|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-J-040 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 30 September – 2 October 2020 |
| **DOCUMENT** |
| **Source:** | TG-Dental Topic Driver |
| **Title:** | TG-Dental: Artificial intelligence for dental image analysis: A guide for authors and reviewers |
| **Purpose:** | Discussion |
| **Contact:** | Falk SchwendickeCharité – Universitätsmedizin BerlinGermany | Tel: +4930450662556Fax: +40304507562556Email: falk.schwendicke@charite.de  |
| **Contact:** | Joachim KroisCharité – Universitätsmedizin BerlinGermany | Tel: +4930450562328Fax: +40304507562932Email: Joachim.krois@charite.de  |

|  |  |
| --- | --- |
| **Abstract:** | The number of studies employing artificial intelligence (AI), specifically machine and deep learning, for dental image analysis is growing fast. The majority of studies shows weaknesses in planning, implementation and reporting, which in turn results in limited robustness and applicability. We propose to discuss and approve a **process to establish a living, non-authoritative guidance for authors and reviewers** on the conception, implementation and reporting of studies on dental image analysis under the roof of the TG Dentistry in the ITU/WHO Focus Group on AI for Health. |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***DRAFT 2020-09-21***

**Artificial intelligence for dental image analysis: A Guide for Authors and Reviewers**

Falk Schwendicke1,2, Tarry Singh2, Jae-Hong Lee2,4, Robert Gaudin4, Joachim Krois1,2

1 Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin, Berlin, Germany

2 ITU/WHO Focus Group AI on Health, Topic Group Dentistry

3 Department of Periodontology, Daejeon Dental Hospital, Institute of Wonkwang Dental Research, Wonkwang University College of Dentistry, Daejeon, Korea

4 Department of Maxillofacial Surgery, Charité – Universitätsmedizin Berlin, Berlin, Germany

Short title: A guide for authors and reviewers on AI in dental image analysis

Keywords: artificial intelligence, checklist, deep learning, dental, machine learning, teeth

Corresponding author:

Prof. Dr. Falk Schwendicke

Oral Diagnostics, Digital Health and Health Services Research Aßmannshauser Str. 4-6

Charité – Universitätsmedizin Berlin

14197 Berlin

Germany

Phone 0049 30 450 62556

Fax 0049 30 450 7562 556

falk.schwendicke@charite.de

Artificial intelligence for dental image analysis: A Guide for Authors and Reviewers

Abstract

Objectives: The number of studies employing artificial intelligence (AI), specifically machine and deep learning, for dental image analysis is growing fast. The majority of studies suffer from limitations in planning, conduct and reporting, resulting in low robustness and applicability. We here present a non-authoritative guide for authors and reviewers to be applied, discussed and further developed.

Methods: Lending from existing reviews in other fields and founded on the principles of evidence-based research practice, a set of guidance items are presented, assisting future scientists, reviewers and editors in planning, conducting, reporting and evaluating studies on AI in dental image analysis. The items have been derived on a discussion basis within the ITU/WHO focus group “Artificial Intelligence for Health (AI4H)”, and the topic group “Dental diagnostics and digital dentistry”, and should be rigorously appraised and adapted.

Results: Thirty-one items on planning, conducting and reporting studies were devised. These involve items on the study’s wider goal, focus, design and specific aims, data sampling and reporting, sample estimation, reference test construction, model parameters, training and evaluation, uncertainty and explainability, performance metrics and data partitions.

Conclusion: Scientists, reviewers and editors should consider this guide when planning, conducting, reporting and evaluating studies on AI for dental image analysis.

Clinical significance: Current studies on AI in dental image analysis show considerable weaknesses, hampering their replication and application. This non-authoritative guide may assist scientists, reviewers and editors to overcome this issue and advance AI research in dentistry as well as facilitate a forward-debate on standards in this fields.

**Introduction**

The term Artificial Intelligence (AI) was coined in the mid 1950s and has evolved from a topic rooted in computer science and applied mathematics into a field that is relevant to almost any scientific domain. From general themes such as learning and perception to very specific ones, such as playing games, writing poetry or music, painting, driving a car, drug discovery and diagnosing diseases. Owing to the universal nature of AI, a precise and unambiguous definitions is hard to find. Russel and Norvig framed AI as thinking and acting humanly or rationally [1]. A more technocratic definition for AI as given by the American National Standard Dictionary of Information Technology [2]: “The capability of a device to perform functions that are normally associated with human intelligence such as reasoning, learning, and self-improvement.”

Applications of artificial AI are entering medicine on a high pace, and one field which is specifically prolific is image analysis (also termed “computer vision”). In particular studies involving convolutional neural networks (CNNs) indicate that detecting diseases or assessing structures at superhuman speed and accuracy seems to be in reach, possibly making diagnostics and treatments safer, more personalized and efficient. Given the importance of imagery in dentistry (dental radiography accounts for the majority of all radiographs taken in medicine) [3], computer vision in dental image analysis certainly has potential. However, there are also doubts growing as to the robustness and generalizability, transparency and replicability as well as ethics, effectiveness and, overall, applicability of the results of studies in this field [4, 5].

Studies employing CNNs in dental image analysis are used to detect structures (teeth, bone) or pathologies (caries, apical lesions), segment images (cut out the area of interest), and classify them (as an enamel caries lesions, or a cyst), show significant weaknesses [6]: Datasets are mostly rather small, with developed AI solutions possibly lacking robustness and stability. The data generation process is oftentimes unclear and not necessarily fitting the question at hand, with both the data sources and the data characteristics not being fully sufficiently representative or reported, but also the annotation strategy (number and characteristics of annotators, instance or pixelwise annotation, independent or joint annotation, definition of reference test from annotations) being oftentimes not fully clear or suited. The choice of model, the training and hyperparameter tuning as well as the validation strategy is often unclear, and the metrics chosen to optimize the model against are not necessarily clinically relevant. Moreover, accuracy scores reported in many studies are often generated in-sample, not on hold-out test datasets or, even better, completely separate independent datasets. It hence often remains unclear if such accuracies can also be reached in other, real-life data, or in prospective sampled groups. Last, it is often not clear if the developed “narrow” application (task specific) is helpful in clinical practice, and which wider impact it has on health, but also further aspects like costs or ease of treatment provision etc.

Hence, there seems great need to raise awareness of scientists planning and conducting AI studies in dental image analysis, but also for authors, reviewers and editors of journals on which aspects to scrutinize when assessing their own or other researchers’ work. Without rapidly evolving from an experimental field into a solid, scientifically grounded and matured discipline, AI in dental research may disappoint or, worse, lead to harmful decisions in clinical practice. There is a need for guidance on how to plan, conduct and report AI studies in dental image analysis. Very limited guidance is available specifically for AI studies [7], usually pertaining on reporting, borrowing from other, more general reporting frameworks (e.g. those published by the EQUATOR network), while the global research audience urgently awaits guidance [5, 8]. There is seldom a focus on planning or conduct, seldom a focus on AI research, and seldom a focus on dental research.

The present narrative review builds on a range of existing studies in other fields, including checklists, and aims to give guidance to authors and journal reviewers dealing with studies on AI for dental image analysis. The authors of this review are topic drivers and members of the ITU/WHO focus group on AI for health (FG AI4H), specifically related to dentistry (<https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Dental.pdf>). They invite all interested parties to participate in this group to further and formally develop the provided guidance. Currently, this review and its content do not claim to be authoritative, but to highlight the issues observed or experienced by the group members and authors as well as to take up the work provided in other fields than dentistry. However, we strongly believe not taking action and waiting until more definite and elaborate guidance becomes available is risky; further resources will be wasted in the meantime, futile research findings will be spread and potentially harmful applications of AI in dental medicine translated into clinical practice [5].

**Methods**

This study was mainly conceived from our experience as researchers and authors in the field as well as reviewers and journal editors. We strongly believe that the quality of AI studies needs to improve fast, and given the exploding number of studies conducted and reported at present, we felt the need to author this document. We want to highlight that the approach taken to come to it was non-systematic: We firstly assess weaknesses on published studies in the field by review and discussion among the authors, who are all members of the topic group on dentistry within the ITU/WHO Focus Group AI for Health (<https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx>). We secondly assessed existing guidance documents, mainly from radiology, specifically a recently published checklist on reporting of AI studies on image analysis in radiology (CLAIM) [7], but also checklists published by the EQUATOR network like STROBE [9], TRIPOD [10], CONSORT [11], STARD [12], RECORD [13]. The group members discussed these and synthesized a list of items which authors, reviewers and journal editors may find useful to consider when planning, conducting and reporting or evaluating studies in AI in dental image analysis. Notably, our study does neither follow a formalized methodological approach nor does it claim to be authoritative. It rather serves to guide the mentioned stakeholders and to serve as discussion base. We invite all interested parties to join the ITU/WHO Focus Group AI for Health and work on standards to improve the level of science in this field.

**Results**

We first provide more general recommendations which may be considered when conceiving and planning studies using AI for dental research. We then provide more detailed and specific guidance on reporting such studies. Notably, authors should consider these as well during the early stages for their research in the sense of “backwards” planning, helping them to make their study design more relevant, robust, implementable and publishable eventually.

Planning and conducting

1. Study goal: Researchers should early on define the relevance, scope and meaning of the AI application they aim to develop or validate. The pitfalls in methodology when using AI methods, especially towards data and technical requirements (see below), should be kept in mind before employing AI techniques just because they are currently *en vogue*. Also, the end-users (patients or dentists) should be borne in mind when designing the study, as should be aspects related not only to the clinical application, but also regulatory requirements, ethics and data protection should the research ever be translated into usage.
2. Study focus: A clear focus on the study aims should be defined, e.g. what goal does the research have (developing a new or validating an existing model, diagnostics or prognostics etc). This focus is relevant to consider appropriate other existing detailed planning tools a priori, but also to lay out an analytic pathway.
3. Data: A major aspect when planning AI studies is data; especially in dentistry, datasets are oftentimes small and imbalanced (i.e. one class, like a pathology, is underrepresented, which makes training any AI models more difficult and needs to be addressed, for example via over-sampling, but also when evaluating the model, see below). Datasets should be as heterogenous as possible for the model to be as generalizable as possible, and this should be planned to be demonstrated (see below). Generally, researchers should consider the target population on which they envisage their model to be applied on, and critically compare the dataset they plan to train and to test their model on against that population.
4. Study aim: Researchers should have a clear idea if their study is exploratory or hypothesis-testing: The former will not necessarily need to be based on formal sample size estimations, but can also later on not claim to demonstrate any value of the model with statistical certainty. Clearly, there is also a need for exploratory studies, which generate hypotheses and open up new avenues; it is just important to define this early on.
5. Reference test: A major difficulty in the field of AI in dental image analysis is the construction of the reference test. Researchers have a number of aspects to consider. First, they need to decide if they aim for a classification model, where the image is instance-annotated (a caries lesions being present somewhere yes or no; this will also be the output of the model), a detection model (where a bounding box is drawn around the area of interest, e.g. the lesion, and later on also the model provides such bounding boxes) or a segmentation model (where annotations are provided pixelwise and the model later on also provides “blobs” on the image to be analysed, indicating for example this area to be a lesion). Different models need different annotation strategies and requirements, are differently computationally intensive, need different efforts when deploying the model later on, but also come with different strategies for applying methods of explainability.
6. Clustering: When having decided this aspect, it is oftentimes difficult to define a hard “gold standard” (e.g. histological assessment). Instead, dental researchers are oftentimes forced to use multiple human annotators independently assessing each image, thereby generating a “fuzzy” gold standard. Depending on the study’s aim, a clear case definition and calibration of annotators might be desired or not. The construction of the single label from this fuzzy data needs further considerations, e.g. researchers may use the majority votes in case of instance-based annotations, or unions or intersections on bounding box or pixelwise annotations. Ideally, if multiple imagery is available or clinical data, this may be used to triangulate the label with, at least for the test dataset: Noise and uncertainty may be to a certain extend acceptable in the training dataset, but hurtful and possible lead to bias in the test dataset.
7. Clustering: When feeding annotated data into the model, it is relevant to consider the specifics of dental imagery data: Often, multiple images are available from the same patient, either from the same time point (e.g. bitewing radiograph pairs, periapical radiographs used for periodontal status) or over different time points (during follow-up). The associated clustering may be used to add information to the model and provide new insights, while it is relevant to not spread imagery from the same patient between training and test dataset to avoid “data snooping bias”.
8. Test dataset: When testing the model, it is most relevant to not only report on validation data (i.e. data the model was exposed to during the training process) but a separate hold-out test dataset which the model has never seen, or even better a completely external dataset (which is the only option to demonstrate generalizability). Researchers should consider this during planning their study. Reporting only data from within-sample validation is insufficient. Validation that follows a k-fold cross validation approach may be taken into consideration but claims of generalizability should be avoided.
9. Computation resource: Researchers should consider, before engaging into AI research in dental image analysis, the computational resources, which are a major constrain especially when dealing with image data with a high resolution, large datasets or complex models. Researchers should be aware that the available computational resources may restrict the resolution of the image that is passed on to the neural network. Hence, features that are clearly identifiable at the original scale may be lost during preprocessing. Further, iterative hyperparameter tuning (see below) is computationally expensive and time consuming, this should at least be considered a priori.
10. Comparator: The model should be compared against relevant alternatives. These could be an independent group of dental examiners, possibly of different experience (to reflect the usefulness of the model in different groups) or against other accepted imagery or clinical tests. For such comparisons, relevant outcomes and outcome metrics should be used. Here, it is important to bridge the gap between dental research and technical disciplines. While the former often reports on accuracy metrics (which are not all useful in imbalanced datasets, see below), the latter considers F1-score or other metrics (which are more robust, but not interpretable from a medical perspective). This aspect becomes even more important when applying object detection or segmentation models. Commonly applied metrics such as (weighted) average precision or intersection over union are very domain specific so that concepts to convey them into the medical/dental research domain are warranted [14]. Ideally, an outcome set which reflects not only on the model’s accuracy, but also further aspects should be considered, again keeping in mind who will employ, commission, receive or pay for the application which may be developed using the model.

Reporting

A range of items which should be reported have been defined. Note that it may not be necessary to expand on each item in each study report, but that all items should be briefly considered and, if not presented, this absence should be justified.

1. Title: The title should clearly lay out that any kind of AI (deep learning, shallow machine learning, or more specifically the type of model like convolutional neural networks or random forests, among others) was used. Moreover, it should mention the study’s focus (diagnostic/prediction, development/validation), the clinical problem (e.g. caries detection on bitewings etc.) and the main outcome metric (accuracy, cost-effectiveness etc.)
2. Abstract: The abstract should present a structured summary of the study’s aim, methods, results, and conclusion. The abstract should stand for itself and should be understandable without reading the main manuscript. That also means that the used data (main characteristics, source of origin, type of sampling, partitioning into training, validation and testing datasets), the model and outcome metrics and the statistical analysis that was performed should be provided. The results section should provide full metrics, including measures of variance, for the primary outcome on the test (not only the training) dataset, and allow for any comparisons against alternatives.
3. Introduction: The introduction should briefly sum up the dental/clinical background of the study, if there is one, and deduce the need for an AI solution as proposed or tested by this study. It should be made clear if there is a clinical, a research or a teaching problem which this study focusses on. The introduction should then lay out the achievements and limitations made in this direction so far to provide a rationale for the study, its goals, and anticipated impact. It should be made clear in this section if the problem is of diagnostic or prognostic nature, if the goal of the study is to model development or validation, and should lay out if the study aims to explore possible applications or demonstrate such applications to have value. The former study type will be different in its setup from the latter type, where finally the introduction will provide one or more hypotheses to be tested and for which a sample size estimation (for the test dataset) will be needed (see above and below).
4. Study design: It is advisable to provide a short overview about the study design. This should include an overview about the study goal (will the model be used for detection, staging, monitoring, surveillance, prediction, or prognosis), data, its origin and sampling (retro- or prospective), modeling techniques, evaluation and scope (exploratory/hypothesis-testing, for the latter: aiming to demonstrate superiority or non-inferiority). If the methods part cannot, due to space restrictions, contain sufficient details for full replication, an appendix may be used and introduced early on. A possible registration of the study should be provided here, as should any definitions or terms used throughout the study. These definitions are relevant to consider given that different terms are used in the dental versus the technical disciplines; it is advisable to clarify them early on (e.g. in dental and, generally, medical research the model would be seen as index test which is tested against a reference test; in the technical disciplines, these terms are not used; similarly, the model aims – classification, detection, segmentation – should be briefly mentioned and defined for clarity reasons). If a reporting checklist or any other guidance document was used and adhered to, this may be reported here, too, to make reviewers, editors and readers aware of it and allow them to consult such a checklist.
5. Data: As data are the main component any AI model, this section is particularly relevant. The source of data for training, validation and testing (primary care, secondary care, general population) including the exact location of each data source, the timeframe of the sampling and in- or exclusion criteria should be defined (see next point). The data should be critically compared against the characteristics of the target population to help the reader gauge the generalizability and applicability (this should be taken up in the discussion). The heterogeneity of the data and potential sources of bias (especially concerning age, sex, ethnicity) should be explored. It should be made clear if any data were used for other studies. Ethical aspects (including ethical approval, informed consent) and data protection aspects (including possibly used strategies for de-identification) should be laid out. Ideally, the data should be provided in a repository. Notably, any code (see below) should so, too (e.g. GitHub, GitLab, etc). Dental datasets are oftentimes relatively small and narrow, which is why dental researchers have used within-sample validation as one means for demonstrating the value of the developed AI models. This should be made very clear; if any kind of hold-out, external or temporally separate test set is available, this should be clarified, as should be the partitioning between training, validation and test dataset, and how they were disjoint. When partitioning data, it should be clarified how repeated imagery from the same patient from the same or different time points was managed, as ideally the disjoint should be on the patient level so that images of the same patient do not appear in each partition. Researchers should indicate if there are any systematic differences between the data in each partition, and if so, why. If any kind of covariates were used for modelling, it should be made clear here which ones, and when and how they were collected and measured.
6. The identification, recruitment and inclusion of eligible imagery should be clarified. Inclusion and exclusion criteria, specifically the case definition (symptoms, characteristics) or criteria related to image type or quality, the data source location and setting, the imagery source (specifically technical characteristics including the machines used to generate images, acquisition parameters, reformat parameters) should be made clear. The study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up, should be specified. The sampling strategy (consecutive, random, or convenience) should be laid out and justified. The number of centers, patients, images and any meta-data should be presented to gauge the representativeness of the data.
7. It is relevant to consider if data protections standards in the US (HIPAA) or the EU (GDPR), or other relevant jurisdictions have been fulfilled, including appropriate Institutional Review if required. De-identification of dental datasets is often considered difficult., e.g. facial profiles can allow identification. Hence, a researcher should specify the means by which de-identification was performed.
8. If any data were missing (e.g. covariates etc.), it should be made clear how they were handled. Researchers should consider the bias that missing, replaced or imputed data might introduce.
9. In almost all studies the raw data needs to be extracted, transposed and loaded into machine-readable formats. Often the data originally stems from patient management systems or image databases (e.g. PACS). Hence, data extraction and preprocessing steps (manual or automated) should be described in detail. Elaborate on the use of normalization, change in bit depth, rescaling, cropping, compression, standardization, anonymization and file types. If applicable describe how inconsistent and missing data and/or wrong data types were handled and define any criteria to remove outliers. Include information (source and version number) on leveraged software, libraries, or any other tools.
10. Reference test: A major difficulty in many dental image analysis studies is the construction of the reference test, as laid out. The reference test setup should match the study aim (classification, detection, segmentation). The case definition and any kind of grading schemes for sub-types should be defined. The test threshold (positive cutoff), if defined, needs to be explained and justified, as it has an impact on the model and possibly also comparative dentists’ accuracy. It should also be made clear if this was all specified upfront or adjusted post hoc. It should be explained whether any clinical information was available to the assessors of the reference standard, or if any pre-annotation by a model was performed (so-called human-in-the-loop approach). Both can significantly impact on the reference test conduct. If any kind of existing label was used (e.g. from free-text imaging reports, electronic health, or existing models), researchers need to lay out how these labels had been generated, and need to gauge the risk of misclassification bias, unmeasured confounding, missing data. If using human annotators, the number of human annotators and qualifications should be specified, as should be any instructions and training given to them, including training materials (handbooks), which may be provided as a supplement. Researchers should describe whether annotations were done independently and how any discrepancies among annotators were resolved, if all annotators assessed each image or not, and which software they used for annotation. If multiple annotators assessed the same image, it needs clear reporting of how the fuzzy labels were translated into a single one, e.g. if for instance-based annotations majority vote schemes were applied, if experts could override the majority, or if for pixel-wise annotations unions etc. were used to come to the final label. These aspects should be especially clarified for the test dataset, as this is used to demonstrate the value of any AI application, but also to showcase dentists’ performance against the AI. If available, triangulation with other data or any efforts to provide a hard gold standard (like histology) should be separately explained here. Researchers should also report oninter- and intra-rater variability, and the steps taken to reduce or mitigate this variability.
11. Sample size: The sample size and how it was determined needs to be fully explained. In the absence of specific methods for AI and image analysis, researcher should use traditional power calculation methods to estimate the required sample size. Specifically, for dental data, researchers should consider clustering effects (lesion being clustered in teeth, teeth in images, images often in the same human, humans in centers etc). Sample size estimation mainly applies to test dataset, as any kind of hypothesis testing will be performed on this dataset. Knowing a prior the sufficient size of the training dataset is difficult. Notably, though, researchers should consider sensitivity analyses on how data drop or addition to their training dataset or sub-setting the training dataset impacts on the model’s performance and generalizability.
12. Model: Reproducibility is the key for a rapid uptake of new techniques; hence a complete and detailed description of the model is warranted. In particular for neural network models inputs, outputs, intermediate layers, pooling, normalization, regularization, and activation should be reported. Cite a reference, if the model was previously published. Further, the structure of the model may be presented in form of a graphical representation in the appendix or in code as supplemental data. Specify the names and version numbers of all software libraries, frameworks, and packages that were used. Further add information on the used hardware, in particular GPU specifications and used platforms (e.g. cloud vs. local cluster vs. on premise). These information may be provided in the appendix.
13. In particular for artificial neural networks the parameter initialization is crucial. Name the applied initialization strategy/distribution (zero, uniform, standard normal, He [15], Glorot [16] etc). If transfer learning is applied, specify the source of the starting weights and if there is a combination of initialization and transfer learning, specify which parts of the model were initialized with which strategies.
14. Training:Describe the training procedures in sufficient detail so that another researcher could reproduce the training process. Describe which data augmentation techniques were applied and e.g., for images the types and ranges of transformations. State how the training process was monitored and which criteria was used for stopping the training. List the values of the hyperparameters, describe the hyperparameter search strategy and provide the ranges of values that were considered. For artificial neural networks, at least the learning rate schedule, optimization method, batch size, dropout rates, regularization parameters (if any) and number of epochs should be provided. Discuss what objective function was applied and why it was selected. If transfer learning was applied, state which model parameters/layers are frozen and the portion of the training (e.g. number of epochs) that was affected.

Describe the method and model metric to select the final model and evaluate it against the held-out test set. If the final algorithm involves an ensemble of models, describe each model in accordance with guidelines outlined above. Describe how the component models are weighted and/or combined.

1. Evaluation: Researchers should describe the outcome and outcome metric(s) used to measure the model’s performance, defining the primary outcome and metric and relating it to the posed problem. Ideally, they should not only report on accuracy, but consider outcomes with relevance for decision making, applicability etc., as well. It should be made clear how any superiority over the current standards or alternatives is demonstrated (or not) and, if available, how the developmental and application costs may be justified thereby. The involvement of the public and patients should be considered when discussing outcome in the absence of any core outcome set.
2. Uncertainty: It needs laying out how uncertainty of the performance metrics values are assessed, how any comparisons between groups was done and how robust these comparisons are, for example by subgroup analyses of tooth groups, dentitions, patient risk groups, or images (from different centers or machinery). If comparing the AI model against individual dentists, their characteristics should be provided here, too.
3. Explainability: If feasible, researchers should lay out how the explainability, trustworthiness, and transparency of the model was assessed. There are an increasing number of applications towards “explainable AI” available [17]. This is also increasingly seen as a regulatory requirement for any kind of clinical application later on.
4. Results: The flow of data, including those in- and excluded, and data partitions into training, validation and test dataset should be clarified; a flowchart may be helpful. The sample should be characterized demographically, but also towards the prevalence of the condition of interest and the population’s risk profile to gauge its representativeness for the target population. If subgroups of severity have been defined, these should be characterized, too.
5. The performance metrics on all data partitions should be provided. The final model’s performance on the test partition should be provided in detail, and benchmarked against current technical standards or individual dentists. Estimates of variance like 95% confidence intervals or nonparametric estimates from bootstrap samples should be reported to gauge uncertainty. If useful, graphical displays like the Receiver Operating Characteristics Curves (ROC) or the Precision Recall curve could be used. Results on subgroups should be presented. If possible, any accuracy estimates should be translated into meaningful measures of decision making (e.g. sensitivity, specificity, positive and negative predictive value), and relative estimates may be translated into absolute ones in this sample. Researchers should finally provide information to understand incorrect predictions, for example by presenting examples of incorrectly classified or segmented cases to help readers better understand the strengths and limitations of the algorithm. Results from any explainability analyses should be presented and explored (e.g. as heatmaps, see above).
6. Discussion: As mostly recommended, we also see four aspects which should be provided; a summary, a strengths and limitations sections, a section on findings and their implications, and one on future directions. The results should be briefly summed up and contrasted against the study aim and hypothesis.
7. The study’s strengths, but more so limitations, especially towards data, the reference test, the applied metrics as well as associated biases, uncertainty, generalizability and robustness should be discussed. Necessary comparison with other studies should be drawn and the findings interpreted, for example as to their relevance for practice, including the potential clinical application of the AI model. Possible subsequent steps that one might take to build upon the provided results should be summed up and aspects which may hamper or facilitate successful translation into practice, research or teaching should be discussed.
8. Other Information: Here, recommendations towards authorship and registration according to the International Committee of Medical Journal Editors (ICMJE) should be provided. ICMJE recommends public registration of clinical trials at or before enrolling the first participant. The full study protocol may be linked in here, as may be data and model code. *S*ources of funding and the role of the funders as well as potential conflicts of interest should be explained.

Discussion

This study uses a narrative review technique to come up with a guide on how to plan, conduct and report studies using AI technology for dental image analysis. Such guide seems warranted given the demonstrated weaknesses of current studies in the field and the limited comparability across studies, hampering robust conclusions as to the overall body of evidence. Notably, our study itself suffers from methodological weaknesses, which have been spelled out. We plan to establish a more formally consented checklist in the future but given the number of studies published each month in this field, felt it necessary to provide some, if not ideal and perfectly robust, guidance for now. Formal consensus on any guidance should admittedly hear the voices of all stakeholders, as pointed out in our own recommendations above, to make sure such guiding document is relevant across disciplines and for all participants in the care process, i.e. patients, providers and payers. The present document does not replace existing checklists, many of which we mentioned, but complements and specifices them. As outlined, there will be checklists on specific study type (e.g. randomized trials) employing AI technologies in the nearer future, while such checklists will not be focused on dentistry but also be narrow in their scope. Given that many studies in the dental arena are rather exploratory and not prospective in nature, such guide on randomized trials using AI will also not be useful for this majority of cases at this time. With the maturation of this field and studies in AI-based dental image analysis increasingly falling into established study characteristics (e.g. randomized trials, observational studies, modelling studies), dental researchers are encouraged to adhere to these checklists as firstline recommendation. Nevertheless, the present document may be useful to prepare any study and to counter-check any final manuscript against it, too. Overall, we see this document as a living one and encourage lively debate. The field is moving fast and so will standards and guide documents need to be. Moreover, scrutiny and a critical view will be needed to move from AI research which currently rather experiments than demonstrates to a truly evidence-based research field.

In conclusion, and within the limitations outlined above, we propose a range of items dental researchers, reviewers and editors should compare their planned, conducted or reported study against to increase its robustness, comprehensiveness and transparency, thereby lifting the field to the standards expected from evidence-based research.

References

[1] S. Russel, P. Norvig, Artificial Intelligence: A Modern Approach, Global Edition, Addison Wesley, Boston, 2016.

[2] ANSDIT, American National Standard Dictionary of Information Technology, Springer, New York, 1996.

[3] HSE, Population Dose from Dental Radiology: 2010 2010.

[4] G. Marcus, Deep Learning: A Critical Appraisal, arXiv.org <https://arxiv.org/abs/1801.00631> (2018).

[5] S. Vollmer, B.A. Mateen, G. Bohner, F.J. Kiraly, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K.S.L. McAllister, P. Myles, D. Granger, M. Birse, R. Branson, K.G.M. Moons, G.S. Collins, J.P.A. Ioannidis, C. Holmes, H. Hemingway, Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, Bmj 368 (2020) l6927.

[6] J. Krois, T. Ekert, L. Meinhold, T. Golla, B. Kharbot, A. Wittemeier, C. Dorfer, F. Schwendicke, Deep Learning for the Radiographic Detection of Periodontal Bone Loss, Scientific reports 9(1) (2019) 8495.

[7] J. Mongan, L. Moy, C.E. Kahn, Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers, Radiology: Artificial Intelligence 2(2) (2020) e200029.

[8] X. Liu, L. Faes, M.J. Calvert, A.K. Denniston, Extension of the CONSORT and SPIRIT statements, The Lancet 394(10205) (2019) 1225.

[9] E. Elm, D.G. Altman, M. Egger, The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, Ann Intern Med 147 (2007).

[10] K.G. Moons, D.G. Altman, J.B. Reitsma, J.P. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration, Ann Intern Med 162(1) (2015) W1-73.

[11] K. Schulz, D. Altman, D. Moher, t.C. Group, CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials, BMC medicine 8(1) (2010) 18.

[12] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, J.G. Lijmer, D. Moher, D. Rennie, H.C. de Vet, H.Y. Kressel, N. Rifai, R.M. Golub, D.G. Altman, L. Hooft, D.A. Korevaar, J.F. Cohen, STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies, Bmj 351 (2015) h5527.

[13] E.I. Benchimol, L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, H.T. Sorensen, E. von Elm, S.M. Langan, The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement, PLoS Med 12(10) (2015) e1001885.

[14] A.G. Cantu, S. Gehrung, J. Krois, A. Chaurasia, J.G. Rossi, R. Gaudin, K. Elhennawy, F. Schwendicke, Detecting caries lesions of different radiographic extension on bitewings using deep learning, Journal of dentistry 100 (2020) 103425.

[15] K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, arXiv 1502.01852 (2015).

[16] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings 9(249-256) (2010).

[17] W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing2019.

Table 1: Items to be considered when planning, conducting and reporting AI studies in dental image analysis.

| Item No | Planning and Conducting |
| --- | --- |
| 1. | Study Goal | Consider relevance, scope and meaning and limitations of the AI application. |
| 2. | Study Focus | Clarify if developing a new or validating an existing model, or if scope diagnostics or prognostics etc.. |
| 3. | Data | Scrutinize the available dataset and mitigate bias, ensuring generalizability. |
| 4.  | Study Aim | Define if study exploratory or hypothesis-testing, and consider implications for study conception. |
| 5. | Reference Test | Decide if study aims for a classification model, detection model, segmentation model etc., and consider when establishing reference test. Decide on a justifiable basis for a method to establish reference test, especially when involving multiple annotators. |
| 6. | Clustering | Consider clustering of teeth and patients in your dataset, for example during data partitioning (“data snooping bias”). |
| 7. | Test Dataset | Report test metrics from independent test dataset. |
| 8. | Computational Resources | Consider resources when working with larger datasets or complex models. |
| 9. | Comparators | Compare your model against relevant comparators (experts, other models) using meaningful metrics. |
|  | Reporting |
| 10. | Title | Define that any kind of AI was used, specify which one and for which focus and problem. |
| 11. | Abstract | Present a structured summary of the study’s aim, methods, results, and conclusion. |
| 12. | Introduction | Sum up the clinical background and need of AI solution; achievements and limitations so far; goal of the study; hypothesis (if needed). |
| 13. | Study design | Assist the reader in understanding your study by providing an overview about the study goal, data characteristics, modeling techniques, evaluation and scope.  |
| 14. | Data | Give details towards the source of data for training and testing, in- and exclusion criteria, sampling framework, fit to target population, heterogeneity, partitioning, and if and where it can be accessed (or why not). |
| 14 a. | Sampling | Provide inclusion and exclusion criteria, case definition, image type and quality, data source(s)/centers, sampling strategy and information towards heterogeneity. |
| 14 b. | Data Protection | Provide information how data protection requirements were fulfilled. |
| 14 c. | Missing data | Explain how missing data was handled. |
| 14. d | Data Processing | Lay out how data processing (extracted, transposed, loaded, preprocessed) was performed. |
| 15. | Reference Test | Explain how the reference test was generated, including case definition, grading schemes, test thresholds and unification strategies for multiple labels.  |
| 16. | Sample Size | If your study is hypothesis-testing, provide information how your arrived at your test dataset sample size. |
| 17. | Model | Provide detail information on model inputs, outputs, intermediate layers, pooling, normalization, regularization, and activation, as well as software packages and hardware used. The structure of the model may be presented. |
| 17. a | Model parameters | Describe how the model parameters were initialized. |
| 18. | Training | Describe the training procedures including data augmentation techniques, criteria was used for stopping the training, hyperparameters and hyperparameter search strategy. For artificial neural networks, at least the learning rate schedule, optimization method, batch size, dropout rates, regularization parameters (if any) and number of epochs should be provided.  |
| 19. | Justify best-performing model | Describe the method and model metric to select the final model and evaluate it against the held-out test set.  |
| 20. | Evaluation | Describe the primary outcome and outcome metric. Consider further outcomes with relevance to the clinical question. |
| 21. | Uncertainty | Describe how uncertainties in the model results (comparisons, subgroups) are reflected on. |
| 22. | Explainability  | Lay out how explainability, trustworthiness, and transparency was assessed. |
| 23. | Results | Provide information on flow of data, including those in- and excluded, and data partitions into training, validation and test dataset. Characterize the dataset. |
| 23 a. | Performance metrics and data partitions | The final model’s performance on the test partition should be provided in detail, and benchmarked against current technical standards. Provide uncertainty estimates. Provide information to understand incorrect predictions and explainability. |
| 24. | Discussion | Provide a summary, a strengths and limitations sections, a section on findings and their implications, and one on future directions. |
| 25. | Other | Provide information towards authorship and registration, study protocol and potential conflicts of interest. |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_