



Un marco de adherencia de reporte para estudios odontológicos de inteligencia artificial: un análisis automatizado de minería de texto

A reporting-adherence framework for dental artificial-intelligence studies: an automated text-mining analysis

Ana Rodríguez Yanes¹  

¹ Universidad de Ciencias Médicas de La Habana, Facultad de Ciencias Médicas “Miguel Enríquez”, Departamento de Estomatología General Integral. La Habana, Cuba.

Submitted: 04-05-2025 | Revised: 02-07-2025 | Accepted: 11-10-2025 | Published: 27-10-2025

Corresponding Author: Ana Rodríguez Yanes 

ABSTRACT

Introduction: Reporting guidelines for artificial intelligence (CLAIM, TRIPOD+AI, STARD-AI) exist but their uptake in dental artificial-intelligence research is unquantified. Scattered assessments in other specialties suggest methodological items are under-reported. **Methods:** Open-access dental artificial-intelligence diagnostic and prognostic studies were retrieved from Europe PMC (2013-2025). The items of CLAIM, TRIPOD+AI and STARD-AI were operationalized into a framework of 22 items across five domains (data, model, validation, performance, transparency), and full texts were mined automatically for each item. **Results:** The framework was applied to 296 studies. Mean study adherence was 73.1% (SD 15.0). Adherence was highest in the data (87.6%) and performance (83.8%) domains and lowest in the model (57.6%) and validation (59.5%) domains. Reporting of the area under the curve reached 99.3%, whereas code availability (33.1%) and external validation (34.1%) were the least reported items. Adherence increased weakly over time (Spearman rho 0.11, p 0.049). **Conclusions:** Performance metrics are reported almost universally in dental artificial-intelligence studies, but methodological and transparency items lag substantially. The five-domain framework provides a reusable, automated instrument for auditing reporting completeness in the field.

Keywords: artificial intelligence; reporting guidelines; dentistry; data mining; reproducibility.

RESUMEN

Introducción: Existen guías de reporte para inteligencia artificial (CLAIM, TRIPOD+AI, STARD-AI), pero su adopción en la investigación odontológica de inteligencia artificial no está cuantificada. Evaluaciones dispersas en otras especialidades sugieren que los ítems metodológicos se reportan de forma deficiente. **Métodos:** Se recuperaron de Europe PMC (2013-2025) estudios odontológicos diagnósticos y pronósticos de inteligencia artificial en acceso abierto. Los ítems de CLAIM, TRIPOD+AI y STARD-AI se operacionalizaron en un marco de 22 ítems en cinco dominios (datos, modelo, validación, rendimiento, transparencia) y los textos completos se minaron automáticamente para cada ítem. **Resultados:** El marco se aplicó a 296 estudios. La adherencia media por estudio fue 73,1% (DE 15,0). La adherencia fue mayor en los dominios de datos (87,6%) y rendimiento (83,8%) y menor en los de modelo (57,6%) y validación (59,5%). El reporte del área bajo la curva alcanzó 99,3%, mientras que la disponibilidad del código (33,1%) y la validación externa (34,1%) fueron los ítems menos reportados. La adherencia aumentó débilmente con el tiempo (rho de Spearman 0,11, p 0,049). **Conclusiones:** Las métricas de rendimiento se reportan casi universalmente en los estudios odontológicos de inteligencia artificial, pero los ítems metodológicos y de transparencia quedan muy por detrás. El marco de cinco dominios ofrece un instrumento automatizado y reutilizable para auditar la completitud del reporte en el campo.

Palabras clave: inteligencia artificial; guías de reporte; odontología; minería de datos; reproducibilidad.

INTRODUCTION

Artificial intelligence has entered dentistry at a pace that has outpaced the mechanisms for judging the quality of the research it produces. Deep-learning models now segment teeth on radiographs, detect caries and periodontal bone loss, estimate age, and predict implant and treatment outcomes, and the applications extend further into intraoral scans, cephalometry, and cone-beam imaging.^(1,2) Each application generates a stream of published studies whose methodological soundness is difficult for readers to assess quickly. The same expansion has occurred across medical imaging more broadly, where machine-learning models for detection, segmentation, and diagnosis now number in the hundreds, and where the imaging specialties have grappled with how to appraise them.⁽³⁻⁵⁾ The result is a literature in which the volume of models is not matched by a shared vocabulary for evaluating whether a model was developed, validated, and reported in a way that supports clinical use. This matters for patient safety as much as for scientific hygiene, because a dentist or oral radiologist who is handed a model without knowing its training data, its validation, or its failure modes cannot judge whether it will behave on the patients in front of them.

A set of reporting guidelines has emerged to fill this gap, but the guidance remains scattered across the methodological literature rather than integrated into a single assessment. CLAIM specifies what an imaging artificial-intelligence study should report, and was updated in 2024.^(6,7) TRIPOD+AI extends prediction-model reporting to machine-learning methods,⁽⁸⁾ STARD-AI addresses diagnostic accuracy studies that use artificial intelligence,⁽⁹⁾ and CONSORT-AI and SPIRIT-AI govern trials and their protocols.^(10,11) Still other checklists have been proposed to cover study designs not addressed by these instruments, and reviews have catalogued how many such guidelines now exist.^(12,13) Each guideline overlaps with the others on core items such as data provenance, model description, and performance metrics, and readers and reviewers are left to select among them without a compact synthesis.⁽¹⁴⁾ The proliferation of instruments, each long and each maintained separately, is itself part of the problem that a consolidated framework is meant to solve. A dental study that trains a convolutional network to detect caries against a radiologist reference standard is simultaneously an imaging study, a prediction-model study, and a diagnostic-accuracy study, so it technically falls under three checklists whose shared core could be reported once. No single tool currently integrates them into a framework that can be applied automatically across a large body of text.

The stakes of incomplete reporting are not cosmetic. When a model is described without its data provenance, preprocessing, or the split between training and evaluation, its performance cannot be reproduced or trusted, and the leakage that inflates reported accuracy goes undetected.^(15,16) Leakage in particular is a failure that only complete reporting can expose, because an inflated accuracy that results from overlapping training and evaluation data looks, in the abstract, identical to a genuine result. Transparency and reproducibility are prerequisites for clinical translation, and their absence has been identified as a persistent weakness of the machine-learning literature in health care.^(17,18) A field that reports its models without the details needed to reproduce them is accumulating demonstrations rather than building tools.

Evidence from other specialties indicates that adherence to these guidelines is incomplete and uneven. Reviews of imaging artificial-intelligence studies have repeatedly found that basic items such as data source and reference standard are reported more often than technical and transparency items such as code availability and external validation.⁽¹⁹⁻²²⁾ A landmark systematic review of deep-learning studies comparing algorithms against clinicians found that only a small minority performed external validation, and that claims of superiority were often made on the basis of internal testing alone.⁽¹⁹⁾ Similar audits in radiology journals and in orthopaedic imaging reach the same conclusion.^(23,24) Journal endorsement of artificial-intelligence reporting guidelines also varies widely, which shapes whether authors are even prompted to follow them, and early evidence suggests that guideline publication alone has produced only modest change.⁽²⁵⁻²⁷⁾ In dentistry specifically, editors have called for dedicated guidance because many dental artificial-intelligence studies lack rigor, transparency, and reproducibility, and no audit has yet quantified the size of that problem across the dental literature.⁽²⁸⁾ What has been missing is a field-specific, automated quantification of the problem across the dental literature. The existing audits in radiology and orthopaedics were carried out study by study by human reviewers, which limits how often they can be repeated and how large a literature they can cover; an automated instrument would make the assessment continuous and comparable across time and specialties.

The aim of this study was therefore to build and apply a reporting-adherence framework for dental artificial-intelligence studies. The framework consolidates the reporting items of CLAIM, TRIPOD+AI and STARD-AI into a text-mineable checklist of 22 items across five domains, and we

applied it automatically to open-access dental artificial-intelligence studies to answer two questions: how often are the required items reported, and has adherence changed over time as the field has grown.

METHODS

Study design and corpus assembly

This was an automated text-mining study of published open-access dental artificial-intelligence research, and the unit of analysis was the individual published article. The corpus was assembled through the Europe PMC REST application programming interface. The search combined artificial-intelligence terms (artificial intelligence, machine learning, deep learning, convolutional neural network, neural network) with dental terms (dental, dentistry, oral, tooth, teeth, gingiva, periodontitis, caries, implant, mandibular, maxillary, cone beam, panoramic, endodontic, orthodontic, cephalometric) in the title. The query was restricted to open-access records with full text available and to a publication year from 2013 to 2025.

Records were then screened automatically to retain diagnostic and prognostic model studies. Records whose title indicated a review, systematic review, meta-analysis, bibliometric study, survey, educational study, opinion, commentary, editorial, letter, protocol, or erratum were excluded. The remaining records were retained when the title or abstract indicated model development or evaluation, defined by the presence of terms such as model, network, segmentation, detection, classification, prediction, diagnosis, accuracy, validation, dataset, or training. The filter is deliberately conservative, erring toward including borderline studies rather than excluding genuine model studies, because the downstream cost of an included non-model study is small whereas dropping a true model study would bias the corpus. Because the number of qualifying studies exceeded the number of full texts that could reasonably be mined, a year-stratified random sample of 296 studies was drawn for full-text analysis, so that each publication year contributed a share of the sample proportional to its size in the qualifying set. The title-based search was chosen over an abstract- or full-text-based search to keep the query precise, because dental terms in the title select studies whose subject is dental rather than studies that merely mention dentistry in passing.

Reporting framework

To build the framework, we consolidated the items of CLAIM, TRIPOD+AI and STARD-AI into a single checklist of 22 reporting items organized into five domains. The data domain covered data source, sample size, inclusion and exclusion criteria, ground truth or annotation, and annotator qualification. The model domain covered preprocessing, data augmentation, model architecture, hyperparameters, and code availability. The validation domain covered train-test splitting, internal validation, and external validation. The performance domain covered the area under the curve, sensitivity and specificity, precision, recall or F1 score, confidence intervals, and comparison with clinicians. The transparency domain covered the ethics statement, funding, conflict of interest, and data availability. Each item was operationalized as a set of case-insensitive keyword patterns that capture its typical expression in the text of a paper; an item was counted as reported when any of its keywords appeared anywhere in the title, abstract, or body. For example, code availability was detected through terms such as GitHub, open source, and source code, and external validation through terms such as external validation, external dataset, independent cohort, and multicenter. The keyword sets were intentionally broad rather than exact, because the same reporting item is expressed in different words across journals; the trade-off is that presence of a keyword is a proxy for reporting rather than a proof of completeness, a limitation we return to in the Discussion.

The choice of five domains reflects the structure shared by the source guidelines rather than an arbitrary grouping. All three guidelines distinguish what was measured (data), how the model was built and tuned (model), how it was evaluated (validation), what it achieved (performance), and how the study was governed and disclosed (transparency). Collapsing their individual items into these domains preserves the distinctions that matter for appraisal while making the framework short enough to apply automatically. The number of items per domain is not weighted, so the domain means should be read as descriptive summaries of item coverage rather than as a formal quality score.

Text extraction and mining

Full-text XML was downloaded for each sampled study through the Europe PMC full-text endpoint. The concatenated text of the title, abstract, and body was extracted, lowercased, and searched for the keyword patterns of each item. Figures and tables were included implicitly through the body text, and reference lists were not searched, so an item counted as reported only when the authors themselves wrote about it. Studies for which no extractable full text was available, defined as fewer than 200 characters, were excluded. For each study, an adherence score was computed as the percentage of the 22 items whose keywords were detected. Domain-level adherence was computed as the mean of the item percentages within each domain, and item-level adherence as the proportion of studies reporting each item.

Statistical analysis

Study adherence was summarized with the mean, median, standard deviation, and range. The temporal trend was assessed with Spearman rank correlation between publication year and study adherence score, a rank-based measure chosen because the study adherence scores are not normally distributed and because a monotonic association, rather than a strictly linear one, is the relevant question. A two-sided alpha of 0.05 was used. Analyses were performed in Python 3.13 with requests, lxml, pandas, NumPy, SciPy, and Matplotlib. The pipeline is reproducible from the public data and is re-runnable, and every reported number is generated directly from the code rather than transcribed, so the results can be recomputed verbatim by anyone with access to the Europe PMC interface.

RESULTS

Corpus characteristics

The Europe PMC search returned 1,000 open-access records, of which 296 studies were sampled for full-text analysis after the diagnostic-model filter. The corpus spanned eight publication years from 2018 to 2025, with the great majority of studies published in the most recent three years (Table 1). The number of studies per year rose steeply, from single studies in 2018 and 2019 to more than one hundred in 2025, mirroring the broader growth of dental artificial-intelligence research. Because the sample was stratified by year, the larger recent cohorts are represented in proportion to their actual prevalence in the qualifying set, so the adherence estimates are not dominated by any single year. The sharp increase in volume also means that any time trend estimated here is driven principally by the contrast between the large recent cohorts and the small early ones.

Table 1. Corpus composition and mean adherence by year.

Year	Studies	Mean adherence (%)
2018	1	50.0
2019	1	77.3
2020	2	88.6
2021	7	72.7
2022	26	76.6
2023	45	68.9
2024	85	70.2
2025	129	75.8

Note: Mean adherence = mean percentage of the 22 reporting items mentioned per study.

The reporting-adherence framework

The consolidated framework comprised 22 items across five domains (Table 2). The five domains describe a natural progression from the data a study used, through how the model was built and validated, to what it achieved and how the study was governed. Applied to the full texts, the framework produced a graded profile of reporting completeness that differed sharply across

domains and across individual items, and this graded profile is the framework's central output: a single view of where the field reports well and where it does not.

Table 2. Reporting adherence for each framework item.

Item	Domain	Reporting (%)
Data source	Data	96.3
Sample size	Data	97.6
Inclusion/exclusion criteria	Data	84.8
Ground truth / annotation	Data	68.9
Annotator qualification	Data	90.2
Preprocessing	Model	61.1
Data augmentation	Model	55.1
Model architecture	Model	78.0
Hyperparameters	Model	60.8
Code availability	Model	33.1
Train/test split	Validation	96.3
Internal validation	Validation	48.0
External validation	Validation	34.1
AUC / ROC	Performance	99.3
Sensitivity / specificity	Performance	66.2
Precision / recall / F1	Performance	98.6
Confidence intervals	Performance	55.7
Comparison with clinicians	Performance	99.0
Ethics statement	Transparency	81.8
Funding	Transparency	60.5
Conflict of interest	Transparency	61.5
Data availability	Transparency	82.1

Note: Reporting (%) = proportion of studies whose text mentioned the item.

Domain-level adherence

Aggregating items into the five domains revealed a clear hierarchy. Adherence was highest in the data domain (87.6%) and the performance domain (83.8%), intermediate in the transparency domain (71.5%), and lowest in the validation domain (59.5%) and the model domain (57.6%) (Table 3, Figure 1). The two weakest domains were also the ones most directly related to whether a model can be reproduced or trusted beyond its original setting, which is the practical significance of this gradient. The ranking of the five domains is stable and internally consistent: the items that journals and reviewers already demand cluster at the top, and the items that require authors to make their work reproducible cluster at the bottom. Within the data domain, studies almost always named their data source and sample size, but ground-truth annotation was described less consistently.

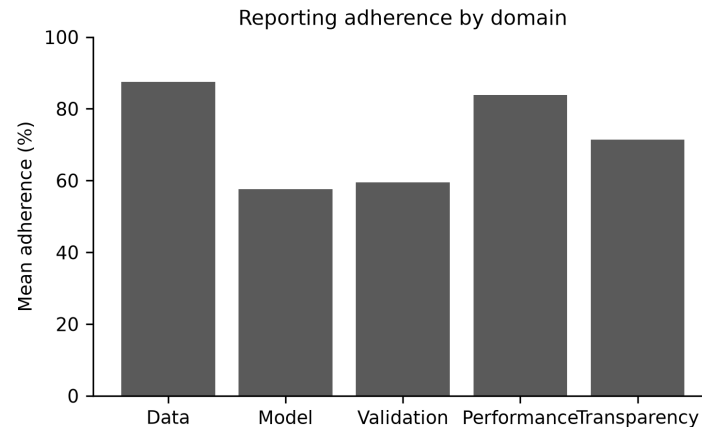
Table 3. Reporting adherence by framework domain.

Domain	Items	Mean adherence (%)
Data	5	87.6
Model	5	57.6
Validation	3	59.5
Performance	5	83.8
Transparency	4	71.5

Note: Mean adherence = mean percentage of items reported within each domain.

The performance domain was held up by a single near-universal item, the area under the curve, which masked the much lower reporting of sensitivity and specificity and of confidence intervals. The model and validation domains were held down by the scarcity of code sharing, internal validation, and external validation.

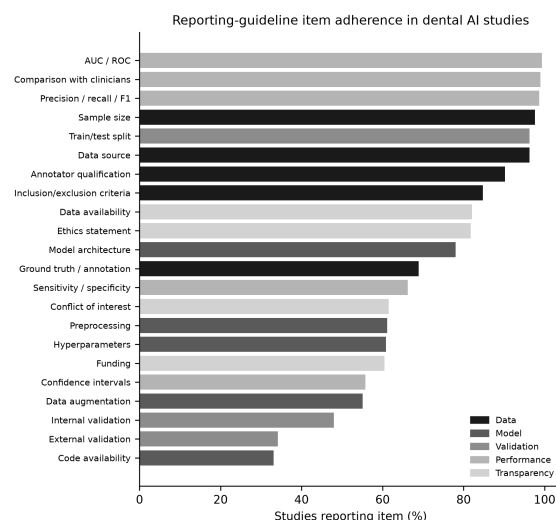
Figure 1. Reporting adherence by framework domain.



Item-level adherence

Within the domains, individual items ranged from near-universal reporting to minority reporting (Figure 2). The area under the curve was the single most reported item at 99.3%, and comparison with clinicians and precision, recall or F1 score were reported at similarly high rates. By contrast, code availability (33.1%) and external validation (34.1%) were reported by only about one third of studies, and internal validation by less than half (48.0%). Ground-truth annotation (68.9%) and confidence intervals (55.7%) also fell below two thirds of studies. Sensitivity and specificity were reported by only 66.2% of studies, a notable gap given the centrality of these measures to diagnostic-accuracy guidelines and to clinicians deciding how to use a model. The pattern within the model domain was equally uneven, with model architecture reported by 78.0% of studies but data augmentation by only about half. Transparency items split into two groups: the ethics statement and data availability were reported by the large majority of studies, whereas funding and conflict-of-interest disclosures were reported by only about three in five. This divergence suggests that statements required by journals and funders are reported more reliably than items left to author discretion.

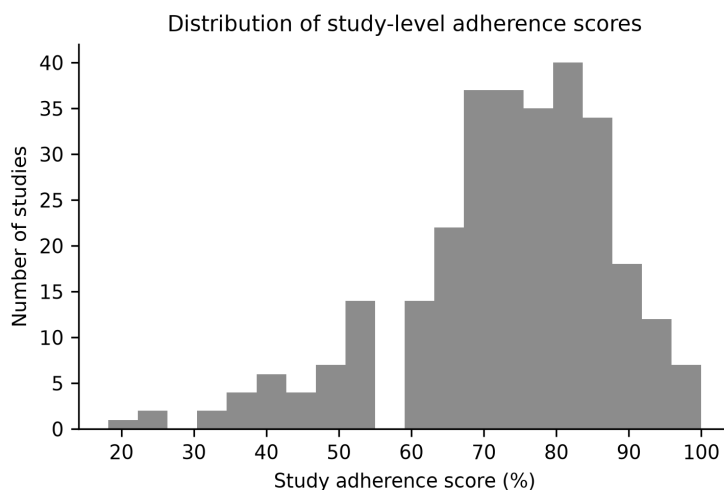
Figure 2. Reporting adherence for each framework item.



Study-level adherence

The mean study adherence score was 73.1% (median 72.7%, SD 15.0%), with a range from 18.2% to 100.0% (Figure 3). The distribution was broad and roughly symmetric around the mean, indicating that reporting completeness varied substantially between individual studies rather than clustering at either extreme. A few studies reported every item, but a comparable number reported only a small fraction, so the field contains both exemplary and largely opaque reporting side by side, and the overall mean is an average of these two rather than a description of any single study.

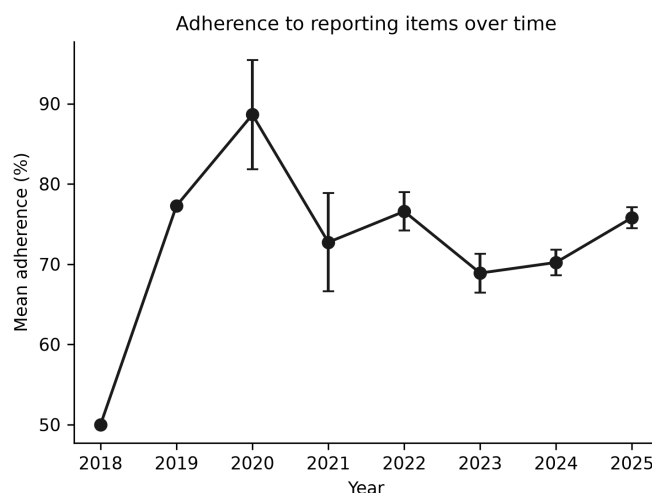
Figure 3. Distribution of study-level adherence scores.



Temporal trend

Adherence increased modestly over the study period, but the improvement was small and barely significant. The Spearman correlation between publication year and study adherence was 0.11 (p 0.049). Mean adherence fluctuated across years without a consistent upward trajectory in the most recent cohorts (Table 1, Figure 4). The years with the largest sample sizes, 2024 and 2025, showed mean adherence near or below the overall average, which argues against a simple story of steady improvement. The rapid growth in the number of studies therefore occurred without a commensurate rise in reporting completeness, and the statistically significant but small correlation reflects this divergence between quantity and quality.

Figure 4. Reporting adherence over time.



DISCUSSION

The principal finding of this study is that dental artificial-intelligence studies report their results almost completely but their methods and transparency only partially. Performance metrics such as the area under the curve and comparison with clinicians are reported by nearly every study, whereas the items that determine reproducibility and generalizability, namely code availability and external validation, are reported by only one third. The five-domain framework exposes this gap as a systematic feature of the field rather than an occasional omission, and the fact that the gap persists at scale, across hundreds of studies and several years, is what makes it actionable rather than anecdotal.

What the framework adds is an integration of guidance that had been scattered across separate guidelines. CLAIM, TRIPOD+AI and STARD-AI each enumerate many items that overlap heavily on data provenance, model description, and performance reporting, but their separate publication has made it easy for authors and reviewers to treat them as optional.(6,8,9)By consolidating them into 22 items in five domains, the framework converts three long checklists into a single graded profile that can be applied to any corpus automatically, and it answers the recurring question of which guideline to choose by combining rather than selecting.(14)This is the contribution the present study makes beyond the guidelines themselves: a compact, operational instrument rather than an aspirational list.

The three source guidelines differ in emphasis even where they overlap in substance. CLAIM is written for imaging studies and foregrounds data provenance and model specification,(6)TRIPOD+AI foregrounds the development and validation of prediction models and the transparent reporting of performance across training and evaluation sets,(8)and STARD-AI foregrounds the diagnostic-accuracy chain from participants through reference standard to test results.(9)A dental study that reports an imaging classifier against a reference standard sits in the overlap of all three, which is why a consolidated framework is more useful for this literature than any single checklist. The framework does not replace the source guidelines; it compresses their shared core into a form that can be applied at the scale of a whole field, and a reader who wants the full item wording is directed back to the original instruments.

The observed pattern matches evidence from other specialties closely. Reviews of imaging artificial-intelligence studies have repeatedly found high reporting of performance measures alongside low reporting of external validation, code sharing, and data availability.(19-22)An audit of a radiology journal found the same gradient, with technical and reproducibility items trailing performance items,(23)and a review of orthopaedic imaging reached the same conclusion.(24)The near-universal reporting of the area under the curve is unsurprising given that it is the currency in which model quality is advertised, but the contrast with the 33.1% code availability and 34.1% external validation illustrates that the field prioritizes demonstrating performance over demonstrating reproducibility. That gap has direct clinical consequences: a model that is not externally validated and whose code is not shared cannot be independently confirmed, which is a prerequisite for translation into practice.(15,16) The dental estimates are of the same order as the compliance rates reported in those earlier audits, which strengthens confidence that the framework is measuring a real and general phenomenon rather than an artifact of this particular corpus.

Why adherence is improving only slowly is itself instructive. The weak time trend (ρ 0.11) suggests that the rapid growth in the number of dental artificial-intelligence studies has not been matched by a rise in reporting quality, even though guidelines have existed since at least 2020.(6)The most plausible explanation is that reporting is driven more by what journals require than by what guidelines recommend. Where journals have not endorsed or mandated the guidelines, authors have little incentive to comply, and meta-research shows that endorsement of artificial-intelligence reporting guidelines remains uneven across journals.(25,26) Early evidence on TRIPOD+AI likewise found only limited change in reporting after the guideline appeared.(27)A clinician facing these models needs the missing details to judge whether a result is real or inflated, and the absence of those details is exactly what a checklist is meant to remedy.(18,29) The asymmetry between performance and methods reporting also mirrors the incentive structure of a fast-moving field: a headline area under the curve advances a line of work, whereas a careful account of preprocessing and validation costs effort without adding novelty. Until that calculus changes, through journal requirements or through the expectations of the clinical readers who are the intended users of these models, the gradient observed here is likely to persist. Reviewers too are part of this system: when a reviewer does not ask for the missing validation or code, the omission is effectively ratified, and the weak time trend suggests that peer review as currently practiced has not closed the gap on its own. Funders, who increasingly require data and code sharing as a condition of awards, are

a further lever that has yet to show up in the reporting of the studies examined here.

Beyond description, the framework has value as a monitoring instrument. It can be re-run against any future corpus to track whether reporting improves after journals adopt mandatory checklists, which is a falsifiable proposition that this study leaves open: if code availability and external validation rise following explicit editorial mandates, that would confirm that reporting quality responds to journal policy rather than to time alone. Such audits matter for value assessment, because a model whose reporting is incomplete cannot be properly evaluated for the cost and benefit claims made about it.⁽³⁰⁾ It therefore functions as both a diagnostic tool and a monitoring instrument for the field.

The choice of an automated, keyword-based approach over manual checklist scoring deserves its own note, because it is both the main strength and the main weakness of the method. Manual scoring by trained reviewers produces richer judgments and can distinguish a complete report from a passing mention, but it does not scale: the dozens of audits published in other specialties have each covered at most a few hundred studies and required substantial reviewer effort.^(19,22) An automated instrument trades that depth for breadth and repeatability, so that a whole literature can be characterized and re-characterized as it grows. The two approaches are complementary rather than competing, and the ideal use of this framework is as a first-pass screen that directs human attention to the items and studies most in need of a closer look. A reviewer could, for example, run the framework on a manuscript and then spend effort only on the domains where keywords are absent, rather than reading for all items equally. A natural future extension is to replace the keyword matching with a semantic scorer, whether a fine-tuned classifier or a large language model, that can distinguish a passing mention from a complete report while retaining the automation and scale of the present approach.

The approach also generalizes beyond dentistry. Its five domains are not dental-specific, and the same keyword instrument could be pointed at any imaging or prediction-model literature to produce a comparable adherence profile, which would allow cross-specialty benchmarking of reporting quality. The dental field is a useful first application because it is growing quickly and because its editors have already flagged reporting quality as a concern,⁽²⁸⁾ but nothing in the method depends on the domain. Cross-specialty comparison would also test whether the gradient observed here, high performance reporting against low reproducibility reporting, is a universal feature of applied artificial-intelligence research or a local one. If other specialties show the same pattern, the finding would argue for a shared, field-agnostic fix rather than discipline-by-discipline patches.

For authors, the five-domain profile is a pre-submission checklist in miniature: before submitting, a group can run their own manuscript against the 22 keywords and see at once whether the model and validation domains are covered, which is where most papers fall short. For reviewers, the profile concentrates attention on the items that matter for reproducibility rather than on the performance numbers that are almost always present. For editors, the framework is an audit instrument that can be applied to a journal's own back catalog or to submissions under review, and it could be embedded in submission systems as a lightweight completeness check before peer review begins. None of these uses requires the full detail of the source guidelines, which remain the reference for a complete assessment.

For the field, the implications are direct. Editors and reviewers can use the five-domain profile to identify at a glance whether a submission reports the methodological and transparency items that determine reproducibility, and the dental community, which has already recognized the problem, now has an instrument sized to it.^(13,28) The weak time trend is a warning that growth alone will not raise quality, and that the next increment of improvement is more likely to come from editorial policy than from the passage of time. The framework makes that improvement measurable, and a clearly measurable and reproducible target is the first step toward actually achieving it.

The interpretation of these results is subject to limitations:

- The analysis is keyword-based rather than semantic. An item was counted as reported whenever its terms appeared, regardless of whether the report was complete or correct. This likely overestimates true adherence, because mentioning an item is weaker than satisfying it. For instance, a study that states that external validation was performed but reports no results for it would be counted as compliant. The magnitude of this inflation is unknown without a reference standard.
- Operationalizing guidelines as keywords inevitably simplifies them. Different keyword choices could shift estimates for individual items even if the overall pattern remains robust.
- The sample is dominated by recent years, so the time-trend estimate rests on relatively few early studies.
- The framework weights items equally even though their importance differs. A study that omits

an essential item such as external validation and one that omits a secondary item such as data augmentation can therefore receive the same score. A weighted version informed by each item's relative impact on reproducibility would sharpen the instrument.

CONCLUSIONS

A five-domain framework consolidating CLAIM, TRIPOD+AI and STARD-AI was applied automatically to a large corpus of open-access dental artificial-intelligence studies. Performance metrics were reported almost universally, while code availability, external validation, and other methodological and transparency items lagged substantially, and reporting adherence improved only weakly over time. The framework provides a reusable, automated instrument for auditing reporting completeness in dental artificial-intelligence research, and it leaves open a specific, testable question: whether editorial mandates, rather than the passage of time, are what will move the field toward the reproducible reporting that clinical adoption requires.

REFERENCES

1. Najeel M, Islam S. Artificial intelligence (AI) in restorative dentistry: current trends and future prospects. *BMC Oral Health*. 2025;25(1):592. doi: 10.1186/s12903-025-05989-1
2. Nordblom NF, Büttner M, Schwendicke F. Artificial Intelligence in Orthodontics: Critical Review. *J Dent Res*. 2024;103(6):577-584. doi: 10.1177/00220345241235606
3. Barragán-Montero A, Javaid U, Valdés G, Nguyen D, Desbordes P, Macq B, et al. Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica*. 2021;83:242-256. doi: 10.1016/j.ejmp.2021.04.016
4. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *npj Digit Med*. 2020;3(1):30. doi: 10.1038/s41746-020-0229-3
5. Drukker L, Noble JA, Papageorgiou AT. Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. *Ultrasound in Obstet & Gyne*. 2020;56(4):498-505. doi: 10.1002/uog.22122
6. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology: Artificial Intelligence*. 2020;2(2):e200029. doi: 10.1148/ryai.2020200029
7. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*. 2024;6(4):e240300. doi: 10.1148/ryai.240300
8. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi: 10.1136/bmj-2023-078378
9. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283-3289. doi: 10.1038/s41591-025-03953-8
10. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Ashrafian H, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*. 2020;2(10):e537-e548. doi: 10.1016/s2589-7500(20)30218-1
11. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, The SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi: 10.1038/s41591-020-1037-7
12. Kwak SG, Kim J. Comprehensive reporting guidelines and checklist for studies developing and utilizing artificial intelligence models. *Korean J Anesthesiol*. 2025;78(3):199-214. doi: 10.4097/kja.25075
13. Sheldermine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Health Care Inform*. 2021;28(1):e100385. doi: 10.1136/bmjhci-2021-100385
14. Klontzas ME, Gatti AA, Tejani AS, Kahn CE. AI Reporting Guidelines: How to Select the Best One for Your Research. *Radiology: Artificial Intelligence*. 2023;5(3):e230055. doi: 10.1148/ryai.230055
15. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804. doi: 10.1016/j.patter.2023.100804
16. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Massive Analysis Quality Control (MAQC) Society Board of Directors, Shraddha T, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586(7829):E14-E16. doi: 10.1038/s41586-020-2766-y
17. Luo X, Wang B, Shi Q, Wang Z, Lai H, Liu H, et al. Lack of methodological rigor and limited coverage of generative artificial intelligence in existing artificial intelligence reporting guidelines: a scoping review. *Journal of Clinical Epidemiology*. 2025;186:111903. doi: 10.1016/j.jclinepi.2025.111903
18. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. *Trans Vis Sci Tech*. 2020;9(2):7. doi: 10.1167/tvst.9.2.7
19. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689. doi: 10.1136/bmj.m689
20. O'Shea RJ, Sharkey AR, Cook GJR, Goh V. Systematic review of research design and reporting of imaging studies applying convolutional neural networks for radiological cancer diagnosis. *Eur Radiol*. 2021;31(10):7969-7983. doi: 10.1007/s00330-021-07881-2
21. Bhandari A, Scott L, Weilbach M, Marwah R, Lasocki A. Assessment of artificial intelligence (AI) reporting methodology in glioma MRI studies using the Checklist for AI in Medical Imaging (CLAIM). *Neuroradiology*. 2023;65(5):907-913. doi: 10.1007/s00234-023-03126-9
22. Koçak B, Köse F, Keleş A, Şendur A, Meşe İ, Karagülle M. Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis. *dir*. 2025;31(5):440-455. doi: 10.4274/dir.2025.243182
23. Kim DY, Oh HW, Suh CH. Reporting Quality of Research Studies on AI Applications in Medical Images According to the CLAIM Guidelines in a Radiology Journal With a Strong Prominence in Asia. *Korean J Radiol*. 2023;24(12):1179. doi: 10.3348/kjr.2023.1027
24. Lans A, Pierik RJB, Bales JR, Fourman MS, Shin D, Kanbier LN, et al. Quality assessment of machine learning models for diagnostic imaging in orthopaedics: A systematic review. *Artificial Intelligence in Medicine*. 2022;132:102396. doi: 10.1016/j.artmed.2022.102396

25. Zhong J, Xing Y, Lu J, Zhang G, Mao S, Chen H, et al. The endorsement of general and artificial intelligence reporting guidelines in radiological journals: a meta-research study. *BMC Med Res Methodol.* 2023;23(1):292. doi: 10.1186/s12874-023-02117-x
26. Koçak B, Keleş A, Köse F. Meta-research on reporting guidelines for artificial intelligence: are authors and reviewers encouraged enough in radiology, nuclear medicine, and medical imaging journals?. *dir.* 2024;0(0):0-0. doi: 10.4274/dir.2024.232604
27. Singh S. Effect of TRIPOD+AI Guidelines on the Reporting Quality of Artificial Intelligence Prediction Models in Orthopaedic Surgery: An 18-Month Bibliometric Study. *Cureus.* 2025;17(11):e97176. doi: 10.7759/cureus.97176
28. Uribe SE, Hamdan MH, Valente NA, Yamaguchi S, Umer F, Tichy A, et al. Evaluating dental AI research papers: Key considerations for editors and reviewers. *Journal of Dentistry.* 2025;160:105867. doi: 10.1016/j.jdent.2025.105867
29. Fountzilas E, Pearce T, Baysal MA, Chakraborty A, Tsimberidou AM. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *npj Digit Med.* 2025;8(1):75. doi: 10.1038/s41746-025-01471-y
30. Fasterholdt I, Naghavi-Behzad M, Rasmussen BSB, Kjølhede T, Skjøth MM, Hildebrandt MG, et al. Value assessment of artificial intelligence in medical imaging: a scoping review. *BMC Med Imaging.* 2022;22(1):187. doi: 10.1186/s12880-022-00918-y

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Ana Rodríguez Yanes.
Writing–review and editing: Ana Rodríguez Yanes.