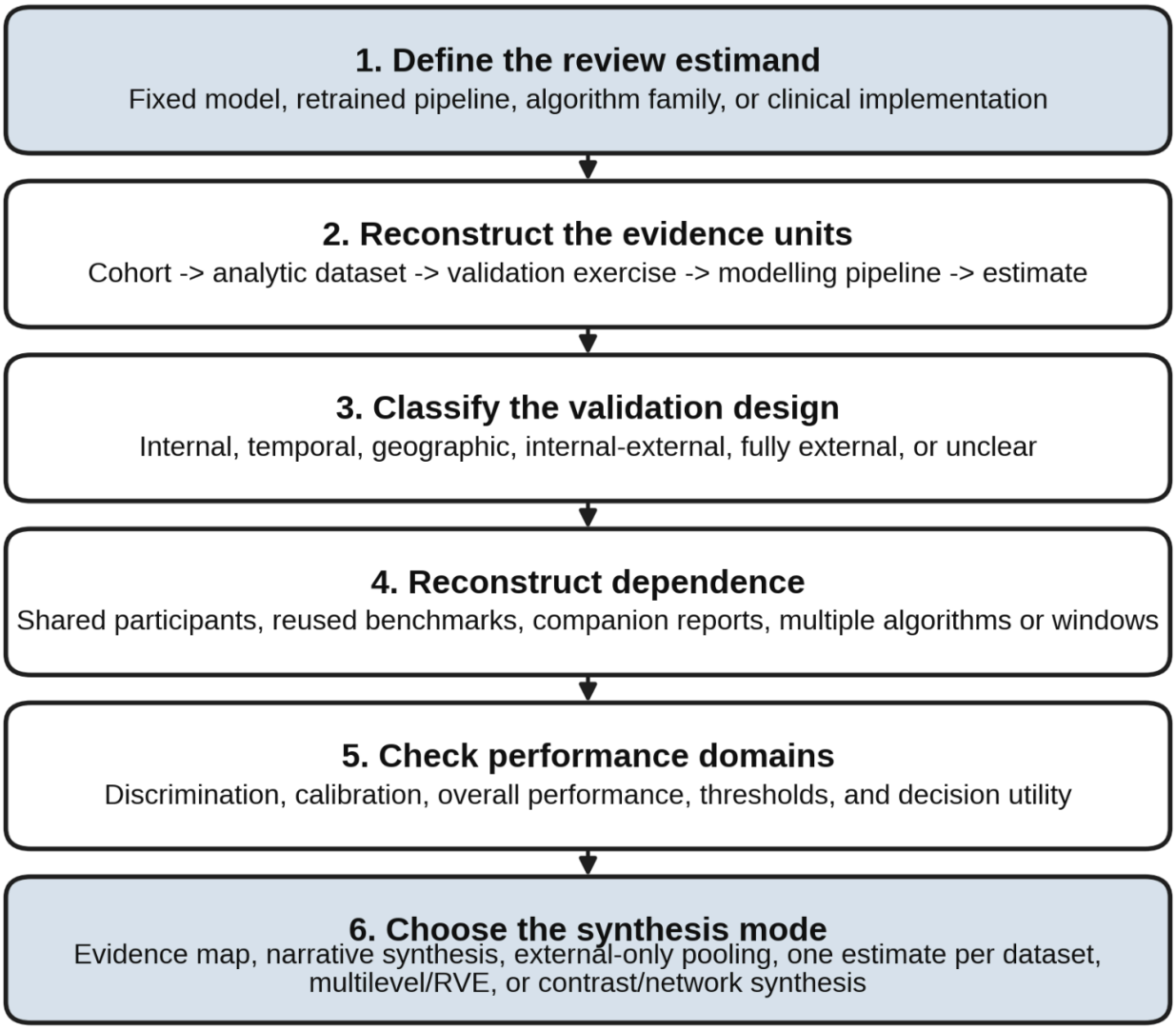


Figure 3

Decision algorithm for dataset-structured synthesis of machine-learning prediction-model performance.

Supplementary Files

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

- Additionalfile1.xlsx
- Additionalfile3.r
- Additionalfile2.xlsx
- Additionalfile4.xlsx