International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

FG-AI4H Output

20 August 2021

TG-Dental Output 1

Artificial intelligence in dental research: A checklist for authors and reviewers

Summary

The checklist for dental artificial intelligence (AI) studies in this document was produced as a collaboration of experts from the International Association for Dental Research (IADR) E-oral Health Network and the ITU/WHO Focus Group on AI for Health.

The number of studies employing artificial intelligence, specifically machine and deep learning, is growing fast. The majority of studies suffer from limitations in planning, conduct and reporting, resulting in low robustness, reproducibility and applicability. This document presents a consensus checklist on planning, conducting and reporting of AI studies for authors, reviewers and readers in dental research.

Lending from existing reviews, standards and other guidance documents, an initial draft of the checklist and an explanatory document were derived and discussed among the members of IADR's e-oral network and the ITU/WHO Focus Group on Artificial Intelligence for Health (AI4H). The checklist was consented by 27 group members via an e-Delphi process.

As a result, 31 items on planning, conducting and reporting studies of AI studies were agreed upon covering: 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.

Current studies on AI in dentistry show considerable weaknesses, hampering their replication and application. Authors, reviewers and readers should consider this checklist when planning, conducting, reporting and evaluating studies on AI in dentistry. This checklist may help to overcome this issue and advance AI research as well as facilitate a debate on standards in this fields.

Keywords

Artificial intelligence; dentistry; dental research; guidelines; checklist; deep learning; machine learning; teeth

Change Log

This document contains Version 1 of the ITU-T Focus Group on AI for Health (FG-AI4H) TG-Dental Output 1 on "*Artificial intelligence in dental research: A checklist for authors and reviewers*" approved by correspondence by the FG-AI4H on 20 August 2021.

Editor:	Falk Schwendicke Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin Germany	Tel: +49 30 450 62556 E-mail: falk.schwendicke@charite.de

Contributors:

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

Tarry Singh	Visiting Faculty AI, University of Dallas, Texas
Jae-Hong Lee	Department of Periodontology, Daejeon Dental Hospital, Institute of Wonkwang Dental Research, Wonkwang University College of Dentistry, Daejeon, KoreaRobert Gaudin Department of Maxillofacial Surgery, Charité – Universitätsmedizin Berlin, Berlin, Germany
Akhilanand Chaurasia	Department of Oral Medicine and Radiology, King George's Medical University.,Lucknow, India
Thomas Wiegand	Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institute
Sergio Uribe	Bioinformatics Research Unit, Riga Stradins University, Riga, Latvia School of Dentistry, Universidad Austral de Chile, Valdivia, Chile
Joachim Krois	Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin, Berlin, Germany

CONTENTS

Page

1 Scope ... 1

2 References ... 1

3 Terms and definitions ... 3

3.1 Terms defined elsewhere ... 3

3.2 Terms defined here ... 3

4 Abbreviations ... 3

5 Conventions ... 3

6 Background ... 3

7 Methods ... 4

7.1 Scoping and derivation of items ... 4

7.2 Delphi process ... 5

7.3 Checklist pilot ... 5

8 Results ... 5

8.1 Planning and conducting ... 5

8.2 Reporting ... 7

9 Discussion ... 11

Appendix I Rationale for the choice of the Delphi technique ... 14

Bibliography ... 15

List of Tables

Page

Table 1 – Items to be considered when planning, conducting and reporting AI studies in dental research ... 12

ITU-T FG-AI4H TG-Dental Output 1

Artificial intelligence in dental research:
A checklist for authors and reviewers

Introduction

The checklist for dental AI studies in this document was produced as a collaboration of experts from the International Association for Dental Research (IADR) E-oral Health Network and the ITU/WHO Focus Group on AI for Health.

Objectives: The number of studies employing artificial intelligence (AI), specifically machine and deep learning, is growing fast. The majority of studies suffer from limitations in planning, conduct and reporting, resulting in low robustness, reproducibility and applicability. This document provides a consented checklist on planning, conducting and reporting of AI studies for authors, reviewers and readers in dental research.

Methods: Lending from existing reviews, standards and other guidance documents, an initial draft of the checklist and an explanatory document were derived and discussed among the members of IADR's e-oral network and the ITU/WHO focus group "Artificial Intelligence for Health (AI4H)". The checklist was consented by 27 group members via an e-Delphi process.

Results: Thirty-one items on planning, conducting and reporting studies of AI studies were agreed on. 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: Authors, reviewers and readers should consider this checklist when planning, conducting, reporting and evaluating studies on AI in dentistry.

Clinical significance: Current studies on AI in dentistry show considerable weaknesses, hampering their replication and application. This checklist may help to overcome this issue and advance AI research as well as facilitate a debate on standards in this fields.

1 Scope

The number of studies employing artificial intelligence (AI), specifically machine and deep learning, is growing fast. The majority of studies suffer from limitations in planning, conduct and reporting, resulting in low robustness, reproducibility and applicability. This document provides a consensus checklist on planning, conducting and reporting of AI studies for authors, reviewers and readers in dental research.

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

[4] 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.

[5] 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.

[6] S.B. Khanagar, A. Al-ehaideb, P.C. Maganur, S. Vishwanathaiah, S. Patil, H.A. Baeshen, S.C. Sarode, S. Bhandi, Developments, application, and performance of artificial intelligence in dentistry – A systematic review, Journal of Dental Sciences 16(1) (2021) 508-522.

[7] F. Schwendicke, W. Samek, J. Krois, Artificial Intelligence in Dentistry: Chances and Challenges, Journal of dental research 99(7) (2020) 769-774.

[8] Y.W. Chen, K. Stanley, W. Att, Artificial intelligence in dentistry: current applications and future perspectives, Quintessence international (Berlin, Germany : 1985) 51(3) (2020) 248-257.

[9] 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.

[10] X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, A.-W. Chan, A. Darzi, C. Holmes, C. Yau, H. Ashrafian, J.J. Deeks, L. Ferrante di Ruffano, L. Faes, P.A. Keane, S.J. Vollmer, A.Y. Lee, A. Jonas, A. Esteva, A.L. Beam, A.-W. Chan, M.B. Panico, C.S. Lee, C. Haug, C.J. Kelly, C. Yau, C. Mulrow, C. Espinoza, J. Fletcher, D. Paltoo, E. Manna, G. Price, G.S. Collins, H. Harvey, J. Matcham, J. Monteiro, M.K. ElZarrad, L. Ferrante di Ruffano, L. Oakden-Rayner, M. McCradden, P.A. Keane, R. Savage, R. Golub, R. Sarkar, S. Rowley, S.-A. The, C.-A.W. Group, A.I. Spirit, C.-A.S. Group, A.I. Spirit, C.-A.C. Group, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, Nature Medicine 26(9) (2020) 1364-1374.

[11] S.C. Rivera, X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension, BMJ 370 (2020) m3210.

[12] 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).

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] F. Schwendicke, T. Golla, M. Dreher, J. Krois, Convolutional neural networks for dental image diagnostics: A scoping review, Journal of dentistry (2019) 103226.

[18] S. Junger, S.A. Payne, J. Brine, L. Radbruch, S.G. Brearley, Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review, Palliative medicine 31(8) (2017) 684-706.

[19] 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.

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

[21] 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).

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

[23] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PloS one 10(3) (2015) e0118432.

3 Terms and definitions

3.1 Terms defined elsewhere

This document does not use any terms defined elsewhere.

3.2 Terms defined here

This document does not define any particular terms.

4 Abbreviations

AI	Artificial intelligence
CREDES	Guidance on conducting and reporting Delphi studies
EQUATOR	Enhancing the Quality and Transparency of Health Research (network)
FG-AI4H	ITU/WHO Focus Group on AI for health
IADR	International Association of Dental Research
ICMJE	International Committee of Medical Journal Editors
ROC	Receiver Operating Characteristics Curves

5 Conventions

None.

6 Background

The term artificial intelligence (AI) was coined in the mid 1950s, while owing to the universal nature of AI, a precise and unambiguous definition is hard to find. Russel and Norvig framed AI as thinking and acting humanly or rationally [1]. A more technocratic definition for AI is 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."

The technology underlying most AI applications in medicine is machine learning, specifically supervised learning, where machines learn from pairs of data and (often human/expert) annotated labels (e.g., "this radiograph contains a carious lesion"). The machine is then able to iteratively learn the inherent statistical patterns underlying this pairing (i.e., model them) and thereby, at some point, to make predictions on unseen and unlabelled data, usually in a test set which is separated and independent from the training dataset, or later on in the "real-world" clinical application.

Applications of artificial AI are entering medicine on a high pace, with the expectations that they may make diagnostics and treatments safer, more personalized and efficient. 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 these studies [3-8]: The datasets used to train AI models are usually 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 sufficiently representative or reported, but also the data annotation (labelling) strategy, e.g., the number and characteristics of annotators, instance or pixelwise annotation, independent or joint annotation, definition of reference test from annotations, being oftentimes not fully clear. As a result, studies – also in dentistry – have been found to suffer from high risks of bias and limited applicability [6]. Moreover, the choice of the AI model, its 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 completely independent data. It often remains unclear if such accuracies can also be reached on real-life data or in prospectively sampled groups. Last, it is often not clear if the developed "narrow", task-specific application is helpful in clinical practice, and which wider impact it has on health, but also costs or ease of treatment provision etc. [3-8]. These limitations, as outlined, apply to dentistry, too, as indicated by recent meta-studies [6-8] and hamper assessment, reproduction and implementation of AI in dental research and practice.

Overall, there seems a need to improve the planning, conducting and reporting of AI studies in dental research. Such guidance has been developed for AI studies in radiology [9], and the EQUATOR network is currently publishing and developing reporting standards or their extensions for AI studies [10,11]. This document mirrors the paper in [AIDCL] presents a checklist that aims to give guidance on planning, conducting and reporting for authors, reviewers, and readers of dental AI studies. The checklist has been reviewed by members of the International Association of Dental Research (IADR) e-oral health network and the ITU/WHO Focus Group on AI for health (FG-AI4H) and it is supported by both groups.

7 Methods

7.1 Scoping and derivation of items

To derive the checklist items for planning, conducting and reporting or evaluating studies in the field, existing guidance documents were assessed, mainly from radiology, specifically a recently published checklist on reporting of AI studies on image analysis in radiology (CLAIM) [6], but also checklists published by the EQUATOR network like STROBE [12], TRIPOD [13], CONSORT [14] and the very recent CONSORT-AI extension [10], STARD [15] and RECORD [16]. A recent systematic scoping review on this matter was further used to identify possible items [17]. Pilot interviews to raise possible additional items with members of the IADR's e-oral health network (https://www.e-oralhealth.org) and the ITU/WHO FG-AI4H, specifically the topic group on Dental Diagnostics and Digital Dentistry (TG-Dental) and the Working Group on Clinical Evaluation (https://itu.int/en/ITU-T/focusgroups/ai4h), were conducted.

A list of items was then synthesized for planning, conducting and reporting or evaluating studies in the field. The list and a guidance document, authored by interested members of the abovenamed groups, was discussed and revised after distributing it among all members of the groups, allowing for items to be suggested for addition.

7.2 Delphi process

Members of the described groups were contacted and were invited to participate in an online Delphi process where they could anonymously vote on the items and suggest additional ones; the group leaders were further asked to support snowballing sampling, inviting further interested parties or individuals. Overall, 72 individuals were contacted, 27 of which eventually participated. The overall consensus groups represented clinicians, researchers from both the clinical and the technical disciplines, methodologists, journal editors and reviewers, regulatory professionals, policy makers, industry representatives and patients. The guidance document itself (excluding the checklist items) was not submitted to any further consensus process.

A one-staged e-Delphi survey was undertaken in December 2020 and January 2021; its reporting follows the Guidance on conducting and reporting Delphi studies (CREDES) [18]. Further details are provided in the appendix. Given our sampling, the Delphi participants had sufficient breadth of expertise and experience and covered a wide geographic range. Some of the experts were familiar to the organizers, and some experts came from the same institution. No adjustments were made for this aspect during the consensus process, as any kind of possible bias introduced by this was assumed to be limited and was accepted, but also as no valid rules for such adjustment are available.

Before the Delphi, participants were given written information about the study. Further demographic details were not inquired. There was the option to not answer single questions (opt-out) and to suggest additional or revised items at the end of the survey.

The Delphi asked for an agreement to each item on a scale of 1-10 (do not at all agree to agree fully). Maximum two stages of the Delphi were planned. Each round was planned to be closed after a two-week period. Two reminders via email were planned for each round. Panellists were allowed to comment on each item. The survey was conducted via a customized online platform; and survey data was analysed descriptively. The following consensus rules applied. (1) Agreement to an item was defined by marking grades 7-10 on the described scale from 1-10. (2) Minimum 70% of all participants needed to agree to an item for this to be consensually accepted. Items which did not meet these criteria after the planned two rounds were dropped. As a stable agreement on all items was achieved in the first round, with all items being agreed on, no second round was needed and hence dropped.

7.3 Checklist pilot

The final checklist was piloted by two independent researchers not involved in the conception and construction of the checklist, evaluating clarity in wording and application. No changes were required after this piloting. The checklist first provides more general recommendations towards conceiving and planning studies using AI for dental research, and then lays out 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, replicable, robust, implementable and publishable eventually.

8 Results

8.1 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, should be kept in mind early on. Also, the end-users (patients or dentists) should be considered when designing the study, as should be regulatory requirements, ethics and data protection.

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.).

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 but also evaluation of AI models more difficult and needs to be addressed). Datasets should be as heterogenous as possible for the model to be as generalizable as possible, and this should be planned to be demonstrated (below it is outlined that this should be done on a hold-out test or a fully independent dataset). Generally, researchers should consider the target population on which they envisage their model to be applied on (e.g., an AI-based caries risk prediction tool for children should be trained and tested in data from children, not adults or seniors), and critically compare the dataset they plan to train and to test their model on against that population (using social and demographic, but also dental covariates like number of teeth or restorations).

4. Study Aim: Researchers should have a clear idea if their study is exploratory or hypothesis-testing: Clearly, there is also a need for exploratory studies, which generate hypotheses and open up new avenues, while a full sample size estimation is required when wanting to demonstrate any value of the model with statistical certainty.

5. Reference Test: A major difficulty is the construction of the reference test (i.e., the labels of the data to be learnt and tested on) in the absence of a hard "gold standard" (e.g., histological assessment). Instead, dental researchers are oftentimes forced to use multiple human annotators to independently label the data, thereby generating a "fuzzy" gold standard (e.g., a pool of existing radiographs is labelled by 3 or 4 independent dental experts, who may not always agree on their verdict). 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 majority votes (e.g., if 3 of 4 experts agree, the majority label is assumed to be "the truth"). Ideally, if multiple data sources are available, they may be used to triangulate (e.g., if in addition to the described radiograph, clinical data is available, these may be used to support labelling). Last, 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.

6. Clustering: When feeding annotated data into the model, it is relevant to consider the specifics of dental data: Often, multiple datapoints (images, clinical assessment) 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 data from the same patient between training and test dataset to avoid "data snooping bias".

7. 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, mainly as using this approach no external and fully independent testing is performed.

8. Computational resource: Researchers should, before engaging into AI research, consider the computational resources, which are a major constrain especially when dealing with image data, large datasets or complex models. Researchers should be aware that the available computational resources may restrict the resolution of the images that are used for training,

with features that are clearly identifiable at the original scale possibly being lost during preprocessing. Further, iterative hyperparameter tuning is computationally expensive and time-consuming, which should be considered, too.

9. 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 researchers and technical disciplines. While the former often report accuracy metrics (which are not all useful in imbalanced datasets, mainly as randomly guessing the more frequent class often leads to high accuracy without indicating usefulness for the clinician), the latter consider 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 (where only the affected pixels are marked). Commonly applied metrics such as (weighted) average precision or intersection over union are very domain specific and hard to interpret from a clinically perspective (is an intersection of 80% of pixels useful or not?), and concepts to convey them into the medical/dental research domain are warranted [19]. 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.

8.2 Reporting

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

10. Title: The title should clearly lay out that any kind of AI (shallow or deep 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.).

11. 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 (e.g., 95% confidence intervals), for the primary outcome on the test (not only the training) dataset, and allow for any comparisons against alternatives (e.g., dental clinicians or established tests).

12. Introduction: The introduction should briefly sum up the dental background of the study, if there is one, and deduce the need for an AI solution. It should be made clear if there is a clinical, a research or a teaching problem. The introduction should then lay out the achievements and limitations in this field 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 and if the study aims to explore new AI applications or validate existing ones.

13. Study Design: It is advisable to provide a short overview about the study design to allow orientation for readers early on. 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), modelling 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 especially when considering that terminology differs between dental and technical disciplines (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 like this one or any other guidance document was used and adhered to, this may be reported here, too.

14. Data: As data are the main component of any AI model, this section is particularly relevant. The source of data for training, validation and testing (primary care, secondary care, general population), the timeframe of sampling and the in- or exclusion criteria should be defined. The data should be critically compared against the characteristics of the target population to help the reader gauge the generalizability and applicability (as outlined above; does the data used to train and test the model fit the population it will be applied to later on?). 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 of these data had been used for other studies before. Ethical aspects (including ethical approval, informed consent) and data protection aspects (e.g., employed 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, as discussed, means that generalizability can not be shown; within-sample validation only demonstrates if the models was able to "learn" the sample data, not to predict on new data. If any kind of hold-out, external or temporally separate test set has been used, this should be clarified, as should be the partitioning between training, validation and test dataset. When partitioning data, it should be clarified how repeated data 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 data of the same patient do not appear in each partition (as otherwise, for example, the same carious teeth from the same patient may appear on two photographs, with one photograph used for training and the other for testing, which may heavily bias the metrics). Researchers should indicate if there are any systematic differences between the data in each partition, and if so, why. A tabularization of the different partitions' characteristics can be helpful. 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.

    a. The identification and recruitment of data should be clarified. Inclusion and exclusion criteria, specifically the case definition (symptoms, characteristics) or criteria related to data type or quality, the data source location and setting, the source (for imagery, the 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 centres, patients, datapoints and any meta-data should be presented to gauge the representativeness of the data.

    b. 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 institutional review if required. De-identification of dental datasets (especially imagery) is often considered difficult., e.g., clinical reports or facial profiles can allow identification and researchers should lay out how they dealt with this challenge.

c. In almost all studies the raw data need to be extracted, transposed and loaded into machine-readable formats. Often the data originally stem from patient management systems or image databases (e.g., PACS). Hence, data extraction and preprocessing steps (manual or automated) should be described. For imagery, elaborate on the use of normalization, change in bit depth, rescaling, cropping, compression, standardization, anonymization and file types. Include information (source and version number) on leveraged software, libraries, or any other tools.

d. 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.

15. Reference test: A major difficulty in AI studies is the construction of the reference test. The case definition and any kind of grading schemes for sub-types should be defined. The test threshold (positive cut-off), 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 record, 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 their qualifications (dental speciality, experience) 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 datapoint or not, and how (software used etc.). If multiple annotators assessed the same datapoint (e.g., an image), it needs clear reporting of how the fuzzy labels were translated into a single one, e.g., if majority vote schemes were applied, if experts could override the majority, etc. 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 on inter- and intra-rater variability, and the steps taken to reduce or mitigate this variability.

16. Sample size: For hypothesis-testing studies, the sample size and how it was determined (sample size estimation) needs to be fully explained. Specifically, for dental data, researchers should consider clustering effects (lesion being clustered in teeth, teeth in humans, in centres, and all of this often in repeated cross-sections). Sample size estimation mainly applies to test dataset, as any kind of hypothesis testing will be performed on this dataset. Knowing a priori 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 impacts on the model's performance and generalizability.

17. Model: A complete and detailed description of the model is warranted, allowing to replicate the employed methods. 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 used. Further add information on the used hardware, in particular GPU specifications and used platforms (e.g., cloud vs. local cluster vs. on premise), possibly in the appendix.

a. In particular for neural networks the parameter initialization is crucial. Name the applied initialization strategy/distribution (zero, uniform, standard normal, He [20], Glorot [21] 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.

18. Training: Describe the training procedures in sufficient detail so that another researcher could reproduce the training process. Describe which data augmentation techniques were applied. State how the training process was monitored, and which criteria were 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 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 were frozen and the portion of the training (e.g., number of epochs) that was affected.

19. Describe the method and model metric (e.g., accuracy, F1-score) to select the final model and evaluate it against the test set. If using an ensemble of models, describe each model in accordance with guidelines outlined above. Describe how the component models were weighted and/or combined.

20. 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 outlined clinical/teaching/research problem. Ideally, they should not only report on accuracy, but consider outcomes relevant for decision making, applicability etc., as well (e.g., treatment consequences, costs, cost-effectiveness, implementation). 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.

21. Lay out how uncertainty of the performance metrics values was assessed, how any comparisons between groups were done and how robust these comparisons were, for example by subgroup analyses of tooth groups, dentitions, patient risk groups, or data sources (from different centres or machinery). If comparing the AI model against individual dentists, the dentists' characteristics should be provided here, too.

22. 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 [22]. This is also increasingly seen as a regulatory requirement for any kind of clinical application later on.

23. 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.

a. The performance metrics on all data partitions (training, test) 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 [23] 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 additionally. Researchers should further provide information to understand incorrect predictions to help readers better understand the strengths and limitations of the model. Results from any explainability analyses should be presented and explored (e.g., as heatmaps, see above).

24. Discussion: As mostly recommended, four aspects are also seen 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.

    a. The study's strengths, but more so limitations, especially towards data, the reference test, the applied metrics as well as associated biases, uncertainty, generalizability, reproducibility 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.

25. Other Information: Here, recommendations towards authorship and registration according to the International Committee of Medical Journal Editors (ICMJE) should be followed. The full study protocol may be linked in here, too. *S*ources of funding and the role of the funders as well as potential conflicts of interest should be explained.

9 Discussion

The checklist described in this document seemed warranted given the outlined weaknesses of studies in the field and the limited comparability across studies, hampering robust conclusions as to the overall body of evidence [3-8]. It may assist authors, reviewers, editors and readers of AI studies in dentistry and thus improve the quality in study conception, conduct and reporting.

Notably, the study itself suffers from methodological weaknesses itself. A limited number of stakeholders, mainly from the expert arena, have been involved, i.e., patients, clinicians and payers have been underrepresented. The items were built on existing guidance documents, reviews and iterations; a new review was not conducted to systematically yield a comprehensive set of possible items, for example based on a newly established risk of bias tabulation of studies in the field. Also, members from two bodies were invited to participate, and while representing a wide range of experiences, expertise and backgrounds, one may argue that a different group of people could have come up with a different list of items. Overall, these caveats were accepted in preparing this checklist, while noting that it does not constitute a clinical guideline, but a voluntary guidance which is perceived as urgently needed to raise the quality in the field. Also, the present document does not replace explicitly existing checklists, many of which are mentioned, but complements and specifies them. As outlined, the EQUATOR network is currently and in a stepwise approach preparing and publishing checklists on different study types employing AI; however, such checklists will not be focused on dentistry and are also narrow in their scope (i.e., on specific study types). Dental researchers remain encouraged to adhere to these checklists, though.

In conclusion, and within the limitations outlined above, this document contains 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.

Table 1 – Items to be considered when planning, conducting and reporting AI studies in dental research

Item No	Planning and conducting		Agreement[1]
1	Study goal	Consider relevance, scope and meaning and limitations of the AI application.	10 (8-10)
2	Study focus	Clarify if developing a new or validating an existing model, or if scope diagnostics or prognostics etc..	9 (7-10)
3	Data	Scrutinize the available dataset and mitigate bias, ensuring generalizability.	10 (5-10)
4	Study aim	Define if the study is exploratory or hypothesis-testing and consider implications for study conception.	9 (6-10)
5	Reference test	Decide on a justifiable basis for a method to establish a reference test, especially when involving multiple annotators.	9 (5-10)
6	Clustering	Consider clustering of teeth and patients in your dataset, for example during data partitioning ("data snooping bias").	8 (5-10)
7	Test dataset	Report test metrics from an independent test dataset.	9 (6-10)
8	Computational resources	Consider resources when working with larger datasets or complex models.	8 (4-10)
9	Comparators	Compare your model against relevant comparators (experts, other models) using meaningful metrics.	9 (5-10)
Reporting			
10	Title	Define that any kind of AI was used, specify which one and for which focus and problem.	10 (7-10)
11	Abstract	Present a structured summary of the study's aim, methods, results, and conclusion.	10 (7-10)
12	Introduction	Sum up the clinical background and need of AI solution; achievements and limitations so far; goal of the study; hypothesis (if needed).	10 (6-10)
13	Study design	Assist the reader in understanding your study by providing an overview about the study goal, data characteristics, modelling techniques, evaluation and scope.	10 (6-10)
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).	10 (8-10)
14.a	Sampling	Provide inclusion and exclusion criteria, case definition, image type and quality, data source(s)/centres, sampling strategy and information towards heterogeneity.	10 (7-10)
14.b	Data protection	Provide information how data protection requirements were fulfilled.	10 (6-10)
14.c	Missing data	Explain how missing data was handled.	10 (6-10)
14.d	Data processing	Lay out how data processing (extracted, transposed, loaded, pre-processed) was performed.	10 (6-10)
15	Reference test	Explain how the reference test was generated, including case definition, grading schemes, test thresholds and unification strategies for multiple labels.	10 (7-10)

[1] The median (min.-max.) agreement on the scale from 1-10 is given.

Item No	Planning and conducting		Agreement[1]
16	Sample size	If your study is hypothesis-testing, provide information how your arrived at your test dataset sample size.	10 (6-10)
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.	10 (7-10)
17.a	Model parameters	Describe how the model parameters were initialized.	9 (6-10)
18	Training	Describe the training procedures including data augmentation techniques, criteria used for stopping the training, hyperparameters and hyperparameter search strategy. For 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.	10 (6-10)
19	Justify best-performing model	Describe the method and model metric to select the final model and evaluate it against the hold-out test set.	10 (7-10)
20	Evaluation	Describe the primary outcome and outcome metric. Consider further outcomes with relevance to your question.	10 (8-10)
21	Uncertainty	Describe how uncertainties in the model results (comparisons, subgroups) are reflected on.	9 (6-10)
22	Explainability	Lay out how explainability, trustworthiness, and transparency were assessed.	9 (6-10)
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.	10 (7-10)
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.	9 (7-10)
24	Discussion	Provide a summary, a strengths and limitations sections, a section on findings and their implications, and one on future directions.	10 (6-10)
25	Other	Provide information towards authorship and registration, study protocol and potential conflicts of interest.	10 (6-10)

Appendix I
Rationale for the choice of the Delphi technique

1. Justification: An online Delphi was employed, allowing for a transparent, anonymous voting. The technique is accepted by the community. By combining the open-ended initial conception and discussion of the items with a Delphi, a systematic and comprehensive consensus process was possible.

Planning and design

2. Planning and process. The consensus rules (see below) were set by the authors and communicated via e-mail before starting the Delphi process. The Delphi asked for an agreement to each item on a scale of 1-10 (do not at all agree to agree fully). Maximum two stages of the Delphi were planned. Each round closed after a 2-week period. Two reminders via email were sent for each round. Panellists were allowed to comment on each item. The survey was conducted via a customized online platform; and survey data was analysed descriptively.

3. Definition of consensus. The following consensus rules applied. (1) Agreement to an item was defined by marking grades 7-10 on a scale from 1-10. (2) Minimum 70% of all participants needed to agree to an item for this to be consensually accepted. Items which did not meet these criteria after the planned 2 rounds were to be dropped.

Study conduct

4. Informational input: The material provided to the panel is described in the main text. Its attainment has been described above.

5. Prevention of bias: A systematic and comprehensive approach under participation of a wide range of experts and two acknowledged international bodies was chosen.

6. Interpretation and processing of results: There was, as discussed, stable agreement to all items after the first round.

7. External validation: Some external validation was sought as the authors have utilized the checklist in recent publications.

Reporting

8. Purpose and rationale: These have been provided.

9. Expert panel: Two acknowledged international bodies invited a comprehensive sample of experts; participation was further open to other interested parties and individuals.

10. Description of the methods: Preparatory steps, conception and authoring of the document, iteration of the checklist, survey rounds have been described.

11. Procedure: The Delphi steps have been described.

12. Definition and attainment of consensus: The following consensus rules applied. (1) Agreement to an item was defined by marking grades 7-10 on a scale from 1-10. (2) Minimum 70% of all participants needed to agree to an item for this to be consensually accepted.

13. Results: The results are reported in the main text.

14. Discussion of limitations: A limited group of people have been invited and came to this consensus, which is a limitation.

15. Adequacy of conclusions: The conclusions reflect the outcomes of the Delphi.

16. Publication and dissemination: The checklist is published in an international journal for dissemination.

Bibliography

[AIDCL] Falk Schwendicke et al., *Artificial intelligence in dental research: A Checklist for Authors and Reviewers*, Journal of Dentistry, Volume 107, April 2021, https://doi.org/10.1016/j.jdent.2021.103610.