Journal of Dentistry 150 (2024) 105318

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Core outcomes measures in dental computer vision studies (DentalCOMS)

Martha Büttner [a], Rata Rokhshad [b], Janet Brinz [c], Julien Issa [d,e], Akhilanand Chaurasia [f], Sergio E. Uribe [c,g,h], Teodora Karteva [b], Sanaa Chala [i], Antonin Tichy [c], Falk Schwendicke [c,*]

[a] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin, Germany
[b] Topic Group Dental Diagnostics and Digital Dentistry, WHO Focus Group AI on Health, Berlin, Germany
[c] Clinic for Conservative Dentistry and Periodontology, LMU Klinikum, Munich, Germany
[d] Department of Diagnostics, Chair of Practical Clinical Dentistry, University of Medical Sciences, Bukowska 70, 60-812, Poznan, Poland
[e] Doctoral School, Poznań University of Medical Sciences, Bukowska 70, 60-812, Poznan, Poland
[f] Department of Oral Medicine and Radiology, King George's Medical University, India
[g] Department of Conservative Dentistry and Oral Health, Riga Stradins University, Riga, Latvia
[h] Baltic Biomaterials Centre of Excellence (BBCE), Headquarters at Riga Technical University, Riga, Latvia
[i] Faculty of Dental Medicine, Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco

ARTICLE INFO

Keywords:
AI
Computer vision
Deep learning
Image analysis
Radiograph

ABSTRACT

Objectives: To improve reporting and comparability as well as to reduce bias in dental computer vision studies, we aimed to develop a Core Outcome Measures Set (COMS) for this field. The COMS was derived consensus based as part of the WHO/ITU/WIPO Global Initiative AI for Health (WHO/ITU/WIPO AI4H).

Methods: We first assessed existing guidance documents of diagnostic accuracy studies and conducted interviews with experts in the field. The resulting list of outcome measures was mapped against computer vision modeling tasks, clinical fields and reporting levels. The resulting systematization focused on providing relevant outcome measures whilst retaining details for meta-research and technical replication, displaying recommendations towards (1) levels of reporting for different clinical fields and tasks, and (2) outcome measures. The COMS was consented using a 2-staged e-Delphi, with 26 participants from various IADR groups, the WHO/ITU/WIPO AI4H, ADEA and AAOMFR.

Results: We assigned agreed levels of reporting to different computer vision tasks. We agreed that human expert assessment and diagnostic accuracy considerations are the only feasible method to achieve clinically meaningful evaluation levels. Studies should at least report on eight core outcome measures: confusion matrix, accuracy, sensitivity, specificity, precision, F-1 score, area-under-the-receiver-operating-characteristic-curve, and area-under-the-precision-recall-curve.

Conclusion: Dental researchers should aim to report computer vision studies along the outlined COMS. Reviewers and editors may consider the defined COMS when assessing studies, and authors are recommended to justify when not employing the COMS.

Clinical significance: Comparing and synthesizing dental computer vision studies is hampered by the variety of reported outcome measures. Adherence to the defined COMS is expected to increase comparability across studies, enable synthesis, and reduce selective reporting.

1. Introduction

Currently, the choice of diagnostic accuracy measures in studies on artificial intelligence (AI) is non-systematic and driven by the AI task, but also by researchers' preferences. A recent systematic review identified a total of 42 different diagnostic accuracy measures being used in dental AI studies [1], some of which are summarized in Table 1. This, in parts, is grounded in the wide range of modeling tasks, as particularly evident in computer vision, i.e. image and video analysis using AI (Fig. 1):

- Classification refers to a modeling problem where a class label is assigned to a given example of an image or video.

[*] Corresponding author at: Clinic for Conservative Dentistry and Periodontology, LMU Klinikum, Munich, Goethestr. 70, 80336, Munich, Germany.
E-mail address: falk.schwendicke@med.uni-muenchen.de (F. Schwendicke).

https://doi.org/10.1016/j.jdent.2024.105318
Received 12 July 2024; Received in revised form 16 August 2024; Accepted 20 August 2024
Available online 27 August 2024

Table 1
Summary of important metrics used to assess machine learning tasks in medicine.

Metric	Calculation	Explanation
Confusion matrix	$\begin{matrix} TP & TN \\ FP & FN \end{matrix}$	Summary of true and false predictions
Sensitivity/ true positive rate/ recall	$\frac{TP}{TP+FN}$	Proportion of lesions that were recognized as such
Specificity/ true negative rate	$\frac{TN}{FP+TN}$	Proportion of healthy that were recognized as such
Precision, positive predictive value	Depending on prevalence: $\frac{SEN \times PRE}{SEN \times PRE + (1-SPE) \times (1-PRE)}$ Independent of the prevalence: $\frac{TP}{TP+FP}$	Predictive power: Number classified as positive which are actually positive
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	Proportion of the total quantity being recognized correctly
F1-Score	$2 \times \frac{P \times SEN}{P+SEN}$	Combination of accuracy and sensitivity (harmonically weighted mean)

- Detection tasks aim to identify and locate objects within an image or video, usually using outlining rectangles (bounding boxes).
- Segmentation tasks outline an object in an image or video, with semantic segmentation classifying each pixel into a particular class (all instances of the same object are classified identically), and instance segmentation providing a unique label to every instance of a particular object in the image.

In addition to this complexity, dental outcomes can be measured on various levels, e.g. on pixel level, site or surface level, tooth level or patient level, resulting in differences in interpretation, and requiring complex statistical considerations given the potential clustering and interdependence of statistical units [2]. Consequently, the results of different dental computer vision studies are not comparable, oftentimes incomplete, with a high risk of reporting bias, low usefulness of individual studies and limited options to synthesize data quantitatively [1].

The problem of variability in outcomes, i.e., what is recorded, also termed domains, and outcome measures, i.e., how is it recorded, also termed instruments, is not restricted to diagnostic accuracy studies, like those around computer vision, and has been discussed intensively in intervention studies [3-9]. Starting within rheumatology, a movement has formed to define agreed collections of outcomes and outcome measures, so called Core Outcomes Sets (COS) and Core Outcome Measures Sets (COMS) for interventions in specific clinical domains. Using COS/COMS is not mandatory, and researchers can employ different outcomes and outcome measures, but this should be transparently and appropriately justified [10]. COS/COMS are supposed to increase reporting consistency and comprehensiveness as well as the usefulness and applicability of interventional research and to reduce reporting bias [9].

We present a COMS for dental computer vision studies to improve reporting completeness, comparability, and reduce bias. The COMS was derived in a consensus based fashion as part of the WHO/ITU/WIPO Global Initiative AI for Health (GIAI4H).

2. Methods

2.1. Scoping and derivation of items

To identify potential COMS, two authors (MB, FS) assessed existing guidance documents of accuracy studies, including CLAIM 2020 [11], TRIPOD [12], STARD [13] as well as a checklist on reporting dental AI studies [2] and the so-far most comprehensive systematic review on this topic [1]. Pilot interviews with experts in the field, e.g., members of the WHO/ITU/WIPO GIAI4H, specifically the Topic Group Dentistry (TG Dental), were also conducted. The resulting list of outcome measures was mapped against modeling tasks, clinical fields, and reporting levels. The resulting systematization focused on providing the most relevant outcome measures for clinicians whilst retaining the most detail for further research (meta-research) and technical replication, laying out statements towards the level of reporting for different clinical fields and tasks, and statements on outcome measures. Two authors (MB, FS) then drafted an accompanying document for the COMS to explain the background and framework for applying the COMS.

The COMS statements and the accompanying document were discussed and revised after being distributed among the members of TG

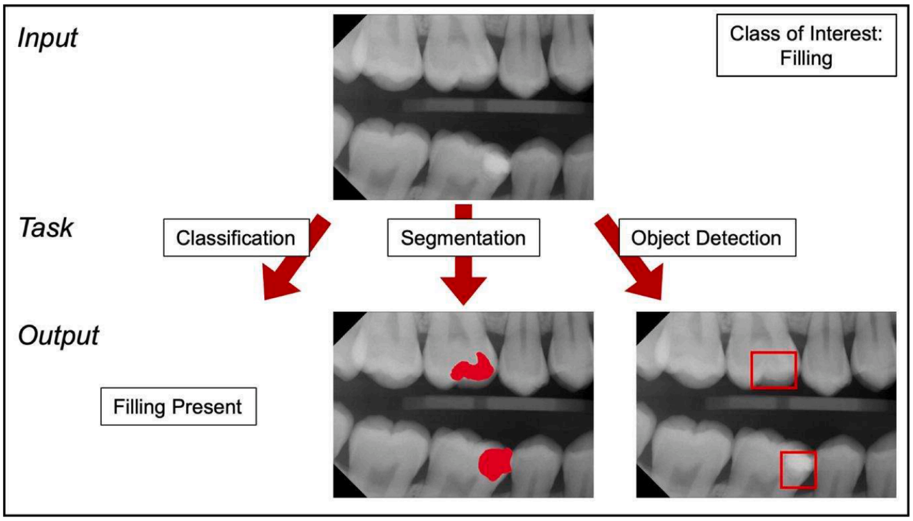

Fig. 1. Visualization of the most common computer vision tasks using deep learning for the example of identifying fillings on a bitewing radiograph.

Dental. The COMS was then consented using an e-Delphi, as laid out below.

2.2. *Delphi process*

The following groups were contacted and invited to participate in an online Delphi process: WHO/ITU/WIPO TG Dental, IADR Diagnostic Sciences Group, IADR e-Oral Health Network, IADR Oral Medicine and Pathology group, American Dental Education Association (ADEA), Asian Association of Maxillofacial Radiology (AAOMFR). Participants could anonymously vote on each statement (see below) and suggest revisions, additions, or deletions. Members were further asked to support snowballing sampling, inviting further interested parties or individuals. The first round was conducted over four weeks in January 2024. The second round was accessible for voting between 19 February and 15 May 2024. Overall, 26 individuals participated. The consensus group represented clinicians, researchers from the clinical and technical disciplines, methodologists, journal editors and reviewers, regulatory professionals, policymakers, industry representatives, and patients.

A two-staged e-Delphi survey was undertaken; its reporting follows the Guidance on conducting and reporting Delphi studies (CREDES) [14]. Further details are provided in the appendix. Given our sampling, the Delphi participants had a wide breadth of expertise and geographic range. Some experts were familiar to the organizers, and three came from the same institution. Before the Delphi, participants were given written information about the study. We did not inquire about further demographic details. There was the option not to 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). A maximum of two Delphi stages, each lasting four weeks, were planned. Two reminders via email were planned for each round. The time distance between the rounds was 11 weeks. The survey was conducted via Microsoft Forms; survey data was analyzed descriptively. Per definition, agreement was given if participants voted seven or higher. A statement was counted as consensually accepted if 70 % or more of all participants agreed to it. An open meeting for discussion was held after the first round. Statements were adapted according to comments and discussions. Statements where no consensus was given were dropped after the second round.

In the following, we first outline the deliberations that determined the development of the COMS, then present the levels of reporting suggested for different clinical fields and tasks, and finally display the COMS.

3. Results

Each computer vision modeling task presents a distinct nature of evaluation. A key issue concerns the level of reporting, which pertains to how models evaluate images. Conclusions and problems for the nature level of reporting are illustrated in Fig. 2.

- For classification tasks, one output is provided for an image or video, which is not always clinically useful. Notably, however,

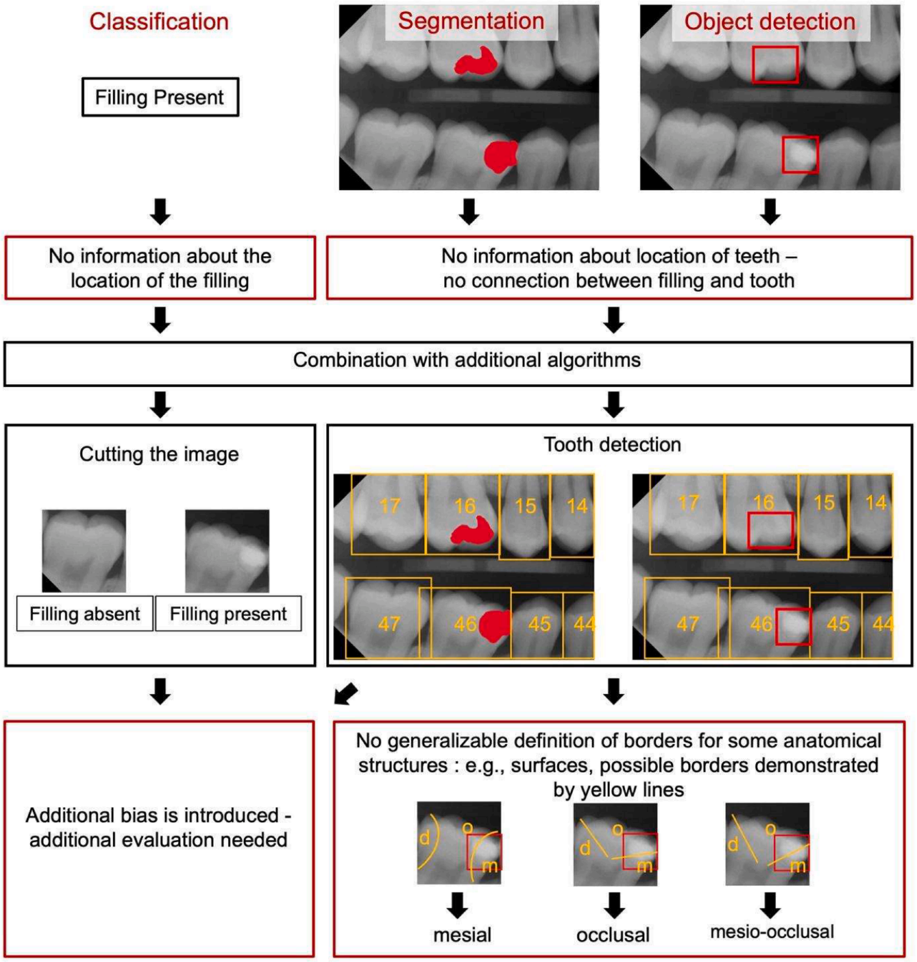
Classification
Segmentation
Object detection
Filling Present
No information about the location of the filling
No information about location of teeth – no connection between filling and tooth
Combination with additional algorithms
Cutting the image
Filling absent
Filling present
Tooth detection
17
16
15
14
47
46
45
44
Additional bias is introduced - additional evaluation needed
No generalizable definition of borders for some anatomical structures : e.g., surfaces, possible borders demonstrated by yellow lines
d
o
m
mesial
occlusal
mesio-occlusal

Fig. 2. Different computer vision tasks require different considerations towards reporting, depending on level of reporting. Black boxes demonstrate conclusions or options, red boxes demonstrate problems and challenges.

classification metrics (accuracy, sensitivity, specificity) for image classification are easy to compute, with the computation not being biased by image-level aspects (see next bullet points). To increase the usefulness, images could be cropped to regions of interest (e.g. smaller images showing a tooth), which would be clinically more useful. At the same time, additional input for cropping is needed. This could come from auxiliary models or humans cropping the images; both would need additional evaluation (Fig. 2).

- For segmentation tasks, each pixel is assigned to a specific class. Evaluating models at this level can introduce bias depending on whether the background class is included or excluded in the analysis. This is specifically relevant in dentistry, where, for most analyses, a major part of the image is considered as background. When the background class is included in the analysis, a correctly identified background improves accuracy – less importance is given to the actual task. However, if the background class is excluded, the evaluation will focus on how well the algorithm identifies classes of interest, but it does not contain information about non-object areas, as true negatives (TN) are no longer defined. Hence, a combination of metrics is necessary to populate a confusion matrix and allow the calculation of metrics like sensitivity and specificity (Fig. 3). Often**,** pixel-wise classification is transformed into a class-wise evaluation (e.g., segmentation), where the intersection of the model's output with the ground truth pixel field is compared. For the latter, a threshold for classification (e.g., minimum 50 % of pixels need to overlap) needs to be decided by the model developer, which may introduce bias. More importantly, pixel-wise classification is limitedly useful for clinicians. To compute clinically relevant measures based on anatomical regions, such as at the tooth or surface level, the precise location of these regions within the image is required to map the model output to these regions (similar to the cropping described above). Once more, auxiliary models or human input would be needed to provide this information, requiring further validation and introducing an additional layer of bias. An alternative approach involves modifying the level of reporting within the outcome classes. Instead of generic classes (e.g., caries), the location could be specifically defined involving the level of reporting (e.g., "caries on tooth 17 mesial"). This, however, would significantly increase the number of required classes and, consequently, the volume of data needed, which is impractical in specialized fields like dentistry, where an average training data set contains 450 images [1].
- For object detection, objects of interest are classified and described using bounding boxes. A detected object can be considered a true positive or false positive, while undetected objects are false negative. Importantly, as an image can contain any number of objects, their maximum is not defined, limiting the option to calculate TN and related metrics like specificity and area under the receiver operating curve (AUC). Hence, further input is required to define the maximum number of objects and thus TN (e.g., the number of teeth, when only one object per tooth can occur) (Fig. 3).

3.1. Computer vision tasks

Table 2 provides a comprehensive overview of the nature of different computer vision tasks, along with the implications and challenges for each evaluation. It concludes that human expert assessment is the only feasible method to achieve clinically meaningful evaluation levels.

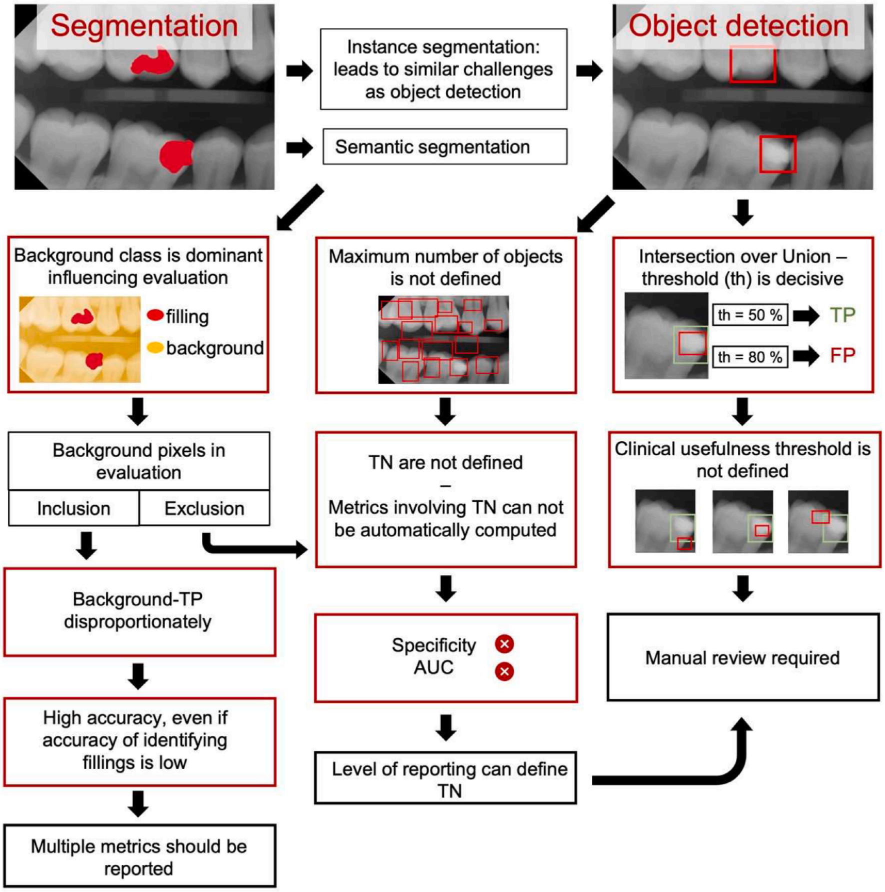
Segmentation
Instance segmentation: leads to similar challenges as object detection
Semantic segmentation
Object detection
Background class is dominant influencing evaluation
filling
background
Maximum number of objects is not defined
Intersection over Union – threshold (th) is decisive
th = 50 %
TP
th = 80 %
FP
Background pixels in evaluation
Inclusion
Exclusion
TN are not defined – Metrics involving TN can not be automatically computed
Clinical usefulness threshold is not defined
Background-TP disproportionately
Specificity
AUC
Manual review required
High accuracy, even if accuracy of identifying fillings is low
Level of reporting can define TN
Multiple metrics should be reported

Fig. 3. Problems of evaluating segmentation or object detection tasks. Black boxes demonstrate conclusions or options, red boxes demonstrate problems and challenges. Segmentation can be either treated as instance segmentation which introduces similar challenges as object detection or as pixelwise segmentation with challenges demonstrated in the left column.

Table 2
The nature of evaluation for different computer vision tasks and its implications, options and challenges.

	Classification	Semantic Segmentation	Object Detection
Level of reporting	Image level	Pixel level	Object level
Model output	(One) label per image	One label per pixel	Label and coordinates (bounding box) for objects of interest
Implication	Image defines level of reporting	No linkage between pixels and anatomical region of interest	No linkage between object and anatomical region of interest
Metrical complications	-	Background class introduces bias	Absence of true negatives complicates metric calculation
Reporting option 1	Cropping images in regions of interest	Auxiliary models or human input needed	Auxiliary models or human input and definition of min./max. possible objects needed
Limitations of option 1	Limitedly useful for clinicians	Auxiliary models or human input introduce own bias; validation needed.	
Reporting option 2		Expanding number of classes (e.g., instead of label “caries”: “caries on tooth 45 mesial”)	
Limitations of option 2		Amount of training images increases dramatically.	

Instance Segmentation is not represented separately as its characterizes are similar to semantic segmentation or object detection.

Table 3
Task and level of reporting mapping. Results of the consensus process are represented with median, minimum (min) and maximum (max) of participants voting.

Statement	Pathology/ Task	Level of reporting	Field of dentistry	Agreement median (min, max)
1	Dental instruments recognition	Object (instrument) level	General	10 (6,10)
2	People identification, age and gender estimation	Patient level	General	10 (5,10)
3	Tooth localization, detection, and segmentation	Tooth level	General	10 (7,10)
4	Angio vascular conditions (e.g., atherosclerotic plaque)	Image level	Oral medicine	9 (2,10)
5	Conditions of oral mucosa (e.g., leukoplakia)	Image and object (lesion) level	Oral medicine	10 (1,10)
6	Salivary gland conditions (e.g., Sjogren’s syndrome)	Gland level	Oral medicine	10 (1,10)
7	Sinuses conditions (e.g., sinusitis)	Patient side level	Oral medicine	9.5 (1,10)
8	Tongue conditions	Image level	Oral medicine	9 (1,10)
9	Tongue landmarks	Object (landmark) level	Oral medicine	10 (1,10)
10	Trabecular landmarks	Object (landmark) level	Oral medicine	9 (1,10)
11	Bone condition (e.g., osteoporosis)	Image level	Oral surgery and implantology	9 (5,10)
12	Implant identification and localization	Object (implant) level	Oral surgery and implantology	10 (6,10)
13	Root morphology classification	Tooth level	Oral surgery and implantology	10 (4,10)
14	Segmentation or detection of anatomic structures (bone, nerve canal, foramen mentale etc.)	Object level	Oral surgery and implantology	10 (7,10)
15	Third molar development	Tooth level	Oral surgery and implantology	10 (7,10)
16	Tooth distance to nerve	Tooth level	Oral surgery and implantology	10 (5,10)
17	Tooth impaction	Tooth level	Oral surgery and implantology	10 (5,10)
18	Tumor, tumor-like diseases, and cysts	Image and object (lesion) level	Oral surgery and implantology	10 (5,10)
19	Angle classification	Image/ patient level	Orthodontics	10 (2,10)
20	Dental cusps classification	Cusps level	Orthodontics	10 (6,10)
21	Facial attractiveness	Image level	Orthodontics	10 (1,10)
22	Need of extraction for orthodontic treatment	Tooth level	Orthodontics	10 (3,10)
23	Need of orthodontic surgery	Patient level	Orthodontics	9 (5,10)
24	Orthodontic landmark detection	Object (landmark) level	Orthodontics	10 (2,10)
25	Skeletal malocclusion	Image level	Orthodontics	10 (2,10)
26	Tooth rotation	Tooth and surface level	Orthodontics	10 (1,10)
27	Attachment loss	Tooth and surface Level	Periodontology	10 (6,10)
28	Dental plaque and biofilm quantification	Patient side and tooth level	Periodontology	10 (5,10)
29	Gingival diseases	tooth level	Periodontology	10 (3,10)
30	Periodontal staging and grading	Image/ patient level	Periodontology	10 (2,10)
31	Teeth brushing quality	Patient side level	Periodontology	9 (1,10)
32	Tooth mobility	Tooth level	Periodontology	10 (5,10)
33	Dental arche condition of edentulous	Jaw level	Prosthodontics	9 (1,10)
34	Apical Lesions	Tooth level	Restorative dentistry and endodontics	10 (5,10)
35	Caries	Tooth and surface level	Restorative dentistry and endodontics	10 (6,10)
36	Debonding probability	Object (restoration) level	Restorative dentistry and endodontics	10 (5,10)
37	Identification of restorations	Tooth and surface level	Restorative dentistry and endodontics	10 (7,10)
38	Need of restoration	Tooth and surface level	Restorative dentistry and endodontics	10 (5,10)
39	Root fractures	Tooth level	Restorative dentistry and endodontics	10 (5,10)
40	Tooth defects (cracks, abrasion, erosion etc.)	Tooth level	Restorative dentistry and endodontics	10 (5,10)

Table 4
Outcome measures. Results of the consensus process are represented with median, minimum (min) and maximum (max) of participants voting.

Outcome measure	Agreement, median (min, max)
Confusion Matrix. A visual and numerical representation that clearly outlines the algorithm's true and false predictions in all categories can provide a comprehensive understanding of its performance. Researchers should present a comprehensive confusion matrix for each task, delineating true positives, true negatives, false positives, and false negatives.	10 (2,10)
Accuracy. To quantify the proportion of total predictions that the algorithm correctly identified.	10 (2,10)
Sensitivity. The algorithm's proficiency in correctly identifying cases with a specific condition	10 (2,10)
Specificity. The algorithm's ability to correctly identify true negatives	10 (2,10)
Precision. Indicative of the exactness of its positive predictions, as it is important in scenarios where false positives bear significant consequences, such as invasive or costly interventions.	10 (1,10)
F-1 score. Representing the harmonic mean of precision and sensitivity, should be provided to convey a balanced view of the algorithm's performance, especially in the presence of uneven class distributions.	9 (1,10)
Area under the receiver operating characteristic curve. Area under the sensitivity and (1-specificity) curve. Evaluates the diagnostic capability compared to random guesses	8 (2,10)
Area under the precision recall curve	8 (1,10)

3.2. Level of reporting

A wide range of computer vision tasks were identified. The consensus group assessed which levels of outcomes should be reported (Table 3). Note that additional levels may be employed and reported, too, or that—depending on the task—another level may be chosen, but this choice should be made clear and justified.

3.3. Core outcome measures

Based on the consensus achieved, studies on computer vision tasks in dentistry should report on eight core outcome measures (Table 4).

4. Discussion

Computer vision has become a mainstream task in dentistry, while the reporting of studies remains highly heterogenous, impeding impairing studies and their synthesis. The development of this COMS will improve the comparability of studies and their clinical usefulness, facilitating the translation of research findings into clinical care and enabling more meaningful meta-analyses. The present study used a structured process to define and consent core outcome metrics for this emanating domain of dentistry. The outlined COMS will support researchers in defining outcome measures of interest, while the consented set of image levels for different dental computer vision tasks will hopefully enable considerate choices. Notably, researchers are not bound to any of the consented statements – it is better to provide an informed and judicious choice of individual metrics and levels of reporting than trying to fit research into standardized boxes. However, if doing so, transparent reporting and justification are needed to enable others to understand and appraise the made decisions. For example, researchers may choose to report the Dice Coefficient, which was not included in the COMS, because it will provide similar results (same in binary tasks) as the F1-score. Both metrics are mathematically similar but interpreted differently. The Dice Coefficient, a similarity metric evaluating the overlap of two area against the union, emerges from segmentation tasks, and comes with less clinical relevance given it being measured on pixel level, as elaborated above. Hence, the F1-score was proposed as core metric, as it provides a harmonic mean between precision and sensitivity.

Our study has some limitations. First, we focused on computer vision – the most prolific, but not only field in dental AI research. Future work may consider natural language processing and predictive dentistry. Second, only a limited number of computer vision tasks, those mainly found in the literature, were considered. Emerging themes in computer vision may not be fully reflected by the present document, while researchers in those new domains are invited to adapt our recommendations to their field of work accordingly. Third, only a limited (albeit diverse) number of experts participated in the Delphi, and it can be expected that further development under broader participation might be needed in the future, particularly given the dynamic nature of the field (in parts reflected by the CLAIM checklist, used by us as one document for our scoping task, has recently been updated) [15]. Fourth, age or gender cannot be evaluated given the consensus being conducted anonymously. Last, we focused on metrics that can be yielded by both diagnostic accuracy studies and clinical trials. Notably, the latter may also report on relevant aspects around the implementation of further impact of computer vision, e.g. costs or diagnostic process or patients' attitudes. These outcomes are not at all reflected, as they are not necessarily to be expected in all studies in the field and hence can be regarded as supplementary. These tasks are of utmost importance to fully characterize AI's impact on clinical care, and research focusing on such outcomes should be highly welcomed.

5. Conclusions

Dental researchers are recommended to report computer vision studies along the outlined COMS. Reviewers and editors may want to consider the defined COMS when assessing studies in the realm, and authors should actively justify when outcome measures were omitted or added. Adherence to the defined COMS is expected to increase comparability across studies, enable synthesis, and reduce selective reporting bias.

Ethical approval and informed consent

No ethical approval or consent was required.

Data availability statement

All data are published in this paper.

CRediT authorship contribution statement

Martha Büttner: Writing – original draft, Visualization, Project administration, Methodology, Formal analysis, Conceptualization. **Rata Rokhshad:** Writing – review & editing, Validation, Investigation. **Janet Brinz:** Writing – review & editing, Validation, Formal analysis. **Julien Issa:** Writing – review & editing, Validation, Methodology. **Akhilanand Chaurasia:** Writing – review & editing, Validation, Methodology, Formal analysis. **Sergio E. Uribe:** Writing – review & editing, Validation, Formal analysis. **Tedy Karteva:** Writing – review & editing, Methodology, Formal analysis. **Sanaa Chala:** Writing – review & editing, Methodology, Formal analysis. **Antonin Tichy:** Participated in the e-consensus Analysed and interpreted the data Revised the manuscript and accepts to be responsible for it. **Falk Schwendicke:** Writing – original draft, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

FS is a cofounder of dentalXrai, a startup focusing on AI-based image analysis in dentistry. The present work was independent from this.

Appendix

Rationale for the choice of the Delphi technique

1. Justification: We employed an online Delphi, 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 to 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: Several 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 to 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.

References

[1] L.T. Arsiwala-Scheppach, A. Chaurasia, A. Müller, J. Krois, F. Schwendicke, Machine learning in dentistry: a scoping review, J. Clin. Med. 12 (3) (2023).

[2] F. Schwendicke, T. Singh, J.H. Lee, R. Gaudin, A. Chaurasia, T. Wiegand, et al., Artificial intelligence in dental research: checklist for authors, reviewers, readers, J. Dent. (2021) 103610.

[3] J.J. Kirkham, S. Gorst, D.G. Altman, J.M. Blazeby, M. Clarke, D. Devane, et al., Core outcome Set-STAndards for reporting: the COS-STAR statement, PLoS Med. 13 (10) (2016) e1002148.

[4] K. Dwan, D.G. Altman, M. Clarke, C. Gamble, J.P. Higgins, J.A. Sterne, et al., Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials, PLoS Med. 11 (6) (2014) e1001666.

[5] K. Dwan, C. Gamble, P.R. Williamson, J.J. Kirkham, Systematic review of the empirical evidence of study publication bias and outcome reporting bias—An updated review, PLoS One 8 (2013).

[6] J. Kirkham, K. Dwan, D. Altman, C. Gamble, S. Dodd, R. Smyth, et al., The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews, BMJ 340 (2010) c365.

[7] J.J. Kirkham, D.G. Altman, P.R. Williamson, Bias due to changes in specified outcomes during the systematic review process, PLoS One 5 (2010).

[8] J.J. Kirkham, K.M. Dwan, D.G. Altman, C. Gamble, S. Dodd, R. Smyth, et al., The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews, BMJ 340 (2010) c365.

[9] J.J. Kirkham, E. Gargon, M. Clarke, P.R Williamson, Can a core outcome set improve the quality of systematic reviews? – a survey of the Co-ordinating Editors of Cochrane review groups, Trials 14 (1) (2013) 21.

[10] P. Williamson, D. Altman, J. Blazeby, M. Clarke, D. Devane, E. Gargon, et al., Developing core outcome sets for clinical trials: issues to consider, Trials 13 (1) (2012) 132.

[11] J. Mongan, L. Moy, C.E. Kahn, Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers, Radiology: Artificial Intellig. 2 (2) (2020) e200029.

[12] K.G. Moons, D.G. Altman, J.B. Reitsma, J.P. Ioannidis, P. Macaskill, E. W. Steyerberg, et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration, Ann. Intern. Med. 162 (1) (2015) W1–73.

[13] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, et al., STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies, BMJ 351 (2015) h5527.

[14] 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, Palliat. Med. 31 (8) (2017) 684–706.

[15] A.S. Tejani, M.E. Klontzas, A.A. Gatti, J.T. Mongan, L. Moy, S.H. Park, et al., Checklist for artificial intelligence in medical imaging (CLAIM): 2024 Update, Radiol Artif Intell 6 (4) (2024) e240300.