narra R

Narrative Review

PROBAST+AI or QUADAS-3? Choosing the appropriate risk-of-bias assessment tool for systematic reviews of AI-based prediction models and diagnostic tests

Muhammad Iqhrammullah[1*], Teuku F. Duta[2] and Eyiuche D. Ezigbo[3]

[1]Postgraduate Program of Public Health, Universitas Muhammadiyah Aceh, Banda Aceh, Indonesia; [2]Faculty of Medicine, Universitas Syiah Kuala, Banda Aceh, Indonesia; [3]Department of Medical Laboratory Sciences, Faculty of Health Sciences and Technology, College of Medicine, University of Nigeria, Enugu, Nigeria

*Corresponding author: m.iqhram@unmuha.ac.id

Abstract

The rapid expansion of artificial intelligence (AI) and machine learning (ML) in healthcare has been accompanied by a growing number of diagnostic systematic reviews and meta-analyses. However, selecting an appropriate risk-of-bias framework remains challenging, particularly when studies combine prediction-model development and validation with diagnostic accuracy evaluation. The aim of this narrative review is to examine the respective scopes of PROBAST+AI and QUADAS-3 and develop practical guidance for selecting and applying these frameworks in systematic reviews of AI-based healthcare research. Two most recently updated appraisal tools used for AI-themed diagnostic systematic reviews were qualitatively assessed, namely PROBAST+AI and QUADAS-3. The assessment includes the scope of both frameworks, which was later used to synthesize practical guidance in selecting and applying them. PROBAST+AI is appropriate when the primary objective is to assess the methodological quality of prediction-model development or the risk of bias and applicability in model evaluation. In contrast, QUADAS-3 is appropriate when the objective is to assess the risk of bias and applicability of diagnostic accuracy estimates from a prespecified AI system evaluated against an appropriate reference standard. The proposed guidance provides a decision pathway for identifying the relevant unit of assessment, distinguishing prediction-model evaluation from diagnostic test-accuracy assessment, and determining when separate assessments using both frameworks are warranted. When a publication includes distinct model-development and diagnostic accuracy components, each component should be assessed using the framework corresponding to its methodological purpose. This guidance provides a conceptual basis for transparent and methodologically appropriate framework selection, although its proposed decision pathway has not yet been empirically validated.

Keywords: Artificial intelligence, machine learning, diagnostic accuracy, systematic review, risk of bias

Copyright: © 2026 by the authors. This is an open access article distributed under the terms and conditions of the CC BY-NC 4.0.

Introduction

The application of artificial intelligence (AI) and machine learning (ML) in healthcare has expanded rapidly, particularly in diagnostic imaging, digital pathology, signal processing, and clinical decision support [1]. This growth has been accompanied by an increasing number of systematic reviews and meta-analyses that aim to summarize the diagnostic performance of AI- and ML-based systems across clinical conditions and data modalities. Recent reviews have

Received: July 11, 2026 - Accepted: October 26, 2026 - Published online: November 10, 2026

evaluated AI-based diagnostic tools in areas ranging from cancer detection and cardiovascular imaging to infectious diseases and emergency care, frequently pooling measures such as sensitivity, specificity, and the area under the receiver operating characteristic curve [2].

As the volume of diagnostic AI evidence continues to increase, rigorous risk-of-bias assessment has become essential for ensuring the credibility and clinical applicability of evidence synthesis [3]. The methodological structure of diagnostic AI studies, however, differs fundamentally from that of conventional diagnostic test-accuracy studies. Unlike conventional diagnostic tests, AI studies frequently combine multiple stages of model development within a single investigation, including feature engineering, model selection, hyperparameter tuning, threshold determination, internal validation, external validation, and diagnostic accuracy evaluation [3]. This hybrid design blurs the traditional boundary between prediction-model research and diagnostic test-accuracy research, making it increasingly difficult for reviewers to determine the most appropriate methodological perspective for risk-of-bias assessment. Reviewers must therefore determine whether the primary object of assessment is a prediction model, a diagnostic index test, or distinct components of a study that serve both purposes.

Historically, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) has been widely used in systematic reviews of diagnostic accuracy studies, including reviews of AI-based systems across different clinical domains [4-6]. Its recently updated version, QUADAS-3, is now recommended for assessing risk of bias and applicability in primary diagnostic test-accuracy studies and introduces several methodological advances, including estimate-level assessment and explicit consideration of synthesis questions [7]. By contrast, many AI studies evaluate multivariable prediction models rather than fully specified and fixed diagnostic tests. For these studies, the Prediction Model Risk of Bias Assessment Tool (PROBAST) [8] and, more recently, PROBAST+AI, provide frameworks specifically designed to assess the risk of bias and applicability of prediction-model development, validation, and evaluation, while incorporating methodological considerations unique to AI-based models [9].

The availability of both PROBAST+AI and QUADAS-3 creates an important methodological question for reviewers of diagnostic AI evidence. An AI model may be regarded as a prediction model during its development and validation, but the same model may later function as an index test when evaluated as a locked model against a reference standard [9]. Therefore, selecting a risk-of-bias tool solely because a study reports diagnostic accuracy, sensitivity, or specificity may be insufficient. Similarly, routinely applying both tools to every study may duplicate judgments, increase assessment burden, and produce inconsistent conclusions because the tools were developed for related but distinct purposes. Framework selection should therefore reflect the methodological purpose of the study or study component, rather than the reported performance measures alone.

Although PROBAST+AI and QUADAS-3 have recently been introduced, practical guidance explaining how reviewers should distinguish between them remains limited [8,9]. Existing publications describe the frameworks individually, but reviewers still need practical guidance on how to select between them when studies combine prediction-model development and diagnostic test-accuracy evaluation [8,9]. In particular, reviewers need guidance on how to identify the appropriate unit of assessment when a publication includes multiple methodological components, determine when one framework adequately addresses the review question, and recognize when separate assessments using both frameworks may be justified. The purpose of this article is not to replace the existing guidance for either framework, but to clarify how their respective scopes can be applied to these situations. This distinction is important for promoting consistent and transparent risk-of-bias assessment across systematic reviews of diagnostic AI.

Therefore, the aim of this review is to provide practical methodological guidance for selecting between PROBAST+AI and QUADAS-3 in systematic reviews and meta-analyses of diagnostic AI and ML studies. Specifically, this review compares the scope and assessment domains of the two tools, clarifies the distinction between an AI prediction model and an AI-based diagnostic index test, identifies circumstances in which either tool should be used alone, and describes situations in which their complementary but separate application may be justified. The guidance is intended primarily for systematic-review authors and evidence-synthesis reviewers assessing diagnostic prediction models and diagnostic test-accuracy evidence, including those with limited experience

with AI prediction-model methodology. It focuses on selecting and applying the two frameworks to diagnostic AI evidence, rather than evaluating all forms of AI research in healthcare. The proposed guidance is conceptual and has not been empirically validated.

This narrative review draws on key publications describing PROBAST+AI and QUADAS-3, together with relevant methodological literature on prediction-model development and evaluation, diagnostic test-accuracy assessment, and evidence synthesis. Relevant publications were identified through the authors' existing knowledge of the field and by examining references cited in key methodological publications. No formal database search, predefined eligibility criteria, or systematic study-selection process was undertaken. The literature was critically examined to compare the intended scope and assessment targets of the two frameworks and to identify methodological situations in which their application may differ or complement one another. The resulting recommendations represent a conceptual methodological synthesis and have not been empirically validated.

Conceptual foundations: Diagnostic prediction models and diagnostic index tests

Diagnostic AI studies are commonly described as either prediction-model studies or diagnostic test-accuracy studies. Although these terms are sometimes used interchangeably, they represent different methodological perspectives and should not be considered synonymous [7,9]. Understanding this distinction is fundamental because it determines the object of evaluation and therefore the most suitable framework for assessing risk of bias.

Diagnostic prediction models combine multiple pieces of information to estimate the probability that a target condition is present in an individual [10]. In diagnostic settings, predictors may include demographic characteristics, symptoms, laboratory results, imaging features, physiological signals, or other routinely collected clinical data. The output may be a continuous probability, a risk category, or a binary classification after application of a threshold [11]. Methodological assessment therefore extends beyond the final classification result and includes how predictors were defined, how missing data were handled, how features were selected, how model complexity was controlled, and whether performance was assessed using data that were independent of model fitting [8]. Diagnostic prediction-model studies are primarily concerned with the development, validation, transportability, and potential clinical usefulness of a multivariable model [9,11].

By contrast, once an AI system is evaluated as a clinical test rather than as a model under development, the methodological focus changes [9]. An AI-based diagnostic index test is evaluated as a test procedure applied to individuals for the purpose of identifying a predefined target condition [12]. At this stage, the AI system should generally be sufficiently specified before testing, including its input requirements, preprocessing steps, model parameters, output interpretation, and decision threshold. Its results are then compared with an appropriate reference standard to determine whether the target condition is present or absent [7,12]. The main concern is not how the AI system was originally constructed, but whether its diagnostic accuracy was estimated without bias in a population representative of the intended clinical setting. Such evaluations are therefore conceptually similar to studies of laboratory assays, imaging tests, or clinical examination procedures, even when the index test is generated by a complex computational system [12].

In practice, however, this distinction is not always straightforward. Many AI and ML studies combine model construction and diagnostic evaluation within a single analytical workflow [9]. A study may train several classifiers, optimize hyperparameters, select the best-performing model, determine a classification threshold, and report sensitivity and specificity using the same dataset partitioning strategy. Although the reported outcomes resemble those of a diagnostic-accuracy study, the evaluated system may still be under development instead of functioning as a prespecified index test [9]. A prespecified AI system is not defined solely by a fixed model architecture. The model version and parameters, preprocessing pipeline, and decision threshold or classification rule should also be established before the diagnostic evaluation to which the results refer. When the architecture is fixed but the decision threshold is selected or adjusted

Iqhramullah et al. Narra Review 2026; 2 (3): e41 - http://doi.org/10.52225/narrarev.v2i3.41

using the evaluation data, the system has not been fully prespecified for that evaluation. Reviewers should examine whether threshold selection used separate development data or the evaluation dataset itself, because data-driven threshold selection on the evaluation dataset can introduce optimism into the reported diagnostic accuracy estimates.

An externally evaluated AI model may remain a prediction model in methodological terms, particularly when the review question concerns calibration, generalizability, or performance across populations [11,13]. The type of performance metric reported does not, by itself, determine whether a study should be treated as prediction-model research or diagnostic-test evaluation.

The position of the study within the model-development pathway and the intended role of the AI system in clinical practice are therefore central to classification. Early-stage studies usually focus on feature engineering, model selection, tuning, and internal validation, whereas later-stage studies assess performance in new institutions, populations, or prospective clinical pathways. The same AI system may also change methodological roles over time: during development it is evaluated as a prediction model, whereas after being finalized it may be evaluated as a diagnostic index test. Review authors should consider whether the system was fixed before evaluation, whether any model modification occurred using the evaluation data, whether the study aimed to assess model validity or test accuracy, and whether the AI output was intended to support, replace, or complement an existing diagnostic process. These considerations establish the conceptual basis for selecting the appropriate risk-of-bias framework in diagnostic AI reviews.

The conceptual distinction between prediction models and diagnostic index tests provides the foundation for selecting an appropriate risk-of-bias framework. This relationship across the clinical AI evidence pathway is presented in **Figure 1**, showing how the methodological role of an AI system determines the appropriate framework for its assessment.

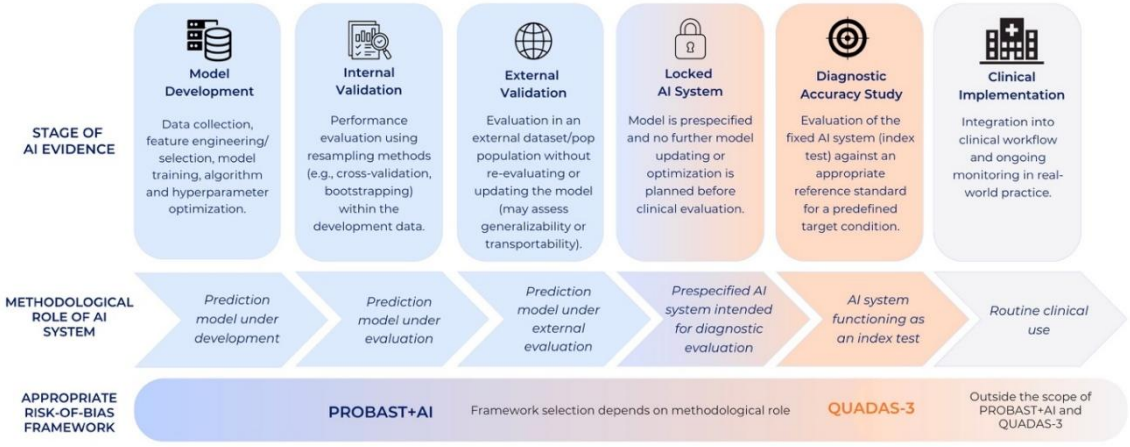


Figure 1. Conceptual framework illustrating the relationship between the clinical AI evidence pathway, the evolving methodological role of AI systems, and the corresponding risk-of-bias assessment framework. The boundaries between stages are conceptual and may overlap in practice. Framework selection should be based on the methodological role of the AI system rather than solely on its chronological stage of model development. Routine clinical implementation lies beyond the intended scope of both PROBAST+AI and QUADAS-3.

Overview of PROBAST+AI

PROBAST+AI is an updated framework for assessing prediction models used in healthcare, regardless of whether they are developed using conventional regression or AI and ML methods. Its scope includes diagnostic and prognostic prediction models developed, updated, validated, or evaluated using clinical variables, medical images, physiological signals, biomarkers, or other health data. The framework is therefore appropriate when the primary object of assessment is the prediction model and the credibility of its reported performance, rather than the diagnostic accuracy of a prespecified clinical test [9].

PROBAST+AI comprises separate components for model development and model evaluation, reflecting the principle that the methodological quality of model development and the risk of bias in model evaluation address related but distinct concerns [9]. Although both

components are organized into four domains—participants and data sources, predictors, outcome, and analysis—they apply different signaling questions according to the study objective. This distinction is particularly important in AI research because methodological limitations introduced during model development, such as inappropriate predictor handling, overfitting, or inadequate validation, may compromise the model's credibility. Model evaluation raises an additional question: whether the reported performance estimates are likely to reflect performance in the intended target population, given the evaluation design, data sources, and analytical methods used. A well-developed model may still produce biased performance estimates if its evaluation is conducted using inappropriate analytical methods or non-representative data [8,9].

Applicability concerns are assessed in relation to the participants and data sources, predictors, and outcome [9]. These judgments indicate how closely the evaluated model corresponds to the systematic review question and its intended clinical context. A study may demonstrate sound methodological conduct yet have limited applicability if its population differs substantially from the target population, if the required predictors are unavailable at the intended point of use, or if the predicted outcome does not correspond to the review question. Applicability should therefore be judged against the intended use of the model rather than regarded as an intrinsic characteristic of the model itself.

PROBAST+AI assessments should be reported at the model level, with separate assessments for model development and model evaluation when both components are present [9]. Models that differ substantially in their predictors, development process, validation strategy, or analytical approach may require distinct judgments. In reporting these assessments, reviewers should distinguish concerns about the conduct of model development from concerns about bias in the evaluation of model performance, while interpreting both in relation to the review question and intended use of the model. An illustrative reporting format is presented in **Table 1**.

A major strength of PROBAST+AI is its explicit coverage of the complete modeling process, including data selection, predictor measurement, model fitting, tuning, validation, and performance assessment [6]. This makes the framework particularly suitable for diagnostic AI studies in which prediction models are developed and evaluated within the same study. Its separation of model development from model evaluation also allows reviewers to distinguish limitations arising from how a model was created from those related to how its performance was estimated. However, the framework requires detailed methodological reporting, which is frequently incomplete in AI studies and may therefore result in unclear judgments despite otherwise well-conducted research. This limitation reflects incomplete reporting rather than a weakness of the framework itself. For this reason, PROBAST+AI is most informative when the review aims to evaluate the methodological credibility and transportability of the prediction model rather than the diagnostic accuracy of a finalized clinical test.

Overview of QUADAS-3

QUADAS-3 is the current version of the Quality Assessment of Diagnostic Accuracy Studies tool and is intended for systematic reviews evaluating the accuracy of tests used for diagnosis, screening, or disease staging [7]. It is designed for primary studies in which one or more index tests are compared with a method used to determine the presence or absence of a target condition. In diagnostic AI reviews, QUADAS-3 is most applicable when an AI or ML system is evaluated as a sufficiently specified diagnostic index test, and the review aims to synthesize accuracy estimates such as sensitivity and specificity. The framework may accommodate prospective or retrospective designs, single or comparative index-test studies, and studies contributing one or more accuracy estimates relevant to a predefined synthesis question.

Application of QUADAS-3 requires reviewers to define the index test, target condition, and method used to establish that condition. The index test refers to the AI-based procedure being evaluated, including its inputs, preprocessing pipeline, output, and decision rule where relevant [7]. The target condition is the disease state or clinical classification that the index test is intended to identify, whereas the reference standard comprises the procedures used to determine whether that condition is present. In AI studies, these elements should be specified carefully because the

reference label may derive from histopathology, imaging interpretation, clinical follow-up, expert consensus, or a combination of sources.

The reference standard is important because it determines the disease-status classification against which the AI system is evaluated. If the reference standard is inaccurate, inconsistently applied, or influenced by the index-test result, patients may be misclassified, potentially biasing estimates of sensitivity and specificity. Reviewers should therefore consider how the reference standard was established and applied, whether it was appropriate for the target condition, and whether the AI index-test result could have influenced the reference classification. The same AI model may also produce multiple outputs or be evaluated at several thresholds; therefore, QUADAS-3 requires reviewers to identify the particular accuracy estimate being assessed rather than assuming that one study necessarily receives a single uniform judgment [7].

Risk of bias is assessed across four domains: participants, index test, target condition, and analysis [7]. These domains evaluate whether the reported diagnostic accuracy estimate is likely to be unbiased for the predefined synthesis question. Applicability is assessed according to the correspondence between the participants, index test, and target condition included in the study and those specified in the review question. Unlike previous versions, QUADAS-3 also requires reviewers to define the characteristics of an ideal test-accuracy study before making domain-level judgments, thereby providing an explicit reference for risk-of-bias assessment.

QUADAS-3 assessments are reported at the level of the diagnostic accuracy estimate rather than the publication itself [7]. A single study evaluating one prespecified AI system at one clinically relevant threshold may contribute one assessment, whereas studies reporting multiple index tests, thresholds, participant groups, or target conditions may require separate assessments for each estimate included in the review. Judgments should remain domain-based and should not be converted into numerical quality scores. An illustrative reporting format is presented in **Table 2**.

A principal strength of QUADAS-3 is its close alignment with the structure of diagnostic-test-accuracy reviews and meta-analyses [7]. Its estimate-level approach is particularly valuable when a study contributes multiple clinically distinct accuracy estimates, while the requirement to define an ideal test-accuracy study helps ensure that judgments are directly linked to the review question. However, QUADAS-3 does not evaluate how an AI model was originally developed, optimized, or internally validated. Consequently, methodological issues such as feature selection, hyperparameter tuning, overfitting, data leakage, and model transportability may remain insufficiently addressed when QUADAS-3 is applied in isolation. For this reason, QUADAS-3 is most informative when the review focuses on the diagnostic accuracy of a prespecified AI system rather than on the methodological robustness of the prediction model itself.

Comparison between PROBAST+AI and QUADAS-3

Intended study type and unit of assessment

The fundamental distinction between PROBAST+AI and QUADAS-3 lies not in their domain structure but in the methodological object they are designed to evaluate. PROBAST+AI is intended for diagnostic and prognostic prediction models, where the primary objective is to assess the methodological quality of model development and the credibility of model evaluation. It therefore applies to studies that combine multiple predictors using regression, ML, or other AI methods to generate individualized predictions or classifications [8,9]. In contrast, QUADAS-3 is designed for diagnostic test-accuracy studies, in which a prespecified index test is evaluated against a predefined target condition using an appropriate reference standard. In this context, the AI system is regarded as a fixed diagnostic procedure rather than a model undergoing development [7].

This distinction directly determines the unit of assessment. Under PROBAST+AI, the prediction model is the principal unit of assessment, and separate judgments may be required when models differ in predictor sets, development strategies, analytical procedures, or validation approaches. Furthermore, model development and model evaluation are assessed independently because they address distinct methodological questions [9].

Table 1. Illustrative reporting of PROBAST+AI assessments

Study	Prediction model(s)	Assessment component	Participants and data sources	Predictors	Outcome	Analysis	Overall judgment	Applicability concern
Study A	Random forest, support vector machine, and XGBoost	Model development	Low concern	Low concern	Low concern	Major concerns	Low model quality	Low
Study A	Ensemble model	Model development	Low concern	Low concern	Low concern	Some concerns	Unclear model quality	Low
Study A	Externally validated ensemble model	Model evaluation	Low risk	Low risk	Low risk	Low risk	Low risk of bias	Low
Study B	Convolutional neural network models 1–3	Model development	Some concerns	Low concern	Low concern	Major concerns	Low model quality	Some concerns
Study C	Random forest	Model evaluation	Low risk	Low risk	Some concerns	Low risk	Some concerns	Low
Study D	Deep-learning classifier	Model evaluation	High risk	Low risk	Low risk	High risk	High risk of bias	High

PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for artificial intelligence

Table 2. Illustrative reporting of QUADAS-3 assessments

Study	Index test or accuracy estimate	Participants	Index test	Target condition	Analysis	Overall risk of bias	Applicability concern
Study A	Locked CNN at prespecified threshold	Low	Low	Low	Low	Low	Low
Study B	Random forest, threshold selected in evaluation data	Low	High	Low	High	High	Low
Study C	AI system for Disease X versus healthy controls	High	Low	Low	Some concerns	High	High
Study C	AI system for Disease X versus alternative diagnoses	Some concerns	Low	Low	Some concerns	Some concerns	Low
Study D	Commercial AI system with indeterminate results excluded	Low	Low	Low	High	High	Some concerns

AI: artificial intelligence; CNN: convolutional neural network

By contrast, QUADAS-3 evaluates the diagnostic accuracy estimate relevant to the predefined synthesis question rather than the publication or AI model itself [7]. A single study may require multiple QUADAS-3 assessments when it reports different index tests, diagnostic thresholds, participant groups, target conditions, or other clinically distinct accuracy estimates.

Importantly, neither framework assumes a one-to-one relationship between a publication and a risk-of-bias assessment. Instead, the number of assessments depends on the methodological object contributing to the evidence synthesis. A single publication may therefore contribute multiple prediction models requiring separate PROBAST+AI assessments or multiple diagnostic accuracy estimates requiring separate QUADAS-3 assessments. Before evaluating individual domains or signaling questions, reviewers should first determine whether the methodological object is a prediction model or a diagnostic accuracy estimate. This principle provides the conceptual foundation for the domain-level comparisons presented in the following sections.

Assessment domains and shared methodological concepts

Although PROBAST+AI and QUADAS-3 both organize their assessments into four methodological domains, similarities in terminology should not be interpreted as conceptual equivalence. Rather than addressing identical sources of bias, the two frameworks address different methodological questions because they were developed for different study designs and different objects of assessment [7,9]. Hence, domains with similar names often serve distinct purposes, and researchers should avoid direct one-to-one correspondence between the two frameworks.

The participants domain illustrates this distinction. In PROBAST+AI, participants and data sources are assessed to determine whether the study population is appropriate for developing or evaluating a prediction model and whether the resulting model is likely to be applicable to the intended target population [9]. In QUADAS-3, the participants domain instead evaluates whether participant selection permits an unbiased estimate of diagnostic accuracy for the predefined synthesis question [7]. Although both domains concern study populations, one focuses on the validity and transportability of a prediction model, whereas the other focuses on the validity of a diagnostic accuracy estimate.

This distinction is illustrated by studies that report both model development and validation. Gulshan *et al.* (2016), for example, developed a deep-learning algorithm for detecting diabetic retinopathy and evaluated its performance using independent validation datasets [14]. Similarly, Wu *et al.* (2024) developed an AI-based model for detecting urothelial carcinoma using urine cytology images and conducted prospective validation in a multicentre diagnostic study [15]. For systematic reviews of prediction-model methodology, PROBAST+AI may be used to assess the development and evaluation components. When the review question specifically concerns diagnostic accuracy, the diagnostic evaluation component may also be assessed using QUADAS-3, provided that the study meets the framework's intended scope. Thus, the appropriate assessment depends on the review question and the component being assessed, rather than on the publication as a whole.

A similar conceptual difference exists between predictors and outcomes in PROBAST+AI and the index test and target condition in QUADAS-3. In PROBAST+AI, predictors are the individual variables used to construct or apply a multivariable prediction model, whereas the outcome represents the condition the model is intended to predict [6]. In QUADAS-3, by contrast, the index test refers to the complete diagnostic procedure—including the finalized AI system, preprocessing pipeline, decision threshold, and interpretation process—while the target condition represents the disease status established using an appropriate reference standard [4]. The entire AI system may therefore constitute the index test in QUADAS-3, whereas its individual inputs remain predictors in PROBAST+AI.

Abràmoff *et al.* (2018) provide an example of an AI system evaluated as a diagnostic index test [16]. Their prospective pivotal trial evaluated an autonomous AI-based system for detecting diabetic retinopathy that had been locked before the trial began, using a reference assessment of retinal disease. This study illustrates the relevance of QUADAS-3 when the review focuses on the diagnostic accuracy of a sufficiently specified AI system. The index test should be understood as

the complete diagnostic procedure, not the algorithm architecture alone. The model version, preprocessing procedures, decision threshold, and interpretation rules are therefore relevant when determining whether the evaluated index test was adequately prespecified.

The analysis domain provides the greatest apparent overlap between the two frameworks, yet their methodological emphasis differs substantially. PROBAST+AI focuses on analytical decisions that influence model development and validation, including predictor handling, feature selection, hyperparameter tuning, overfitting, missing-data management, internal validation, and performance estimation [9]. QUADAS-3 instead evaluates whether diagnostic accuracy was estimated without bias, considering issues such as prespecification of the index test, participant exclusions, handling of indeterminate results, and the appropriateness of statistical analyses [7]. Similar domain labels should therefore not be interpreted as methodological redundancy but rather as reflecting different perspectives on study validity.

The distinction is particularly relevant when the evaluated model or its operating threshold changes during a study. Kelly *et al.* (2026) reported retrospective and prospective technical-feasibility evaluations of AI for breast cancer screening, including adjustment of the operating threshold during prospective evaluation [17]. This example highlights the need to distinguish changes to the model itself from changes to the threshold used to classify results. If the model parameters remain unchanged but the threshold is selected or adjusted using evaluation data, reviewers should consider how this decision may affect the credibility of the reported performance estimates and whether the index test was prespecified for the relevant analysis. Threshold adjustment alone does not establish that the model was retrained or that one framework must replace the other. The assessment should instead reflect the study component and the review question.

The same principles apply when a publication compares multiple AI models or combines model development with prospective diagnostic evaluation. For example, Maiter *et al.* (2026) conducted an independent head-to-head comparison of seven commercial AI devices for lung cancer detection using the same set of chest radiographs [18]. In studies comparing several models, reviewers should identify the models and their respective evaluation procedures as the relevant assessment units, while considering whether model development or tuning could compromise the reported comparison. When model development and prospective diagnostic evaluation are reported in one publication, distinct components may require different assessments: PROBAST+AI can address model development and prediction-model evaluation, whereas QUADAS-3 can address diagnostic accuracy when the evaluation falls within its intended scope.

These comparisons demonstrate that apparent similarities between PROBAST+AI and QUADAS-3 arise because both frameworks aim to support trustworthy clinical evidence. However, they address different aspects of study validity. PROBAST+AI assesses the credibility of a prediction model throughout its development and evaluation, whereas QUADAS-3 assesses the credibility of a diagnostic accuracy estimate generated from a prespecified index test. When a publication contains multiple methodological components, the two frameworks may be applied to different components when justified by the review question. Their use should not be determined solely by the AI label, algorithm type, or performance metrics reported. These differences provide the conceptual basis for understanding why the two frameworks may legitimately produce different judgments for the same AI study. The practical implications for framework selection are summarized in **Table 3**.

Table 3. Practical guide for selecting a risk-of-bias assessment framework in diagnostic AI reviews

Review question or methodological component	Recommended framework	Practical considerations
Model development, updating, or evaluation of predictive performance is the primary focus.	PROBAST+AI	Assess model-development quality and/or risk of bias in model evaluation, as appropriate to the study component and review question.
Diagnostic accuracy of a sufficiently specified AI system against an appropriate reference standard is the primary focus.	QUADAS-3	Define the index test, target condition, reference standard, and accuracy estimate relevant to the synthesis question.

Review question or methodological component	Recommended framework	Practical considerations
A publication includes both model development and a distinct diagnostic accuracy evaluation.	Consider separate assessments using PROBAST+AI and QUADAS-3.	Assess each component according to its methodological purpose. Apply both frameworks only when each component falls within the relevant framework's scope.
A study compares multiple prediction models or reports multiple diagnostic accuracy estimates.	Apply the relevant framework to each eligible model or estimate, as appropriate.	Distinguish models or estimates when differences in development procedures, thresholds, target conditions, participant groups, or analytical methods affect the assessment.
The review primarily addresses clinical effectiveness, implementation, fairness, or other outcomes beyond prediction-model validity and diagnostic accuracy.	Consider another or an additional assessment framework.	PROBAST+AI and QUADAS-3 may not adequately address the full review question.

PROBAST+AI: Prediction Model Risk of Bias Assessment Tool and artificial intelligence; QUADAS-3: Quality Assessment of Diagnostic Accuracy Studies, version 3

Model development, validation, and fixed-test evaluation

The distinction between PROBAST+AI and QUADAS-3 becomes particularly important when AI studies progress through successive stages of model development and clinical evaluation. Although these stages are often presented as a linear pathway, the appropriate methodological perspective is determined not simply by the stage of the study but by the methodological role of the AI system at that stage [7,9,11].

During model development, the AI system remains a prediction model undergoing construction, optimization, or refinement. Activities such as feature engineering, predictor selection, hyperparameter tuning, threshold optimization, and internal validation are intended to improve model performance before clinical implementation [9,11]. Because these processes directly influence the methodological credibility and transportability of the resulting model, the primary object of assessment remains the prediction model itself.

External validation represents a transitional stage within the AI evidence pathway and should not automatically be regarded as diagnostic test-accuracy research. Many external validation studies continue to evaluate the generalizability, calibration, or transportability of a prediction model without treating the AI system as a finalized diagnostic test [8,9,11]. The methodological object therefore remains the prediction model, even when evaluation is performed in independent datasets or different clinical settings, provided that the AI system continues to function as a prediction model rather than as a prespecified diagnostic index test.

The methodological perspective changes once the AI system is prespecified and evaluated as a fixed diagnostic procedure against an appropriate reference standard in the intended clinical setting. At this stage, the AI system functions as an index test, and the primary objective shifts from evaluating model validity to obtaining an unbiased estimate of diagnostic accuracy [7]. The focus of assessment shifts from the methodological credibility of the prediction model to the validity of the diagnostic accuracy estimate generated by the index test.

This transition does not necessarily occur at a single point in the AI evidence pathway. A model may undergo repeated external evaluations while remaining a prediction model, whereas another model may become a diagnostic index test immediately after being finalized for prospective clinical evaluation. Reviewers should not rely solely on study labels such as development, validation, or diagnostic accuracy. Instead, framework selection should be guided by how the AI system functions at the time of evaluation. Together, these considerations explain why PROBAST+AI and QUADAS-3 may legitimately produce different judgments for the same study.

Complementary and divergent judgments

Although PROBAST+AI and QUADAS-3 are sometimes viewed as alternative risk-of-bias frameworks, they should not be expected to produce identical judgments because they evaluate different methodological objects. Apparent disagreement between the two frameworks therefore does not necessarily indicate inconsistency; instead, it reflects differences in the methodological aspects of the study being assessed [7,9].

For example, an AI model may be developed using inadequate sample size, inappropriate predictor selection, extensive hyperparameter optimization without appropriate validation, or other practices that increase the risk of overfitting. Even if the finalized model is subsequently evaluated prospectively using an appropriate reference standard and a rigorous diagnostic study design, PROBAST+AI may identify important methodological concerns related to model development, whereas QUADAS-3 may judge the diagnostic accuracy study to have a low risk of bias because the evaluation itself was conducted appropriately [7-9].

In contrast, a prediction model may be developed and externally validated using robust methodological practices, resulting in a favorable PROBAST+AI assessment. However, if the same model is later evaluated as a diagnostic test in a study with inappropriate participant selection, non-prespecified index-test interpretation, an inadequate reference standard, or biased statistical analyses, QUADAS-3 may identify a high risk of bias despite the underlying model having been developed appropriately [7,9].

These examples illustrate that PROBAST+AI and QUADAS-3 frequently provide complementary rather than conflicting information. PROBAST+AI evaluates whether the prediction model was developed and evaluated in a methodologically robust manner, whereas QUADAS-3 evaluates whether the reported diagnostic accuracy estimate represents an unbiased assessment of the model's clinical performance. Divergent judgments may legitimately coexist because they reflect different sources of bias arising at different stages of the AI evidence pathway.

Rather than attempting to reconcile apparently different judgments, reviewers should interpret each framework according to the methodological question it is intended to address. Different risk-of-bias judgments should not be interpreted as contradictory but as reflecting evaluation of different methodological objects within the same study. Together, PROBAST+AI and QUADAS-3 provide a more comprehensive evaluation of diagnostic AI studies by distinguishing limitations originating from model development from those arising during diagnostic test evaluation.

When both frameworks are applied to different components of the same publication, reviewers should retain separate, framework-specific judgments and document the methodological reasons underlying each judgment. Differences should be interpreted in relation to the object of assessment, the review question, and the evidence available for the relevant study component. Reviewers should not average, combine, or force the judgments into a single overall rating, because the two frameworks assess different methodological questions.

For systematic reviews, risk-of-bias assessments should preferably be conducted independently by at least two reviewers for each relevant study component or diagnostic accuracy estimate. Disagreements should first be discussed and resolved by reference to the study methods and the relevant signaling questions. If consensus cannot be reached, a third reviewer may adjudicate. Reviewers should document the rationale for consequential judgments, particularly when the frameworks yield different conclusions or when the same publication requires separate assessments. This process supports transparent and reproducible decision-making without implying that the judgments from the two frameworks must agree.

AI-specific methodological issues across both frameworks

AI introduces methodological challenges that extend across multiple stages of the evidence pathway, from model development to clinical evaluation. Issues such as data leakage, overfitting, dataset shift, and threshold optimization can influence the validity, reproducibility, and generalizability of AI-generated evidence, highlighting the need for careful methodological appraisal rather than reliance on reported performance alone [9,12,19]. These challenges are not unique to either PROBAST+AI or QUADAS-3. Instead, they illustrate how the same methodological issue may assume different importance depending on the methodological objective of the study.

The differing perspectives of the two frameworks become apparent when the same methodological issue is considered in different study contexts. During prediction-model development and validation, concerns such as data leakage and overfitting primarily threaten the credibility of reported model performance because they may produce overly optimistic estimates before the model is evaluated in clinical practice [9,11,20]. Once an AI system is assessed as a

prespecified diagnostic test, however, attention shifts towards the validity of diagnostic accuracy estimates. At this stage, participant selection, the quality of the reference standard, and the conduct and interpretation of the index test become the principal considerations because they determine whether reported diagnostic performance is likely to reflect clinical performance under real-world conditions [7,12,21]. Although the methodological issue itself may remain unchanged, its implications depend on the type of evidence being evaluated.

External validation illustrates this principle particularly well because it is frequently regarded as equally relevant to both frameworks. Although this perception is understandable, it overlooks an important methodological distinction. In prediction-model research, external validation examines whether a model maintains its performance in an independent population, with the primary emphasis placed on model transportability and generalizability [3,9]. In diagnostic accuracy studies, a similar study design evaluates the performance of a locked AI system against an appropriate reference standard in a new clinical setting, where the objective is to obtain an unbiased estimate of diagnostic accuracy [7,12,21]. The study design may appear similar, but the methodological question being addressed differs, leading to different appraisal priorities.

These examples show that AI-specific methodological issues should be interpreted within the methodological context in which they arise rather than as framework-specific concepts. Their importance depends on the type of evidence being evaluated and the methodological question the study seeks to answer. This perspective also explains why framework selection should be guided by the methodological objective of the investigation rather than by the presence of AI alone. The following section builds on this principle by translating it into practical recommendations for selecting the most appropriate risk-of-bias framework.

Selecting and applying risk-of-bias frameworks in diagnostic AI reviews

General principles of framework selection

The selection of an appropriate risk-of-bias framework should begin with the systematic review question rather than the characteristics of the AI algorithm or the reported performance metrics. Before selecting a framework, reviewers should identify the methodological object being synthesized, the intended purpose of the included studies, and the clinical role of the AI system. These considerations determine whether the primary focus is the credibility of a prediction model, the validity of a diagnostic accuracy estimate, or another form of clinical evaluation.

Several commonly reported study characteristics should not be used as the primary basis for framework selection. Terms such as AI, ML, deep learning, or diagnostic model do not determine whether a study should be evaluated using PROBAST+AI or QUADAS-3. Likewise, the type of algorithm (e.g., neural network, support vector machine, or random forest) and the reported performance metrics (e.g., sensitivity, specificity, area under the receiver operating characteristic curve, or F1 score) may appear in both prediction-model studies and diagnostic accuracy studies. These characteristics, although informative, provide limited guidance for framework selection when considered in isolation.

To make framework selection transparent and reproducible, reviewers should follow a structured decision process. First, they should define the review question, including the intended clinical use of the AI system and the type of evidence to be synthesized. Second, they should identify the methodological object of assessment: a prediction model undergoing development, updating, or evaluation; an AI system functioning as a sufficiently specified diagnostic index test; or distinct components of a publication that serve both purposes. Third, each eligible study or study component should be classified according to its methodological purpose and the role of the AI system, based on the reported methods rather than the terminology used by the study authors. Fourth, reviewers should select the framework that corresponds to the assessment objective. PROBAST+AI is appropriate when the review concerns model-development quality or the risk of bias and applicability of prediction-model evaluation. QUADAS-3 is appropriate when the review concerns the risk of bias and applicability of a diagnostic accuracy estimate generated by an AI index test against an appropriate reference standard. When a publication contains distinct model-

development and diagnostic-accuracy components, separate assessments using both frameworks may be justified if each component falls within the relevant framework's scope. Similarly, multiple models or accuracy estimates should be distinguished when their methodological characteristics are relevant to the review question.

Finally, reviewers should document the rationale for each framework selection, including the assessment target, the methodological purpose of the study component, and any relevant limitations or uncertainties. If neither framework adequately addresses the primary review objective, reviewers should consider a design-appropriate or complementary assessment tool rather than forcing the study into an unsuitable framework. The resulting decisions should be reported transparently so that other reviewers can understand and reproduce the selection process. This decision process is summarized in **Figure 2**.

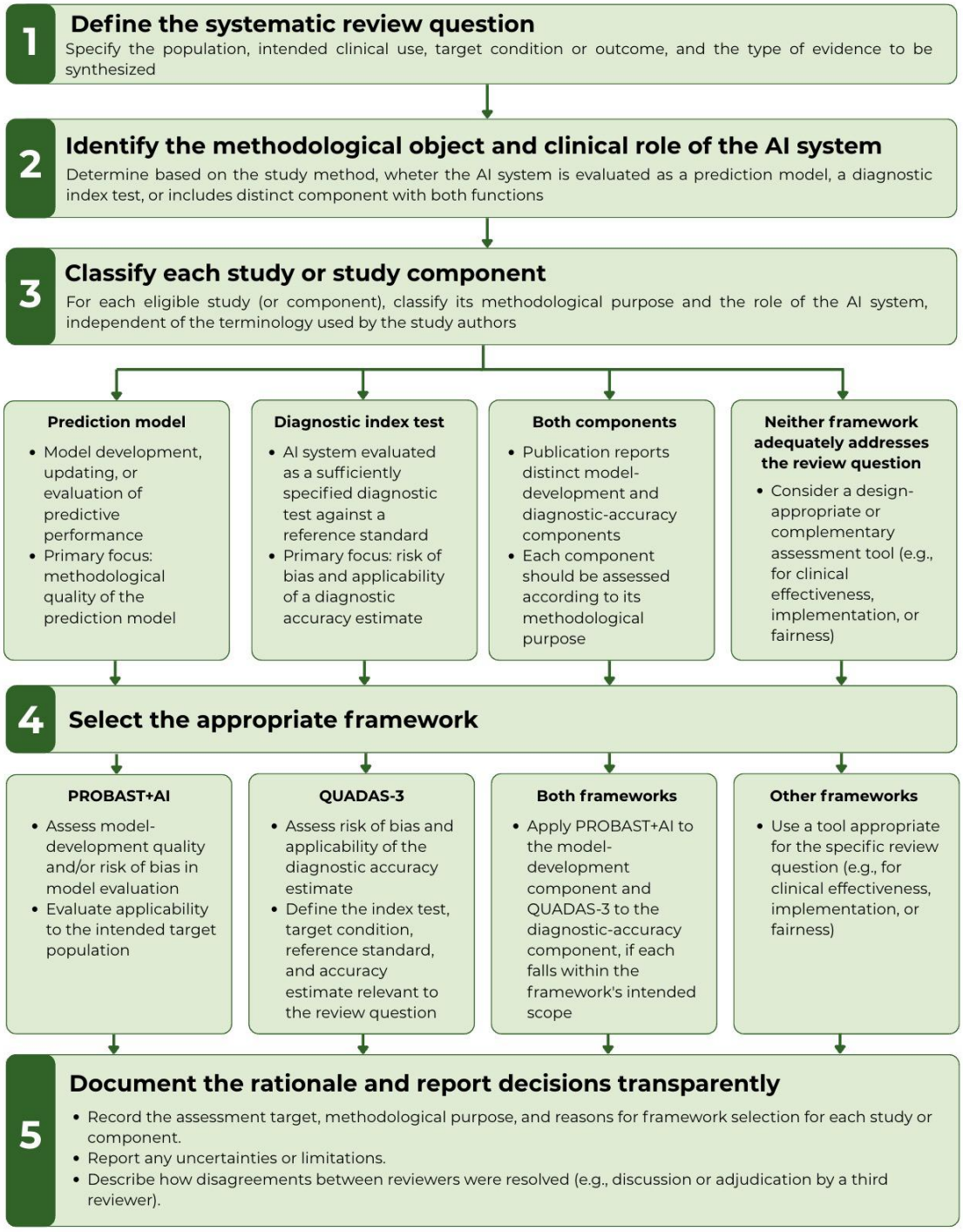


Figure 2. Conceptual decision pathway for selecting a risk-of-bias assessment framework in systematic reviews of diagnostic AI.

When to use PROBAST+AI alone

PROBAST+AI should be selected when the primary objective of the systematic review is to evaluate the methodological quality and credibility of diagnostic prediction models rather than the diagnostic accuracy of a fixed clinical test. It is therefore appropriate for reviews of studies involving model development, updating, internal or external validation, or comparative evaluation of prediction models developed using statistical or machine-learning approaches [8,9].

Studies involving model development, optimization, recalibration, or validation before clinical implementation generally fall within the scope of PROBAST+AI. Although these studies frequently report diagnostic performance measures such as sensitivity, specificity, or the area under the receiver operating characteristic curve, these metrics alone do not indicate that the AI system is being evaluated as a prespecified diagnostic index test. For instance, Idroes *et al.* (2024) developed an explainable LightGBM model to predict differentiated thyroid cancer recurrence using clinical and pathological features [22]. Because the study focused on predicting a clinical outcome rather than evaluating a fixed diagnostic test against a reference standard, it illustrates the type of research for which PROBAST+AI is more appropriate. The presence of diagnostic accuracy measures should not, by itself, justify selecting QUADAS-3.

When to use QUADAS-3 alone

QUADAS-3 should be selected when the primary objective of the systematic review is to evaluate the diagnostic accuracy of a prespecified AI-based index test. It is therefore appropriate for diagnostic test-accuracy reviews in which a finalized AI system is evaluated against an appropriate reference standard under clinically relevant conditions [7].

Typical examples include prospective or external clinical evaluations of locked AI systems, commercially available diagnostic software, or established AI tools assessed alongside clinicians or existing diagnostic tests. In these studies, the primary concern is whether the reported diagnostic accuracy estimates are unbiased, rather than how the model was originally developed. A detailed evaluation of the model development process is generally outside the scope of the review. For instance, Abràmoff *et al.* (2018) evaluated an autonomous AI-based diagnostic system for diabetic retinopathy in a prospective pivotal trial, comparing its diagnostic performance with grading by an independent reading center [16]. This illustrates a setting in which QUADAS-3 would be appropriate when the review focuses on the accuracy of the finalized system rather than its development.

When both frameworks may be needed

Some systematic reviews may legitimately require both PROBAST+AI and QUADAS-3 because they include methodologically distinct types of evidence. This situation commonly arises when reviews combine prediction-model studies with diagnostic test-accuracy studies evaluating finalized AI systems [7,9]. In such cases, no single framework is sufficient to address all included evidence.

Reviewers should therefore classify studies according to their methodological purpose before selecting the appropriate framework. Studies focused on model development, updating, or validation should be assessed using PROBAST+AI, whereas studies evaluating a prespecified AI system as a diagnostic index test should be assessed using QUADAS-3 [7,9]. In some cases, a single publication may require separate assessments if it reports both model development and a subsequent diagnostic accuracy evaluation.

This approach has been illustrated in previous systematic reviews. Niazai *et al.* (2025) used PROBAST for prognostic and risk-prediction models and QUADAS-2 for diagnostic accuracy studies in cardiovascular medicine [23], while Folasole *et al.* (2025) applied the same two frameworks to prediction-model and diagnostic accuracy studies in cancer imaging [24]. Although these reviews used earlier versions of the tools, they demonstrate the practical value of matching the appraisal framework to the methodological purpose of the included evidence. Accordingly, PROBAST+AI and QUADAS-3 are recommended for contemporary reviews addressing these respective study types.

Applying both frameworks should be regarded as complementary rather than redundant. The objective is not to compare the judgments generated by the two frameworks but to ensure

that each component of the evidence is evaluated using the most appropriate methodological approach.

Situations beyond the scope of both frameworks

Not all AI-related systematic reviews fall within the intended scope of PROBAST+AI or QUADAS-3. Both frameworks were developed to evaluate specific methodological questions, namely, the credibility of prediction models and the validity of diagnostic accuracy estimates. Reviews addressing different research questions may therefore require alternative appraisal tools.

Examples include systematic reviews evaluating the clinical effectiveness of AI-assisted care, implementation strategies, workflow efficiency, resource utilization, or patient outcomes. For instance, a study evaluating whether AI-supported triage reduces emergency department waiting times addresses the effects of implementing an AI system in clinical practice rather than diagnostic accuracy alone [25]. Such studies may therefore require risk-of-bias tools appropriate to intervention or implementation studies. PROBAST+AI or QUADAS-3 may still be appropriate if prediction-model validity or diagnostic accuracy is a separate component of the review.

The selection of an appropriate framework therefore should be guided by the review question and methodological object, not by the presence of AI itself. When neither prediction-model validity nor diagnostic accuracy is the primary focus of the evidence synthesis, reviewers should consider alternative risk-of-bias tools specifically designed for the study designs included in the review.

Applying different frameworks within the same systematic review

Systematic reviews of diagnostic AI increasingly include studies representing different stages of the AI evidence pathway, such as model development, external validation, and prospective diagnostic accuracy evaluation [7-9]. Reviewers should avoid applying a single risk-of-bias framework uniformly across all included studies merely for methodological consistency. Instead, studies should first be classified according to their methodological objective before selecting the most appropriate assessment framework.

When prediction-model studies and diagnostic test-accuracy studies coexist within the same review, each study—or, when appropriate, each relevant study component—should be assessed using the framework corresponding to its methodological role. Separate assessments may therefore be necessary when different components of a single publication address distinct methodological questions.

Risk-of-bias judgments obtained using different frameworks should be reported and interpreted independently rather than combined into a single overall assessment. Because PROBAST+AI and QUADAS-3 evaluate different methodological objects, their judgments are not directly comparable and should not be averaged, merged, or used to derive a unified quality score [7-9]. Instead, reviewers should present framework-specific assessments transparently and discuss how each contributes to the interpretation of the evidence synthesis.

The rationale for framework selection should also be specified a priori in the systematic review protocol and described clearly in the Methods section [21]. When multiple frameworks are used, reviewers should explain how studies were classified, why different frameworks were applied, and how the resulting assessments informed evidence synthesis. Transparent reporting of these decisions facilitates reproducibility and promotes consistent, methodologically appropriate risk-of-bias assessment across heterogeneous diagnostic AI evidence.

Limitations and future directions

Although this review provides practical methodological guidance for selecting risk-of-bias frameworks in diagnostic AI systematic reviews, several limitations should be acknowledged. First, the distinction between a prediction model and a diagnostic index test is not always explicit in published studies. Many AI investigations combine model development, internal validation, external validation, and diagnostic accuracy evaluation within a single publication, while reporting practices often provide insufficient methodological detail to determine whether an AI system should be regarded as a prediction model or as a prespecified diagnostic index test. Reviewers may need to classify individual study components rather than applying a single

framework to the entire publication. The proposed guidance should therefore be interpreted as practical methodological guidance rather than as a prescriptive standard.

Second, the recommendations presented in this review are derived from a conceptual synthesis of the current methodological literature rather than from formal empirical validation or expert consensus. Although the proposed framework is intended to improve consistency and transparency in framework selection, some studies will inevitably continue to exhibit methodological characteristics spanning both prediction-model development and diagnostic test-accuracy evaluation. In such situations, reviewer judgment remains essential, and the proposed guidance should be interpreted alongside the review question, study objectives, and the methodological role of the AI system.

Third, this review focuses specifically on PROBAST+AI and QUADAS-3 because these frameworks currently represent the principal risk-of-bias tools for prediction-model and diagnostic test-accuracy studies involving AI. Other reporting guidelines and methodological frameworks were considered only insofar as they informed the conceptual distinction between these two approaches and therefore fall outside the primary scope of this review.

Future research should prospectively evaluate the usability, reproducibility, and inter-reviewer agreement of the proposed guidance in systematic reviews of diagnostic AI. Formal consensus methods, such as Delphi studies involving methodological experts, together with empirical evaluation across diverse clinical applications, would further strengthen and refine this guidance as AI methodologies and risk-of-bias frameworks continue to evolve.

Conclusion

PROBAST+AI and QUADAS-3 should be regarded as complementary rather than competing risk-of-bias frameworks because they evaluate different methodological objects. The appropriate framework should be selected according to the objective of the systematic review and the methodological role of the AI system, rather than the presence of AI, algorithm type, or reported performance metrics. In reviews including heterogeneous evidence, prediction-model studies and diagnostic test-accuracy studies may legitimately require separate assessments using different frameworks. The proposed guidance, including the conceptual distinction between prediction-model studies and diagnostic test-accuracy studies, is intended to promote more consistent, transparent, and methodologically appropriate risk-of-bias assessment, thereby strengthening the methodological rigor and interpretability of systematic reviews of diagnostic AI.

Acknowledgments

The authors have no additional acknowledgments to declare.

Competing interests

All authors declare that they have no competing interests

Funding

This study received no external funding.

Underlying data

This narrative methodological review did not generate or analyze original datasets.

Declaration of artificial intelligence use

AI-assisted tools were used during manuscript preparation to support language refinement, improve clarity and structure, and help present methodological concepts. The authors critically reviewed and revised all AI-generated suggestions, making the final decisions and interpretations. The authors take full responsibility for the accuracy, integrity, and reliability of the published work.

How to cite

Iqhramullah M, Duta TF, Ezigbo ED. PROBAST+AI or QUADAS-3? Choosing the appropriate risk-of-bias assessment tool for systematic reviews of AI-based prediction models and diagnostic tests. Narra Rev 2026; 2 (3): e41 - http://doi.org/10.52225/narrarev.v2i3.41.

References

1. Takita H, Kabata D, Walston SL, *et al*. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med 2025;8(1):175.
2. Baklola M, Elmahdi RR, Ali S, *et al*. Artificial intelligence in disease diagnostics: A comprehensive narrative review of current advances, applications, and future challenges in healthcare. Ann Med Surg 2025;87(7):4237.
3. Debray TP, Damen JA, Snell KI, *et al*. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.
4. Zhan J, Zhang J, Zhu S, *et al*. Diagnostic performance of ultrasound characteristics-based artificial intelligence models for thyroid nodules: A systematic review and meta-analysis. Front Oncol 2025;15:1614603.
5. Alshammari RD, Haliqa WMA, Majzoub RY. Diagnostic accuracy of AI or DL-enhanced technologies in the diagnosis of diabetic retinopathy: A systematic review. J Med Imaging Interv Radiol 2025;12(1):39.
6. Iqhrammullah M, Abdullah A, Hermansyah, *et al*. Accuracy and interpretability of smartwatch electrocardiogram for early detection of atrial fibrillation: A systematic review and meta-analysis. J Arrhythm 2025;41(3):e70087.
7. Whiting PF, Tomlinson E, Rutjes AW, *et al*. QUADAS-3: A revised tool for the quality assessment of diagnostic test accuracy studies. Ann Intern Med 2026; 179(4):548-555.
8. Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58.
9. Moons KG, Damen JA, Kaul T, *et al*. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025;388:e082505.
10. Moons KG, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for Individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med 2015;162(1):W1-W73.
11. Steyerberg EW, Nieboer D, Debray TP, *et al*. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Stat Med 2019;38(22):4290-4309.
12. Sounderajah V, Guni A, Liu X, *et al*. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med 2025;31(10):3283-3289.
13. Collins GS, Moons KG, Dhiman P, *et al*. TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378.
14. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402-2410.
15. Wu S, Shen R, Hong G, *et al*. Development and validation of an artificial intelligence-based model for detecting urothelial carcinoma using urine cytology images: A multicentre, diagnostic study with prospective validation. EClinicalMedicine 2024;71:102566.
16. Abràmoff MD, Lavin PT, Birch M, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1(1):39.
17. Kelly CJ, Wilson M, Warren LM, *et al*. Diagnostic accuracy, fairness and clinical implementation of AI for breast cancer screening: Results of multicenter retrospective and prospective technical feasibility studies. Nat Cancer 2026;7(3):494-506.
18. Maiter A, Taylor J, Metherall P, *et al*. Independent head-to-head comparison of commercial artificial intelligence devices for lung cancer detection on chest radiographs. Radiology 2026;319(2):e252205.
19. Finlayson SG, Subbaswamy A, Singh K, *et al*. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283-286.
20. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020;323(4):305-306.
21. Deeks JJ, Bossuyt PM, Leeflang MM, *et al*. Cochrane handbook for systematic reviews of diagnostic test accuracy. Chichester: John Wiley & Sons; 2023.
22. Idroes GM, Noviandy TR, Idroes GM, *et al*. Prognostication of differentiated thyroid cancer recurrence: An explainable machine learning approach. Narra X 2024;2(3):e183-e183.

23. Niazai A, Jamil H, Hameed M, *et al*. Artificial intelligence in cardiovascular diagnostics: A systematic review and descriptive analysis of clinical applications and diagnostic performance. BMC Cardiovasc Disord 2025;25(1):849.
24. Folasole A, Noah GU, Akangbe B, *et al*. Leveraging machine learning and artificial intelligence in cancer diagnostics imaging: A systematic review. Cureus 2025;17(12).
25. Boğa E. Impact of artificial intelligence-supported triage systems on emergency department management: A comparison of infermedica, emergency severity index, and Manchester triage system. West J Emerg Med 2026;27(2):257.