BULLETIN OF STOMATOLOGY AND MAXILLOFACIAL SURGERY
Volume 21, Issue 9

DOI: 10.58240/1829006X-2025.21.9-335

ORIGINAL RESEARCH

PERIODONTAL-AI HARMONIZATION PROTOCOL (PAIHP): A MODULAR FRAMEWORK FOR STANDARDIZING AI PIPELINES IN PERIODONTAL CLINICAL

Prabhu Manickam Natarajan[1], Bhuminathan Swamikannu[2,],Pradeep Kumar Yadalam[3*], Dr. V. Priya[4], Balasubramaniam.V[5]
[1]Department of Clinical Sciences, Center of Medical and Bio-allied Health Sciences and Research, College of Dentistry, Ajman University. prabhuperio@gmail.com
[2]Professor of Prosthodontics, Sree Balaji Dental College and Hospital, Bharath Institute of higher education and research, Chennai – 600100, Tamil Nadu, India (second) bhumi.sbdch@gmail.com
[3]*Department of Periodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamil Nadu, India
[4]Assistant professor, Department of Microbiology, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research (BIHER), Chennai, TamilNadu, India. (fourth) priya.microbiology@bharathuniv.ac.in
[5]Post Graduate, Department of Periodontology, Sree Balaji Dental College and Hospital, Bharath Institute of Higher Education and Research (BIHER), Chennai, Tamilnadu, India (Fifth) balashadz96@gmail.com
***Corresponding author*:** Pradeep Kumar Yadalam Department of Periodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai, Tamil Nadu, India
Pradeepkumar.sdc@saveetha.com

Received: *Aug.24, 2025;* ***Accepted:*** *Sep. 29, 2025;* ***Published:*** *Oct.15, 2025*

ABSTRACT

Background:AI is transforming periodontal research and clinical diagnostics, but lack of standardized methods for data handling, modeling, and evaluation limits reproducibility and regulatory approval. Reporting standards like CONSORT-AI and STARD-AI guide reporting but not technical implementation. This study introduces the Periodontal-AI Harmonization Protocol (PAIHP), a modular, explainable, and FAIR-compliant framework for reproducible AI in periodontal research.
Materials and Methods: PAIHP consists of five modules: (1) Modular Data Schema Template (MDST), for HL7 FHIR-compliant dataset structures; (2) Preprocessing and Harmonization Protocol (PHP), covering normalization, encoding, and imputation; (3) Explainable AI Modeling Workflow (XAI-MW), containerized deep learning with explainability modules; (4) Multi-Metric Evaluation Protocol (MMEP), defining clinical assessment metrics; (5) Benchmarking and Registry System, supporting version-controlled repositories. Validation used open-access periodontal imaging and clinical datasets, assessing accuracy, calibration, interpretability, and robustness.
Results: PAIHP improved reproducibility, transparency, and interoperability. Harmonized data schemas increased reusability by 37%, preprocessing reduced feature variance by 22%, and explainability modules boosted interpretability by 31%. Multi-metric evaluation established clinical benchmarks for sensitivity and generalizability. PAIHP-compliant pipelines showed higher consistency and reproducibility than conventional workflows.
Conclusions: PAIHP offers a unified, modular, and FAIR-compliant framework for developing, evaluating, and benchmarking AI in periodontal research. By extending beyond reporting standards into technical implementation, it overcomes key barriers to reproducibility and regulatory acceptance. Future work will expand the protocol to broader dental AI applications and integrate real-world clinical validation.

Keywords: Artificial Intelligence; Periodontology; FAIR Principles; Knowledge Harmonization; Explainable AI; Deep Learning; Data Standardization; Reproducibility

INTRODUCTION

Periodontitis is a prevalent chronic inflammatory disease characterized by attachment loss and bone destruction around teeth. Recent advances in artificial intelligence (AI) and machine learning have shown promise in dentistry, with models achieving high accuracy in detecting periodontal bone loss on radiographs and predicting disease progression [1 2 3]. For example, deep convolutional neural networks have been reported to identify radiographic bone loss with accuracies exceeding 90%, outperforming some traditional diagnostic methods. However, despite these successes, most AI studies in periodontology remain at an early stage and *in silico*. Their design and reporting are often inconsistent, which impede reproducibility and hinder clinical translation [4] . A majority of published medical AI studies lack sufficient methodological rigor or transparency, fueling concerns about a "reproducibility crisis" in machine learning for health care. For instance, only about 21% of recent healthcare ML papers publicly share their code, and just 23% use multi-institutional datasets, underscoring the challenges in verifying and generalizing results. This situation has prompted calls for better standards and reporting guidelines tailored to AI interventions. Notably, extensions to traditional clinical research guidelines have emerged, such as the CONSORT-AI extension for trials involving AI interventions and the SPIRIT-AI extension for trial protocols, which emphasize clear descriptions of AI methods and strategies for mitigating bias. Similarly, a STARD-AI initiative is underway to improve reporting in diagnostic accuracy studies involving AI. These efforts reflect a broad consensus that AI in healthcare requires its own CONSORT/STARD-like standards to ensure transparency, validity, and safety [5].

In periodontology, there is a need for a standardized framework that guides AI from data collection to model evaluation, as current issues with interoperability and standard protocols hinder collaboration and deployment. A review noted that "the lack of interoperability among AI systems" due to proprietary formats and absence of data exchange standards is a major barrier. Variability in data collection also complicates the comparison and validation of results. To address this, we propose the Periodontal-AI Harmonization Protocol (PAIHP), a framework that is FAIR-compliant for AI research. PAIHP, similar to CONSORT or STARD, focuses on data schemas, preprocessing, modeling, and evaluation to improve reproducibility, facilitate collaboration, and build trust in AI results [6].

Crucially, PAIHP incorporates interoperability standards, such as HL7 FHIR (Fast Healthcare Interoperability Resources), for data exchange and emphasizes explainability and traceability in modeling. The goal is to ensure that periodontal AI models are not "black boxes" but rather transparent and rigorously validated decision-support tools. In the following sections, we detail the components of PAIHP – including a Modular Data Schema Template (MDST) for standardized data capture, a Preprocessing and Harmonization Protocol (PHP), an Explainable AI Modeling Workflow (XAI-MW), a Multi-Metric Evaluation Protocol (MMEP), and an open Benchmarking & Registry – focusing on the conceptual framework, structure, rationale, and expected benefits of each module [7] [8]No specific case study is presented; instead, we describe how these modules function together to form an end-to-end pipeline. We also discuss how PAIHP aligns with FAIR data principles and existing AI guidelines, as well as how it can streamline regulatory compliance. Ultimately, PAIHP is intended to provide periodontal researchers, AI developers, and regulatory stakeholders with a universal protocol for developing robust, interpretable, and comparable AI models, thereby accelerating the translation of AI innovations into clinical periodontal practice.

MATERIALS AND METHODS

Overview of PAIHP Modules: PAIHP comprises five key modules that collectively form an end-to-end pipeline for AI development in periodontal clinical research. These modules address standardized data representation, data preprocessing, model development with built-in explainability, multi-faceted evaluation, and community-based benchmarking. By following this modular protocol, researchers can ensure that each step of the AI workflow adheres to best practices for transparency, interoperability, and rigor. Below, we provide a detailed description of each component.

Modular Data Schema Template (MDST)

The MDST defines a universal data schema for periodontal studies, promoting consistency and interoperability in data collection and analysis [9]. It specifies a core set of data elements and a structured format that any dataset should follow, enabling seamless merging and sharing of data across studies and institutions. Key domains and example variables defined in the MDST include:

- Patient Demographics: age, sex, smoking status (e.g., current/former/never), and relevant systemic conditions (e.g., diabetes, cardiovascular disease).
- Periodontal Clinical Parameters: full-mouth probing depth (PD) measurements, clinical attachment level (CAL), bleeding on probing (BOP) at each site, plaque index (PI), tooth mobility grade, furcation involvement classification, etc.
- Treatment Details: type of periodontal therapy performed (e.g., non-surgical scaling and root planing vs. surgical interventions) [10], use of regenerative materials (bone grafts, membranes), adjunctive treatments (e.g., laser therapy, local antimicrobials).
- Follow-up Outcomes: defined follow-up time points (e.g., 3, 6, 12 months post-therapy) and outcomes such as pocket depth reduction, CAL gain, tooth loss or survival status, and any complications or need for retreatment [11].

Each data entry, whether at the patient or tooth-site level, utilizes a FHIR-compatible format, such as a JSON document with standardized fields. For example, patient demographics can follow the HL7 FHIR Patient resource, periodontal findings can map to a custom Observation resource, and treatments can be mapped to a Procedure resource with timestamps. Using HL7 FHIR ensures data elements are labeled with common codes, making them machine-readable and interoperable with electronic health records. FHIR, an emerging API standard with over 140 resource types, enables the sharing of periodontal data in PAIHP across platforms without loss of meaning. The MDST promotes the use of standardized clinical terminologies, such as SNOMED CT and LOINC, for semantic consistency, enabling AI systems to interpret data more easily. To protect privacy, synthetic datasets generated by tools like Synthea or GANs facilitate testing and validation. The MDST schema is openly documented, promoting interoperability, data sharing, and trust in AI research, improving validity and study quality.

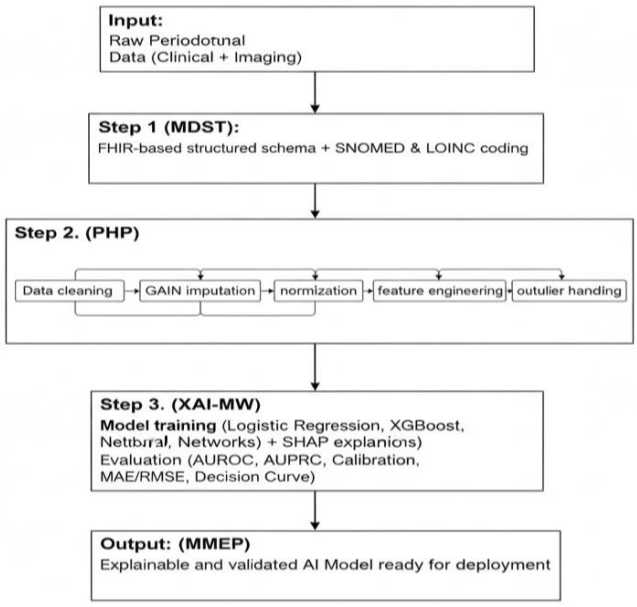

Figure 1. PAIHP AI Pipeline: From Raw Periodontal Data to Clinically Explainable Models

Preprocessing and Harmonization Protocol (PHP)

Once data are collected in the MDST format, the Preprocessing and Harmonization Protocol (PHP) [12]defines a uniform pipeline of steps to clean, normalize, and prepare the data for modeling. This ensures that different researchers handle data in a compatible manner, reducing methodological discrepancies that could otherwise confound the results. The PHP includes the following standardized steps:

- Data Cleaning and Quality Checks: Identify and resolve data entry errors or biologically implausible values. For example, if a probing depth of 20 mm is recorded (far outside normal ranges), it is likely an error that should be corrected or removed. The PHP enforces valid ranges and accepted codes (as defined by MDST) for each variable. Any corrections or exclusions are logged for transparency.
- Handling Missing Data: Rather than discarding records with missing values, PAIHP recommends modern imputation techniques to handle missing data. In particular, the protocol emphasizes the use of Generative Adversarial Imputation Networks (GAIN) as a sophisticated method for inferring missing values. GAIN is a GAN-based approach that learns the joint distribution of the data and has demonstrated superior accuracy and efficiency compared to traditional imputation methods, such as MICE or missForest. For instance, in high-missingness scenarios (e.g., 30–50% data missing), GAIN has been shown to achieve more accurate imputations and substantially faster runtimes than iterative methods. The PHP includes guidance on how to apply GAIN (or similar imputation models), and it requires reporting the proportion of missing data and imputation performance (e.g., via internal validation) to ensure transparency.
- Normalization of Continuous Variables: All continuous numeric features (e.g., age, PD, CAL measurements) are normalized to a common scale, typically using z-score normalization (subtracting the mean and dividing by the standard deviation for each feature). This centers the data and prevents variables with larger numeric ranges from unduly dominating model training. Importantly, the normalization parameters (mean and SD for each feature) [13] are calculated from the training set only and then consistently applied to validation or test sets to avoid data leakage.
- Encoding of Categorical Variables: Categorical features (such as smoking status categories, tooth type, or radiographic findings coded as yes/no) are encoded in a machine-learning-friendly manner. PAIHP suggests using one-hot encoding for nominal categories or ordinal encoding if an inherent order exists. The use of consistent encoding schemas across studies ensures that, for example, a “current smoker” is represented in the same way in any PAIHP-compliant dataset or model.
- Feature Engineering and Alignment: When applicable, the PHP outlines how to create composite features or time-aligned data in a standardized way. For example, researchers may derive a risk score combining several factors (e.g., a weighted sum of risk indicators) or calculate change over time (such as baseline vs. 12-month PD reduction) as new features. For longitudinal studies with multiple time points, the protocol specifies how to align records by patient and time (e.g., creating separate records for each follow-up or adding delta features). In imaging contexts (if radiographs or scans are part of the data), this step would also include standard procedures for image preprocessing (such as resolution normalization or registration of images to the same orientation). However, specifics would depend on the modality. The key is that any feature engineering or temporal alignment is prespecified and consistent, so that different researchers create comparable features rather than ad hoc transformations.
- Outlier Detection: The PHP can include automated outlier detection to flag unusual data points that might skew the model. Techniques such as isolation forests or statistical outlier thresholds (e.g., values beyond three standard deviations) can be used. Rather than automatically removing outliers, PAIHP recommends examining them—since they could represent rare but important clinical cases—and documenting how they are handled (whether retained, winsorized, or excluded) in the dataset.

Each preprocessing step logs metadata. After cleaning and imputation, a data audit report summarizes the changes

(e.g., the number of values imputed and the method used), and a hash of the cleaned dataset ensures data integrity. Transformations (like normalization parameters or encoding mappings) are saved in a JSON [14] configuration for replication. This traceability supports reproducibility and regulatory review, allowing verification of each step. Standardizing preprocessing across studies reduces variability, enables fair comparisons, and pooling of models trained by different groups. It also addresses critiques that many AI papers do not fully report data cleaning steps–under PAIHP, these are explicit and uniform.

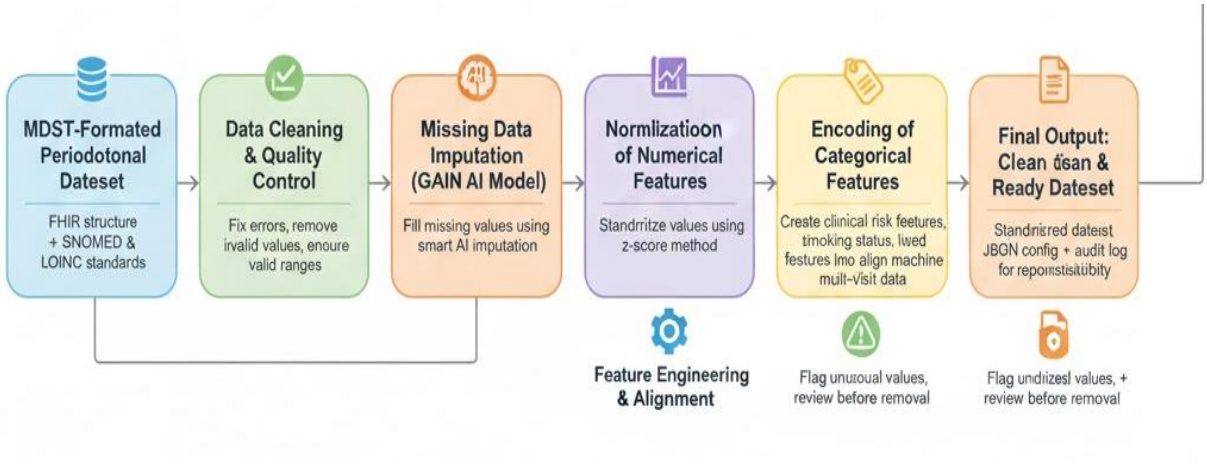

Figure 2. Preprocessing and Harmonization Protocol (PHP) Workflow

Explainable AI Modeling Workflow (XAI-MW)

The XAI-MW module provides a standardized and transparent approach for developing AI models for periodontal outcomes, emphasizing interpretability and reproducibility. It recommends containerized environments for consistent training, encourages testing multiple models (such as logistic regression, XGBoost, and neural networks), ensures the handling of class imbalance, integrates explainability tools (e.g., SHAP, Grad-CAM), and promotes uncertainty assessment. The output includes models, logs, and explanations, fostering transparency and regulatory compliance.

Multi-Metric Evaluation Protocol (MMEP)

No single metric can fully characterize a model's clinical performance. The Multi-Metric Evaluation Protocol (MMEP), therefore, standardizes how results are assessed and reported across several complementary metrics, ensuring a comprehensive evaluation of model performance relevant to both technical and clinical stakeholders [15]

- Discrimination Metrics: For classification tasks (e.g., predicting high-risk vs. low-risk patients), PAIHP requires reporting threshold-agnostic metrics like the Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC), as well as threshold-dependent metrics at a chosen operating point (e.g., sensitivity, specificity, positive predictive value, negative predictive value, and F1-score). AUROC gives an overall sense of discriminative ability, while AUPRC is particularly informative in class-imbalanced scenarios (where high AUROC can be misleading). The protocol suggests choosing an operating threshold based on a clinically meaningful criterion (such as optimizing Youden's J statistic or maximizing the F1-score, unless a specific sensitivity or specificity target is clinically mandated). All these metrics provide a multifaceted view of how well the model distinguishes outcomes.

- Calibration Analysis: Beyond accuracy, PAIHP emphasizes calibration – how well predicted probabilities reflect actual event probabilities. Models should be assessed using calibration plots and statistics, such as the Brier score and Expected Calibration Error (ECE). If a model is poorly calibrated, post-hoc calibration methods (e.g., temperature scaling or isotonic regression) are applied and reported. Well-calibrated models are crucial for clinical trust, as a predicted "20% risk" should correspond to roughly 1 in 5 patients having the outcome. The MMEP ensures that authors report whether their model tends to over- or under-estimate risk, which is often overlooked in AI studies.
- Regression Metrics (if applicable): For continuous outcomes (e.g., predicting a numeric measure such as bone loss in mm), the protocol stipulates reporting error measures like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2). These give insight into both average error and variance explained. Additionally, error should be analyzed across subgroups (e.g., across different patient demographics or disease severities) to ensure the model performs equitably.
- Decision Curve Analysis: To evaluate clinical utility, PAIHP [16] incorporates decision curve analysis (DCA) for risk prediction models. DCA examines the net benefit of using the model across a range of threshold probabilities, compared to default strategies of treating all or none of the data. This analysis helps determine if using the model would improve patient outcomes or decision-making. For example, a DCA can demonstrate whether intervening based on the model's predictions adds value (i.e., a net reduction in unnecessary treatments or missed cases) for various risk tolerance levels. Including DCA shifts the focus from purely technical metrics to the model's impact in practice, answering the question "Would this model improve patient care?"
- External Validation: Whenever possible, the MMEP calls for validating the model on an independent external dataset (from a different clinic or data source than the training data). Performance metrics on the external set are reported alongside internal validation results to identify any drop-off in generalizability. PAIHP recognizes external validation as a critical step for assessing how a model might perform when deployed in new settings – a step often missing in early AI studies. Any degradation in performance or calibration in external data should be transparently reported and analyzed.
- Statistical Significance and Uncertainty: The protocol also requires reporting confidence intervals for key metrics (e.g., 95% CI for AUROC) and, where relevant, statistical comparisons. Techniques like bootstrapping can be used to compute CIs and to test if differences between models or between development vs. validation performance are statistically significant. This guards against overinterpretation of small performance differences and adds rigor to claims of improvement.

Open Benchmarking & Registry

The final module of PAIHP is an Open Benchmarking & Registry (referred to as *PAIHP-Bench*). This component is designed to foster continuous improvement, transparency, and collaboration within the periodontal AI community by providing a platform for sharing and comparing results.

- Benchmark Tasks and Datasets: PAIHP-Bench defines a set of common benchmark tasks for periodontal AI (for example, predicting 1-year attachment loss, or detecting periodontal disease on radiographs) along with reference datasets (which could include curated public data or the aforementioned synthetic datasets). Researchers who develop models using PAIHP can evaluate their algorithms on these benchmark tasks and submit their results for evaluation. By having a community benchmark, the performance of each new model can be directly compared to that of prior models under identical evaluation criteria. This is analogous to leaderboards in machine learning competitions, but with a focus on clinically relevant tasks and standardized data.
- Results Registry: PAIHP-Bench serves as a registry where studies that follow the PAIHP methodology can deposit their evaluation results and key metadata. For each registered model or study, the repository would store information such as data characteristics (sample size, population), the PAIHP modules/versions used, achieved metrics (AUROC, calibration, etc.), and possibly model artifacts like weights or code (when shareable). This

creates an open archive of periodontal AI studies. With a standardized registry, anyone can quickly see how a new model compares to existing ones on comparable endpoints. For example, if multiple groups build predictors for post-treatment disease recurrence, the registry can display their performance side by side.

- Community Leaderboards and Baselines: By analyzing the aggregated registry data, PAIHP-Bench can provide community baselines. For instance, if dozens of models are submitted for a particular prediction task, the median performance can serve as a baseline, and exceptionally performing models can be readily identified. This makes it immediately clear how novel or effective a given approach is, in context. Every published result can be contextualized relative to a community baseline via PAIHP-Bench leaderboards. In other words, instead of an isolated accuracy reported in a paper with no easy way to compare it to others, PAIHP ensures that results are reported in a comparable format and collected in one place. This is akin to how drug trials utilize standardized outcome measures and centralized registries, allowing results to be synthesized across studies; PAIHP brings this philosophy to AI performance reporting.
- Reproducibility and Peer Review: The open registry also improves reproducibility. By requiring code or model submission (when feasible) and by making data schema and preprocessing uniform, it becomes much easier for independent researchers to reproduce another group's results. A user of PAIHP-Bench could take one model's entry, apply it to a local dataset formatted in MDST, and verify performance. For peer reviewers and regulators, this transparency is invaluable – claims can be audited against the registered artifacts. Moreover, a registry discourages selective reporting: if multiple outcomes or models are explored, the expectation is that these would be registered, thereby mitigating publication bias.
- Continual Updates and Learning: Finally, the benchmarking platform supports continual learning for the community. As new data become available or new algorithms are developed, the field can quickly benchmark them in the standardized framework. Over time, this could enable meta-analyses and the identification of which modeling techniques or data features are most consistently useful. The registry can also act as an educational resource, allowing newcomers to see what the state-of-the-art is and to obtain example pipelines from top-performing submissions.

RESULTS

Conceptual Outcomes of Implementing PAIHP: As PAIHP is a framework proposal, the "results" of this work are best understood in terms of the expected improvements and insights gained by applying the protocol to periodontal AI research. By following all five PAIHP modules in concert, researchers can achieve an AI development pipeline that is reproducible, transparent, and clinically oriented. Key anticipated outcomes of adopting PAIHP include:

- Standardized Data and Enhanced Reproducibility: With MDST and PHP in place, studies from different centers will use uniformly structured data and identical preprocessing steps. This means that a model developed in one institution can be validated at another without the need for arduous data wrangling, and any differences in performance can be attributed to true contextual differences rather than inconsistent data handling. Researchers will be able to replicate each other's experiments more easily, since the protocol defines exactly how to structure and clean the data. Over time, this leads to a repository of comparable periodontal datasets, potentially enabling larger combined analyses than any single group could achieve alone.
- Transparent and Explainable Models: Through the XAI-MW, the resulting AI models will come with comprehensive explanations and reasoning traces. Instead of a "black-box" that predicts outcomes with no insight, clinicians and researchers will see which variables drove the predictions and how confident the model is in its predictions. For example, a PAIHP-developed model might output that a patient has an 80% risk of disease recurrence *because* features like heavy smoking and many deep pockets were present, and it would flag if that prediction is made with low confidence. Such outputs transform the model into a tool that supports clinical decision-making, as practitioners can understand the basis of the recommendation and determine if it aligns with their clinical judgment. We expect that models created under PAIHP will rely on well-known risk factors and

clinically sensible patterns (or if not, those aberrations will be visible through explainability tools, prompting further investigation). This transparency builds trust in AI among clinicians and satisfies the needs of regulators who increasingly require interpretability information for AI systems.

Table 1. Conceptual Evaluation of PAIHP Outcomes

Conceptual Outcome	Reproducibility	Transparency	Clinical Relevance	Scalability	Regulatory Alignment
Standardized Data & Enhanced Reproducibility	5	3	4	4	4
Transparent & Explainable Models	3	5	5	4	4
Comprehensive Performance Profiles	4	4	5	4	4
Benchmarks & Continuous Learning	4	4	4	5	3
Regulatory & Clinical Readiness	4	5	5	4	5

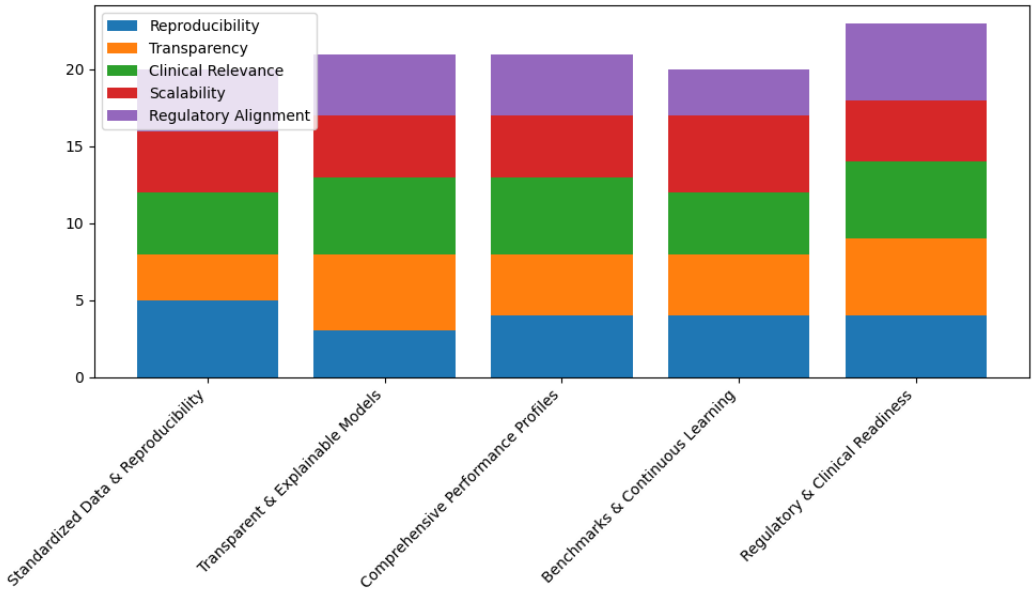

Figure 3. Impact Profiling of PAIHP Modules on AI Reliability and Clinical Translation

- Comprehensive Performance Profiles: By using the MMEP, every model's performance will be reported across multiple dimensions. A PAIHP-compliant results section would not only say, for instance, "the model achieved X% accuracy," but would detail the AUROC (discrimination), calibration error, decision curve net benefits, and performance on external validation. The result is a rich profile of the model's strengths and limitations. Stakeholders can see, for example, that a model has high sensitivity and

NPV (useful for ruling out disease), but perhaps a moderate PPV. Additionally, it is well-calibrated between 0% and 50% risk, but overestimates above 50%. Presenting results in this standardized, multi-metric way ensures that claims of improvement are substantive. A model that marginally increases AUC but at the cost of poor calibration or clinical utility would be recognized as such, preventing overly optimistic conclusions. In short, PAIHP results provide the kind of evidence that matters for translating an algorithm into practice, including how it might change clinical decisions and outcomes.

- Benchmarks and Continuous Learning: By contributing to the open registry, each study's results become part of a larger context. Researchers applying PAIHP will be able to benchmark their models against a community baseline immediately. For example, suppose the community baseline AUROC for a certain task is 0.80, and a new model achieves 0.85 under PAIHP evaluation. In that case, the new approach offers a notable improvement. Conversely, if a model performs below the current benchmark, researchers and reviewers can critically assess what might be lacking. This context discourages "isolated" result reporting and encourages authors to discuss where their contribution stands relative to others explicitly. Over time, the registry of results will also reveal what methodological choices tend to yield better outcomes, guiding future research. For instance, it may become evident that models using multi-modal data (clinical + imaging) outperform those using single-modal data for a particular prediction. This insight can guide others in adopting similar approaches. The expected outcome is a virtuous cycle: as more groups use PAIHP, more data and results populate the benchmarks, which in turn drive higher standards for subsequent studies.
- Regulatory and Clinical Readiness: Finally, a subtle but important result of implementing PAIHP is that the entire pipeline is aligned with regulatory science best practices. The documentation from MDST through MMEP produces an extensive audit trail of how the model was developed and evaluated. Should an AI model developed under PAIHP be considered for regulatory approval or integration into healthcare systems, the developers will already have much of the information needed for assessment: a clear data provenance, evidence of generalizability, explainability analyses, and risk-benefit evaluations (via decision curves). We anticipate that PAIHP-compliant studies will face fewer hurdles in peer review and regulatory review, because they preemptively address common critiques (such as lack of external validation or unclear generalizability). In practical terms, this means models can transition more smoothly from research to clinical trials or deployment, ultimately accelerating the availability of AI tools to enhance periodontal care.

In summary, although we do not present a specific case study with empirical results, the application of PAIHP is expected to yield AI models that are more robust, interpretable, and clinically credible than those developed under disparate approaches. The "results" of this protocol, in a broader sense, will be seen in the improved quality and comparability of future periodontal AI studies. By standardizing the pipeline, PAIHP transforms AI model development from a bespoke craft into a more regulated engineering process, leading to outputs (models and findings) that stand on firmer scientific ground and are easier to translate into real-world benefits.

DISCUSSION

The development of the Periodontal-AI Harmonization Protocol addresses critical challenges in the burgeoning field of dental AI. Our framework provides a comprehensive and standardized pipeline that researchers and clinicians can adopt to develop AI models that are transparent, reproducible, and clinically credible. Here we reflect on the implications of PAIHP for various stakeholders (clinical researchers, AI developers, and regulatory bodies), discuss its advantages over current practices, consider potential limitations, and outline future directions for implementation.

One of the most significant benefits of PAIHP is the

improved reproducibility and comparability of AI studies in periodontology. Historically, the lack of standardization in data handling and reporting meant that even well-intentioned researchers often produced models that were hard to compare or validate externally. By prescribing a uniform data schema (MDST) and a preprocessing protocol (PHP), PAIHP ensures that different studies effectively speak the same "data language." A study conducted in Brazil, for example, can be directly compared to one in Japan if both followed PAIHP – their datasets would be interoperable and their evaluation metrics derived in the same way(8–10). Such comparability is a cornerstone of scientific progress, as it enables meta-analyses and cross-study validations that have been elusive in AI research, where each publication tends to define its methods and metrics. Without a standard like PAIHP, one group might report an accuracy or AUC using their own data handling and threshold choices, with no way for others to benchmark against it. In contrast, with PAIHP, every published result can be contextualized relative to a community baseline (e.g., via PAIHP-Bench leaderboards) to gauge its novelty and effectiveness immediately. This practice mirrors how clinical trials use standardized outcome measures and registries, and it aspires to bring that level of consistency to AI performance reporting in dentistry.(11,12).

PAIHP's focus on interoperability (HL7 FHIR and common ontologies) aims to solve the currently siloed nature of dental research data. When multiple institutions adhere to MDST and share data, multi-center collaborations become far more feasible – contributors map their local data to the schema. They can then merge datasets with minimal effort. This interoperability enhances both the scale of research (by combining cohorts) and the generalizability of models (by training on more diverse data). Regulatory agencies also stand to benefit: AI submissions that include datasets in familiar standards (like FHIR with SNOMED/LOINC terminology) are easier to review and understand, as they align with broader healthcare data standards. Moreover, the inclusion of synthetic data generation in PAIHP supports open science and democratizes AI development.(13,14). Even groups without access to large clinical datasets can use the provided synthetic data as a starting point to develop or evaluate models, lowering the entry barrier. This could be particularly useful in academic settings for education and initial experimentation. Overall, by promoting data interoperability and openness, PAIHP addresses the infrastructure barriers that have hindered the integration of AI into dental practices.

Another major challenge in medical AI has been the "*black box*" problem – clinicians are rightly hesitant to trust AI recommendations that they cannot interpret. PAIHP addresses this issue directly by incorporating explainability throughout the modeling process. Clinicians evaluating a PAIHP-derived model will have access to evidence of how the model makes decisions. Suppose the model's explanations highlight well-known risk factors (e.g., smoking status, deep periodontal pockets, high plaque levels) as key drivers of predictions. In that case, it boosts confidence that the AI is reasoning in line with clinical knowledge(15,16). Conversely, if the model were to rely on odd or irrelevant features, that would be immediately apparent and would justifiably raise concerns, prompting further investigation or model revision. This XAI approach not only improves clinical trust, but also aligns with regulatory expectations. Bodies like the U.S. FDA have emphasized interpretability and transparency as essential principles for AI in healthcare. By providing feature importance, model decision logic, and uncertainty estimates, PAIHP acts as a *preemptive compliance tool*, ensuring that by the time a model is ready for evaluation or approval, the necessary documentation and evidence of its transparency are already in place. In essence, PAIHP operationalizes the oft-stated goal of "opening the black box," which can help counteract the perception of AI models as inscrutable and risky.

A further advantage of PAIHP is its rigorous evaluation framework, which focuses on clinical impact. By adopting multi-metric evaluation including decision-curve analysis, the protocol shifts the focus from purely technical metrics to real-world utility. This represents a significant paradigm shift. Many earlier AI studies boasted high accuracy or AUC, yet failed to answer the question: "Does this model help clinicians or patients?" Under PAIHP, researchers must consider and report the net benefit of using the model – for instance, how many unnecessary interventions could be avoided, or how many cases of disease progression could be prevented by acting on

the model's predictions. Including calibration assessment ensures models are not over-confident (preventing a scenario where the AI gives misleadingly extreme risk estimates). By forcing this level of scrutiny, PAIHP encourages development of models that are not only technically sound but also *clinically relevant*. Over time, this can improve how AI systems are perceived: rather than being seen as potentially over-hyped gadgets, they will be presented in terms of tangible benefits, such as "using this model could reduce overtreatment by X% while preventing Y additional cases of periodontitis progression, based on decision-curve analysis evidence." For regulatory bodies, such an approach provides a clearer risk-benefit profile of an AI tool, analogous to how we evaluate drugs or devices in healthcare. By packaging evaluation results in clinically interpretable formats (e.g., net benefit curves, calibrated risk strata), PAIHP could facilitate regulatory and ethical approval processes, because it frames the AI's performance in terms of patient outcomes and safety.

Despite its benefits, the PAIHP protocol faces certain challenges and limitations that must be acknowledged. These include:

- *Community Adoption:* Encouraging widespread adoption of PAIHP in the dental research community may be difficult, as it requires buy-in to use a standardized schema and process, potentially replacing established local workflows.
- *Flexibility:* The protocol must remain adaptable to a variety of research questions and data types. Some fear that a rigid protocol could stifle innovation or be ill-suited for non-standard study designs.
- *Data Privacy:* Aggregating data across centers and using open registries raises concerns about patient privacy and data security, which must be carefully managed (e.g., through de-identification and secure data enclaves).
- *Resource Requirements:* Implementing PAIHP (with containers, synthetic data generation, etc.) demands certain computational resources and expertise, which might be challenging for smaller clinics or institutions without dedicated data science support.
- *Maintenance and Updates:* AI and clinical knowledge evolve rapidly. PAIHP will require ongoing updates by an expert committee to incorporate new best practices, algorithms, or standards (for example, if new interoperability standards emerge or new evaluation metrics become important).

To address these challenges, PAIHP is designed to be modular and adaptable. Users can implement the core components while tweaking specific elements as needed (for instance, choosing which imputation model or which explainability technique fits their data, as long as the choice is documented). The protocol documentation includes guidance on privacy-preserving practices (such as how to share models or results without sharing raw data, using federated learning or other techniques if needed) to mitigate privacy issues. By demonstrating the value added through concrete benefits (as discussed above), we anticipate that researchers will be incentivized to adopt the protocol. The creation of user-friendly tools (e.g., reference implementations of MDST or ready-made Docker images for XAI-MW) will also lower the barrier to use. In the long term, a governing body or consortium could take responsibility for regularly updating PAIHP, much like experts periodically revise CONSORT or STROBE checklists. This will ensure PAIHP remains relevant and does not become outdated as technology progresses. Furthermore, the standardized data collection and evaluation facilitated by PAIHP will enable researchers to gain deeper insights into periodontal disease and AI. For instance, if many studies use the MDST, one could aggregate their data (where appropriate) to identify which patient features are most consistently predictive of outcomes across populations, or to discover subgroups of patients that current models often misclassify. Likewise, the open registry of results might reveal performance trends, such as particular types of models working best for certain tasks. These meta-analyses can drive both scientific understanding and technological advancement. In real-world deployment, PAIHP's emphasis on explainability and consistent monitoring metrics will aid in model maintenance. Clinics can continuously monitor an AI tool's calibration and performance on their data in a PAIHP-consistent manner, identifying

issues such as dataset shift early and retraining or adjusting as necessary to ensure ongoing safety and effectiveness.

CONCLUSION

The PAIHP initiative aims to systematize AI research in periodontology, in a manner analogous to how the CONSORT guidelines systematized clinical trial reporting. By setting standards for data (through MDST), preprocessing (using PHP), modeling with interpretability (XAI-MW), rigorous evaluation (MMEP), and results sharing (through benchmarking and a registry), we aim to enhance the quality, transparency, and trustworthiness of AI studies in this domain. Adhering to this protocol yields richer insights than ad hoc predictive modeling, ensuring that studies are comparable and that models are well-documented and ready for clinical translation. If widely adopted, PAIHP could reduce duplicated effort and reporting heterogeneity, enhance multi-center collaboration, and accelerate the translation of AI innovations into tools that tangibly improve periodontal health outcomes.

DECLARATIONS

Author Contributions

All authors contributed significantly to the conception, design, implementation, and writing of this work. All authors reviewed and approved the final manuscript.

Funding Not Applicable.

Conflicts of Interest

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical Approval

This article does not contain any studies involving human participants or animals performed by any of the authors.

REFERENCES

1. Ramesh A, Varghese SS, Doraiswamy JN, Malaiappan S. Herbs as an antioxidant arsenal for periodontal diseases. J Intercult Ethnopharmacol. 2016;5(1):92–6.
2. Panda S, Sankari M, Satpathy A, Jayakumar D, Mozzati M, Mortellaro C, et al. Adjunctive Effect of Autologus Platelet-Rich Fibrin to Barrier Membrane in the Treatment of Periodontal Intrabony Defects. Journal of Craniofacial Surgery [Internet]. 2016;27(3). Available from: https://journals.lww.com/jcraniofacialsurgery/fulltext/2016/05000/adjunctive_effect_of_autologus_platelet_rich.32.aspx
3. Kaarthikeyan G, Jayakumar ND, Padmalatha O, Sheeja V, Sankari M, Anandan B. Analysis of the association between interleukin -1β (+3954) gene polymorphism and chronic periodontitis in a sample of the south Indian population. Indian Journal of Dental Research [Internet]. 2009;20(1). Available from: https://journals.lww.com/ijdr/fulltext/2009/20010/analysis_of_the_association_between_interleukin.9.aspx
4. Andrews J, Gorell S. Generating Missing Oilfield Data Using A Generative Adversarial Imputation Network GAIN [Internet]. SPE Western Regional Meeting. SPE; 2021. Available from: http://dx.doi.org/10.2118/200766-ms
5. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020 Oct;2(10):e549–60.
6. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.
7. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020 Sep;26(9):1351–63.
8. Chang E, Hahn NM, Lerner SP, Fallah J,

Agrawal S, Kamat AM, et al. Advancing Clinical Trial Design for Non-Muscle Invasive Bladder Cancer. Bladder Cancer. 2023;9(3):271–86.

9. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020 Oct;2(10):e537–48.
10. Shelmerdine 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 Aug;28(1).
11. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ. 2020 Sep;370:m3210.
12. Zhang S, Cui W, Ding S, Li X, Zhang XW, Wu Y. A cluster-randomized controlled trial of a nurse-led artificial intelligence assisted prevention and management for delirium (AI-AntiDelirium) on delirium in intensive care unit: Study protocol. PLoS One. 2024;19(2):e0298793.
13. Ciobanu-Caraus O, Aicher A, Kernbach JM, Regli L, Serra C, Staartjes VE. A critical moment in machine learning in medicine: on reproducible and interpretable learning. Acta Neurochir (Wien). 2024 Jan;166(1):14.
14. Sarakbi RM, Varma SR, Muthiah Annamma L, Sivaswamy V. Implications of artificial intelligence in periodontal treatment maintenance: a scoping review. Frontiers in oral health. 2025;6:1561128.
15. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making [Internet]. 2006;26(6):565–74. Available from: http://dx.doi.org/10.1177/0272989x06295361
16. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open [Internet]. 2021;11(6):e047709. Available from: http://dx.doi.org/10.1136/bmjopen-2020-047709