



# Machine learning models for dental caries risk prediction: a systematic review with risk-of-bias appraisal

## Modelos de aprendizaje automático para predecir el riesgo de caries dental: una revisión sistemática

Yivelin Fernandez Queija<sup>1</sup>  

Ernesto Learsys Campello Trujillo<sup>1</sup>  

Lourdes Hernandez Cuetara<sup>1</sup>  

<sup>1</sup> Universidad de Ciencias Médicas de la Habana, La Habana, Cuba.

Submitted: 30-06-2025 | Revised: 06-08-2025 | Accepted: 20-09-2025 | Published: 25-09-2025

**Corresponding Author:** Ernesto Learsys Campello Trujillo 

### ABSTRACT

**Introduction:** Dental caries is preventable, but its risk is multifactorial and hard to quantify in the clinic. We synthesized the performance and risk of bias of machine-learning models that predict caries risk or incidence from clinical, behavioral, and social data, appraised with PROBAST. **Methods:** We searched PubMed, OpenAlex, and Crossref for studies published from 2013 to 2025, screened 295 records after deduplication, and included prediction-model studies with a quantitative performance measure. Discrimination was the primary outcome, calibration and comparative performance secondary. Risk of bias was appraised with PROBAST. **Results:** Eleven studies were included, most in children and most using tree-based models. Discrimination was inconsistently reported; where the area under the curve was given, it ranged from 0.77 to 0.98, with the highest value coming from a model trained on simulated data and the lowest from a large national survey. Calibration was reported by only a minority of studies. PROBAST risk of bias was high or unclear in most studies, driven mainly by small samples, internal validation, and incomplete reporting of predictors and outcomes. **Conclusions:** Machine-learning models predict caries risk with reported discrimination that is highly variable and frequently overestimated by design, and the evidence is weakened by a high risk of bias, so these models are not yet ready to replace conventional risk assessment in routine practice.

**Keywords:** dental caries; machine learning; risk assessment; prediction model; PROBAST.

### RESUMEN

**Introducción:** La caries dental es prevenible, pero su riesgo es multifactorial y difícil de cuantificar en la clínica. Sintetizamos el rendimiento y el riesgo de sesgo de los modelos de aprendizaje automático que predicen el riesgo o la incidencia de caries a partir de datos clínicos, conductuales y sociales, evaluados con PROBAST. **Métodos:** Buscamos en PubMed, OpenAlex y Crossref estudios publicados de 2013 a 2025, cribamos 295 registros tras eliminar duplicados e incluimos estudios de modelos predictivos con una medida de rendimiento cuantitativa. La discriminación fue el desenlace principal; la calibración y el rendimiento comparativo, secundarios. El riesgo de sesgo se evaluó con PROBAST. **Resultados:** Se incluyeron once estudios, la mayoría en niños y con modelos basados en árboles. La discriminación se comunicó de forma inconsistente; cuando se dio el área bajo la curva, osciló entre 0,77 y 0,98, con el valor más alto procedente de un modelo entrenado con datos simulados y el más bajo de una gran encuesta nacional. La calibración solo se comunicó en una minoría de estudios. El riesgo de sesgo PROBAST fue alto o incierto en la mayoría, por muestras pequeñas, validación interna y comunicación incompleta de predictores y desenlaces. **Conclusiones:** Los modelos de aprendizaje automático predicen el riesgo de caries con una discriminación comunicada muy variable y con frecuencia sobreestimada por el diseño, y la evidencia está debilitada por un alto riesgo de sesgo, por lo que estos modelos aún no están listos para sustituir la evaluación de riesgo convencional en la práctica.

**Palabras clave:** caries dental; aprendizaje automático; evaluación de riesgo; modelo predictivo; PROBAST.

## INTRODUCTION

Dental caries is the most common chronic disease of childhood and one of the most prevalent conditions of adulthood, yet it is almost entirely preventable, which makes identifying who is at risk the central clinical question in its management. The disease arises from a prolonged imbalance between demineralization and remineralization driven by diet, oral hygiene, fluoride exposure, and the oral microbiota, and these factors interact with social and behavioral determinants in ways that make individual risk hard to judge by inspection alone.<sup>(1,2)</sup> This multifactorial etiology is both the reason prevention works and the reason prediction is hard. The relevant risk factors are known in outline, frequent sugar intake, inadequate fluoride, poor plaque control, low socioeconomic status, and prior caries experience are all established, but their relative weights and their interactions differ across ages, populations, and settings, and the clinician must integrate a dozen partial signals into a single judgment about whether, and how aggressively, to intervene before the disease appears. The global burden is enormous and unequally distributed, with untreated caries in permanent teeth among the most prevalent conditions worldwide and early childhood caries concentrated in disadvantaged populations, so a reliable, low-cost way to stratify risk would be valuable across every health system.<sup>(3)</sup> The burden is also shifting: as populations age and retain their teeth longer, root caries and caries in older adults are rising, which broadens the case for risk prediction beyond childhood and into the whole life course, and the adult models included in this review are an early response to that demographic pressure, even though they remain outnumbered by the childhood studies. Conventional caries risk assessment exists for this purpose, in the form of instruments such as Cariogram, CAMBRA, and the American Academy of Pediatric Dentistry's caries-risk assessment tool, which combine a small set of clinical and behavioral indicators into a categorical risk judgment.<sup>(4,5)</sup> These tools are widely recommended but have a known and modest predictive performance, and the evidence on whether they actually predict future caries better than chance is, at best, mixed, with systematic reviews of their measurement properties and their validity finding limited and inconsistent evidence.<sup>(6-8)</sup> The tools also demand time and training, which limits their uptake in busy and under-resourced settings, and their categorical outputs are a blunt instrument for a risk that is continuous. The gap between the clinical need for risk stratification and the limited accuracy and usability of the tools currently used is exactly where machine learning has been proposed to help.

Machine learning offers the promise of learning risk directly from data, capturing interactions and non-linearities that a small hand-built score cannot, and a growing literature has applied it to caries risk and incidence prediction using clinical, behavioral, social, and increasingly biological predictors.<sup>(9,10)</sup> The broader artificial-intelligence literature in dentistry has grown explosively in the same period, spanning image-based caries detection, diagnostics, and treatment planning, and it has generated a steady stream of reviews and commentaries, which testifies to the field's momentum but also to its fragmentation.<sup>(11-13)</sup> Much of that literature concerns the detection of existing caries lesions on radiographs or intraoral images, a task that is clinically different from predicting future disease, and the two are frequently conflated, which is one reason the predictive literature remains harder to assess than it should be.<sup>(14-16)</sup> The predictive task also overlaps with early-childhood-caries management and with pediatric dentistry, where the case for early risk stratification is strongest and where several narrative and scoping reviews have argued for artificial intelligence as a path to more equitable prevention.<sup>(17-19)</sup> The appeal is straightforward: if a model can predict which children will develop caries before it happens, prevention can be targeted to those who need it, and resources can be concentrated where they have the most effect. The field sits at the intersection of two concerns that matter differently to different audiences. For the clinician, the question is discrimination and calibration, whether the model identifies the right patients and produces probabilities that are trustworthy. For the methodologist, the question is whether the studies that report those numbers are biased, because prediction-model studies are known to be vulnerable to a specific set of biases, small samples, internal validation, selective predictor reporting, and optimistic handling of missing data, that inflate apparent performance.<sup>(20)</sup> These two questions are not independent: a study that reports a high area under the curve from an internal split has answered neither of them, because the estimate itself is untrustworthy, and much of the apparent disagreement in this literature dissolves once the two are kept apart. These two concerns correspond to two distinct outcomes, and they should not be confused: a model can discriminate well and still be useless if its calibration is wrong, and a study can report an excellent area under the curve that is an artifact of its design. The tool that formalizes the methodologist's concern is PROBAST, a structured, domain-by-domain assessment of risk of bias and applicability developed specifically for prediction-model studies, and its application is what distinguishes a critical review

from a catalogue of claimed accuracies.(20)

The objective of this review was to synthesize the discrimination, calibration, and risk of bias of machine-learning models that predict dental caries risk or incidence, in order to judge whether they are ready to support risk-based prevention in practice. The review separates the primary outcome, discrimination, from the secondary outcomes of calibration and comparative performance against conventional risk assessment, and from exploratory analyses such as feature importance, and it appraises the risk of bias of the evidence with PROBAST, the standard tool for prediction-model studies.(20)The question is whether machine-learning models predict caries risk or incidence accurately enough, and with low enough risk of bias, to be used in practice, and the review answers it by weighing the primary outcome most heavily and the exploratory analyses with explicit caution. The expectation, stated before the search, is that the literature will show a familiar pattern: promising discrimination reported on small or internal datasets, a high risk of bias, and little external validation, and the value of the review lies in quantifying that pattern rather than in celebrating the headline numbers. The hierarchy of outcomes is not a stylistic choice but a discipline: it forces the synthesis to foreground the measure that matters, discrimination, and to report calibration and comparison honestly rather than letting a favorable accuracy dominate the narrative while the model's probabilities go unexamined, and that discipline is what this review tries to apply throughout.

## METHODS

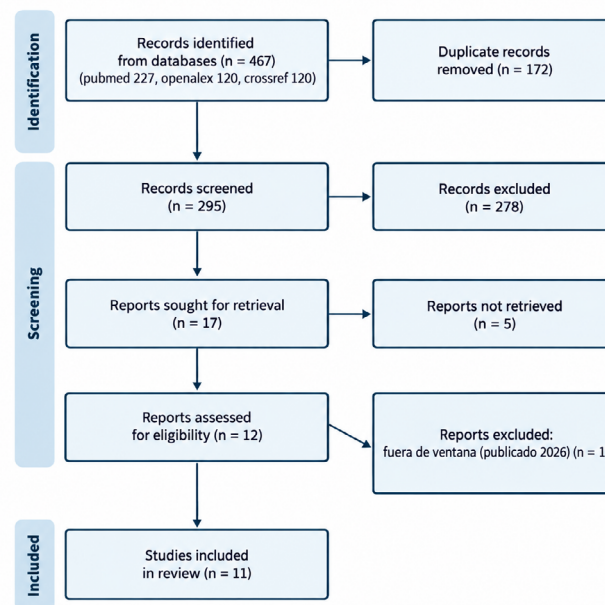
The review followed PRISMA 2020 for the reporting of systematic reviews and CHARMS for the extraction of prediction-model studies, and it was not registered prospectively, which is stated rather than concealed. The primary outcome was prespecified as discrimination, expressed as the area under the receiver operating characteristic curve where reported, and, where it was not reported, as the model's accuracy, sensitivity, or specificity, because the field has not converged on a single metric. The secondary outcomes were calibration, reported as a calibration plot, intercept, or slope, and comparative performance against a conventional risk-assessment tool or a logistic-regression baseline. Exploratory analyses were feature importance, subgroup performance, and any sensitivity analysis reported by the authors, and these were treated as hypothesis-generating only. Studies were eligible if they developed or evaluated a machine-learning model that predicted dental caries risk, incidence, or a caries-related outcome from clinical, behavioral, or social predictors, and reported at least one quantitative performance measure; the window was 2013-2025 and the language English. We excluded image-based caries detection, studies of other oral conditions, unsupervised clustering or trajectory studies that did not produce a prediction, reviews, and non-English reports. PubMed, OpenAlex, and Crossref were searched with terms for machine learning ("machine learning", "artificial intelligence", "deep learning", "prediction model", "random forest"), for the condition ("dental caries", "caries", "early childhood caries", "tooth decay"), and for the task ("risk", "prediction", "incidence"), and the complete search log is reproducible from the review records. A representative string combined these groups with Boolean operators, and the searches were run over the full 2013-2025 window. Records were pooled, deduplicated, and screened by title and abstract and then by full text where open access permitted, with screening performed by a single reviewer assisted by artificial intelligence and then verified, and disagreements were resolved by rereading the record against the eligibility criteria. For each included study we extracted the population, sample, predictors, model type, outcome, and the reported performance measures, and no field was completed without evidence. Risk of bias was appraised with PROBAST across its four domains, participants, predictors, outcome, and analysis. The participants domain was judged high risk when the sample was a small convenience set or the study used a case-control design, the predictors domain when predictors were measured after the outcome or defined inconsistently, the outcome domain when the caries outcome was not clearly defined or not determined by a standardized examination, and the analysis domain when the model was evaluated on the same data used for development without an external or prospective validation, when calibration was omitted, or when overfitting was not addressed. The category "unclear" was used liberally and low risk was never assigned on the basis of missing information. A meta-analysis was planned but not performed because the studies reported different metrics, outcomes, and populations, and any pooled estimate would have been uninformative; the synthesis is narrative and hierarchical, weighted toward the primary outcome, with the secondary and exploratory results reported but given proportionally less interpretative weight.

## RESULTS

### Study selection and characteristics of the models

The searches identified 467 records, 172 duplicates were removed, and 295 records were screened (Figure 1). Title and abstract screening excluded 278 records, the large majority of which were reviews, image-based caries detection studies rather than risk prediction, or studies of other oral conditions, and this distribution reflects both the breadth of the original search and the rapid growth of the image-detection literature that the predictive task is easily confused with. A further 17 reports were sought; 5 could not be retrieved in open access, and 12 were assessed in full text. One report was excluded at full-text assessment because its publication date fell outside the 2013-2025 window, leaving 11 studies in the review (Table 1).

**Figure 1. PRISMA 2020 flow diagram of study selection.**



Source: PRISMA 2020 flow diagram generated from the review's search and screening records.

**Table 1. Characteristics of the included models.**

| Study (year)               | Country        | Population         | Sample                 | Models                  | Outcome          |
|----------------------------|----------------|--------------------|------------------------|-------------------------|------------------|
| Park et al. (2021)         | South Korea    | Children 1-5 years | 4195 (national survey) | XGB, RF, LightGBM, LR   | ECC              |
| Pang et al. (2021)         | China          | Teenagers 13 years | 1055                   | ML with genetic markers | Caries incidence |
| Sadegh-Zadeh et al. (2022) | United Kingdom | Children <5 years  | 780                    | Ten classifiers         | Caries risk      |
| Ciftci et al. (2024)       | Turkey         | Adults 18+         | 2000                   | NB, LR, SVM, others     | DMFT risk        |
| Sadegh-Zadeh et al. (2024) | United Kingdom | Children ≤7 years  | 356                    | LR, DT, RF              | Caries risk      |
| Ogwo et al. (2024)         | United States  | Young adults       | Iowa cohort            | LASSO, GBM, XGB         | Caries at age 23 |
| Hasan et al. (2025)        | Bangladesh     | Children <6 years  | 724 mother-child       | RF, XGB, SVM, MLP       | ECC              |
| Vayrynen et al. (2025)     | Finland        | Adolescents 13-17  | 218                    | ML with oversampling    | Caries lesions   |
| Atni et al. (2025)         | Malaysia       | Adults             | 3000                   | Stacking ensemble       | Caries risk      |
| Zhang et al. (2025)        | China          | Preschool 3-5      | multi-region           | LASSO + LR              | ECC              |
| Sachelarie et al. (2025)   | Romania        | General            | simulated              | Feed-forward NN         | Caries risk      |

Note: LR = logistic regression; RF = random forest; XGB = extreme gradient boosting; MLP = multilayer perceptron; NN = neural network.

The 11 included studies were published between 2021 and 2025 and spanned Asia, Europe, and North America, with 7 studies in children or adolescents and 4 in adults. The populations reflect where caries risk prediction is most clinically urgent: early childhood caries in preschool children was the single most common target, studied in Bangladesh, China, Korea, and the United Kingdom, while the adult studies addressed general caries risk in Turkey and Malaysia and long-term caries outcomes in a United States cohort.(21-27) Tree-based ensembles dominated: random forest, XGBoost, and LightGBM appeared in most studies, compared against a logistic-regression baseline, and neural networks were the exception rather than the rule.(21,23,28) The predictors were overwhelmingly the classic caries determinants, sugar intake, oral hygiene, fluoride exposure, socioeconomic status, and prior caries experience, with a minority of studies adding biological predictors such as the salivary microbiome or genetic markers.(29) Sample sizes ranged from a few hundred to over four thousand, with the largest being a national survey and the smallest a convenience sample of adolescents, and the imbalance between these two designs is itself a source of the heterogeneity in the reported performance.(23,30) One study was not a conventional clinical prediction model at all but a hybrid mathematical simulation feeding an artificial neural network, and it is retained but weighted lightly.(31)

### Primary outcome: discrimination

Discrimination was the most reported and the least consistently reported outcome, and the two facts are related. The area under the curve was reported explicitly by only a minority of studies, and where it was given it ranged from 0.77 to 0.98, a spread that spans from mediocre to near-perfect and that is explained by the data rather than by the models (Table 2). The lowest and most credible figure came from the largest and most representative study, a national survey of 4195 Korean children in which the four models reached areas under the curve between 0.774 and 0.785 and did not differ significantly from one another, a result that is consistent with the limited predictive information contained in the conventional risk factors.(23) The highest figure, an area under the curve of 0.98, came from the hybrid simulation study, whose synthetic data and literature-derived labels are exactly the conditions that flatter a model, and it should be read as an upper bound of method rather than as evidence of real-world performance.(31) Most studies reported accuracy, sensitivity, and specificity instead of discrimination, and these figures were uniformly high, but accuracy is a poor measure for a low-prevalence outcome such as caries incidence, because a model that predicts everyone as low-risk can achieve high accuracy while identifying no one at risk. The children's models illustrate the point: the classification models built on convenience samples reported accuracy in the eighties and nineties, the stacking ensemble built on three thousand records reported a similar figure, and the national-survey models reported the same area under the curve whether they were trees or a simple regression, which suggests the apparent differences between studies owe more to the data than to the method.(23,24,26)

**Table 2. Reported performance of the included models.**

| Study (year)               | Metric                       | Reported value |
|----------------------------|------------------------------|----------------|
| Park et al. (2021)         | AUC (four models)            | 0.774-0.785    |
| Pang et al. (2021)         | AUC                          | —              |
| Sadegh-Zadeh et al. (2022) | Accuracy                     | —              |
| Ciftci et al. (2024)       | Accuracy                     | —              |
| Sadegh-Zadeh et al. (2024) | AUC                          | —              |
| Ogwo et al. (2024)         | AUC                          | —              |
| Hasan et al. (2025)        | Accuracy                     | —              |
| Vayrynen et al. (2025)     | Accuracy                     | —              |
| Atni et al. (2025)         | Accuracy                     | —              |
| Zhang et al. (2025)        | Calibration + decision curve | —              |
| Sachelarie et al. (2025)   | AUC / Acc                    | 0.98 / 91.2%   |

Note: AUC = area under the receiver operating characteristic curve; Acc = accuracy; Se = sensitivity; Sp = specificity; — = not reported in the extractable text.

The adolescent study addressed the low-prevalence problem directly with oversampling and found that the choice of handling the imbalance mattered as much as the classifier, and the longitudinal cohort study found that the predictive value of early caries experience dominated that of the later risk factors, both of which are more informative findings than any single accuracy. (27,30) The adult study based on a large university sample likewise reported that the tree models and the simpler classifiers performed within a few points of one another, and the adult study from Malaysia showed that a stacking ensemble improved on its base learners but not by a margin that would change practice. (25,26) The studies that reported discrimination, the primary outcome of this review, therefore do not support a single summary estimate, and the field's failure to converge on the area under the curve is itself a finding.

### **Secondary outcomes: calibration and comparison**

Calibration, the property that tells a clinician whether a predicted probability can be trusted as a probability, was the most neglected outcome in the literature. Only a small minority of the included studies reported any calibration measure, and the single study that reported calibration and decision-curve analysis was also one of the few to use a transparent, regression-based model rather than a black-box ensemble, which may explain why it is the exception. (22) The pattern is not accidental: calibration requires a probabilistic output and a principled comparison of predicted against observed risk, whereas accuracy requires only a threshold, and the field has gravitated toward the easier and less informative measure. The consequence is that the literature tells the reader which patients a model would classify, but not whether the probability attached to that classification is one that should be believed. The absence of calibration is not a minor omission but a structural one: a model can have excellent discrimination and still produce probabilities that are systematically wrong, and a clinician who uses a miscalibrated probability to decide whether to apply fluoride or refer a child for treatment will be misled by the very tool meant to help. For a low-prevalence outcome, the practical consequence of a miscalibrated model is that the number of patients flagged as high risk, and therefore the number referred for prevention, is wrong by a factor that no amount of discrimination can correct. Comparative performance against conventional risk assessment was reported in even fewer studies, and no included study benchmarked its model head-to-head against Cariogram or CAMBRA on the same population, so the central clinical claim, that machine learning improves on the tools already in use, was asserted in the introductions of these papers but not tested in their analyses. The one comparison that was made, against a logistic-regression baseline, repeatedly showed only trivial differences, with the tree-based ensembles matching rather than clearly beating the simpler model, which echoes the wider finding that, for tabular clinical data, flexible models add little over well-specified regression. (23) This is not an argument against machine learning but a statement about where the information lives: when the predictors are a modest set of well-understood variables, the modeling choice is a second-order concern, and the field's energy would be better spent on better predictors and better evaluation than on more elaborate models. The secondary outcomes therefore do not rescue the primary one: the evidence for calibration and for superiority over existing tools is largely absent.

### **Additional and sensitivity analyses**

The exploratory analyses in the included studies were dominated by feature importance, computed predominantly through SHAP or permutation importance, and they converged on a familiar and unsurprising set of predictors: sugar intake frequency, oral hygiene, fluoride exposure, and prior caries experience ranked consistently as the strongest determinants. (21,28) This convergence is reassuring for face validity but does not constitute new knowledge, because these are the same factors that the conventional risk tools have used for decades, and the machine-learning studies largely rediscovered them rather than uncovering new ones. The feature-importance rankings also differed in detail between studies, with some weighting baseline caries experience most heavily and others weighting diet or hygiene, which reflects the different ages and populations studied and cautions against treating any single ranking as general. Where studies added biological predictors, such as the salivary microbiome or genetic markers, the incremental value was modest and the studies were small, so no conclusion can be drawn about whether these novel predictors improve discrimination. (29) Sensitivity analyses were sparse, and only a few studies tested the stability of their results under alternative variable-selection or resampling choices, which is a further sign of the field's early stage. The special consideration of missing data was handled poorly across the literature, with most studies either excluding incomplete records without reporting the impact or not describing their approach at all, and this is one of the specific sources of bias that PROBAST is

designed to capture. The handling of class imbalance, a separate but related problem given the low prevalence of incident caries in many cohorts, was addressed explicitly by only a handful of studies, and where it was, the choice of oversampling or weighting changed the reported performance, which means the results are sensitive to decisions that were in most cases left unreported. The sensitivity of the conclusions to these unreported choices is a further reason that the headline accuracies in this literature should be read with reserve, because a number reported without its preprocessing decisions is a number that cannot be reproduced or compared.

Risk of bias

The risk-of-bias appraisal found a high or unclear risk of bias in the large majority of studies, and the overall risk of bias was judged high (Table 3). The domain that failed most was analysis, because internal random splits, the absence of external validation, and the omission of calibration were the rule rather than the exception, and these are precisely the practices that inflate apparent performance. The participants domain also failed in several studies, because small convenience samples and single-institution recruitment were common, and the predictors domain was rated unclear in most reports because the measurement and timing of predictors were incompletely described. The outcome domain was the least problematic, because caries is a well-defined clinical outcome when measured, but even here the use of different caries indices and thresholds made the studies hard to compare. The simulated-data study failed every domain, as it should, because its inputs and its labels were both synthesized rather than observed, and its near-perfect performance is the clearest illustration in the review of why risk of bias and reported accuracy move in opposite directions.(31)The single study based on a national survey and the single prospective longitudinal study were the only ones that approached a low risk of bias, and they reported the least impressive discrimination, which is itself the most important result of the appraisal.(23,29) The applicability of the models was also limited in several cases, because a model developed on a specific age band, a specific region, or a specific care pathway may not apply to another, and this applicability concern, while distinct from risk of bias, points in the same direction of caution about using any of these models beyond their original setting.

Table 3. PROBAST risk of bias across the included studies.

| Study (year)               | Participants | Predictors | Outcome | Analysis |
|----------------------------|--------------|------------|---------|----------|
| Park et al. (2021)         | L            | L          | L       | U        |
| Pang et al. (2021)         | L            | U          | L       | U        |
| Sadegh-Zadeh et al. (2022) | U            | U          | U       | H        |
| Ciftci et al. (2024)       | U            | U          | U       | H        |
| Sadegh-Zadeh et al. (2024) | U            | U          | U       | H        |
| Ogwo et al. (2024)         | L            | L          | L       | U        |
| Hasan et al. (2025)        | U            | U          | L       | H        |
| Vayrynen et al. (2025)     | H            | U          | U       | H        |
| Atni et al. (2025)         | U            | U          | U       | H        |
| Zhang et al. (2025)        | U            | U          | L       | U        |
| Sachelarie et al. (2025)   | H            | H          | H       | H        |

Note: H = high risk; U = unclear risk; L = low risk, across the four PROBAST domains. Source: PROBAST appraisal of the included studies.

DISCUSSION

The primary outcome, discrimination, shows a field that has not yet produced a reliable estimate of how well caries risk can be predicted, and this is the main finding of the review. The area under the curve ranges from 0.77 to 0.98, but the range is an artifact of the data and the design, with the low end coming from a large, representative national survey and the high end from simulated data, and the honest summary is that on real, population-representative data the discrimination of these models is modest and does not clearly exceed a logistic regression.(23)

This is consistent with the prior systematic reviews of the topic, which have likewise found that machine-learning models for caries prediction report good accuracy on their own datasets but that the evidence is heterogeneous and the methodological quality is low, and with the broader literature on clinical prediction models, which repeatedly finds that flexible machine learning adds little over regression when the predictors are a modest set of tabular variables.(9,32) The scoping and narrative reviews of artificial intelligence for caries risk and for early childhood caries reach the same conclusion from a different angle, emphasizing the breadth of activity but stopping short of claiming that any model has been shown to outperform the conventional tools in a head-to-head comparison.(10,17) The agreement across reviews conducted with different searches and different inclusion decisions is itself a form of robustness, and it means a reader can be confident in the broad verdict, that the field is active but not yet clinically validated, without needing to arbitrate between individual studies. The reason is not mysterious: caries risk is dominated by a handful of well-known behavioral and social factors, and once those are in the model, there is little residual structure for a random forest or a boosting ensemble to find that a regression cannot, a point that the near-identical performance of the tree models and the logistic baseline in the largest study makes concretely.(23) This pattern is not specific to caries but is the general finding of the clinical prediction-model literature, in which flexible models rarely beat well-specified regression for tabular data of modest dimensionality, and it should temper the expectation that a change of algorithm will transform risk prediction.(20) The comparison with conventional risk assessment is the more relevant benchmark, and the one study in this review that did compare methods found the established instruments and the machine-learning models in broad agreement, which again suggests that the predictive ceiling is set by the information content of the risk factors rather than by the modeling.(33) A model built on the same sugar, hygiene, and fluoride variables that Cariogram already uses cannot outperform it by much unless it either finds interactions the score misses or incorporates predictors the score lacks, and the included studies provide no evidence that either of those things has happened at a scale that matters.

The secondary outcomes deepen the caution rather than relieving it. Calibration is the property that makes a predicted probability usable, and its near-total absence from the literature means that even a model with good discrimination cannot be translated into a treatment decision, because a probability of 0.4 that should be 0.2 will drive the same preventive action to the wrong patient.(20) The absence of head-to-head comparison against Cariogram, CAMBRA, or the AAPD tool is the more consequential gap, because the clinical question is not whether machine learning can predict caries, but whether it predicts better than the instruments already in use, and no included study answered that question.(4,7) The comparison that was made, against logistic regression, is the right internal baseline but the wrong external one, because a clinician does not choose between a random forest and a regression, they choose between a model and the risk tool they already have.(5) A recent comparative study of conventional methods found that CAMBRA, Cariogram, and a simpler semaphore disagreed substantially in their risk assignments, which suggests that the conventional tools themselves do not define a single gold-standard risk judgment against which a model could be benchmarked, and this is a complication that any future comparison must confront rather than ignore.(33) Until a study benchmarks a machine-learning model against a conventional risk-assessment instrument on the same children, the claim of superiority remains a hope rather than a result.

The risk of bias is the decisive finding, because it explains both the inflated figures and the difficulty of interpreting them. The PROBAST appraisal found a high risk of bias in most studies, driven by the analysis domain, internal validation, no calibration, no external testing, and this is a known and correctable pattern in prediction-model research that has been documented across medicine and that the caries literature reproduces faithfully.(20) The direction of the bias is uniformly toward optimism, so the published accuracies are best read as upper bounds, and the studies that are least biased, the national survey and the prospective cohort, are also the ones that report the least impressive discrimination, which suggests that the truth lies closer to the modest end of the range.(23,29) This inverse relationship between methodological quality and reported performance is itself diagnostic: it is the signature of a field in which the impressive results are, to a substantial degree, artifacts of the evaluation rather than properties of the models. The message for the clinician is not that machine learning is useless for caries, but that the evidence currently available cannot tell how useful it is, a verdict that commentaries on the topic have reached from the outside and that the present appraisal confirms from within.(34)

A further distinction runs through this literature and is worth stating explicitly, because it bears on where the field can realistically go. The included models use clinical and behavioral predictors, but a parallel body of work has asked whether biological predictors, the salivary and oral

microbiome, genetic markers, and metabolic indicators, carry incremental information for caries prediction, and the answer so far is mixed.(35-37)Some studies report that microbiome composition predicts early childhood caries onset, and others find that clinical and delivery factors outweigh the microbiome, while a meta-analysis concluded that microbiome studies could improve prediction but remain inconsistent.(35-37)Nutritional and anthropometric predictors have been explored in the same spirit, with machine learning used to link early childhood caries to malnutrition and anemia in a way that highlights the social and biological entanglement of the disease in disadvantaged populations.(38) The single included study that incorporated genetic markers found them to add value, but on a modest sample and without external validation, so the biological direction is promising but unproven.(29)The prediction of a specific caries outcome, such as distal caries in second molars associated with impacted third molars, has also been attempted and shows that narrowly defined prediction tasks can achieve high accuracy, but such tasks are not generalizable risk stratification.(39)Separately, the image-based detection literature, which detects existing lesions rather than predicting future ones, is far larger and better funded, and its methods are not interchangeable with risk prediction, a point that reviews of both tasks have made.(12,40) A model that finds a cavity that is already present answers a different question from a model that says who will develop one, and conflating the two has inflated expectations for the predictive task.

The generalizability of the models across health systems is also unexamined, and it deserves more attention than the literature gives it. The included studies span high-, middle-, and low-income settings, from national surveys in Korea and Finland to community samples in Bangladesh and China, but each model was developed for a single population, and no study tested whether a model built in one setting transferred to another.(21-23)Caries risk is strongly shaped by diet, water fluoridation, dental-service access, and social conditions, all of which vary far more between countries than between the patients of a single institution, so a model's predictors and its coefficients cannot be assumed to travel. A model built where water is fluoridated may overweight a predictor that is uninformative where it is not, and a model built on a high-caries population may not calibrate to a low-caries one. The value proposition of machine learning for caries is strongest precisely in the settings where conventional risk assessment is least available, but those are the settings least represented in the development data, and closing that gap, by building and validating models in low- and middle-income populations, is a distinct priority that the current literature barely addresses.

The review has limitations. The search was restricted to three databases and to English, and full-text retrieval was limited to open access, which left 5 otherwise relevant reports unassessed and may have biased the included set toward certain venues and toward studies that report their methods openly, which, if anything, would make the included set look better than the excluded one. Screening was performed by a single reviewer assisted by artificial intelligence rather than by two independent reviewers. The heterogeneity of outcomes, metrics, and populations prevented a meta-analysis, so the synthesis is qualitative and cannot offer a pooled discrimination with a confidence interval, and the hierarchy of outcomes, while prespecified, is a judgment that others might draw differently. These limitations do not change the direction of the findings but they bound their precision. The restriction to studies that reported a quantitative performance measure also means that models developed but never evaluated, of which there are likely many in a field this active, are invisible to this review, and the same is true of models embedded in commercial products that are never published at all. The review therefore describes the published evidence, not the full landscape of what exists, and a clinician's impression of the field from this synthesis should be tempered by that gap. The heterogeneity of the outcome definitions, caries incidence versus prevalence versus treatment need, is a further limit on synthesis, because these are related but not identical targets, and a model that predicts treatment need is not directly comparable to one that predicts new cavitated lesions.

The implications follow the hierarchy. For practice, the models are not ready to replace conventional caries risk assessment, because their discrimination is unproven on representative populations, their calibration is unreported, and their superiority to existing tools is untested. For research, the field needs what prediction-model research in general needs: prospective or external validation, the reporting of both discrimination and calibration, head-to-head comparison against Cariogram or CAMBRA, and honest handling of missing data, and none of these requires a technical advance, only a change in study design.(20)There is also a case for reporting standards specific to caries, because the current diversity of outcomes, indices, and thresholds makes studies incomparable, and a minimal common set, the area under the curve, a calibration measure, and a clearly stated caries endpoint, would let the next review of this topic do the meta-analysis that this one could not. The most valuable next study is not a better model but a better evaluation of the models

already built, on a representative population, against the tool a clinician would actually use, and until that study exists the promise of machine learning for caries risk prediction should be stated in the conditional. In the meantime, the most defensible use of these models is exploratory, as a complement to clinical judgment in settings where they can be evaluated locally, rather than as a replacement for the risk assessment a clinician already performs. The ingredients of such a study are not exotic: a consecutive or population-based sample, a prespecified set of predictors measured before the outcome, an external or prospective validation, a report of both discrimination and calibration, and a head-to-head comparison against Cariogram or CAMBRA, all of which are within reach of the groups already working in the field. A field that already has national-survey data and longitudinal cohorts could produce that study with existing resources, and it is the absence of such a study, more than any technical limitation, that currently separates the promise from the evidence.

## CONCLUSIONS

Machine-learning models for caries risk prediction report discrimination that ranges from modest on representative national data to near-perfect on simulated data, and the field has not yet established a reliable estimate of real-world performance. The evidence is weakened by a high risk of bias across most studies, driven by small samples, internal validation, and the absence of calibration and of comparison against conventional risk-assessment tools, and the studies that are least biased are also the ones that report the least impressive discrimination. On the current evidence, these models are not ready to replace Cariogram, CAMBRA, or clinical judgment in routine practice, and the next step is not a better algorithm but a better evaluation, prospective, external, and head-to-head against the tools already in use. The promise is real and the pathway to realizing it is clear, but the field has so far produced more demonstrations of feasibility than evidence of clinical value, and the distinction between the two should govern how these models are described and adopted.

## REFERENCES

1. Selwitz RH, Ismail AI, Pitts NB. Dental caries. *The Lancet*. 2007;369(9555):51-59. doi: 10.1016/s0140-6736(07)60031-2
2. Pitts NB, Zero DT, Marsh PD, Ekstrand K, Weintraub JA, Ramos-Gomez F, et al. Dental caries. *Nat Rev Dis Primers*. 2017;3(1):17030. doi: 10.1038/nrdp.2017.30
3. Kassebaum NJ, Smith AGC, Bernabé E, Fleming TD, Reynolds AE, Vos T, et al. Global, Regional, and National Prevalence, Incidence, and Disability-Adjusted Life Years for Oral Conditions for 195 Countries, 1990–2015: A Systematic Analysis for the Global Burden of Diseases, Injuries, and Risk Factors. *J Dent Res*. 2017;96(4):380-387. doi: 10.1177/0022034517693566
4. Featherstone JDB, Crystal YO, Alston P, Chaffee BW, Doméjean S, Rechmann P, et al. A Comparison of Four Caries Risk Assessment Methods. *Front Oral Health*. 2021;2:656558. doi: 10.3389/froh.2021.656558
5. Ng TCH, Luo BW, Lam WYH, Baysan A, Chu CH, Yu OY. Updates on Caries Risk Assessment—A Literature Review. *Dentistry Journal*. 2024;12(10):312. doi: 10.3390/dj12100312
6. Tellez M, Gomez J, Pretty I, Ellwood R, Ismail A. Evidence on existing caries risk assessment systems: are they predictive of future caries?. *Comm Dent Oral Epid*. 2013;41(1):67-78. doi: 10.1111/cdoe.12003
7. Cagetti MG, Bontà G, Cocco F, Lingstrom P, Strohmer L, Campus G. Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review. *BMC Oral Health*. 2018;18(1):123. doi: 10.1186/s12903-018-0585-4
8. Gao X, Di Wu I, Lo ECM, Chu CH, Hsu CYS, Wong MCM. Validity of caries risk assessment programmes in preschool children. *Journal of Dentistry*. 2013;41(9):787-795. doi: 10.1016/j.jdent.2013.06.005
9. Reyes LT, Knorst JK, Ortiz FR, Ardenghi TM. Machine Learning in the Diagnosis and Prognostic Prediction of Dental Caries: A Systematic Review. *Caries Res*. 2022;56(3):161-170. doi: 10.1159/000524167
10. Bhatia S, Gupta VK, Kumar S, Mishra G, Malhotra S, Arif K, et al. Artificial intelligence based techniques for caries risk prediction and assessment: A scoping review. *Journal of Oral Biology and Craniofacial Research*. 2025;15(6):1497-1507. doi: 10.1016/j.jobcr.2025.08.027
11. Anil S, Porwal P, Porwal A. Transforming Dental Caries Diagnosis Through Artificial Intelligence-Based Techniques. *Cureus*. 2023;15(7):e41694. doi: 10.7759/cureus.41694
12. Albano D, Galiano V, Basile M, Di Luca F, Gitto S, Messina C, et al. Artificial intelligence for radiographic imaging detection of caries lesions: a systematic review. *BMC Oral Health*. 2024;24(1):274. doi: 10.1186/s12903-024-04046-7
13. Dettori M, Lamloum D, Lingström P, Campus G. Artificial Intelligence and Innovation in Oral Health Care Sciences: A Conceptual Review. *Healthcare*. 2025;13(24):3327. doi: 10.3390/healthcare13243327
14. Asiri FYI. Automated Classification of Enamel Caries from Intraoral Images Using Deep Learning Models: A Diagnostic Study. *JCM*. 2025;14(24):8959. doi: 10.3390/jcm14248959
15. Bansal RK, Arya A, Singh B, Singla M, Gupta S. Role of Artificial Intelligence and Machine Learning in Conservative Dentistry and Endodontics: A Review. *Cureus*. 2025;17(7):e88515. doi: 10.7759/cureus.88515
16. Mallineni SK, Sethi M, Punugoti D, Kotha SB, Alkhayal Z, Mubarak S, et al. Artificial Intelligence in Dentistry: A Descriptive Review. *Bioengineering*. 2024;11(12):1267. doi: 10.3390/bioengineering11121267
17. Nayak PP, Shetty V, S S, Zacharias L, Gore I. AI-driven approaches in the management of early childhood caries: A path toward global oral health. *Journal of Oral Biology and Craniofacial Research*. 2025;15(5):1134-1140. doi: 10.1016/j.jobcr.2025.07.022
18. Tanna DA, Bhandary S, Hegde KS. Tech Bytes—Harnessing Artificial Intelligence for Pediatric Oral Health: A Scoping Review. *Int J Clin Pediatr Dent*. 2024;17(11):1289-1295. doi: 10.5005/jp-journals-10005-2971
19. Liang Y, Li D, Deng D, Chu CH, Mei ML, Li Y, et al. AI-Driven Dental Caries Management Strategies: From Clinical Practice to Professional Education and Public Self Care. *International Dental Journal*. 2025;75(4):100827. doi: 10.1016/j.identj.2025.04.007
20. 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(1):51-58. doi: 10.7326/m18-1376
21. Hasan F, Tantawi ME, Haque F, Foláyan MO, Virtanen JI. Early childhood caries risk prediction using machine learning approaches in Bangladesh. *BMC Oral Health*. 2025;25(1):49. doi: 10.1186/s12903-025-05419-2
22. Zhang Z, Dou Z, Zhang X, Wang D, Abudourehman K, Yan Y, et al. Development and validation of a novel public prediction platform for deciduous caries in preschool children: an observational study from Northwest China. *BMC Pediatr*. 2025;25(1):600. doi: 10.1186/s12887-025-05763-w
23. Park YH, Kim SH, Choi YY. Prediction Models of Early Childhood Caries Based on Machine Learning Algorithms. *IJERPH*. 2021;18(16):8613. doi: 10.3390/ijerph18168613
24. Sadegh-Zadeh SA, Rahmani Qeranqayeh A, Benkhalifa E, Dyke D, Taylor L, Bagheri M. Dental Caries Risk Assessment in Children 5 Years Old and under via Machine Learning. *Dentistry Journal*. 2022;10(9):164. doi: 10.3390/dj10090164
25. Çiftçi BT, Aşantoğrol F. Utilization of machine learning models in predicting caries risk groups and oral health-related risk factors in adults. *BMC Oral Health*. 2024;24(1):430. doi: 10.1186/s12903-024-04210-z
26. Atni MHM, Rosdy NMMNM, Tajudin MAAM, Rusly AA, Raob NAA, Sabri BAM. Development and evaluation of a multi-model stacking approach for caries risk assessment in adults using supervised machine learning. *Br Dent J*. 2025. doi: 10.1038/s41415-025-9105-5
27. Ogwo C, Brown G, Warren J, Caplan D, Levy S. Predicting dental caries outcomes in young adults using machine learning approach. *BMC Oral Health*. 2024;24(1):529. doi: 10.1186/s12903-024-04294-7

28. Sadegh-Zadeh SA, Bagheri M, Saadat M. Decoding children dental health risks: a machine learning approach to identifying key influencing factors. *Front Artif Intell.* 2024;7:1392597. doi: 10.3389/frai.2024.1392597
29. Pang L, Wang K, Tao Y, Zhi Q, Zhang J, Lin H. A New Model for Caries Risk Prediction in Teenagers Using a Machine Learning Algorithm Based on Environmental and Genetic Factors. *Front Genet.* 2021;12:636867. doi: 10.3389/fgene.2021.636867
30. Väyrynen E, Tirkkonen O, Tiensuu H, Suutala J, Anttonen V, Laitala ML, et al. A Machine Learning Algorithm With an Oversampling Technique in Limited Data Scenarios for the Prediction of Present and Future Restorative Treatment Need: Development and Validation Study. *JMIR Med Inform.* 2025;13:e75117. doi: 10.2196/75117
31. Sachelarie L, Scrobota I, Cristea RA, Hodişan R, Pantor M, Ciavoi G. Mathematical and AI-Based Predictive Modelling for Dental Caries Risk Using Clinical and Behavioural Parameters. *Bioengineering.* 2025;12(11):1190. doi: 10.3390/bioengineering12111190
32. Wang X, Zhang P, Lu H, Luo D, Yang D, Li K, et al. Risk prediction models for dental caries in children and adolescents: a systematic review and meta-analysis. *BMJ Open.* 2025;15(3):e088253. doi: 10.1136/bmjopen-2024-088253
33. Català-Benavent I, Iranzo-Cortés JE, Almerich-Torres T, Márquez-Arrico CF, Almerich-Silla JM, Montiel-Company JM. Comparative Study of Methods for Caries Risk Evaluation: CAMBRA, the Cariogram, and Caries Risk Semaphore. *JCM.* 2025;14(15):5378. doi: 10.3390/jcm14155378
34. Bennett L, Tiplady L, Taylor G. Can AI find the cavities in caries prediction and diagnosis?. *Evid Based Dent.* 2025;26(3):139-140. doi: 10.1038/s41432-025-01181-0
35. Butcher MC, Short B, Veena CLR, Bradshaw D, Pratten JR, McLean W, et al. Meta-analysis of caries microbiome studies can improve upon disease prediction outcomes. *APMIS.* 2022;130(12):763-777. doi: 10.1111/apm.13272
36. Grier A, Myers JA, O'Connor TG, Quivey RG, Gill SR, Kopycka-Kedzierawski DT. Oral Microbiota Composition Predicts Early Childhood Caries Onset. *J Dent Res.* 2021;100(6):599-607. doi: 10.1177/0022034520979926
37. Wu TT, Xiao J, Manning S, Saraithong P, Pattanaporn K, Paster BJ, et al. Multimodal Data Integration Reveals Mode of Delivery and Snack Consumption Outrank Salivary Microbiome in Association With Caries Outcome in Thai Children. *Front Cell Infect Microbiol.* 2022;12:881899. doi: 10.3389/fcimb.2022.881899
38. Fasna K, Khan SY, Ahmad A, Sharma MK. Exploring the association between early childhood caries, malnutrition, and anemia by machine learning algorithm. *J Indian Soc Pedod Prev Dent.* 2024;42(1):22-27. doi: 10.4103/jisppd.jisppd\_50\_24
39. Hur SH, Lee EY, Kim MK, Kim S, Kang JY, Lim JS. Machine learning to predict distal caries in mandibular second molars associated with impacted third molars. *Sci Rep.* 2021;11(1):15447. doi: 10.1038/s41598-021-95024-4
40. Al-Namankany A. Influence of Artificial Intelligence-Driven Diagnostic Tools on Treatment Decision-Making in Early Childhood Caries: A Systematic Review of Accuracy and Clinical Outcomes. *Dentistry Journal.* 2023;11(9):214. doi: 10.3390/dj11090214

**FUNDING**

No financing

**CONFLICT OF INTEREST**

None

**AUTHORSHIP CONTRIBUTIONS**

Drafting – original draft: Yivelin Fernandez Queija, Ernesto Learsys Campello Trujillo, Lourdes Hernandez Cuetara.

Writing–review and editing: Yivelin Fernandez Queija, Ernesto Learsys Campello Trujillo, Lourdes Hernandez Cuetara.