American National Standard/
American Dental Association

Technical Report No. 1109

Dentistry—Evaluation of Dental Image Analysis Systems Using Augmented/ Artificial Intelligence

Standards Consensus Body 12—AI and Knowledge Management

ADA American Dental Association®

2025

AMERICAN DENTAL ASSOCIATION TECHNICAL REPORT NO. 1109 FOR DENTISTRY – EVALUATION OF DENTAL IMAGE ANALYSIS SYSTEMS USING AUGMENTED/ ARTIFICIAL INTELLIGENCE

ADA Consensus Body 12 on AI and Knowledge Management has approved ADA Technical Report No. 1109, *Dentistry – Evaluation of Dental Image Analysis Systems Using Augmented/Artificial Intelligence.*

The scope of ADA Consensus Body 12 on AI and Knowledge Management is:

> *Development of standards deliverables for nomenclature and requirements for quality, integrity, aggregation, organization and analysis of patient-centric information, knowledge representation and augmented intelligence/artificial intelligence for dentistry.*

ADA Consensus Body 12 has representation from appropriate interests in the United States in the standardization of products and technologies within its scope.

Approval of ADA Technical Report No. 1109 was granted by ADA Consensus Body 12 on AI and Knowledge Management on January 21, 2025.

This is the first edition of ADA Technical Report No. 1109.

ADA Consensus Body 12 on AI and Knowledge Management thanks the members of Working Group 12.7 and the organizations with which they were affiliated at the time the standard was developed:
Robert Faiella (chair), Individual Expert, Cotuit, MA;
Veeratrishul Allareddy, University of Iowa College of Dentistry, Iowa City;
Kenneth Aschheim, Office of The Chief Medical Examiner for the City of New York, NY;
Scott Benjamin, Advanced Integration & Mentoring, Hancock, NY;
Arnaud Berthier, Dental Monitoring SAS, Austin, TX;
Sonia Bordin-Aykroyd, Visage Laser Institute, Grapevine, TX;
Carolyn Brown, MouthWatch LLC, Ponte Vedra Beach, FL;
Michael Colvard, University of Illinois at Chicago;
Claudia Cotca, Washington Institute for Dentistry & Laser Surgery, Chevy Chase, MD;
William Duncan, University of Florida, Newberry;
Josephine Esquivel-Upshaw, University of Florida College of Dentistry, Gainesville;
Adam Foresman, Videa Health, Inc., Chicago, IL;
Paul Glassman, California Northstate University, Greenbrae, CA;
Stephen Glenn, Individual Representative, Tulsa, OK;
Gary Goodell, Individual Representative, Colorado Springs, CO;
Gary Guest, University of Texas San Antonio;
Kimberly Harding, Individual Representative, Rockville, MD;
Ray Jeter, Augusta University - The Dental College of Georgia;

Mark Jurkovich, Health Partners Institute, Minneapolis, MN;
Reza Khosravi, Dental AI Association, Palm Desert, CA;
Karen Panetta, Tufts University School of Dental Medicine, Medford, MA;
Margaret Scarlett, Scarlett Consulting International, Atlanta, GA;
Joel White, Marquette University School of Dentistry, Half Moon Bay, CA; and
Gregory Zeller, Individual Representative, Lutherville Timonium, MD.

AMERICAN DENTAL ASSOCIATION TECHNICAL REPORT NO. 1109 FOR DENTISTRY – EVALUATION OF DENTAL IMAGE ANALYSIS SYSTEMS USING AUGMENTED/ ARTIFICIAL INTELLIGENCE

Foreword

(This Foreword does not form a part of ADA Technical Report No. 1109 for Dentistry – Evaluation of Dental Image Analysis Systems Using Augmented/Artificial Intelligence).

Artificial and augmented intelligence for dental image analysis is progressing rapidly. By understanding key principles of machine learning training and validation, and by asking key questions about intended use and system performance, clinicians can ensure the best patient care while reaping the benefits of an advancing technology.

This technical report was prepared by ADA Standards Working Group 12.7 on Augmented Intelligence in Dentistry. The working group chair is Robert Faiella. The working group wishes to acknowledge the assistance of Joel Karafin and Margaret Scarlett in preparing this report, along with Karen Panetta and others, which was done at the request of Gary Guest, chair of ADA Consensus Body 12 on AI and Knowledge Management.

AMERICAN DENTAL ASSOCIATION TECHNICAL REPORT NO. 1109 FOR DENTISTRY – EVALUATION OF DENTAL IMAGE ANALYSIS SYSTEMS USING AUGMENTED/ ARTIFICIAL INTELLIGENCE

Introduction

2D imaging in the form of dental radiographs, capture radiographic findings in the form or radiographic radiolucencies and radiopacities results of radiographic tests. Dentists utilize radiographic tests and then go through a process of radiographic interpretation, what they think it is in the mouth. Dentists then confirm the radiographic interpretations by completing a dental examination to derive findings, conditions and make diagnosis of health and disease.

2D AI predications referred to in this technical report are in the context of aiding the dentist to do radiographic interpretations of radiographic radiolucencies and radiopacities which are confirmed by clinical examination to determine findings, conditions and make a diagnosis. Radiographic interpretation, combined with clinical examination, intraoral photographs/scans, find ings, conditions, diagnosis and procedures, are all beneficial for AI predictions.

It should also be noted that images can be processed to enhance the precision and accuracy of the radiographic images. Raw images allow for AI predictions that include errors of technique as well as artifacts. Images can be processed to enhance the precision and accuracy of the radiographic images. These processed images allow for better predictions and improved AI available to the clinician.

2D imaging are most commonly dental radiographs (dental x-rays) and optical images (photographs). The latter may be used alone or in combination with radiographic images. Radiographs may also include 2D images obtained from 3D radiographs. This technical report is applicable to all 2D imaging modalities.

Today, clinical decision support with Artificial or Augmented Intelligence (AI) is available for dental practitioners for automated image analysis of dental images of common oral diseases and conditions, notably dental caries and periodontal diseases.

This technical report (TR) is intended for AI vendors, with an overview of key principles and methods for use of AI as clinical decision support, including external validation of AI algorithms for two-dimensional images. A secondary audience is dental clinicians and prospective vendors of AI products. This report is limited to AI products with static algorithms and single mode, not multi-modal products.

Purpose

The purpose of this Technical Report is to:

a) Discuss external validation criteria for use cases of AI that automate interpretation of two-dimensional images with AI, while ensuring privacy and security of patients as the source of images;
b) Describe general characteristics of the development of an independent database for validation of two-dimensional images;
c) Provide methods to clarify specific use cases by vendors, while retaining autonomy of clinical judgement by clinicians, ensuring patient safety and efficacy; and

d) General considerations for a simple labeling of products, with explainability, by AI vendors for various use cases.

Background

While AI development issues for developing training and internal validation test data are standardized in the engineering field, external validation of AI is needed for this emerging clinical decision support tool. Creating and maintaining a validation database should be representative of the population at large, with additions to make this more representative over time. A prior ADA white paper on AI has outlined some key issues. [1]

A model for construction of an external validation dataset is proposed with establishment of "ground truth." By external, we mean one that is not obtained from a manufacturer of an AI product from their dataset. Additionally, considerations for clear collection, annotation, and collation will be described in general, with a proposed label for AI products using two-dimensional images for specific use cases. A logical framework for image collection will be described. The image set used by the manufacturer to create the use case may include a specific population and it is important for the clinician to understand whether the images in the dataset in a particular AI software are applicable to a specific patient for this particular use case, and other objective information about system's performance and its labeling.

Any labeling would include a dashboard of criteria for use cases so clinicians can easily compare various AI products for interpretation of two-dimensional images with specific use-case claims. Clear and simple labeling by manufacturers would be expected to span Food and Drug Administration (FDA) approved software as a medical device (SaMD) categories Class I, II and III, a risk-based approach. Since FDA only requires specific labeling of Class II and III products, labeling would also include exempt Class I devices. Considerations for labeling would be weighted towards explainability of the algorithms for various clinical decision support tools for two dimensional images.

The need and characteristics of an independent third party to validate AI tools for two-dimensional images is described. A logical framework for how an independent database for validation would be collected and organized from various sources is presented. While the clinician is expected to retain autonomy of decision-making for diagnosis and a treat/no treat decision, explainability of AI products is important, with the method for assessment as important as the "answer" provided by any AI Product.

Clinicians find that AI provides an opportunity to educate patients about their own individual issues within specified parameters; for patients, this can provide information about comparative data provided by AI for clinical decision support to practitioners, and the opportunities and limitations of AI dental use cases. Simple labeling of AI SaMD for each use case claim can assist clinicians in assessing 'what" and "how" a product is utilized. Simple labeling, similar to Nutrition Facts for food, can be utilized so that clinicians can easily compare and contrast various use cases for different AI products for clinical decision support, and assess risks and benefits. A label for a specific use claim is ideal, while addressing multiple claims in a single label for a single product should be usable.

The Appendices provide detailed information for clinicians on types of AI machine learning tools and development of AI products, as well as how the FDA views clinical decision support and overall evaluation. Key principles of Machine Learning (ML) at AI's core will be described, including its training and internal validation. Finally, standard methods for evaluation of AI will be described, including considerations of risks and benefits. These ML models do not currently include Large Language Models (LLM). As a regulatory body for SaMD and risk determination of AI tools, the Food and Drug Administration (FDA) resources on AI will be described; further information for product clearance and approvals or product exemption are on the FDA website. The Appendices and Glossary are provided to assist clinicians in providing the best patient care, while reaping the benefits of this advancing technology.

While this TR is limited to two-dimensional image validation, the principles could be applied to three-dimensional images through other standards in the future.

AMERICAN DENTAL ASSOCIATION TECHNICAL REPORT NO. 1109 FOR DENTISTRY – EVALUATION OF DENTAL IMAGE ANALYSIS SYSTEMS USING AUGMENTED/ARTIFICIAL INTELLIGENCE

Dedication

This ADA Technical Report is dedicated to Joel Karafin, who was the primary author, and who contributed greatly to the understanding of AI in dentistry and its relationship to the quality and warehousing of two-dimensional images for external validation of various AI products.

AMERICAN DENTAL ASSOCIATION TECHNICAL REPORT NO. 1109 FOR DENTISTRY – EVALUATION OF DENTAL IMAGE ANALYSIS SYSTEMS USING AUGMENTED/ARTIFICIAL INTELLIGENCE

1 Scope

This Technical Report describes general principles of AI and a rationale for proposed methods for external validation. These include safety and performance of the AI tool, avoidance of bias, and privacy and security. Included in the report is a logic framework for independent external validation of two-dimensional images with AI for common dental diseases, conditions, and disorders by an independent body.

2 General Principles and Definitions of AI in Healthcare

AI in healthcare is defined as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments in healthcare, health systems or by health clinicians."[2] These are further described in the ADA White Paper No. 1106.[1] Key principles for guidance on use of AI in healthcare are described. For example, the World Health Organization (WHO) has described six core principles for use of AI in healthcare as: (1) protect autonomy; (2) promote human well-being, human safety and the public interest; (3) ensure transparency, explainability, and intelligibility; (4) foster responsibility and accountability; (5) ensure inclusiveness and equity; (6) promote AI that is responsive and sustainable."[2] In particular, WHO emphasizes the need for privacy and security of data and methods to avoid bias. On the issue of bias, WHO states, "the data used to train AI may be biased, generating misleading or inaccurate information that could pose risks to health, equity and inclusiveness."[2] Additional background information for both developers and clinicians on various types of AI is included in Appendix A.

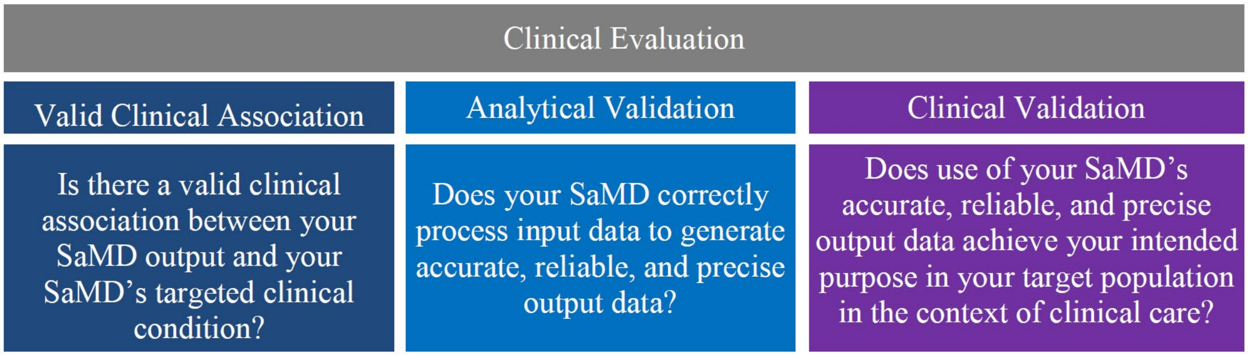
2.1 US Regulatory Agency is Food and Drug Administration (FDA) for AI in healthcare, including dentistry

While the European Union (EU) has already adopted regulatory standards for AI for implementation in 2024, emerging AI tools in the US lag behind the EU in the development of a general regulatory framework based on risks to users for AI in general, including healthcare. The FDA is the regulatory agency responsible for AI in healthcare and dentistry, regulating marketing claims of any AI product. FDA has established overall methods of evaluation by FDA and evaluation of AI clinical decision support tools, described in Appendix B for ease of access. The current approach is mostly a sectoral approach to AI at the time of this writing; however, it is anticipated that as a dynamic field, there will be many rapid additions.

However, some have noted that there is a knowledge gap between US FDA Clearance and how clinicians use AI Algorithms. [3] This makes a strong case for construction of an external validation dataset.

2.1.1 Clinical Evaluation is considered by FDA in the following ways:

Clearance for marketing by FDA for Software as a Medical Device (SaMD), which includes dental image AI systems, has been outlined in ADA White Paper, 1106.[1] The FDA describes three activities which should be part of an ongoing life cycle process by a system's developer. Additional information about questions on this is included in Appendix B.



Clinical Evaluation

Valid Clinical Association

Is there a valid clinical association between your SaMD output and your SaMD's targeted clinical condition?

Analytical Validation

Does your SaMD correctly process input data to generate accurate, reliable, and precise output data?

Clinical Validation

Does use of your SaMD's accurate, reliable, and precise output data achieve your intended purpose in your target population in the context of clinical care?

Figure 1. Types of clinical evaluation for software as a medical device (SaMD) by FDA.
Source: FDA

2.1.2 Regulations for Ensuring Transparency of AI Software as a Medical Device Emphasized by the National Coordinator for Health IT (NCHIT)

The National Coordinator for Health IT (NCHIT) is the governmental entity in the Department of Health and Human Services (HHS) that regulates use of health Information technology (health IT) across various parts of the health system. For AI in healthcare, there is insufficient transparency in the quality and scope of the internal validation datasets currently being used. External data from sources other than those obtained for model training and testing data are needed beyond that for model creation. In fact, performance on datasets other than the original dataset may result in poorer performance.[4] The question is whether performance of a particular model will be consistent across different or more representative datasets. In the past, validation datasets for AI/ML algorithms were rarely over 1000 patients (9/118) and the problem of transparency and lack of adequate sample size for internal validation data is noted in a NCHIT regulatory document of December 2023.[5] As of 2024, this entity is housed in the Assistant Secretary for IT Policy. (ASTP). in HHS.

A review of AI algorithms notes gaps in validation.[6] While in more recent AI models or algorithms, much larger datasets are used now with thousands of dental images, whether the AI models or resulting algorithms can be generalized to all populations is unknown. Therefore, the creation of a dataset that represents all populations, including insured and uninsured, by gender, age, race/ethnicity, etc., should be representative of the population as a whole to be generalizable.

It is possible that there are many companies and entities working alone, and in some cases collaborating with universities, to develop AI algorithms that seemingly work well for their reference dataset. However, the claimed effectiveness falls short when the same software algorithm is applied to another dataset. Without an effective independent assessment, the claims purported cannot be refuted or supported.

This gives rise to the need for an independent database to which all AI software designed for analytics with 2D dental radiographs may be evaluated. There is a need to establish a reference point, or gauge, so as to validate the effectiveness of each software developers' claims. This requires an independent database with known diagnosis from validated sources, perhaps with clinical confirmation, perhaps with pathology and biopsy reports, and confirmed through a panel of experts such as board certified oral maxillofacial radiologists.

A similar process now exists for the evaluation of digital intra-oral radiographic systems as set out in ANSI/ADA Standard No. 1094, Quality Assurance for Digital Intra-oral Radiographic Systems and ANSI/ADA Standard No. 1099, Dentistry — Quality Assurance for Digital Panoramic and Cephalometric Radiographic Systems. [7,8] Using the quality assurance protocol set out in ANSI/ADA Standard Nos. 1094 and 1099 and a standardized phantom, any digital intraoral radiographic system may be evaluated for effectiveness in terms of latitude, contrast perceptibility and spatial resolution while preventing blooming or clipping of data. In the publication, "Evaluation of image quality parameters of representative intraoral digital radiographic systems" by Udupa et. al.,[9] eighteen different intraoral radiographic systems were evaluated using the manufacturer own imaging software, thereby removing any bias.

A similar approach to AI is possible with an independent test dataset for the evaluation of AI programs, which would allow all users, developers and approval agencies to compare each proposed AI algorithm for accuracy and specificity, along with false positives, and false negatives.[9] After the proposed AI algorithm has been compared to the test dataset, it would not be returned to the software developer, preventing any AI learning that may result from exposure to the independent test dataset. This is the only way to preserve the integrity of the reference test dataset.

For AI to develop into a more robust and useful tool for clinicians, the accuracy and applicability must become more robust and stand against effective test protocols such as an independent dataset. Such an independent dataset could be kept, stored and secured by a third party that is not in the business of developing AI algorithms, such as the ADA, the FDA, the American Academy of Oral and Maxillofacial Radiology (AAOMR), or even an international organization.

2..1.3 Federal Agency Guidance for Reducing Bias in AI in Healthcare

Guidance for reducing bias in AI has been issued by another Federal agency, the Agency for Healthcare Research and Quality.[10] The report notes that bias in algorithms has impacted other sectors, including housing, banking and education, noting that healthcare can use key principles to reduce bias and to use AI SaMD to improve patient outcomes and reduce costs. The goal is to ensure that biases in the currently available data, which may only be for less than half the population in any given year, are not inadvertently allowed to be perpetuated without some method to mitigate them within emerging AI validation.

AI vendors should consider key principles and methods to implement these. Some of these are described below:

2.1.4 Principles of AI for Vendors to Consider:

1. Understand basic AI principles and regulatory pathways for product development (see Figure 9 and Appendices);
2. Establish a method for data governance, including privacy and security. This may include anonymization, encryption, or character alterations to assure privacy and de-identification of data at all times;
3. Address transparency and communicate the explainability of AI to the clinician; and
4. Develop a plan for continuous monitoring of privacy, including HIPAA, HiTrust and security.
5. Devise methods for communicating update of products with additions to database or addition of additional populations (such as age, gender, race/ethnicity).

3 Ensuring Validity of AI and a Call for External Validation of AI Algorithms in Dentistry

At this stage of AI development and deployment, validation of safety and of performance for accuracy should be assessed, similar to phase I and phase II clinical drug trials. [11] In fact, human subjects research, including patient informed consent, should be similar for the development of AI for all healthcare, including dentistry. Even small changes in data between the AI product training dataset and clinical evaluation leads to detectable errors in safety and accuracy, possibly leading to harm.[11, 12] Therefore, this TR is an urgent call for external validation of AI algorithms in dentistry. It is proposed that an independent organization house an independent dataset for use in external validation, which is described in Appendix D

3.1 System Validation

3.1.1 Why is System Validation Needed for AI in Healthcare?

One reason for system validation is to ensure that various models are trusted by both clinicians and patients. In late December 2022, a national sample of US adults was conducted in which over 60% of Americans reported that they would be uncomfortable with use of AI for their own health care. However, one benefit that was reported was avoiding bias.[11, 12]

Therefore, avoiding bias is an important principle when evaluating detection support models. A recent assessment of AI among physicians determined that physicians had difficulty in assessing systematically biased clinical decisions supporting AI imaging tools.[12] Data are not available summarizing best practices for avoiding bias by any health provider, including dentists, nor are specific methods to both assess and minimize bias in AI.[13] (Note: The Code of Federal Regulations (21 CFR 820) defines device specification or use case as "…establishing by objective evidence that device specifications conform with user needs and intended use(s))." 21 CFR 820 by final rule will be coordinated with ISO 13485:2016, Medical devices – Quality management systems – Requirements for regulatory purposes.

3.1.2 Rationale for External Validation of AI tools used in Dentistry

When a clinician's diagnostic ability is being evaluated, the evaluator can confirm that the clinician understands the diagnostic task, make sure that the clinician understands the features of the test case and question the clinician about the reasoning being used. In general, none of these are possible in evaluating a neural network. (See Appendix A for definition and description.) Though the network may report a test case as part of a cluster in feature n-space, it has no understanding of the diagnostic meaning of that cluster. Though the network may encode certain features in its hidden nodes, we generally have no way to know what those features are or what they mean to the network's processing. And though a node in the network's final layer may report "yes" for a suggested finding, there is no way to question the network as to how it arrived at its conclusion. This makes objective external validation of a neural network all the more important.

In managing the introduction of self-driving systems in cars, there is much attention given as to whether a human driver must be present, whether the human must have hands on the wheel and whether the human or the system is being relied on for safe driving. These are issues of intended use. Are driving systems like previous technologies that warn of data drift merely an aid to the driver? Or is the system the responsible actor? Similarly, the provider of a machine learning system in dentistry should be clear and specific about the tasks included in the system's intended use and should be clear and specific about where diagnostic responsibility lies. If human drivers are still responsible for driving safely, they must stay awake in the driver's seat with their hands on the wheel. And if human clinicians are still responsible for diagnosis, they must be sure to treat a machine learning system merely as an instrument in their armamentarium.

Of course, a machine learning system is a complex and sophisticated instrument. So how will the clinician have confidence in what the system reports? Against what standard will the system be evaluated?

For a system to be validated, there must be a reference standard to which it's held. For a human clinician, that standard may be the opinion of teachers or of a review board. But for devices, validation is typically achieved though testing against a **Validation Dataset** of test cases. And because the system cannot be interrogated as to its methods, the only way to evaluate the system is by its effectiveness at those test cases. In the future, biologic measures, such as comparisons of radiographs with objective biologic criteria, such as micro-CT scans or periodontal histology, may provide higher reliability and validity of these datasets. This paper acknowledges the subjective nature of expert opinion.

CONCLUSION: Therefore, confidence in a Dental Image AI System should be limited by confidence in its Validation Dataset.

4 Establishment of Ground Truth

Ground Truth is a critical component of external validation dataset. While proprietary AI tools in dentistry use their own training and testing datasets for internal validation, limitations in data or use of an algorithm on a population or group of individuals not included in the original dataset may cause an error in assessment. It is therefore critical, for each validation test case, that the expected findings be correct; that the Ground Truth for each case be well established.

It would be ideal to have, for each test image, conclusive determination of the disease status of the imaged area. But histological or micro CT examination is not practical for cases other than extracted teeth. Post-treatment notes about cases may record actual findings, but they may be difficult to obtain. And not every case requiring treatment may have had treatment and associated notes. Analysis by oral radiologists is generally highly regarded, but their participation in tagging may be difficult to obtain. In practice, most validation image sets are tagged by dental clinicians; tagging that is easier to obtain but can lead to falsely identifying disease more frequently.

For now, we have agreed to use "expert opinion" for external validation, until such time as more sensitive and specific measures are available. An example in the future might be different consensus on "ground truth." For example, this might be an accurate and reliable test for active vs. inactive caries, which would be used with micro CT data from radiographs for a more precise dataset.

If the system is ever to enhance the clinician's ability ground truth must be determined by a method more reliable than clinicians' tagging. If the system's intended use is to augment and not to replace the diagnostician, then a clinician's tagging in a validation image may be acceptable. As long as the system is not expected to detect findings which the clinician cannot, then the tagging accuracy can be no less than the clinician's ability.

The validation dataset must be sequestered from the system's training and testing processes. Such sequestration is the only way to validate that the system's learning has been generalized beyond the specific samples in the training and test datasets.

4.1 Characteristics of Validation Dataset and Scope as Part of Determination of Ground Truth

AI tools should measure what they purport to measure. This is called being *valid*. A further distinction is drawn between internal and external validity in scientific studies. An internally valid study is tightly controlled and can provide confidence in results as far as study participants are concerned. Externally validity results apply not only to subjects in the initial study, but also to others in the broader population. The same is true for evaluation of AI.

Ideally, any AI product should rarely miss a specific condition of its use case. In other words, it should avoid false negative results, called highly *sensitive.* Sensitivity is expressed by a percentage. So, 99.9% sensitivity is what we want. It should make sure that any specific cases it does detect are real cases and should also avoid false positive results. When it rarely determines that a non-case is a case, it's said to be highly *specific*. Specificity is also expressed by a percentage; again, the higher the better. Whether that number is 75% or 90% is not known. Finally, an AI product is said to be *reliable* if there is a high correlation between repeated use of the same image. A correlation is *expressed as a percentage. Accuracy means that a product is free of error or mistakes, or even bias. Validity* is when the product measures what it purports to measure.

External validation datasets are important to ensure that there are objective measures for the broader population. It should be representative of the population as a whole. External validation can best be achieved by randomized prospective clinical trials, prospective longitudinal data sets, and, less reliably, by retrospective analysis of clinical data, such as that captured in an electronic health record. The latter data may contain treatment payment biases based on claims, and the absence of diagnostic codes. In the near future, use of ICD-10 by dental professionals will be advanced so that diagnosis of oral diseases and conditions are matched to appropriate treatment.

Comparisons of AI predictions to expert opinion, including clinician documentation of findings, conditions, diagnosis and treatment captured in patients' electronic record utilizing purposive sampling, can be used for external validation. [14]

The Validation Dataset Scope must be adequate:

- To test for the various findings the system will be expected to detect, including findings of no disease.
- To test the system's ability to analyze images of poor quality, such as those with cone cuts or under-exposure.[7,8]
- To test the system's performance among subpopulations, including patients of varying ethnicity, sex, age, and socioeconomic status.
- To test the system's reaction to novel images, such as images with electronic noise, images flipped from left to right, images of atypical restorations, or images that are not radiographs at all.

Ideally, all machine learning systems would be validated against the same validation dataset, allowing a direct comparison between systems. However, once the dataset became public, designers could train their networks to provide correct findings for just that data, ignoring the very large variety of real-world cases. It would be like allowing students to study from the test's answer book. Optimizing or "tuning "networks to a specific dataset leads the AI/ML not to learn the disease, but rather only that data.

This drawback could be overcome by a widely trusted and well-funded source for images and tagging. If the source could periodically publish a fresh validation dataset of sufficient quality and scope, then designers could periodically be required to validate their systems against test cases never before encountered.

Ideally, the dataset would be housed in an independent organization, without ties to commercial entities. The independent organization would house a collection of raw images, which are collected, collated, and annotated. This would also be iterative. These images in a data warehouse would be secure and protected. The independent organization would assess the demographics representation for relevant populations, where possible, and add to its collection to avoid bias. Sustainability of this trusted source of external validation would be possible by assessing a reasonable fee from AI vendors for certification with external validation. In addition, some methods for local validation from similar sources should be sought.[15]

4.2.1 External Validation Dataset Acquisition for Determining Ground Truth

The methods for a valid external validation dataset have not been described previously. Appendix D, Figure 10 proposes a consortium that would collect and continuously update the external validation data for various populations segmented by age, gender, race/ethnicity, etc. An independent body would be designated to keep this data secure and confidential, and accessible for external validation of vendor AI products with 2 dimensional images. A consortium of images could be collected from private practitioners, dental education, and from patients who do not access dental services, to comprise the external validation data set.

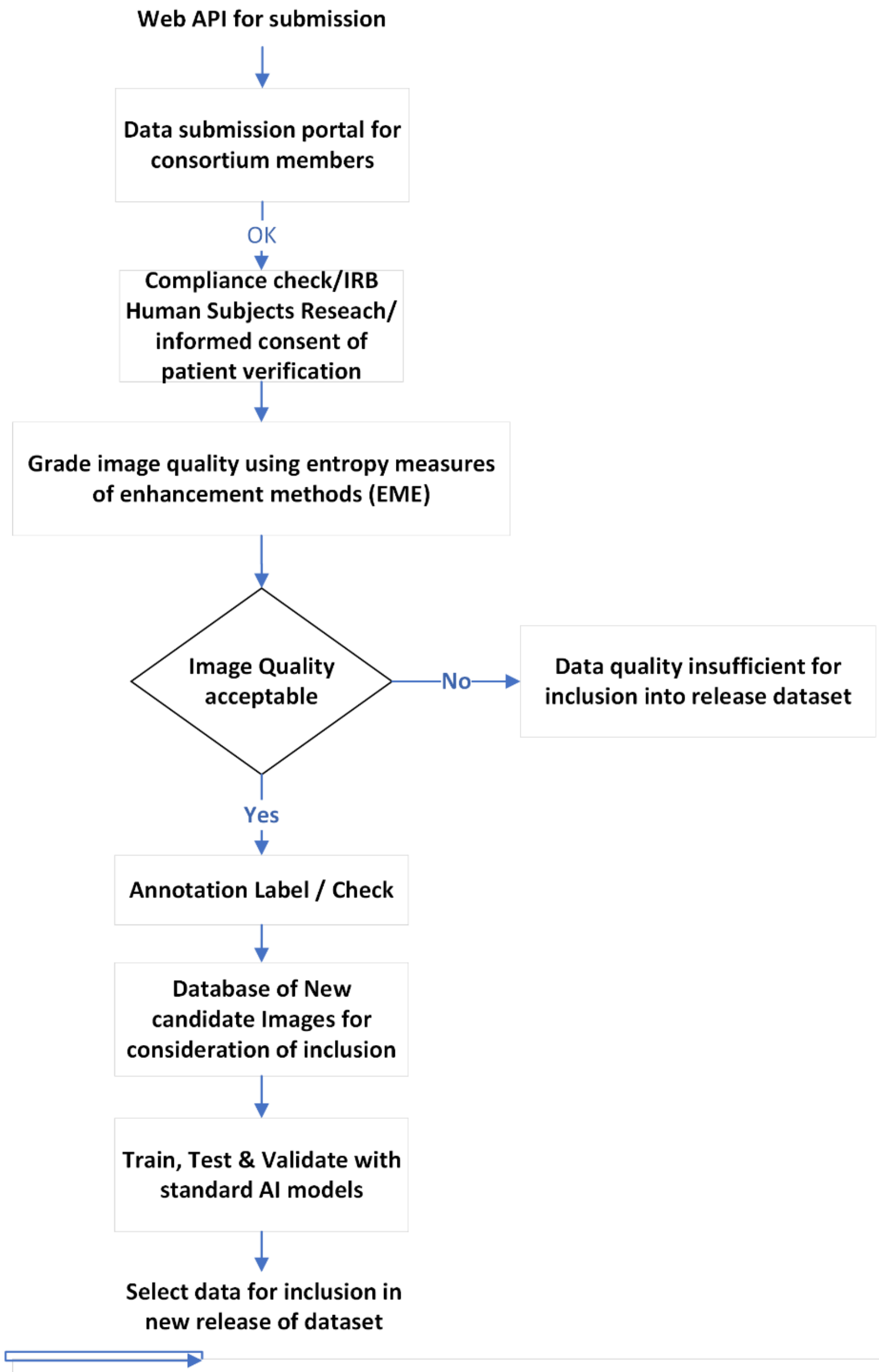
Appendix D, Table 1 describes verification of image metadata accuracy and/or the minimum image metadata required to be included in the external validation database. Missing data elements such as age/race/gender would be annotated on images. The ideal dataset would ensure that vendors would be validated on a subset of the database to meet the design intent or use case for their product or products. For example, if the use case is only for children, then external validation would include only children's image data to ensure that verification of the image metadata is correct. Ideally, annotation of image metadata would include characteristics of ground truth and the characteristics beyond binary presence or absence of findings.

Annotation of additional descriptive factors could also be considered as part of the description of ground truth on a voluntary basis, especially for Class I AI devices which do not require FDA labeling; or the voluntary simple labeling described above could be used.

5 Data Architecture for External Validation Dataset

5.1 Proposed Logic Framework for Data Architecture:

Data architecture for the creation of an external validation dataset is shown in Figure 2.
A web application programming interface (API) would ensure proper authentication and authorization and other components of human subjects' research, including institutional review boards (IRBs), and informed consent. Since most academic centers have IRBs, the requirement would be fairly easy for academic dentistry but more of a challenge for private practitioners. However, the emergence of community IRBs may be a more economical and representative group to review any protocols in a local community for private practicing dentists.


Web API for submission
Data submission portal for consortium members
OK
Compliance check/IRB Human Subjects Reseach/ informed consent of patient verification
Grade image quality using entropy measures of enhancement methods (EME)
Image Quality acceptable
No
Data quality insufficient for inclusion into release dataset
Yes
Annotation Label / Check
Database of New candidate Images for consideration of inclusion
Train, Test & Validate with standard AI models
Select data for inclusion in new release of dataset

This diagram describes the process architecture to collect new data and images from different sources to create a standard dataset to train and benchmark assistive artificial intelligence (AI) algorithms for dental medicine. Data will be submitted via a secure web portal by authorized collaborators/contributers. The output of this process produces a robust dataset that is frequently updated with new test cases to increase the knowledge base and ensure that the AI is learning the diseases. The resulting dataset establishes a common baseline to train and evaluate the accuracy of any new assistive AI system for dental medicine.

Figure 2. Overview of data and image collection

Source: Dr. Karen Panetta, Tufts

5.2 Overcoming Current Challenges for Construction of External Validation Datasets with 2- Dimensional Images

Several questions remain on the construction of a validation dataset for two-dimensional radiographs.

Validation dataset scope and quality: Vendors are required to validate test accuracy, but currently there is no standard for validation datasets.

- There is no consensus on the acceptability of various methods of determining Ground Truth. Most validation datasets rely on general clinicians to tag images, yet there is no consensus for credentials expected of taggers, for tagging procedures, or for required demographic information.
- There is no consensus on how many samples should be used in validating dental image AI systems.
- There is no consensus on the proper way to construct sets of test cases based on image quality, intended uses, and patient subpopulations.
- Deep fakes, those generated by AI, could either enhance imaging or ensure that algorithms fail to generate accuracy similar to submission for training datasets used to obtain regulatory clearance in the US.[16] Therefore, machine learning in dentistry would benefit from having a standard for the content of validation datasets, the processes by which samples are collected and the mechanisms for establishing ground truth. System designers, regulators and users could then have appropriate confidence in the validated systems. This is why the ADA standards program has approved a work item for the development of such a standard.

As stated above, machine learning in dentistry would benefit from a widely trusted and well-funded source for images and tagging. With periodically published trusted validation datasets of sufficient quality and scope, the stage would be set for open and fair competition; and system designers would have a verifiable basis for claiming continual improvement. Perhaps a consortium of dental schools, augmented with data from general dentists, will take up the challenge of providing such a service. Data collected solely from dental schools might have a drawback of data collection from a non-representative population that would be expected to be more urban and have specialized populations with both accessibility and affordability concerns.

In addition, human subjects research protocols will be operational, including institutional review boards (IRBs). This may be supplemented in various locations with community-based IRBs. How data from these sources will avoid bias has yet to be determined, with updating to ensure representative sub-populations needed.

The proposed data architecture in Figure 2 demonstrates a pathway by which data collection could occur. However, ensuring that collated images are collected from representative sources will be an ongoing challenge in the process. Beyond claims-based dental school image collection, methods for a standard for creating the most representative external validation database will remain a challenge. Ensuring representative population based data collection in an external validation dataset, and independence of the organization that houses this data, will be an especially challenging task, parameters for which should be established by the standard. The parameters for the organization should be independent of corporate affiliations or influences. Maintenance of the database would be funded by reasonable fees charged to AI developers and could be utilized in any simple labeling of AI systems that analyze two dimensional images.

5.3 Ethical Use of AI with Regulatory Analogy to Human Subjects Research for Clinical Trials with Protection of Data, Privacy and Security for Any External Validation Dataset Warehouse

As mentioned, AI development tools are similar to pharmaceutical clinical trials in evaluating both safety and performance, if not efficacy. Both professional and governmental rulemaking, in coordination with each other, should ensure that AI is ethical, safe, and effective in its use case. In 1991, the Common Rule, or the Federal Policy for the Protection of Human Subjects, was updated to include the protection of human subjects in research. HHS adopted this as a regulation codified in 45 CFR part 46.[17] This includes basic provisions for IRBs, as well as truly informed consent from patients, and institutional Assurances of Compliance.

6 Other Methods for Explainability of AI Algorithms and Product Labeling

6.1 Product Labeling

Medical devices, including SaMDs, are evaluated for risk and assigned by the FDA to one of three classes. Class I devices generally pose the lowest risk to the patient and/or user, while Class III devices pose the highest risk. Class II and Class III systems require labeling. However, labeling requirements for dental SaMDs have not been established. One challenge is that, unlike many medical tests and systems, dental image AI systems purport to be suitable for a variety of intended uses, making more than one claim and reporting on a variety of conditions. Therefore, it may be important for the label to report accuracy for each separate claim for each separate condition or use case.

In late 2024, the FDA put out a request for comment on whether or not generative AI should be regulated by FDA. Anticipatory guidance is expected on this issue in 2025., with meetings on whether or not manufacturers should be required to go back to the FDA following enhancements of this rapidly progressing technology (see also: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024).

Any labeling should include the version of the SaMD utilized so that the provider knows the version and date of last update, like any other software. Regardless of the FDA's deliberations on use of regulatory oversight of generative AI, which is outside the scope of this TR, consensus on the need for continued post-marketing surveillance of AI healthcare products is emerging to ensure patient safety and highest quality care.[18]

Clear labeling is a benefit to patient outcomes and for clinician decision support. (See Appendix B on FDA's clinical decision support scheme). Whether labeling is available prior to purchase on a non-paywall website or is provided as a standalone document can be determined by manufacturers. Labeling for comparison purposes ensures that the generalizability of the data upon which the algorithm is based is, in fact, true. It may also be important for a label to report facts about the dataset used to validate the system: Validation Dataset subpopulations, number of samples from each subpopulation and accuracy within each subpopulation, ,date of collection of data and date of last update.

While we might all agree that a dataset with all women might be biased, no standards exist for the "ideal" representative data samples. In national surveys, minority populations are oversampled to ensure that there is accuracy, and this may be considered for adding to any test datasets or to dynamic algorithms in the future. For now, such information would empower the clinician to provide optimal patient care for the best patient outcomes, including reduction of prevalence and incidence of common oral diseases, such as caries, periodontal disease, oral cancers, etc.

The clinician could use this information to:

- Compare AI systems with clinicians;
- Compare AI systems with each other;
- Assess an AI system's appropriateness for the clinician's patient population; and
- Assess an AI system's clinical applicability to a particular patient.

6.2 Clinician Responsibility

As clinicians are responsible for diagnosis and treatment planning, they must guard against becoming over-reliant on machine learning systems. For example:

- For a system intended to identify tooth numbers and automatically mount radiographs, false findings may not be significant.
- For a system intended to discover lesions, and having low false positives but high false negatives, the clinician might have high confidence in the identified lesions but must still scan the entirety of each radiograph lest a finding be missed.
- For a system intended to discover lesions, and having high false positives but low false negatives, the clinician might have high confidence that all lesions have been identified but must guard against the system over-diagnosing lesions.
- Initial lesions confined to enamel might be identified but might not be targeted for restoration; rather, remineralization may be considered along with other preventive measures, such as fluorides.

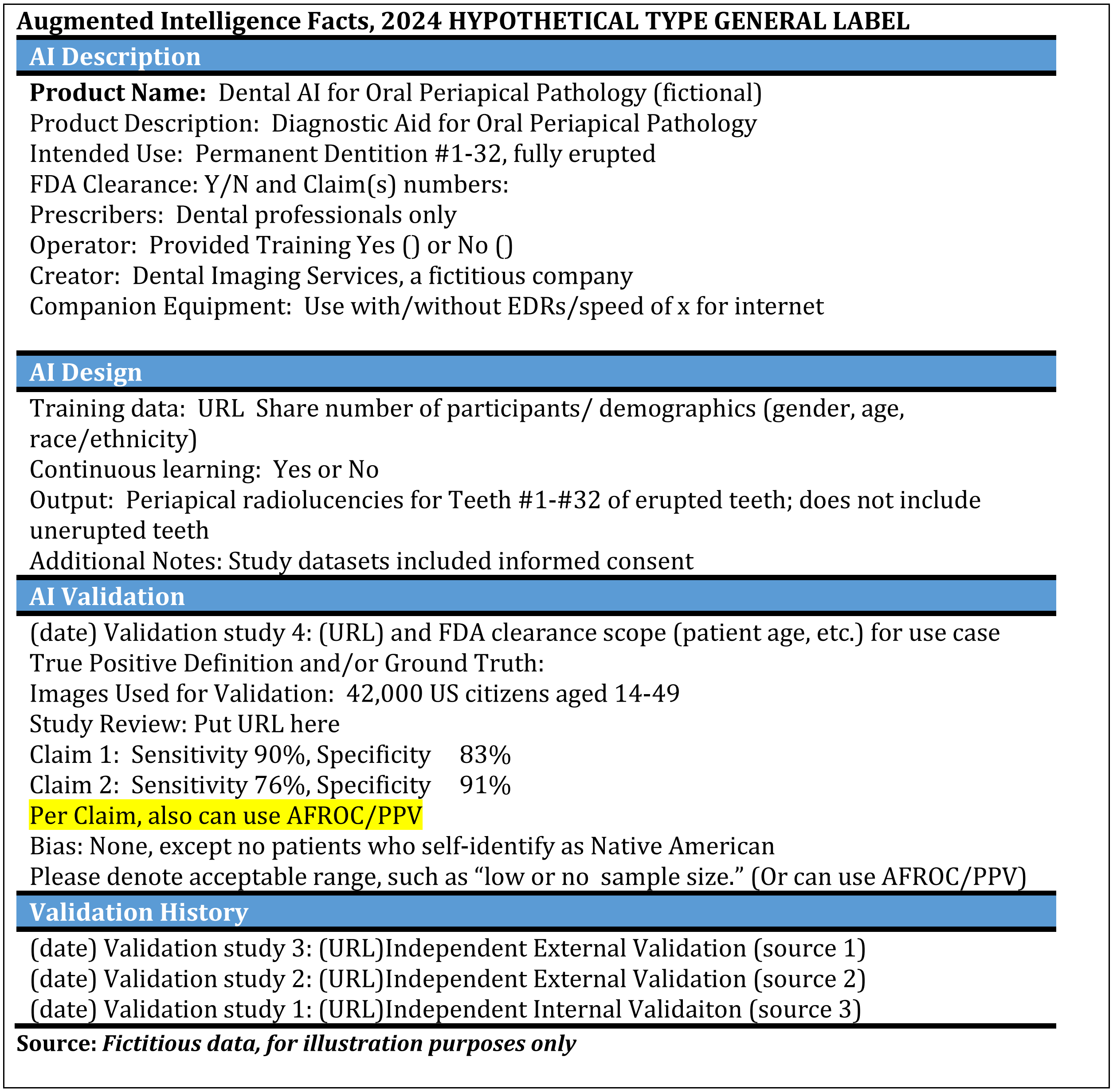
6.3 Use Case Labeling

Specific products have specific claims, and it is important for clinicians to understand specific use cases. Vendors should know that clinicians want to examine:

- Use case;
- Training data;
- How the product was developed/type AI;
- Accuracy of model;
- Validation Data safety/efficacy;
- Assessments for fairness and bias;
- Any variability race/ethnicity/gender.

6.3.1 Example of Use Case Labeling by Claim for Two Dimensional Image

Ensuring that a clinician can compare and contrast various use claims by a vendor of AI tools for clinical decision support is important (Appendix C). Since FDA characterizes SaMD as a Class I, II and III, only Class II and III, which have some degree of risk, require labeling. AI tools that are Class I are exempt from labeling. Ideally, an interactive web page that would allow a clinician to search for training data, and the population(s) that they serve, could be searched without a paywall. One of the problems is the specifications for labeling by FDA are detailed and may not be read by clinicians. Figure 3 provides a sample label that could be considered a simple "Nutrition Fact" type food label that could be used for each specific claim.



Augmented Intelligence Facts, 2024 HYPOTHETICAL TYPE GENERAL LABEL

AI Description

Product Name: Dental AI for Oral Periapical Pathology (fictional)
Product Description: Diagnostic Aid for Oral Periapical Pathology
Intended Use: Permanent Dentition #1-32, fully erupted
FDA Clearance: Y/N and Claim(s) numbers:
Prescribers: Dental professionals only
Operator: Provided Training Yes () or No ()
Creator: Dental Imaging Services, a fictitious company
Companion Equipment: Use with/without EDRs/speed of x for internet

AI Design

Training data: URL Share number of participants/ demographics (gender, age, race/ethnicity)
Continuous learning: Yes or No
Output: Periapical radiolucencies for Teeth #1-#32 of erupted teeth; does not include unerupted teeth
Additional Notes: Study datasets included informed consent

AI Validation

(date) Validation study 4: (URL) and FDA clearance scope (patient age, etc.) for use case
True Positive Definition and/or Ground Truth:
Images Used for Validation: 42,000 US citizens aged 14-49
Study Review: Put URL here
Claim 1: Sensitivity 90%, Specificity 83%
Claim 2: Sensitivity 76%, Specificity 91%
Per Claim, also can use AFROC/PPV
Bias: None, except no patients who self-identify as Native American
Please denote acceptable range, such as "low or no sample size." (Or can use AFROC/PPV)

Validation History

(date) Validation study 3: (URL)Independent External Validation (source 1)
(date) Validation study 2: (URL)Independent External Validation (source 2)
(date) Validation study 1: (URL)Independent Internal Validaiton (source 3)

Source: ***Fictitious data, for illustration purposes only***

Figure 3: Fictitious label for Use Case claims explainability

The static hypothetical label would ideally be available without purchase, so that clinicians could review basic data used to construct AI models, with training data demographics with specific populations noted.

6.3.2 Other Considerations for Explainability, including Training and Updating

Vendors and prospective vendors may benefit from ensuring explainability for their products at all times. While proprietary systems are maintained by vendors, it is in the best interest of clinicians and patients to provide decision makers with as much information about explainability of product use as possible. While regulatory authorities, including FDA, have oversight of vendors based on risk, improvement of patient outcomes for common oral diseases is proposed as a voluntary strategy for producers who provide SaMD to providers. If possible, comparisons to current patient outcomes can be emphasized as a part of training for use of various products. Informing clinicians about updates on static models and how that impacts any new claims is important part of training. In addition, ensuring informed consent and privacy and security of data inputs at all times requires constant vigilance.

7 Summary

This TR outlines the parameters for development of an external validation database by an independent organization for use cases with two dimensional radiographs. The paper also presents a proposed data architecture structure for collection, annotation, storage, and security. Ensuring that use cases for AI are clearly defined, as well as specifying the human oversight, holds promise for greater accuracy in precision diagnostics and treatments. Simple labeling of all AI products will assist clinicians to compare and contrast different products and product use claims for each claim. Labeling should include all Class I, II and III SaMD, noting that FDA does not currently require AI labeling for Class I or for clinician beta testing. Future standards development will focus on these areas in greater detail.

Appendix A

Development of Algorithms for AI and Testing and Training Data by AI Vendors

Systems with Human Encoded Algorithms

One typically thinks of a computerized system as following the instructions of human programmers. So, of course, one would expect that programmers could describe how the system arrives at a conclusion. For a system doing classification, one would expect programmers could describe how the system arrives at its classification findings. The term augmented intelligence/artificial intelligence has been applied to a variety of classification systems for which this is true.

One class of such systems encodes human expert knowledge into explicit rules. Called ***Expert Systems***, they define their feature spaces and cluster regions by relying on human expert knowledge. Such systems can stand alone within computers or can be embedded into hardware systems that conduct measurements, e.g. heart monitors.

Systems with Machine Derived Algorithms

Humans are extremely good at recognizing human faces. We take in very large amounts of detailed visual input, extract a large number of facial features, and recognize patterns to draw a conclusion – that is, to classify a face. But which of us can explain how we do it? And if we cannot explain our own process, a system designer must find a way to build a system without encoding any rules.

A revolution in AI began when developers began creating successful pattern recognition systems for which they could not provide a description of how the system did its classification. Most such systems are based on ***Artificial Neural Networks*** (ANN). And although mathematics can provide an explanation of how these networks learn, and results can be measured for success, it is typically not possible to determine what patterns the network might be recognizing, or how the network uses those patterns to provide a finding.

Artificial Neural Networks

The developers of ANN systems were Inspired by biological systems. After all, if human experts use biological neural networks to recognize patterns and classify results, perhaps computer systems could replicate network abilities by simulating their processes.

So designers mimicked the behavior of neurons in animal brains. In these computer programs, there is a node playing the part of a metaphorical neuron, individually modeled to have inputs, an activation process, and (to the extent activated) outputs to other nodes (neurons). Each node belongs to a layer, with the network having three or more layers of varying size. The nodes of each layer provide input for nodes in the next layer. The first layer receives the individual inputs from a sample, and the last layer reports results (Figure 4).

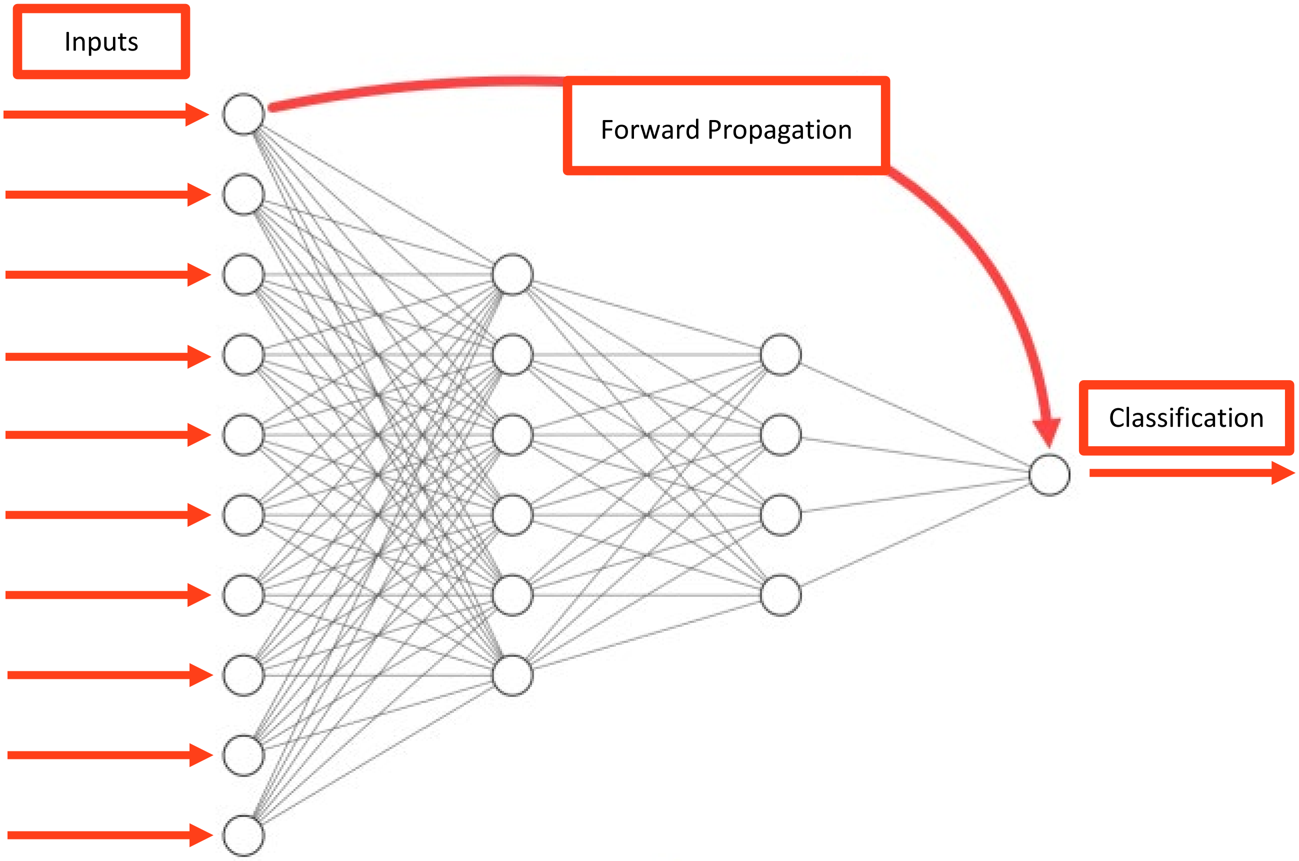


Figure 4. Artificial neural network

It is common that, as the network propagates information from the inputs toward a result, the number of nodes per layer decreases. This hierarchical structure often represents recognition of patterns in the input, which maps to clusters in feature space, which in turn supports classification.

Classification is usually accomplished by considering various features of an object, situation, or case. These features might be directly measured (tooth shape, position, brightness) or otherwise sourced (age, hygiene practices, general health). Combinations of such features are often determinative of a finding. So, just as a clinician might, a computer program can associate a certain combination of features with a certain finding; then it can suggest that finding when it subsequently encounters that feature combination.

Figure 5 provides a simplified example. Consider a volume dotted with examples of interproximal radiolucencies. Each radiolucency is represented by a point in the figure and has been tagged as being either an interproximal lesion or an artifact. A point's position along the x-axis denotes location on the tooth. The point's position along the y-axis denotes how triangular the morphology is. And the point's position along the z-axis denotes the radiolucency's intensity. This is a feature space of three features, a 3-space. The result might be a cluster of points in one part of the 3-space, with almost all of the points representing lesions, and a different cluster elsewhere with almost all of the points representing artifact. Other points might fall outside the two clusters, representing unusual examples of these conditions, representing other conditions, or representing no recognizable condition at all. Now consider a new case not yet tagged with any finding. By determining the new case's position in this feature 3-space, a system can suggest a possible finding.

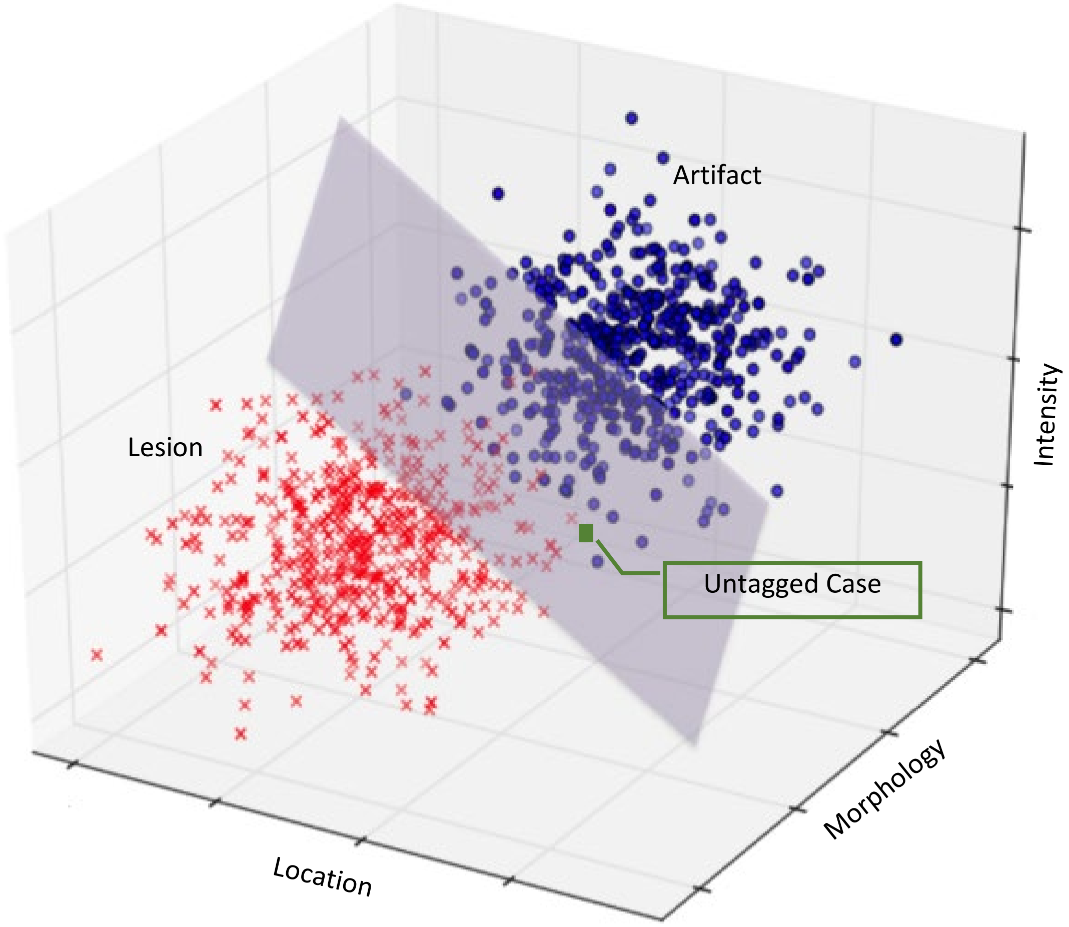


Figure 5. Feature 3-space

Of course, more than three features are often required for a particular diagnostic or other task. In these cases, feature spaces must have more than three dimensions. Mathematicians call such constructs n-spaces, but the idea is the same: In a feature n-space, identify clusters of tagged cases associated with particular findings; then, for each newly encountered un-tagged case, if its point is in a cluster, report the cluster's associated finding as the system's suggestion.

Inputs to each node has weights, and an activation function establishes a threshold that must be met to produce a signal "yes" to the next layer of nodes (Figure 6).

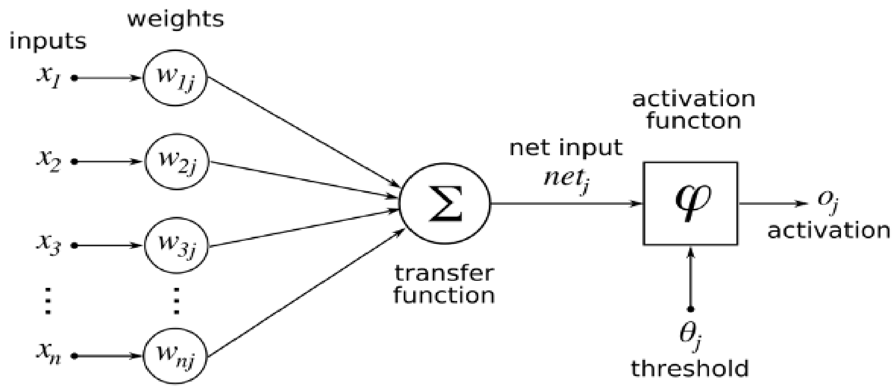


Figure 6. Schematic of multiple processing layers of neural networks.

As an example, consider a neural network for processing an intraoral radiograph. The first layer is a metaphorical retina; each retinal node takes input from its associated pixel and based on the node's activation process, determines what it will send to nodes in the next layer. The late layer may contain a node for a diagnostic cluster; when such a late-layer node's inputs and activation process result in activation, it signals a particular diagnostic finding. Thus, one late-layer node may signal whether a lesion is present, another whether recurrent decay is present and a third, whether over-sharpening is present. For any given region of the radiograph, one would expect either none or one of these late-layer nodes to signal "yes." With this late-layer output, the system can suggest a finding for the given region. Introduced in 1995, an ANN was for the first time used as the basis for a caries detector.

The first network layer receives input, and the last network layer provides results. Layers in between are called *Hidden Layers*. The network in Figure 6 has two hidden layers.

The first artificial neural networks had three layers (with a single hidden layer), and this was satisfactory for certain tasks. But more subtle or complex tasks require more layers; and training these structures can require much more data and computing power. Fortunately, computing power increased to accommodate these advances. The resulting *Deep Learning* systems are simply neural networks which contain at least two hidden layers.

One benefit to many-layered networks was the ability to use certain layers for certain purposes. For instance, in a *Convolution Neural Network* (CNN), an early layer is dedicated to recognizing patterns in the incoming data. Certain structures in a frog's retina may be a good metaphor. By being sensitive to motion over a range of retinal cells, these retinal structures pre-process large amounts of input into a feature, i.e., "it moves." Then a simple signal can be passed from the frog's retina to its brain indicating movement.

In a CNN, an early-layer node (or set of nodes) mathematically models a possibly relevant feature, such as "triangle" or "less-bright-region." So, rather than thousands of original inputs such as "pixel 1 has intensity 255" and "pixel 999 has intensity 37," the next layer may receive merely dozens of signals such as "triangle = yes" and "less-bright-region = yes." This greatly reduces the information which must be passed to and processed by the next node layer; and fewer nodes and connections means faster, more reliable training.

In some situations, such as analyzing for bone loss or for progression of disease, a sequence of images must be considered. A *Recurrent Neural Network* (RNN) is designed for this purpose. In such a network, outputs from a later layer become inputs for an earlier layer. In this way, features from a previous sample become inputs for a subsequent sample. So, the network has, in a way, a memory from previous images. The network might therefore train on sequences of tagged images. Then, once trained, the network might accept sequences of new, untagged images as input, recognizing patterns of change to suggest a finding.

Neural Network Learning Methods

To the extent neural networks express intelligence, the intelligence is encoded in the details of each node's activation process and in the weight given to each node's activated output to later-layer nodes. *Neural Network Training* is a process of adjusting connection weights. Each input sample is tagged with a desired result and contributes to this training. For each sample, the network applies the weights in a forward propagation to create a result. This result is contrasted with the sample's tagged result, and the difference is mathematically sent backward through the network to adjust the inter-neural weights. This is called *Back-Propagation* (Figure 7). Theoretically, this minimization of differences yields a network that generates results close to the set of tagged training samples.

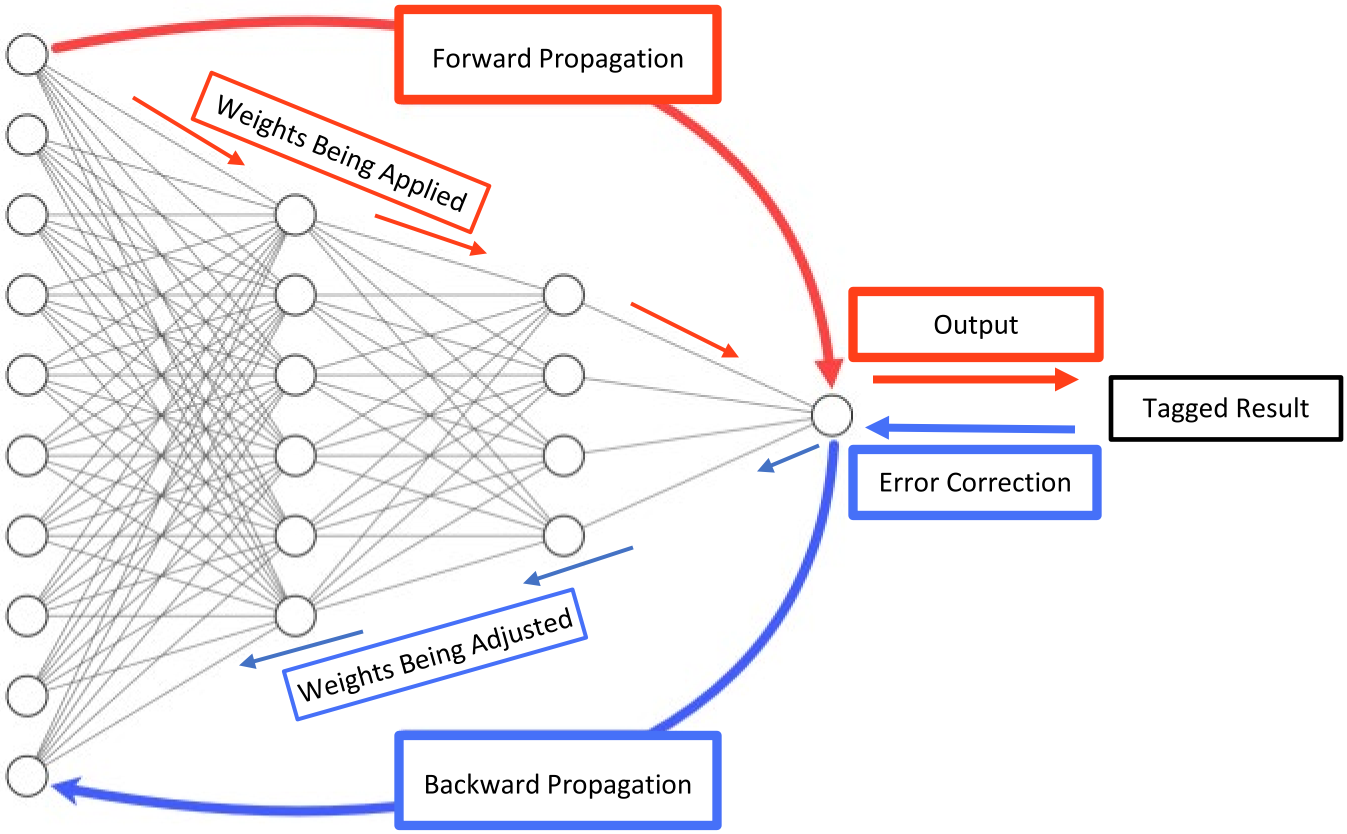


Figure 7. Training using back propagation

An intriguing aspect of machine learning is its ability to discover features in sample input. Designers have found that networks with more than one hidden layer can discover more complex features. However, it is usually immensely difficult to determine what those features are. For example, one would not expect to be able to say, "The network discovered triangles in the input data, and those triangles are an important feature." This is part of a key issue with neural networks: It was the network's training process, rather than a designer, which set the network weights; and it is generally not possible to know why network weights are as they are.

If training is fundamental to the performance of the network, then the *Training Dataset* of samples is also fundamental. In dental imaging, the training set is typically a collection of dental images, such as intraoral radiographs that have been annotated/labeled/tagged into correct diagnosis categories by experts. This establishes a ground truth for training and validation of the models. The samples in the dataset will provide examples of the kinds of findings the network is to detect. For instance, the sample radiographs might have a variety of already-labeled class II lesions. It is then hoped the resulting network detects those lesions as effectively as the humans who originally identified them. For every kind of finding the network is to detect, the training dataset must have a sufficient number and variety of samples of that finding. It is also important for the network to properly detect a finding of absence of disease; the training dataset therefore may need to contain images without any disease. It is essential that training images include both diseased and disease-free teeth.

Human identification of findings in a dataset is called *Data Labeling or Annotating.* . Training to match annotating is called *Supervised Learning*, and this is the most common form of training for neural networks in dentistry. Since the point of such training is to mimic the behavior of the annotators, the network generally cannot improve on the annotator's expertise. Therefore, with supervised learning, one should not expect the network to detect findings the taggers could not. For example, if a lesion is undetectable by any of the annotators, it may be undetectable by the resulting neural network, because the ground truth dataset does not contain a sample of these unseen conditions.

However, images are obtained from sensors that contain information that is not visible to the human eye. The AI may very well learn from its training and "see" features from the ground truth dataset that the human may not see. These cases may classify an image as one with a lesion, but without subsequent analysis and inspection by the human annotators, and subsequent updating of the ground truth to reflect these scenarios, most outputs will be considered erroneous by an end-user who is not reliant on the updated ground truth dataset. This is also due to the fact that the AI can't explain what features it saw that met the criteria, i.e., explainability. In general, if an artifact deceives most annotators into believing there is a lesion, then the neural network may be similarly deceived because the data set doesn't contain examples of these scenarios. As the capability becomes more refined, the "discovery" of new lesions may be beyond what the human eye will detect; therefore, we may need to explore other methods with the agreement of experts to establish ground truth.

Certain types of machine learning are possible without human tagging of sample data. Using *Unsupervised Learning*, a neural network can independently discover patterns in a set of training images. These patterns may or may not be associated with relevant findings. However, they may still be useful. When shown the set of images that formed the pattern, a human expert might notice a feature, or even a finding, that can be part of the next network design. So, unsupervised learning can be an aid to network development, but it cannot alone recognize patterns and effectively suggest findings. Some external information is needed to define success and to provide the necessary feedback to train the network.

For some tasks, it is best to take a step by step (or “Markovian”) approach to reaching a finding. For example, in professional landmarking of 3D cephalometric studies, professional operators’ attention first focuses on the global features of an image, then moves to the local regions of interest to catch the local features for final annotation. Neural networks can learn such behaviors with *Reinforcement Learning*. In this example, for each 3D image, the network makes guesses as to how to traverse the image. Landmarks tagged by professionals provide feedback on which the network can train.

Another example of reinforcement learning is a *Generative Adversarial Network* (GAN). Here one network is pitted against another. An example in imaging might be to generate realistic-looking images that appear to suggest certain findings. The first network takes a real sample image and generates a simulated image meant to display the same finding. The second network is a standard classifier network that determines whether the simulated image displays the finding, discriminating between good and bad simulations. This feedback is used by the generative network to improve its simulations – to train. As the generating network learns, it generates better and better simulations. Finally, if the discriminating network can no longer tell the difference between real and simulated images, the generative network can be used to create effective simulations.

As important as the training dataset is the *Test Dataset*. During the design and training process, it is important for developers to periodically test the neural network. Such testing can lead the developers to alter the network design, training regimen, or data inputs. To maximize the value of such testing, it is important that the developers sequester the test dataset from the training process. Presenting the network with distinct input is generally required to see if the network has generalized its training.

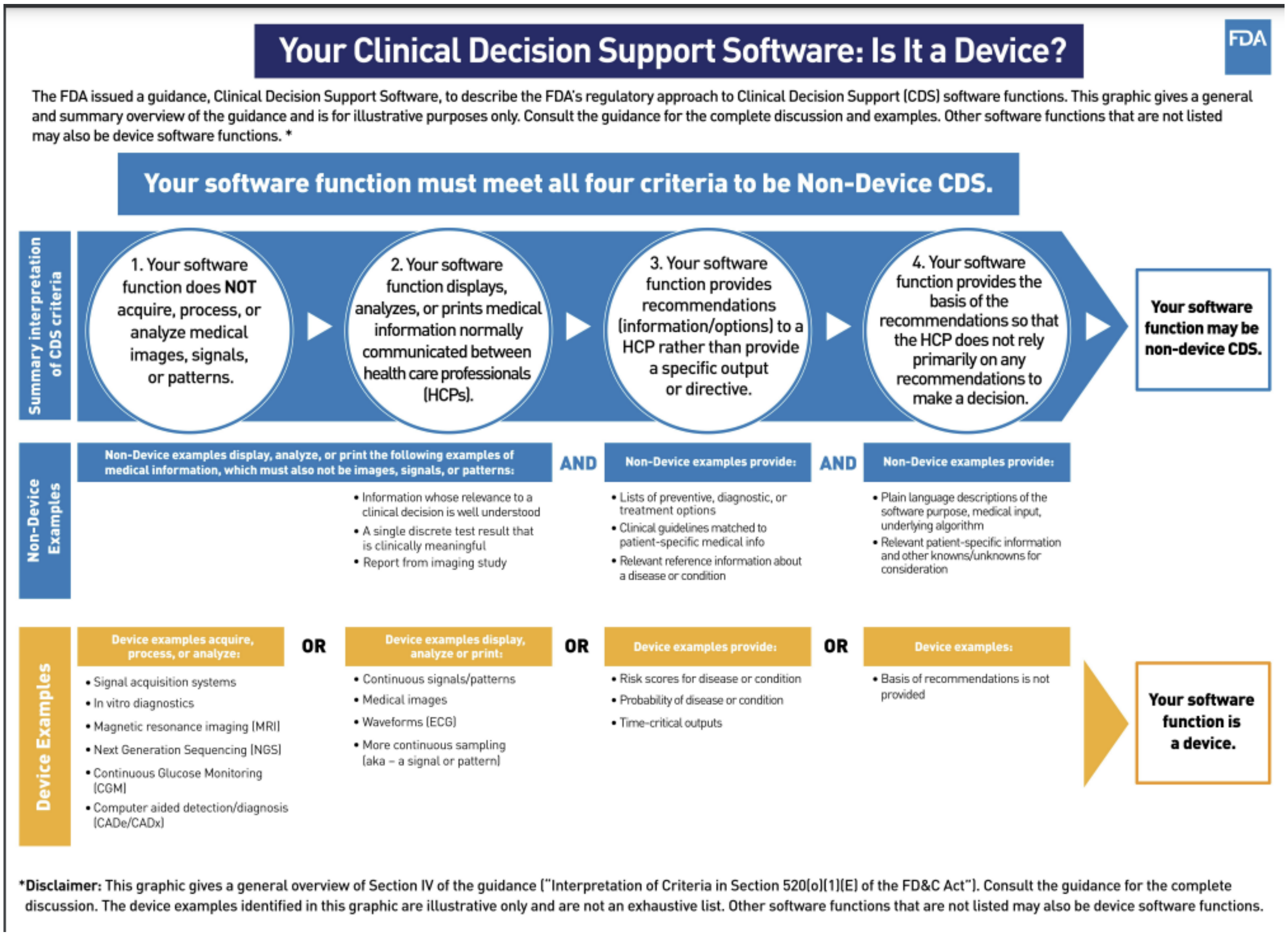
One consideration in testing is *Novelty Detection*. If both the training and test data have only typical inputs, then the resulting network may be unable to deal with highly atypical input. In dentistry, these might be images with images flipped left to right, images with electronic noise, or images which are not radiographs at all. Without proper testing, a network might catastrophically fail to properly analyze such real-world input, suggesting highly inappropriate findings.

Whenever reasonable, clinicians use multiple kinds of information to inform their findings. This might include manual probing, lab results, multiple modalities of imaging, or information from a patient’s chart. Similarly, a network might improve its performance by accessing some of the same kinds of information. Of course, gathering such input at image analysis time might be challenging, requiring access into databases and natural language processing.

Appendix B

FDA model for Clinical Decision Support

This model is provided by the FDA for prospective vendors and clinicians.[19]



FDA

Your Clinical Decision Support Software: Is It a Device?

The FDA issued a guidance, Clinical Decision Support Software, to describe the FDA's regulatory approach to Clinical Decision Support (CDS) software functions. This graphic gives a general and summary overview of the guidance and is for illustrative purposes only. Consult the guidance for the complete discussion and examples. Other software functions that are not listed may also be device software functions. *

Your software function must meet all four criteria to be Non-Device CDS.

Summary interpretation of CDS criteria

1. Your software function does **NOT** acquire, process, or analyze medical images, signals, or patterns.

2. Your software function displays, analyzes, or prints medical information normally communicated between health care professionals (HCPs).

3. Your software function provides recommendations (information/options) to a HCP rather than provide a specific output or directive.

4. Your software function provides the basis of the recommendations so that the HCP does not rely primarily on any recommendations to make a decision.

Your software function may be non-device CDS.

Non-Device Examples

Non-Device examples display, analyze, or print the following examples of medical information, which must also not be images, signals, or patterns:

- Information whose relevance to a clinical decision is well understood
- A single discrete test result that is clinically meaningful
- Report from imaging study

AND

Non-Device examples provide:

- Lists of preventive, diagnostic, or treatment options
- Clinical guidelines matched to patient-specific medical info
- Relevant reference information about a disease or condition

AND

Non-Device examples provide:

- Plain language descriptions of the software purpose, medical input, underlying algorithm
- Relevant patient-specific information and other knowns/unknowns for consideration

Device Examples

Device examples acquire, process, or analyze:

- Signal acquisition systems
- In vitro diagnostics
- Magnetic resonance imaging (MRI)
- Next Generation Sequencing (NGS)
- Continuous Glucose Monitoring (CGM)
- Computer aided detection/diagnosis (CADe/CADx)

OR

Device examples display, analyze or print:

- Continuous signals/patterns
- Medical images
- Waveforms (ECG)
- More continuous sampling (aka – a signal or pattern)

OR

Device examples provide:

- Risk scores for disease or condition
- Probability of disease or condition
- Time-critical outputs

OR

Device examples:

- Basis of recommendations is not provided

Your software function is a device.

***Disclaimer:** This graphic gives a general overview of Section IV of the guidance ("Interpretation of Criteria in Section 520(o)(1)(E) of the FD&C Act"). Consult the guidance for the complete discussion. The device examples identified in this graphic are illustrative only and are not an exhaustive list. Other software functions that are not listed may also be device software functions.

Figure 8. Your Clinical Decision Support Device: Is it a Device?

Appendix C

Considerations for Evaluation of AI Products by Users for Safety and Performance

Stages of Evaluation and Questions for Consideration in Development of An External Validation Dataset

When evaluating a diagnostic test or system, it's important to consider a variety of points of view, and the goals associated with those points of view. Lijmer et.al.[20] suggested a phased approach for system evaluation (Figure 9).

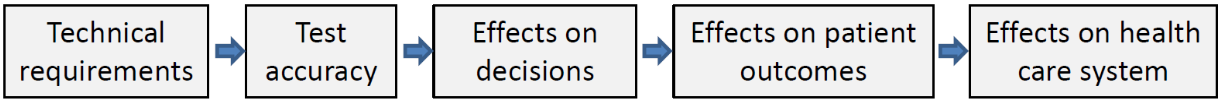


Figure 9. Phased stages of evaluation

Note that Dental Imaging AI, as a young field, needs evaluation in the early phases. Yet, given the speed with which the systems are likely to become ubiquitous, it is not too early to consider questions in all the phases. In the Technical Requirements phase, system specifiers and engineers are concerned with proper construction and implementation of the system. For dental imaging AI, great progress has been made in this phase, but there is not yet consensus on certain questions:

- What quality of images can be used for training? Or can it be used for system evaluation?
- Should only the least processed ("raw") images be utilized, or should adjusted (filtered) images be allowed, even AI-adjusted images? Evaluation of least processed images versus adjusted (filtered) images is needed to answer this question.
- What metrics should be used to evaluate the quality of an image?
- What are reasonable and achievable intended uses?
- What information about the images should be used for training and/or evaluation?
- What patient data can be used for training? Or can be used for evaluation?

In the Test Accuracy phase, questions of clinical evaluation arise. How often does the system falsely lead to treatment? Or falsely fail to lead to treatment? More generally:

- How should evaluations be conducted?
- What level of performance is acceptable?
- How large and how varied must sample sizes be?
- How can system evaluation be conducted in a way to promote fairness and equity across populations?

In the Effects on Decisions phase, clinicians' responses to the new technology are considered.

- How will AI systems affect decision making by clinicians who implement the systems, but who poorly understand the systems? The trust that clinicians place on these systems needs to be measurable and explainable.
- How will AI systems affect decision making by clinicians who are employed by organizations which encourage the use of those AI systems?
- Will clinical decision makers become dependent on the systems?
- Will implementation of AI systems cause a shift in responsibility and liability for clinical decisions?

In the Effects on Patient Outcomes phase, questions arise about the impact on patient care.

- Will AI systems promote patient acceptance of appropriate treatment? Of inappropriate treatment? How do we measure patient trust in AI, which may be very different than the trust in AI from the Clinician's perspective?
- Will AI systems promote payor acceptance of appropriate treatment?
- Will AI systems improve diagnostic quality through machine learning insights?
- Will AI systems degrade the diagnostic ability of the clinician due to any over-reliance on the technology and deprecating human insight? This may include other factors such as oral hygiene or past caries experience, etc., not included in current algorithms.

Finally, in the Effects on Healthcare System phase, broader implications are considered.

- Will AI systems lessen the cost of dental care?
- Will AI systems widen the availability of dental care?
- Will AI systems cause care inequity across different patient populations?
- Will AI systems allow other holistic information to be included in oral health care for better prognosis, intervention, and more effective treatment plans?

As clinicians are introduced to new Dental Image AI Systems, and regulators are tasked with their evaluation, they are confronted first with issues of Test Accuracy. How are they to decide, for each claim made by the systems, whether they are safe, effective, and appropriate for the intended use? In sum, how should clinical evaluation be conducted?

Clinical Evaluation

Valid Clinical Association datasets are necessary within an external validation dataset. Several lines of inquiry are posed.

What feature is the system detecting? And is that feature associated with a clinical condition the clinician intends to treat? For example, consider a system claiming to detect when a level of enamel demineralization suggests the need for a restoration. then one must confirm that such demineralization is appropriately associated with the clinical diagnosis, i.e., caries penetrating the dental-enamel junction. For a more challenging example, consider a system using the crown-to-root ratio to suggest the need for extraction. Though the system may accurately measure the ratio, and a certain ratio may be accepted as suggestive, other non-imaging factors may also have to be considered, such as the anatomy and histologic health of the periodontium.

Analytic Validation. Given the stated intended uses, and the specific claims within them, how well does the system perform?

- How reliable is the system at reporting a condition when it is actually present? This is the *Sensitivity* of a system. When a system fails to report a condition which is actually present, this is a *False Negative*. The more frequent false negatives are, the lower a system's sensitivity.
- How reliable is the system at not reporting a condition when it is actually absent? This is the *Specificity* of a system. When a system reports a condition which is actually absent, this is a *False Positive*. The more frequent false positives are, the lower a system's specificity.
- How reliable is the system overall? This is *Test Efficiency* of a system. The more frequent false reports are, the lower a system's test efficiency.

- The *Predictive Value* of a system also considers the population to which the patient belongs. If the population has a 90% prevalence of the condition, then even an 85% sensitivity may be disappointing. Similarly, if the population has a 10% prevalence of the condition, then even an 85% specificity may be disappointing. This highlights the importance of validating across a variety of populations, indeed across a wide variety of populations, based on ethnicity, sex, age, socioeconomic status, etc.

Clinical Validation. Therefore, validation of results, and the results' implications, depends on context. Predictive value shows the importance of prevalence within the population. But there are many kinds of context, many ways an individual's diagnosis and treatment plan could be affected. For example:

- Membership in populations varying by ethnicity, sex, age, socioeconomic status, etc.
- Individual health history.
- Other indicators of current health.
- Likelihood that follow-up diagnosis and/or treatment will occur.

Therefore, a Dental Image AI System is best used only as an instrument in the armamentarium. And, to evaluate a dental image AI system, one must understand, for each intended use, the set of images against which it was validated. Did the image set variety cover the specific population – the specific patient – being diagnosed? This question demonstrates the importance of a system's performance and its labeling.

System Performance and Labeling

As for many medical devices, a machine learning system's performance cannot be expressed as pass or fail. It is more useful to rate the system's performance for each of its claims. For example, how well does the system identify tooth numbers? How well does it identify a widened PDL space? Or bone loss? Or how often will the system over-diagnose a lesion? How often will it under-diagnose pulp involvement?

The needs of the user may be relevant to the questions to be asked.

- If the intended use of the system is to motivate an irreversible treatment, then the system's specificity may be most relevant: Is the rate of false positives very low?
- If the intended use for the system is radiographic screening, or clinician education, then the system's sensitivity may be most relevant: Is the rate of false negatives very low?
- If the intended use of the system is triage, then the system's test efficiency may be most relevant: Is the overall percentage of correct findings high?

Machine learning systems can be sensitive to such considerations. In the last layer of a neural network, the activation process of each node determines whether it classifies the case as a "yes" or "no" for the finding the node represents. System designers, by modifying the last-layer activation processes, can adjust the *Classification Threshold* for each finding. These thresholds can be tuned to minimize false positives, minimize false negatives, or achieve some combination of both of which the designer considers optimal.

In this way, a system can be tuned to meet goals related to maximizing patient care, clinical productivity, or business goals. Some designers even allow the user to adjust such thresholds, adapting the system for the task at hand.

Key Questions for Users: Clinicians' Guidance for Using AI for Clinical Decision Support

With an understanding of AI's principles, individuals can ask key questions to evaluate a system's applicability for a particular clinical practice.

On the validation of a machine learning system:

1) What tasks are claimed as part of the system's intended use?
2) How was it established that the validation dataset had enough images for each classification task claimed as part of intended use?
3) How was it established that the validation dataset had sufficient variety in ethnicity, sex, age, and socioeconomic status? And how was it established that there were enough images for each subpopulation?
4) Was the validation dataset sequestered from the training and testing processes?

On the clinical use of a machine learning system:

5) Is the system or the clinician responsible for diagnosis and treatment planning?
6) Based on the system's validation dataset, is it appropriate for use with the clinician's patient population?
7) Are the system's findings compatible with the clinician's own?
8) Does the system report levels of confidence in its results and explainability of how the decisions were made?
9) Will the system fail catastrophically if novel input is encountered?

On the performance of a machine learning system:

10) If the system's intended use includes treatment planning, what rate of false positives should be expected?
11) If the system's intended use includes radiographic screening, what rate of false negatives should be expected?
12) To accommodate specific intended uses, can thresholds be configured to adjust these rates?

By posing these questions, and pressing system providers to answer them, individual clinicians can help guide advancing technology toward both better patient care and higher efficiencies.

Appendix D

A Primer on Validity

How should a clinician judge the data provided for various use cases of AI that uses two dimensional images? What is the proper way to view the labeling for use claims, the data used to create the AI model and how to distinguish the use of internal validity (from the manufacturer of the AI product) vs the external validity (from an independent source.) So, here is an explanation.

Any test that has a dichotomous result (i.e., caries present or absent, or periodontal disease present or absent) should be accurate in 3 major ways: sensitivity, specificity and validity,

Sensitivity

Shreffler and Huecker[21] define sensitivity as, "the proportion of true positives tests out of all patients with a condition." In other words, it is "the ability of a test or instrument to yield a positive result for a subject that has that disease." The ability to correctly classify a test is essential. Shreffler and Huecker cite the following equation for sensitivity:

Sensitivity=(True Positives (A))/(True Positives (A)+False Negatives (C))

Sensitivity does not allow providers to understand individuals who tested positive but did not have the disease. False positives are a consideration through measurements of specificity and positive predictive value.

Specificity

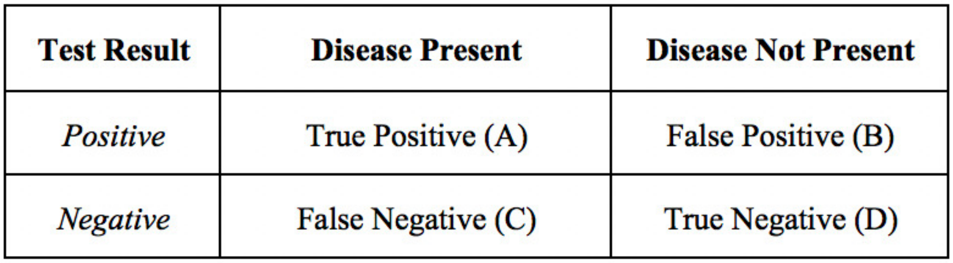
Shreffler and Huecker[21] define specificity as, "the percentage of true negatives out of all subjects who do not have a disease or condition." In other words, it is "the ability of the test or instrument to obtain normal range or negative results for a person who does not have a disease."[21]

Shreffler and Huecker cite the following equation for specificity:

Specificity=(True Negatives (D))/(True Negatives (D)+False Positives (B))

Sensitivity and specificity are inversely related: as sensitivity increases, specificity tends to decrease, and vice versa. Highly sensitive tests will lead to positive findings for patients with a disease, whereas highly specific tests will show patients without a finding having no disease. Sensitivity and specificity should always merit consideration together to provide a holistic picture of a diagnostic test. Diagnostic testing accuracy may be displayed as shown in Table 1.

Table 1. Diagnostic Testing Accuracy



Test Result	Disease Present	Disease Not Present
Positive	True Positive (A)	False Positive (B)
Negative	False Negative (C)	True Negative (D)

Source: Huecker and Shreffler[21]

Validity

Thirdly, results should be consistent across repeated administrations. High correlations between scores from two separate administrations of the test indicate that the test is *reliable*. A correlation is expressed as a percentage. Second, the tests should measure what they purport to measure. This is called being *valid*. Evidence of validity consists of high correlations between scores on the test of the construct of interest and measures of one or more other variables that are rationally connected. For example, valid self-reported pain scores should correlate with a dentist's ratings of periodontal disease, which may be or may not be correlated with the 5 stages of disease agreed upon by American Academy of Periodontology (AAP).

These are standard scientific measures. We should use the same measures for AI. AI manufacturers and FDA agree on the metrics, calculation method, sample size, and acceptance pass/fail criteria for these 3 types of tests. Labeling with these and various use cases in a simple way either on paper or before any pay walls on a website are important.

It should be noted that these reported metrics and methods are not the same across AI platforms. It may be helpful to review using Return on Investment (ROI) & AI view rate data and any design validation/clinical study data because that is more standardized across AI manufacturers. The type of population, source of data and how it was obtained to create the algorithm may also be important.

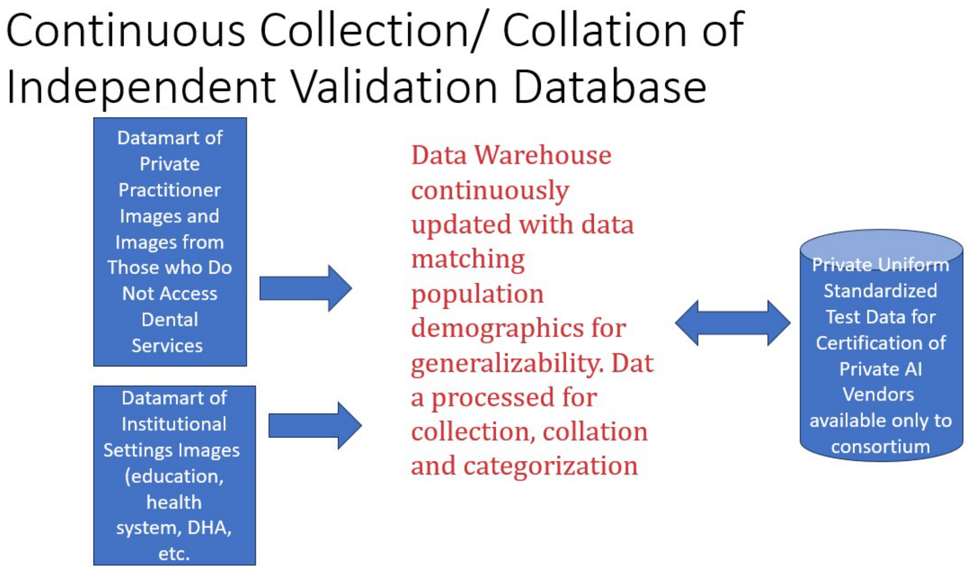
Internal validity for an AI product is different than external validity

A further distinction is drawn between internal and external validity in scientific studies. An internally valid study is tightly controlled and can provide confidence in results as far as study participants are concerned. Externally valid study results apply not only to subjects in the initial study, but also to others in the broader population.

AI applications in dentistry could involve either dichotomous diagnoses or continuous measures of variables such as ability to chew. With training, AI diagnostics should become quite sensitive and specific. Machines are inherently more reliable (consistent) than people. However, the validity of AI decisions must be assessed by comparing them with other logically related or analytically important variables.

There is one very important factor that should be considered by clinicians and manufacturers alike: ***Know the type of AI Model used.*** There are different types of models in which AI is developed. What type of model are they using? In medicine, deep learning models are proven to be the best for scalp disease by 99 %**,** Alzheimer disease by 96%, thyroid disease by 99%, 96% in skin disease, 99.37% in case of Arrhythmia disease, 95.7% in diabetic disease, while machine learning models achieved 89% in diabetic disease, 88.67% in tuberculosis, 86.84% in Alzheimer disease, etc. Like cars, it is helpful to know the model, but you don't have to know everything under the hood to drive the car. But you do need to know what kind of model it is for you to understand its strengths and limitations, or even exceptions.

Criteria for a valid external validation dataset have not been described previously. Appendix D, Figure 10 proposes a consortium that would collect and continuously update the external validation data for various populations segmented by age, gender, race/ethnicity/, etc. An independent body would be designated to keep this data secure and confidential, and accessible for external validation of vendor AI products with 2 dimensional images. A consortium of images could be collected from private practitioners, dental education, and from patients who do not access dental services, to comprise the external validation data set.



Continuous Collection/ Collation of Independent Validation Database

Datamart of Private Practitioner Images and Images from Those who Do Not Access Dental Services

Datamart of Institutional Settings Images (education, health system, DHA, etc.

Data Warehouse continuously updated with data matching population demographics for generalizability. Dat a processed for collection, collation and categorization

Private Uniform Standardized Test Data for Certification of Private AI Vendors available only to consortium

Figure 10. Consortium of images collected from practitioners, educators and patients

Glossary

Artificial Intelligence (AI) – Intelligence demonstrated by machines as opposed to natural intelligence displayed by humans. Some AI textbooks define the field as the study of any system that perceives its environment and takes actions that maximize the chance of achieving its goals.

Augmented Intelligence (AuI) – Sometimes referred to as intelligence amplification, AuI plays a similar role to AI except that it keeps human intelligence elements in its procedure. Rather than performing an assignment for a clinician like AI might do, AuI acts as a tool to assist the clinician in the task. One aspect of The American Medical Association House of Delegates' definition emphasizes that AuI's design *enhances* human intelligence rather than replacing it.

Clinical Decision Support (CDS) – Clinical decision support provides timely information, usually at the point of care, to help inform decisions about a patient's care. CDS tools and systems help clinical teams by taking over some routine tasks, warning of potential problems, or providing suggestions for the clinical team and patient to consider.

Ground Truth (also referred to as gold standard classification) – In machine learning, the term "ground truth" refers to the accuracy of the training set's classification for supervised learning techniques. This is used in statistical models to prove or disprove research hypotheses. It is critical for each validation test case that expected findings be correct and that the *ground truth* for each case is well established. In the case of dental imaging, analysis by oral and maxillofacial radiologists (OMR) to set this *ground truth* is highly regarded, but their participation in establishing this ground truth for any specific product or service is not guaranteed.

Machine Learning (ML) – IBM defines machine learning as a branch of AI and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually and automatically improving its accuracy.

Testing Dataset – After an algorithm is created using the training dataset and validation dataset, the testing dataset (also known as a "holdout dataset" because it is a set of data never before seen by the algorithm) may be used to verify the algorithm's ability to perform on new data.

Training Dataset – In dental imaging, the training dataset is typically a collection of dental images, such as intraoral radiographs. The samples in the dataset will provide examples of the kinds of finding the network is to detect. For instance, the sample radiographs might have a variety of already labeled class II lesions. It is then hoped the resulting network detects those lesions as effectively as the humans who originally identified them.

Validation Dataset – For a system to be validated, there must be a reference standard to which it's held. For a human clinician, that standard may be the opinion of teachers or of a review board. But for a software system, validation is typically achieved though testing against a validation dataset of test cases, which operates as a gold standard. And because the system cannot be interrogated as to its methods, the only way to evaluate the system is by its effectiveness in those test cases.

Bibliography

1. ADA SCDI White Paper No. 1106 for Dentistry – Overview of Artificial and Augmented Intelligence Uses in Dentistry. American Dental Association, 2022.

2. World Health Organization (WHO). Ethics and governance of artificial intelligence for health. WHO Guidance. Geneva: 2021. ISBN: 9789240029200.) The need for transparency and reduction of bias, if not elimination, is emphasized in US government efforts to date. (Feb 2024)

3. Lin M. What's Needed to Bridge the Gap Between US FDA Clearance and Real-world Use of AI Algorithms. Acad Radiol. 2022 Apr;29(4):567-568. doi: 10.1016/j.acra.2021.10.007. Epub 2021 Nov 20. PMID: 34794879; PMCID: PMC8903084.

4. Scott I, Carter S, and Colera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health & Care Informatics, 28(1), 2021.

5. Dept of Health and Human Services, National Coordinator for Health IT, Health Data, Technology, and Interoperability. RIN 0955-AA03 45 CFR Parts 170, 171 Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Washington DC: Dec 2023. https://www.healthit.gov/sites/default/files/page/2023-12/hti-1-final-rule.pdf

6. Ebrahimian S, Kalra MK, Agarwal S, Bizzo BC, Elkholy M, Wald C, Allen B, Dreyer KJ. FDA-regulated AI Algorithms: Trends, Strengths, and Gaps of Validation Studies. Acad Radiol. 2022 Apr;29(4):559-566. doi: 10.1016/j.acra.2021.09.002. Epub 2021 Dec 27. PMID: 34969610.)

7. ADA SCDI, ANSI/ADA Standard 1094, Quality Assurance for Digital Intra-oral Radiographic System

8. ADA SCDI. ANSI/ADA Standard 1099. Quality Assurance for Digital Panoramic and Cephalometric Radiogrphic Systems.

9. Udupa H, Mah P, Dover SB and Mc David WD. Evaluation of image quality parameters of representative intraoral digital radiographic systems. Oral Surgery, oral medicine, oral pathology and oral radiology. 116. 774-83. 10.1016/). 2013.Agency for Healthcare Research & Quality. Rockville. Guiding principles help healthcare community address potential bias resulting from algorithms. Guiding Principles Help Healthcare Community Address Potential Bias Resulting from Algorithms | Agency for Healthcare Research and Quality (ahrq.gov) (2023).   https://www.ahrq.gov/news/newsroom/press-releases/guiding-principles.html.

10. Jayakumar, S., Sounderajah, V., Normahani, P. *et al.* Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *npj Digit. Med.* **5**, 11 (2022). https://doi.org/10.1038/s41746-021-00544-y

11. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021) and

12. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).

13. Pew Research, Relying on AI in Their Own Health Care. Washington DC: Feb 22, 2023. Online at https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/)

14. Kearney VP, Yansane AM, Brandon RG, Vaderhobli R, Lin GH, Hekmatian H, Deng W, Joshi N, Bhandari H, Sadat AS, White JM. A generative adversarial inpainting network to enhance prediction of periodontal clinical attachment level. J Dent. 2022 Aug; 123:104211. doi: 10.1016/j.jdent.2022.104211. Epub 2022 Jun 26. PMID: 35760207.

15. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nat Med. 2023 Nov;29(11):2686-2687. doi: 10.1038/s41591-023-02540-z. PMID: 37853136.)

16. Kearney VP, Yansane AM, Brandon RG, Vaderhobli R, Lin GH, Hekmatian H, Deng W, Joshi N, Bhandari H, Sadat AS, White JM. A generative adversarial inpainting network to enhance pre diction of periodontal clinical attachment level. J Dent. 2022 Aug;123:104211. doi: 10.1016/ j.jdent.2022.104211. Epub 2022 Jun 26. PMID: 35760207.

17. HHS Policy for Protection of Human Subjects , 45 CFR Part 46 https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

18 Warraich HJ, Tazbaz T, Califf RM. FDA Perspective on the Regulation of Augmented intelligence/artificial intelligence in Health Care and Biomedicine. JAMA. Published online October 15, 2024. doi:10.1001/jama.2024.21451

19 Food and Drug Administration. FDA's Clinical Decision Support. Accessed on 2/26/2024: https://www.fda.gov/medical-devices/software-medical-device-samd/your-clinical-decision-support-software-it-medical-device.

20 Lijmer JG, Leeflang M, Bossuyt PMM. Proposals for a phased evaluation of medical tests. Medical Tests—White Paper Series. Agency for Healthcare Research and Quality. Rockville: MD.

21 Shreffler J, Huecker MR. Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios. [Updated 2023 Mar 6]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557491/

ADA American Dental Association®
America's leading advocate for oral health

211 East Chicago Avenue, Chicago, Illinois 60611
T 312.440.2500 F 312.440.7494 www.ada.org