EUROPEAN JOURNAL OF EMERGING DENTAL RESEARCH

An Audit of Publicly Accessible Dental Imaging Datasets for Artificial Intelligence Applications

Dr. Nazeera K. Halvani
Department of Oral Biology, Lirona Dental Institute of Health Sciences, Muscat, Oman

Dr. Emelio D. Travanik
Faculty of Restorative Dentistry, Norbelia University School of Dental Medicine, Valletta, Malta

V0LUME01 ISSUE01 (2024)
Published Date: 17 December 2024 // Page no.: - 28-33

ABSTRACT

The integration of artificial intelligence (AI) into dentistry holds the promise of revolutionizing diagnostics, treatment planning, and patient care. However, the development of robust, equitable, and unbiased AI models is critically dependent on the availability of large, diverse, and meticulously documented datasets. This article provides a comprehensive systematic review of the current state of publicly available dental image datasets intended for AI research. We conducted an extensive search across academic databases, data repositories, and AI challenge platforms to identify and evaluate existing datasets. The evaluation was based on the FAIR (Findable, Accessible, Interoperable, and Reusable) guiding principles, metadata completeness, and current best practices for responsible data documentation. Our findings reveal a significant scarcity of high-quality, large-scale dental imaging datasets, particularly when compared to other medical fields like radiology and ophthalmology. The 16 unique datasets identified are predominantly from a few countries, feature a limited number of imaging modalities, and focus heavily on tooth segmentation tasks. Crucially, many existing datasets lack standardized metadata, clear licensing, comprehensive documentation, and transparent reporting of ethical approval, which severely limits their utility and hampers the development of generalizable AI models (1, 21). Furthermore, issues of poor data sharing compliance and the high potential for inherent demographic and technical biases within these datasets present significant challenges to the field (3, 7). This review highlights the urgent need for a concerted, global effort within the dental community to create, curate, and share high-quality, ethically sourced, and openly accessible datasets. Establishing robust data infrastructure and mandating adherence to data documentation standards, such as data cards and the Croissant format, are essential steps to accelerate innovation and ensure the trustworthy and equitable development of AI in dentistry (2, 10, 12).

Keywords: artificial intelligence, dentistry, medical imaging, big data, FAIR principles, datasets, machine learning, data sharing, trustworthy AI, health equity.

INTRODUCTION

Artificial intelligence (AI) is poised to become an integral and transformative force in modern dentistry. Its potential applications span the entire clinical workflow, from enhancing diagnostic accuracy in radiographic interpretation to enabling personalized treatment planning and predicting therapy outcomes (17, 25). The successful development, validation, and clinical translation of these powerful AI tools, however, are fundamentally reliant on a single, critical resource: vast, diverse, and well-annotated datasets (11). While the concept of "data dentistry"—leveraging large-scale data analytics to reshape clinical care and research—is rapidly gaining traction (15), the field confronts a significant bottleneck: a pronounced and debilitating lack of publicly available, high-quality dental image datasets.

This "data-drought" in dentistry stands in stark contrast to other medical specialties. Fields like radiology and ophthalmology have witnessed a rapid proliferation of large-scale public datasets (e.g., ChestX-ray8, UK Biobank, numerous datasets for diabetic retinopathy), which has fueled a surge in AI innovation and regulatory approvals (8, 11). Dentistry, however, has lagged significantly behind (18). This scarcity creates a critical barrier that not only impedes research progress but also introduces a substantial risk of developing AI models that are brittle, non-generalizable, and inequitable (10). It is a well-established principle in machine learning that AI systems trained on limited, homogenous, or poorly documented data are prone to significant performance failures and algorithmic biases, which can perpetuate or even exacerbate existing healthcare disparities (3). The often-opaque, "black box" nature of complex deep learning models further complicates their clinical adoption, making trustworthy, transparent, and ethical development a paramount concern (10).

A foundational framework for addressing these challenges is provided by the FAIR Guiding Principles, which advocate for

scientific data to be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (22).

- **Findability** ensures that data can be discovered by the wider research community through persistent identifiers and indexed repositories.
- **Accessibility** dictates that the data can be retrieved through well-defined, open, and standardized protocols.
- **Interoperability** requires that data and metadata use common formats and vocabularies, allowing them to be combined and analyzed with other data sources.
- **Reusability** demands that datasets are richly described with clear provenance and have an explicit license, enabling their use in future studies.

Adherence to these principles is essential for maximizing the scientific and societal value of research data and for fostering a collaborative, innovative, and reproducible research ecosystem (5, 20). Unfortunately, studies within the broader biomedical sciences have shown that compliance with data sharing statements is often poor, links to data frequently become inaccessible, and the quality of available data is highly inconsistent (7, 20).

This systematic review, therefore, aims to provide the first comprehensive audit of the global landscape of publicly available dental image datasets for AI. We seek not only to identify and catalogue these resources but also to critically evaluate their characteristics, their geographic and demographic diversity, and their quality against the FAIR principles and modern standards for dataset documentation. By creating a centralized overview, we aim to highlight the key challenges and strategic opportunities for advancing data-driven research and building a foundation for trustworthy AI in dentistry.

2. Methods

This observational study was designed and conducted to systematically identify, characterize, and evaluate all publicly available dental image datasets relevant to AI research. The study protocol was registered with the Open Science Framework (OSF) and adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines.

2.1. Search Strategy and Data Sources

A comprehensive and systematic search was performed to identify relevant datasets. The search was not limited to traditional academic databases, as AI-related datasets are often shared across a variety of platforms. The search, conducted between September 2022 and January 2024, included the following sources:

- **Academic Databases:** PubMed, IEEE Xplore.
- **Preprint Servers:** arXiv, medRxiv.
- **General Data Repositories:** Zenodo, Mendeley Data, Figshare, Open Science Framework (OSF).
- **Coding and Collaboration Platforms:** GitHub.
- **Specialized AI/Data Science Platforms:** Kaggle, Google Dataset Search, Grand Challenge, OpenDataLab CN.

The search strategy employed a broad range of keywords and MeSH terms related to dentistry, imaging modalities, and AI. Search terms included combinations of ("dental" OR "oral" OR "tooth" OR "maxillofacial") AND ("dataset" OR "database" OR "images") AND ("radiograph" OR "panoramic" OR "CBCT" OR "intraoral scan" OR "cephalometric") AND ("artificial intelligence" OR "machine learning" OR "deep learning"). The search of PubMed was extended back to 2011 to ensure comprehensive coverage. The search was executed independently by six investigators, with a final review by the lead author to ensure consistency and completeness.

2.2. Inclusion and Exclusion Criteria

Datasets were included in the review if they met the following criteria:

1. Contained dental or maxillofacial imaging data (e.g., radiographs of any type, cone-beam computed tomography [CBCT], intraoral scans, clinical photographs).
2. Were publicly accessible for download, either directly or via a registration process.
3. Contained a minimum of 50 images or cases to ensure a baseline level of utility for machine learning experiments.
4. Were described in a scientific publication, preprint, or a dedicated website/repository page.

Datasets were excluded if they: (1) contained fewer than 50 images; (2) were focused on non-dental data; (3) comprised only text or numerical data without corresponding images; or (4) were only available "upon request." This final exclusion criterion was based on evidence showing that such offers to share data have an extremely low rate of author responsiveness, effectively rendering the data inaccessible (7). The screening process involved two independent reviewers for each potential dataset, with any disagreements resolved by a third senior investigator.

2.3. Data Extraction and Evaluation Framework

A standardized data extraction form was developed and refined through three training sessions with the review team to ensure consistency. For each included dataset, the following categories of information were extracted, with the dataset repository considered the definitive source in cases of

conflicting information:

- **General Characteristics:** Dataset name, year of publication, associated publication DOI, country of origin, data collection period, and primary research focus (e.g., segmentation, classification).
- **Imaging and Technical Details:** Imaging modality, equipment manufacturer/model, image format (e.g., DICOM, PNG, JPEG), image resolution, and any reported image processing or manipulation (e.g., cropping, normalization).
- **Ethical and Legal Information:** Presence and details of ethical approval, patient informed consent practices, and the dataset's license type (e.g., Creative Commons, MIT, or unspecified).
- **Dataset Content:** Total number of patients and images, patient inclusion/exclusion criteria, and the source of data acquisition (e.g., dental clinic, university hospital).
- **Annotation and Ground Truth:** Presence and type of annotations (e.g., bounding boxes, pixel-level segmentation, labels), description of the annotation software used, and the methodology for establishing the ground truth (e.g., expert consensus, majority vote).
- **Annotator Information:** Number of annotators, their experience level, whether they were calibrated, and the process for resolving disagreements between them.
- **Demographic Data:** Reported gender and ethnicity of the patient cohort.

2.4. Dataset Quality and FAIR Assessment

The quality and utility of each dataset were evaluated using a multi-faceted framework. The primary component was an assessment against the FAIR principles using the 41 FAIRSFAIR Data Object Assessment Metrics (v0.5), a standardized tool for quantitatively measuring FAIRness (5, 20). This involved scoring each dataset on metrics related to its findability, accessibility, interoperability, and reusability. Additionally, we assessed the completeness of the extracted metadata as a proxy for overall dataset quality and transparency, with a particular focus on elements critical for responsible AI development, such as ethical reporting, licensing, and ground-truth documentation (12).

2.5. Data Synthesis and Analysis

The unit of analysis was the individual dataset. Data were synthesized descriptively. We used R Software (v4.1.2) (13) to generate summary statistics, frequency distributions, and visualizations (e.g., bar charts, world maps) to provide a comprehensive overview of the landscape of available datasets, their characteristics, and their metadata completeness.

3. Results

The comprehensive search process initially identified 131,028 records across all platforms. After screening and removal of duplicates, 121 records underwent full review. From this, a final set of **16 unique publicly available dental imaging datasets** met all inclusion criteria. These datasets were hosted on a variety of platforms, with Kaggle (18.8%), GitHub (12.5%), Google Datasets (12.5%), Mendeley (12.5%), PubMed-linked repositories (12.5%), and Zenodo (12.5%) being the most common sources.

3.1. General Dataset Characteristics

The publication of these datasets is a recent trend, with a clear increase over time: one dataset was published in 2020, two in 2021, six in 2022, and seven in 2023. The majority of datasets (68.8%) were associated with a peer-reviewed scholarly publication.

Geographic Distribution: The datasets originated from 13 different countries, but the distribution was highly skewed. China was the largest contributor by image volume (2,413 images), followed by Switzerland (2,332 images), and a combination of France and Belgium (1,800 images). The United States also contributed a significant number of datasets. Notably, there was a complete absence of datasets from Oceania and only one from Africa, highlighting significant gaps in global representation.

Imaging Modalities and Research Focus: Panoramic radiography was the most prevalent imaging modality, appearing in 58.8% of the datasets. This was followed by CBCT (11.8%) and intraoral photographs (11.8%). Other modalities like cephalometric radiographs, periapical radiographs, and 3D intraoral scans were rare, each appearing in only one dataset (5.9%). The primary research tasks supported by these datasets were anatomical or lesion segmentation (62.5%) and tooth numbering/labeling (56.2%). Datasets for other important clinical tasks, such as caries classification or periodontal disease assessment, were less common. One dataset stood out for its multimodal approach, incorporating CBCT, panoramic, and intraoral images from the same patient cohort (9).

3.2. Metadata Completeness and Quality

The reporting of crucial metadata was highly inconsistent and often incomplete, revealing significant quality issues across the board.

- **Ethical and Legal Reporting:** This was a major area of weakness. Only **31.2%** of datasets reported having received ethical approval. Even more concerning, a mere **5.9%** explicitly stated that patient consent had been obtained. The legal basis for reuse was also ambiguous, as **56.3%** of datasets did not specify any license at all, leaving researchers in a state of uncertainty regarding usage rights.

- **Annotation and Ground Truth:** While 75% of the datasets contained some form of annotation, the process for creating this ground truth was often opaque. The method for establishing ground truth was explained in only 56.2% of cases, with methods varying between a single expert's decision, a majority vote of multiple experts, or being entirely undescribed. Information about the annotators themselves was sparse: only 53.8% of annotated datasets provided any information about the annotators, just 18.8% reported on annotator calibration, and only 16.7% described how disagreements between annotators were handled.
- **Patient and Technical Data:** Basic demographic information was frequently absent, with only 18.8% of datasets reporting patient sex distribution and none reporting on patient ethnicity. Technical details, such as the imaging equipment used, were reported in about half of the datasets. Anonymization strategies, which are critical for patient privacy, were only described in 43.8% of cases.

3.3. FAIR Principles Assessment

The evaluation of datasets against the FAIR principles showed varied but generally suboptimal performance. Intraoral radiograph datasets scored the highest overall, particularly on findability. In contrast, CBCT datasets, despite their clinical richness, scored the lowest across all FAIR categories. Panoramic radiograph datasets, the most common type, had FAIRness scores that ranged from initial to advanced. A significant limitation across most datasets was the lack of a globally unique and persistent identifier (like a DOI), which severely impacts findability. Reusability was consistently the lowest-scoring principle, primarily due to the widespread lack of clear licensing and detailed provenance documentation. While the FAIRness scores were generally higher than those reported in previous studies of general dental research data (20), they still fall short of the standards required for robust and reproducible AI research.

4. Discussion

This systematic review provides the first comprehensive audit of publicly available dental imaging datasets for AI, revealing a landscape that is both nascent and fraught with significant challenges. The findings underscore a critical infrastructure gap that impedes the progress of trustworthy AI in dentistry.

4.1. The Pervasive "Data-Drought" and Its Consequences

The identification of only 16 unique datasets, totaling just over 10,000 images, confirms a severe "data-drought" in dentistry. This figure is dwarfed by the resources in medical radiology, where datasets can contain millions of images from tens of thousands of patients (11). This disparity is a primary reason for the slower pace of AI innovation in dentistry compared to other fields. The consequences are far-reaching. Firstly, it stifles innovation and democratized research, concentrating progress within a few well-resourced institutions that hold large private datasets and creating high barriers to entry for other researchers. Secondly, and more critically, it elevates the risk of developing biased and non-generalizable AI models (3, 10). Models trained on small, geographically and demographically homogenous datasets are unlikely to perform reliably across diverse patient populations, different clinical settings, and varied imaging equipment. This can lead to AI tools that fail silently in certain populations, potentially widening existing health disparities (3).

4.2. The Crisis of Quality: Metadata, Licensing, and Trust

Beyond the sheer quantity of data, our findings highlight a profound crisis in data quality and documentation. The inconsistent and incomplete state of metadata is a major threat to the development of trustworthy AI.

- **Ethical Ambiguity:** The lack of transparent reporting on ethical approval and patient consent is deeply concerning. It raises questions about the ethical soundness of these foundational datasets and creates risks for researchers who use them.
- **Legal Uncertainty:** The absence of clear licensing in over half of the datasets places them in a legal gray area. This ambiguity around data reuse rights discourages their adoption and integration into larger studies, thereby limiting their value.
- **Untrustworthy Ground Truth:** The reliability of an AI model is fundamentally dependent on the quality of its ground truth. The opaque and inconsistent methods for annotation found in our review—with little information on annotator expertise, calibration, or consensus-building—introduce unknown levels of label noise and uncertainty (6, 19). This makes it impossible to reliably evaluate model performance or compare results across studies.

4.3. The Path Forward: A Call for a Global Data Ecosystem

Overcoming these challenges requires a concerted, multi-stakeholder effort to build a robust and ethical data ecosystem for dental AI. We propose a multi-pronged strategy:

1. **Establishment of a Centralized, FAIR Repository:** There is an urgent need for a centralized, searchable repository or federation of repositories dedicated to dental AI datasets. This platform should be built on the FAIR principles, ensuring that datasets are easy to find, access, and integrate.
2. **Mandating Standards for Documentation:** The dental AI community must move towards adopting standardized documentation practices. Initiatives like **Data Cards** (12),

which provide structured summaries of a dataset's composition, collection process, and intended uses, and the **Croissant format** (2), a machine-readable metadata format for datasets, should become the norm. These tools provide the transparency needed to assess a dataset's fitness for a specific purpose and to audit for potential biases.

3. **Incentivizing Data Sharing and Curation:** Journals, funding agencies, and academic institutions have a critical role to play. They should implement and enforce policies that mandate the sharing of data and code as a condition of publication or funding (16, 21). Furthermore, the significant effort required to curate and document a high-quality dataset must be recognized as a valuable scholarly contribution.

4. **Fostering Global Collaboration:** Addressing the severe geographic imbalance in data requires global collaboration. Initiatives like the Medical AI-ready Datasets Alliance (MAIDA) (14) and the WHO's Global Initiative on AI for Health (23) provide excellent models for how to establish frameworks for international data sharing that respect privacy and governance. The dental community should actively participate in and adapt these frameworks.

4.4. Limitations

This study, while comprehensive, has several limitations. Our search was restricted to publicly indexed sources and likely missed datasets held within private institutional collaborations. The landscape of data availability is dynamic, and some datasets may have become available or inaccessible since our search concluded. Finally, while we assessed metadata completeness as a proxy for quality, we did not perform an in-depth analysis of the internal quality or potential biases within each dataset's images or labels. Such an analysis is a critical area for future research.

5. Conclusion

The advancement of artificial intelligence in dentistry is being critically held back by a severe scarcity of large, diverse, and high-quality public imaging datasets. Our comprehensive review reveals a landscape characterized by a small number of datasets that are geographically skewed and frequently fail to meet foundational standards for metadata reporting, ethical transparency, and legal reusability. This "data-drought" not only slows the pace of innovation but also poses a serious risk of creating biased, inequitable, and untrustworthy AI tools. To unlock the immense potential of AI to improve oral health for all, the global dental community—including researchers, clinicians, academic institutions, industry partners, and scientific journals—must prioritize a collaborative and urgent effort to build, curate, and share large-scale, well-documented, and ethically sourced datasets. Adopting rigorous standards for data quality and transparency is not merely a technical prerequisite; it is an ethical imperative for building a future where AI in dentistry is both powerful and trustworthy.

References

1. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, McKay F, et al. 2023. The value of standards for health datasets in artificial intelligence-based applications. Nat Med. 29(11):2929–2938.

2. Benjelloun O, Simperl E, Marcenac P, Ruyssen P, Conforti C, Kuchnik M, van der Velde J, Oala L, Vogler S, Akthar M, et al. 2024. Croissant format specification. Croissant site; [accessed 2024 Mar 7]. https://mlcommons.github.io/croissant/docs/croissant-spec.html.

3. Celi LA, Cellini J, Charpignon M-L, Dee EC, Dernoncourt F, Eber R, Mitchell WG, Moukheiber L, Schirmer J, Situ J, et al. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities-a global review. PLOS Digit Health. 1(3):e0000022.

4. Chrimes D, Kim C. 2022. Review of publically available health big data sets. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, pp. 6625–6627.

5. Devaraju A, Huber R. 2021. An automated solution for measuring the progress toward FAIR research data. Patterns (N Y). 2(11):100370.

6. Dumitrache A, Inel O, Timmermans B, Ortiz C, Sips R-J, Aroyo L, Welty C. 2021. Empirical methodology for crowdsourcing ground truth. Semant Web. 12(3):403–421.

7. Gabelica M, Bojčić R, Puljak L. 2022. Many researchers were not compliant with their published data sharing statement: mixed-methods study. J Clin Epidemiol. 150:33–41.

8. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, Keane PA, Sebire NJ, Burton MJ, Denniston AK. 2021. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health. 3(1):e51–e66.

9. Liu J, Hao J, Lin H, Pan W, Yang J, Feng Y, Wang G, Li J, Jin Z, Zhao Z, et al. 2023. Deep learning-enabled 3D multimodal fusion of cone-beam CT and intraoral mesh scans for clinically applicable tooth-bone reconstruction. Patterns (N Y). 4(9):100825.

10. Ma J, Schneider L, Lapuschkin S, Achtibat R, Duchrau M, Krois J, Schwendicke F, Samek W. 2022. Towards trustworthy AI in dentistry. J Dent Res. 101(11):1263–

1268.

11. Mongan J, Halabi SS. 2023. On the centrality of data: data resources in radiologic artificial intelligence. Radiol Artif Intell. 5(5):e230231.

12. Pushkarna M, Zaldivar A, Kjartansson O. 2022. Data cards: purposeful and transparent dataset documentation for responsible AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22); June 21–24, 2022; Seoul, Republic of Korea. New York (NY): Association for Computing Machinery. p. 1776–1826.

13. R Core Team. 2021. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing [accessed 2024 Feb 26]. http://www.R-project.org/.

14. Saenz A, Chen E, Marklund H, Rajpurkar P. 2024. The MAIDA initiative: establishing a framework for global medical-imaging data sharing. Lancet Digit Health. 6(1):e6–e8.

15. Schwendicke F, Krois J. 2022. Data dentistry: how data are changing clinical care and research. J Dent Res. 101(1):21–29.

16. Schwendicke F, Marazita ML, Jakubovics NS, Krois J. 2022. Big data and complex data analytics: breaking peer review? J Dent Res. 101(4):369–370.

17. Schwendicke F, Samek W, Krois J. 2020. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 99(7):769–774.

18. Sengupta N, Sarode SC, Sarode GS, Ghone U. 2022. Scarcity of publicly available oral cancer image datasets for machine learning research. Oral Oncol. 126:105737.

19. Sylolypavan A, Sleeman D, Wu H, Sim M. 2023. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med. 6(1):26.

20. Uribe SE, Sofi-Mahmudi A, Raittio E, Maldupa I, Vilne B. 2022. Dental research data availability and quality according to the FAIR principles. J Dent Res. 101(11):1307–1313.

21. Venkatesh K, Santomartino SM, Sulam J, Yi PH. 2022. Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study. Radiol Artif Intell. 4(5):e220081.

22. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3:160018.

23. World Health Organization. 2023. Global Initiative on AI for Health [accessed 2024 Feb 27]. https://www.who.int/initiatives/global-initiative-on-ai-for-health