



Documentation completeness and demographic representativeness of public dental imaging datasets: a cross-sectional analysis of repository metadata

Integridad de la documentación y representatividad demográfica de los conjuntos de datos públicos de imagen odontológica: un análisis transversal de metadatos de repositorios

Diana Belkis Moreno Benítez¹

Roberto Antonio Vega Bonilla¹

Flor M Escalante Freitez²  

¹ Universidad Nacional de Chimborazo, Facultad de Ciencias de la Salud. Riobamba, Ecuador.

² Universidad de Carabobo, Venezuela

Submitted: 12-04-2026 | Revised: 08-06-2026 | Accepted: 16-08-2026 | Published: 20-08-2026

Corresponding Author: Flor M Escalante Freitez 

ABSTRACT

Introduction: Public dental imaging datasets are the fuel of dental artificial intelligence, but their documentation and their demographic composition are rarely described in the studies that use them. Methods: A cross-sectional sample of public dental imaging datasets was assembled from a public model repository, and each dataset's metadata was assessed for documentation completeness and for the presence of demographic information in its description. Results: Among 13 datasets, 5 (38.5%; 95% CI 17.7 to 64.5) carried a license and 9 (69.2%) carried a readme, while only 5 (38.5%) mentioned any demographic variable, and none mentioned sex or race or ethnicity. Conclusions: Public dental imaging datasets are documented incompletely and almost never describe the demographic composition of their subjects, which limits their reuse and obscures the representativeness of the models trained on them.

Keywords: datasets; dental imaging; documentation; demography; artificial intelligence.

RESUMEN

Introducción: Los conjuntos de datos públicos de imagen odontológica son el combustible de la inteligencia artificial odontológica, pero su documentación y su composición demográfica rara vez se describen en los estudios que los usan. Métodos: Se conformó una muestra transversal de conjuntos de datos públicos de imagen odontológica a partir de un repositorio público de modelos, y se evaluaron los metadatos de cada conjunto en cuanto a la integridad de la documentación y a la presencia de información demográfica en su descripción. Resultados: Entre 13 conjuntos de datos, 5 (38,5%; IC 95% 17,7 a 64,5) llevaban licencia y 9 (69,2%) llevaban archivo léame, mientras que solo 5 (38,5%) mencionaban alguna variable demográfica, y ninguno mencionaba el sexo ni la raza o etnia. Conclusiones: Los conjuntos de datos públicos de imagen odontológica están documentados de forma incompleta y casi nunca describen la composición demográfica de sus sujetos, lo que limita su reutilización y oscurece la representatividad de los modelos entrenados con ellos.

Palabras clave: conjuntos de datos; imagen odontológica; documentación; demografía; inteligencia artificial.

INTRODUCTION

Public dental imaging datasets are the raw material of dental artificial intelligence, because a model that detects caries, classifies panoramic pathology or segments teeth is trained on exactly such a dataset, and the model inherits whatever the dataset contains and conceals, so the dataset is the model's source of both its knowledge and its blind spots, and the documentation that accompanies the dataset is the only window through which a reader can see which of the two it is. The dental deep-learning literature has grown rapidly on these datasets, and studies of pediatric caries detection, panoramic object recognition and multimodal dental recognition have demonstrated that the models work, as the works of Rahimi et al.,⁽¹⁾ Cao et al.⁽²⁾ and Huang et al.⁽³⁾ document.

The datasets themselves are numerous and small rather than few and large, because dental imaging is produced in many clinics and shared through many repositories, and the result is a fragmented field of datasets whose individual documentation, rather than their collective size, is what determines whether they can be combined and reused. The fragmentation is the reason the documentation question is worth asking, because a field that shared one large, well-documented dataset would have answered the question once, whereas a field that shares many small datasets must answer it for each one, and the answer, as the present study shows, is frequently not available.

The present study is therefore a measurement of the field's documentation habit rather than of any single dataset, and it is offered as a baseline against which the habit's improvement can later be judged, so that the field can see, in a later repetition of the same measurement, whether its documentation practice has moved toward the standard that the FAIR and fairness literatures describe, and the repetition is made easy by the fact that the measurement is a short script against a public interface. What the studies rarely report is the documentation that accompanies the dataset and the demographic composition of the subjects whose images it contains, and the omission matters because a model trained on an undocumented or demographically narrow dataset cannot be judged for its reuse or for its fairness, and a dental artificial intelligence system that cannot be judged on those two grounds is a system whose evidence is incomplete even when its accuracy is high.

The construct examined here has two dimensions. The first is documentation completeness, which is the extent to which a dataset carries the metadata that a user needs to reuse it, and it decomposes into a license, which states the legal terms of reuse, a readme, which states what the data are, and a description, which states what the data contain.

The second is demographic representativeness, which is the extent to which a dataset describes the demographic composition of its subjects, and it decomposes into age, sex, race or ethnicity, and the source population. These two dimensions are the dataset-level counterpart of the principles that the FAIR data literature has articulated for scientific data generally, as the founding statement of Wilkinson et al.⁽⁴⁾ and the later elaborations of Götz et al.⁽⁵⁾ and van Swieten et al.⁽⁶⁾ describe, and they are the dimensions that a user would need to assess before trusting or reusing a dataset.

The FAIR principles are usually stated as findability, accessibility, interoperability and reusability, and the two dimensions examined here are the reusability part of that quartet, the license and the description being the properties that make a dataset reusable, while the demographic representativeness is the part that makes the reuse fair, and the two parts are therefore the bridge from the FAIR principles to the fairness principles that the artificial-intelligence ethics literature has added to them, a bridge that the dental imaging field has not yet crossed.

The two dimensions are distinct but complementary, because documentation states what a dataset is and where it may be used, while demographic representativeness states who is in it, and a dataset can be strong on the first and silent on the second. Both dimensions are measurable from the public metadata alone, which is the methodological point that makes the present study possible, because a dataset's license, readme, description and demographic mention are all properties that a repository exposes without the need to download the images themselves, and a metadata-level measurement is therefore a complete measurement of what a user would see before deciding to reuse the dataset.

The measurement is thus the measurement of the dataset's public face rather than of its private contents, and it is exactly the measurement that a prospective user, rather than a privileged insider, would be able to take, which is what makes it a fair test of the field's documentation rather than a test of anything the datasets hide.

The dimensions matter for two reasons that are specific to dental artificial intelligence. First, an undocumented dataset is a dataset that cannot be reused with confidence, because a user who cannot see the license or the description cannot know whether reuse is permitted or what the images show, and the licensing of FAIR data is the subject of Labastida et al..⁽⁷⁾

Second, a demographically undocumented dataset is a dataset whose fairness cannot be assessed, because a model trained on it inherits its demographic composition, and the bias and fairness literature in medical imaging, from the cardiac-imaging bias study of Gorijavolu et al.(8) to the dental-imaging bias review of Suganya et al.,(9) has shown that demographic composition is a property that a dataset either reports or conceals.

The two reasons converge on a single practical point, that a dental artificial intelligence model is only as trustworthy as the dataset it was trained on is knowable, and a dataset that is neither licensed nor demographically described is, for the purposes of trust, largely unknowable, which is a property that no amount of model accuracy can remedy, because the accuracy itself is measured on the same unknown data that the trust question is about.

The study therefore measures the knowability of the dental imaging datasets, and it does so on the two axes, legal and demographic, on which knowability turns, and the two axes are carried through the methods, the results and the discussion in the same order, so that the reader can follow each dimension from its definition to its finding.

The point is the same one that the fairness literature makes in general terms, and it is applied here to the specific and under-examined case of the dental image, which is a case in which the demographic structure of the disease is well established and the demographic structure of the data is, as the results will show, almost entirely unrecorded, a gap between what is known and what is reported that the present study measures.

The objective of this study is to measure documentation completeness and demographic representativeness in a sample of public dental imaging datasets, so that the gap between what these datasets contain and what they report becomes measurable rather than anecdotal, and so that the gap can be stated with a number and a confidence interval rather than with an impression.

The objective is served by reporting, for each dimension, the count and proportion of datasets that meet the dimension's criterion, and the two dimensions are reported in the same order throughout, so that the reader can follow the same structure from the methods through the results to the discussion.

METHODS

Documentation completeness

Documentation completeness was defined operationally as the presence of three items in a dataset's repository record: a license, a readme file, and a non-empty description. The instrument was the dataset's public metadata, read from the repository's application programming interface, and each item was recorded as present or absent, with a license recorded as present when the metadata carried a license tag, a readme recorded as present when the dataset's file list carried a readme file, and a description recorded as present when the metadata carried a non-empty description text.

The analysis was descriptive, reporting the count and proportion of datasets carrying each item, and the proportion carrying a license was reported with a Wilson 95% confidence interval computed with the Statsmodels library, so that the precision of the estimate was visible alongside its value. The three items were treated as separate indicators rather than as a composite score, because a license and a readme answer different questions and a dataset that carries one does not carry the other, and the separate reporting preserves the distinction that a composite would collapse.

Demographic representativeness

Demographic representativeness was defined operationally as the mention, in a dataset's description, of any of four variables: age, sex, race or ethnicity, and the source population. The instrument was a deterministic keyword search of the description text, and a variable was recorded as mentioned when the description contained a keyword from that variable's list, with the age list including the words age, years old, adult, child and pediatric, the sex list including sex, gender, male and female, the race list including race, ethnic and ethnicity, and the population list including patient, cohort, population, institution and hospital.

The analysis was descriptive, reporting the count and proportion of datasets mentioning each variable, and the proportion mentioning any demographic variable was reported with a Wilson 95% confidence interval. The four variables were treated as separate indicators rather than as a single

composite, because age, sex, race and population answer different questions about a dataset's subjects, and a dataset that mentions one does not mention the others, as the results show.

Sample

The sample was assembled from a public model repository by searching for dental imaging terms, including dental radiograph, panoramic dental, dental X-ray, dental image, dental caries dataset, dental CBCT, dental radiography, dental panoramic, intraoral image and dental segmentation, and every dataset returned by any of the searches was retained, so the sample is the complete set of dental imaging datasets in that repository as of the retrieval date. No sampling or exclusion was applied, and the dataset identifiers, the licenses and the descriptions were read from the public metadata without modification, so the sample reflects the repository as it was rather than a curated selection.

The retrieval date is recorded, and a reader who repeats the search on a later date would obtain a possibly larger sample, which is the natural growth of the repository, and the reproducibility of the sample is therefore a matter of repeating the stated search terms against the repository at the stated date. The choice of a single repository was a deliberate simplification, and its consequence is stated in the limitations.

RESULTS

Documentation completeness

In the dimension of documentation completeness, the full sample of 13 datasets presented an incomplete profile (Table 1). Descriptively, 5 of the 13 datasets (0.4%) carried a license, 9 (0.7%) carried a readme, and 7 (7.0%) carried a non-empty description, so the three items were present in declining order and the license was the item provided in the fewest datasets. Inferentially, the license proportion of 0.4% carried a 95% confidence interval of 0.18-0.64, which was wide because the sample was small and which therefore locates the true license rate between roughly one in six and two in three datasets. In relation to the other items, the readme was present more often than the license, and the description more often than the license but less often than the readme, so the three items did not co-occur uniformly and a dataset that carried one item did not necessarily carry the others.

Table 1. Documentation completeness of the dental imaging datasets.

Documentation item	n	percent
License present	5	38.5
README present	9	69.2
Description present	7	53.8

Demographic representativeness

In the dimension of demographic representativeness, the same full sample of 13 datasets presented an almost complete absence of demographic information (Table 2). Descriptively, 5 of the 13 datasets mentioned age, 0 mentioned sex, 0 mentioned race or ethnicity, and 1 mentioned the source population, so the four variables were mentioned in descending order and sex and race were mentioned in none. Inferentially, 5 of the 13 datasets (0.4%) mentioned any demographic variable, and the proportion carried a 95% confidence interval of 0.18-0.64, which was wide for the same reason as before and which locates the true demographic-mention rate between roughly one in six and two in three. In relation to the individual variables, the only demographic information that any dataset reported was age, which appeared in the descriptions of the five documented datasets, and the two variables most relevant to fairness, sex and race or ethnicity, were absent from every description in the sample.

Relations between the dimensions

In relation to the two dimensions taken together, documentation and demography were associated in the direction that the FAIR framing would predict (Table 3). The datasets that mentioned a demographic variable were concentrated among those that carried a description, because a demographic mention is itself part of a description, and the five datasets that mentioned any demographic variable were all among the seven that carried a description, so the demographic information, where it existed, was carried entirely within the documented rather than the undocumented datasets.

Table 2. Demographic variables mentioned in the dataset descriptions.

Demographic variable	n	percent
Age mentioned	5	38.5
Sex mentioned	0	0.0
Race or ethnicity mentioned	0	0.0
Population or source mentioned	1	7.7
Any demographic variable	5	38.5

For this variable, the relation was therefore structural rather than incidental, because a dataset cannot mention a demographic variable without first having a description in which to mention it, and the absence of a description entailed, in every case, the absence of demographic information. Within this component, the per-dataset table makes the relation visible at the level of the individual dataset, and it shows that the three datasets that carried a license and a description and a readme were also the datasets that mentioned age, which is the pattern of a documentation practice that extends, in a minority of datasets, from the legal to the demographic.

Table 3. Documentation and demographic mention for each dataset.

Dataset	License	README	Description	Age	Sex	Race	Population
ismaelportog/Panoramic_Radiographs_for_Dental_Condition	no	yes	yes	no	no	no	no
darmasmrmz/radiograph_dental	yes	yes	yes	no	no	no	no
liodon-ai/dental-panoramic-xray-yolo	yes	yes	yes	yes	no	no	no
hetusuki/dental-panoramic-xray-yolo	yes	yes	yes	yes	no	no	no
DAnllili/dental-panoramic-xray-yolo	yes	yes	yes	yes	no	no	no
InfoBayAI/Xray-OPG-Dental-Radiology-Reports-Without-Findings-Dataset	no	yes	yes	yes	no	no	yes
sudhakark4227/dental-xray-dataset	yes	yes	no	no	no	no	no
sudhakark4227/dental-xray-segmentation	no	no	no	no	no	no	no
reza362/dental-xray-caries	no	no	no	no	no	no	no
sidiki07/dental-xray-segmentation	no	no	no	no	no	no	no
DeKenny/dental_image_dataset	no	yes	no	no	no	no	no
pangaribuan/dental-caries-dataset	no	no	no	no	no	no	no
ekacare/IntraOral_Gingivitis_Image_Captioning	no	yes	yes	yes	no	no	no

DISCUSSION

The objective of this study was to measure documentation completeness and demographic representativeness in public dental imaging datasets, and the measurement now shows that both dimensions are deficient, with the license missing from the majority of datasets and the demographic composition missing from almost all of them. The two deficiencies are related but distinct, because a dataset can be documented without being demographically described, and the present sample shows that even the documented datasets stop short of reporting who their subjects are.

In the dimension of documentation completeness, the finding that only 0.4% of datasets carried a license is the most consequential, because a dataset without a license is a dataset whose legal reuse is undefined, and the FAIR literature treats the license as the entry condition of reuse, as Labastida et al.(7) and Kim et al.(10) argue from the reuse side.

The license was also the least frequently provided of the three items, which is the more striking because a license is the easiest item to provide, being a single choice from a short list rather than a written text, and its absence is therefore the least excusable of the three gaps.

The ease of providing a license is a point that the repository itself could exploit, by making a license a required field at upload time, and the present finding is evidence that such a requirement, if it exists in principle, is not being met in practice, which is itself a finding about the repository's enforcement rather than only about the authors' diligence.

The finding is consistent with the broader observation of Bishop et al.(11) that data discoverability and metadata fitness are the barriers to reuse, with the repository-level analyses of Kerfant et al.(12) and Schwarzfuchs et al.,(13) which found that FAIR compliance is uneven even where the principles are endorsed, and with the convergence analysis of Papadiamantis et al.,(14) which located the documentation gap at the point where the FAIR principles meet artificial-intelligence ethics, so the present finding in dental imaging is a specific instance of a gap that the wider data literature has already identified and not yet closed.

The readme was the item provided in the most datasets, which is encouraging, but a readme without a license is documentation without legal standing, and the two items are not interchangeable, because a readme tells a user what the data are while a license tells them whether they may use the data, and the second is the one that a legal objection would invoke. The practical consequence is that a user who downloads a dental imaging dataset from this repository and wishes to reuse it in a model, a publication or a product cannot, in most cases, determine from the repository alone whether the reuse is permitted, and the burden of resolving that uncertainty falls on the user, which is the exact burden that a license is meant to remove.

The burden is not only an inconvenience, because an unresolved license question can prevent a dataset from being used at all, and a dataset that cannot be used is a dataset that cannot be built upon, so the missing license is a missing link in the chain of dental artificial intelligence rather than a cosmetic omission, and the chain is only as strong as its least documented link. The contrast with the dental dataset papers that do document their license, such as those of Hill et al.(15) and Lv et al.,(16) shows that the omission is not forced by the repository or by the field, and the difference between a documented and an undocumented dental dataset is therefore a difference in the care of its authors rather than in the medium that carries it, a difference that the present study has now made visible and countable.

The benchmark-construction studies of García-Gascó et al.(17) and Schür et al.(18) make the same point from the other direction, by treating the documentation as a deliverable of the benchmark rather than as an optional extra, and the present sample is the evidence of what happens when the documentation is treated as optional.

In the dimension of demographic representativeness, the finding that sex and race or ethnicity were mentioned in none of the datasets is the clearest result, and it means that the demographic composition of the subjects in public dental imaging datasets is, in this sample, entirely unreported.

The finding matters because a model trained on these datasets inherits their demographic composition, and the bias literature in medical imaging, from the cardiac-imaging echo-chamber study of Gorijavolu et al.(8) to the dental-imaging fairness review of Suganya et al.(9) and the fairness evaluations of Ganta et al.(19) and Meng et al.,(20) has established that an unreported demographic composition is the condition under which demographic bias goes undetected, and the acquisition-site and projection-bias studies of Dehkharghanian et al.(21) and Yayan et al.(22) show the same principle, that a dataset carries demographic and acquisition structure even when its documentation does not report it, so the structure is present whether or not it is visible.

The absence of sex and race from every description is therefore not a neutral fact but an active gap, and the mitigation literature, from Gu et al.(23) and Yata et al.(24) on bias mitigation to

the governance framework of Shaikh et al.,(25) presupposes the very demographic documentation that these datasets lack, so the gap between the datasets and the mitigation tools is a gap that the field has not yet closed, and the present study documents its current width.

The gap is all the more consequential because dental imaging is not demographically neutral, with caries and periodontal disease known to vary by age and by socioeconomic position, and a dataset that does not report age therefore cannot support the age-specific evaluation that the clinical question would demand, and a dataset that does not report sex or race cannot support the fairness evaluation that the deployment question would demand. The finding that age was the only demographic variable mentioned, and that it appeared in a minority of datasets, is a partial exception that proves the rule, because the datasets that did mention age were the documented minority, and the sex and race variables that fairness most needs were the variables that no dataset supplied.

The pattern suggests that the field's documentation habit, where it exists at all, is oriented toward the clinical descriptor of age rather than toward the fairness descriptors of sex and race, and the habit therefore needs to be extended in a specific direction rather than simply increased in amount. The specific direction is the addition of sex and race or ethnicity to the dataset description, because those are the two variables whose absence most directly blocks a fairness assessment, and they are the variables that the present sample shows to be universally missing, a universality that is the clearest single result of the study.

In relation to the two dimensions together, the reading is that the dental imaging dataset field has not yet adopted the documentation practices that the FAIR and fairness literatures recommend, and the consequence is that the dental artificial intelligence built on these datasets is built on a foundation whose license and whose demographic composition are, in most cases, unknown.

The dental dataset papers of Bahaa et al.,(26) Hill et al.(15) and Lv et al.(16) each describe a specific dataset with, in some cases, a documented license and a described population, which shows that the practice is possible, and the augmentation study of Khan et al.(27) and the multimodal study of Huang et al.(3) show that the datasets are used even where their documentation is thin.

The benchmark-construction literature, from Engdahl et al.(28) to Xiang et al.(29) and the chest-radiograph dataset of Pham et al.,(30) describes the standard that a documented dataset should meet, and the same standard is applied in the open-benchmark studies of García-Gascó et al.(17) and Schür et al.,(18) which treat the documentation as part of the benchmark rather than as an afterthought, and the present sample falls short of that standard.

The joint reading is therefore that the gap is one of practice rather than of possibility, because the standard exists and some dental datasets meet it, and the field's task is to extend the meeting to the datasets that do not yet reach it. The reading is not a call for more data but for better data about the data, and it is a reading that the present study can support because the measurement it reports is itself the kind of documentation that the field lacks, a measurement of the datasets' knowability that the datasets themselves do not provide.

The task is not technical, because attaching a license and stating the subjects' age, sex and race are editorial acts rather than computational ones, and the field's failure to perform them is therefore a failure of habit rather than a failure of capability, which is the kind of failure that a measured, documented gap is well placed to correct, because a habit can be changed by a standard that the field agrees to follow, and the standard is already written in the FAIR principles and in the datasheet proposals that the machine-learning literature has produced.

The correction is the more feasible because it can be made at the moment of publication, when the dataset's authors are still in possession of the information that the license and the demographic description would record, and a later correction would require reconstructing information that the authors alone had and may no longer have, which is why the recommendation is directed at the moment of release rather than at a later audit.

The timing point is a practical one for the field, because it means the documentation gap is best closed by changing the habit of authors at the moment they release a dataset, and a retrospective effort to document already-released datasets would be both harder and less likely to succeed, which is why the present measurement matters now rather than after the field has grown further.

The limitations of the study are stated before its implications. The sample is small, with 13 datasets from a single repository, so the proportions carry wide confidence intervals and the absence of a variable in this sample does not establish its absence in the field as a whole, and a larger sample across multiple repositories would be needed to confirm the proportions.

The assessment of demographic representativeness is based on the dataset description alone, so a dataset whose demographic composition is recorded elsewhere, in a linked paper or a separate file, would be scored as not mentioning it here, and the measurement therefore describes

what the repository metadata reports rather than what the dataset contains, which is the property that a user browsing the repository would actually see.

The keyword search is deterministic and may miss demographic mentions that use different vocabulary, and a manual review would be needed to bound that miss rate, although the absence of sex and race from every description is unlikely to be a keyword artefact. Each limitation bounds the claim, and none changes the central finding that the license and the demographic composition were, in this sample, mostly unreported.

The implications of the finding are concrete and they name the parties that can act on them. A researcher who publishes a dental imaging dataset should attach a license, because the license is the single item whose absence most limits reuse, and should state the age, sex and race or ethnicity of the subjects in the dataset's description, because those are the variables on which fairness is judged, and the two steps together would raise the dataset from unknowable to reusable.

A user who downloads a dataset should treat the absence of a license and of demographic information as a signal to proceed cautiously rather than as a reason to assume the worst, and should seek the missing information from the dataset's linked paper or its authors before training a model on it, because the missing information, once sought, may be found. A reviewer who evaluates a dental artificial intelligence study should ask what dataset the model was trained on and whether its license and its demographic composition were reported, because the model's evidence is incomplete without that answer, and the question is the cheapest available check on the reproducibility and the fairness of the model.

These implications are the practical yield of the measurement, and they follow directly from the two dimensions that the study set out to measure, so the measurement is not an end in itself but a means to the practices that it recommends, and the practices are the point at which the study's descriptive finding becomes an actionable one for the field.

CONCLUSIONS

Public dental imaging datasets are documented incompletely, because most of them lack a license, and they almost never describe the demographic composition of their subjects, because sex and race or ethnicity are entirely absent from every description in the sample, which limits their legal reuse and leaves the representativeness of the models trained on them impossible to assess.

REFERENCES

1. Rahimi H, Naeimi SM, Darvish S, Nazemi Salman B, Razzaghi P, Luchian I, et al. Detection of Pediatric Dental Caries in Panoramic Radiograph Using Deep Learning: A Benchmark Study on MD-OPG. *Sensors*. 2026;26(8):2481. doi: 10.3390/s26082481
2. Cao Z, Feng C, Shen Y, Ye G, Xu H, Wu J, et al. DPML: Prior-guided multitask learning for dental object recognition on limited panoramic radiograph dataset. *Expert Systems with Applications*. 2024;254:124446. doi: 10.1016/j.eswa.2024.124446
3. Huang Y, Liu W, Yao C, Miao X, Guan X, Lu X, et al. A multimodal dental dataset facilitating machine learning research and clinic services. *Sci Data*. 2024;11(1):1291. doi: 10.1038/s41597-024-04130-1
4. FAIR principles for data stewardship. *Nat Genet*. 2016;48(4):343-343. doi: 10.1038/ng.3544
5. Götz A. The fair principles: Trusting in fair data repositories. *Open Access Government*. 2023;39(1):262-263. doi: 10.56367/oag-039-10749
6. van Swieten MMH, Haselgrove C. Editorial: Navigating the landscape of FAIR data sharing and reuse: repositories, standards, and resources. *Front Neuroinform*. 2024;18:1387758. doi: 10.3389/fninf.2024.1387758
7. Labastida I, Margoni T. Licensing FAIR Data for Reuse. *Data Intellegence*. 2020;2(1-2):199-207. doi: 10.1162/dint_a_00042
8. Gorijavolu R, Jaiswal N, Pratap JS, Madapati K, del Pilar Arias López M, Leite M, et al. Echo Chambers: Bias and Representation in Cardiac Imaging Datasets for Artificial Intelligence. *Curr Cardiol Rep*. 2026;28(1):85. doi: 10.1007/s11886-026-02406-z
9. Suganya P, Dupada P, Sruthi KG, Panda P, Mohanty JR. Bias, fairness, and equity in artificial intelligence systems used in dental imaging: A systematic review. *International Journal of Medical Informatics*. 2026;214:106433. doi: 10.1016/j.ijmedinf.2026.106433
10. Kim Y, Lee JH, Shim W. The impact of perceived FAIR principles on data reuse intentions among biomedical scientists. *Journal of Librarianship and Information Science*. 2026;09610006261496883. doi: 10.1177/09610006261496883
11. Bishop BW, Hank C, Webster J, Howard R. Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. *Proceedings of the Association for Information Science and Technology*. 2019;56(1):21-31. doi: 10.1002/pr2.4
12. Kerfant C, Ruiz-Pérez J, García-Granero JJ, Lancelotti C, Madella M, Karoune E. A dataset for assessing phytolith data for implementation of the FAIR data principles. *Sci Data*. 2023;10(1):479. doi: 10.1038/s41597-023-02296-8
13. Schwarzfuchs EN. FAIR data, fair data: data management principles and data ownership in citizen science. *Information, Communication & Society*. 2026;29(11):3387-3403. doi: 10.1080/1369118x.2025.2604664
14. Papadiamantis AG, Lynch I, Melagraki G, Afantitis A. FAIR data principles and AI ethics: Exploring convergence and gaps. *FAIR Connect: Empowering Data Stewardship*. 2025;3:2949799X251403108. doi: 10.1177/2949799X251403108
15. Hill E, Albarella U. An international, open-access dataset of dental wear patterns and associated broad age classes in archaeological cattle mandibles. *Sci Data*. 2024;11(1):713. doi: 10.1038/s41597-024-03377-y
16. Lv H, Haq I, Du J, Ma J, Zhu B, Dang X, et al. A benchmark multimodal oro-dental dataset for large vision-language models. *Sci Data*. 2026;13(1):990. doi: 10.1038/s41597-026-07342-9
17. García-Gascón C, Bas-Bolufier J, Castelló-Pedrero P, García-Manrique JA. An open benchmark dataset for machine learning and intelligent trajectory optimization in fixed-wing unmanned aerial systems. *Sci Data*. 2026;13(1):364. doi: 10.1038/s41597-026-06716-3
18. Schür C, Gasser L, Perez-Cruz F, Schirmer K, Baity-Jesi M. A benchmark dataset for machine learning in ecotoxicology. *Sci Data*. 2023;10(1):718. doi: 10.1038/s41597-023-02612-2
19. Ganta T, Kia A, Parchure P, Wang MH, Besculides M, Mazumdar M, et al. Fairness in Predicting Cancer Mortality Across Racial Subgroups. *JAMA Netw Open*. 2024;7(7):e2421290. doi: 10.1001/jamanetworkopen.2024.21290
20. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep*. 2022;12(1):7166. doi: 10.1038/s41598-022-11012-2
21. Dehkharghanian T, Bidgoli AA, Riasatian A, Mazaheri P, Campbell CJV, Pantanowitz L, et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn Pathol*. 2023;18(1):67. doi: 10.1186/s13000-023-01355-3
22. Yayan J. Projection-Related Bias in the Detection of Thoracic Abnormalities: A Large-Scale Analysis of the NIH ChestX-Ray14 Dataset. *J Imaging*. 2026;12(5):187. doi: 10.3390/jimaging12050187
23. Gu T, Pan W, Yu J, Ji G, Meng X, Wang Y, et al. Mitigating bias in AI mortality predictions for minority populations: a transfer learning approach. *BMC Med Inform Decis Mak*. 2025;25(1):30. doi: 10.1186/s12911-025-02862-7
24. Yata VK, Jena S, Indracanti M, Chitta S, Kolliputi N. Adversarial debiasing for age-equitable diabetes prediction: performance-fairness trade-offs and partition dependency in machine learning. *Front Digit Health*. 2026;8:1816806. doi: 10.3389/fdgh.2026.1816806
25. Shaikh M, Ebrahimi A, Moradbeiki P, Wiil UK. FairGuard: Blockchain-Enforced Continuous Fairness Governance for Demographically Equitable LLM-Based Emergency Triage Decision Support. In: *Studies in Health Technology and Informatics*. IOS Press; 2026. p. 403-408. doi: 10.3233/shti260872
26. Bahaa S. A Histopathological Image Dataset for Oral Squamous Cell Carcinoma and Normal Tissue Analysis. *Melba*. 2026;2026(MICCAI Open Data 2026):758-763. doi: 10.59275/j.melba.2026-273d
27. Khan S, Mukati A. Dataset Augmentation for Machine Learning Applications of Dental Radiography. *IJACSA*. 2020;11(2). doi: 10.14569/ijacsa.2020.0110258

28. Engdahl I. Agreements 'in the wild': Standards and alignment in machine learning benchmark dataset construction. *Big Data & Society*. 2024;11(2):20539517241242457. doi: 10.1177/20539517241242457
29. Xiang A, Andrews JTA, Bourke RL, Thong W, LaChance JM, Georgievski T, et al. Fair human-centric image dataset for ethical AI benchmarking. *Nature*. 2025;648(8092):97-108. doi: 10.1038/s41586-025-09716-2
30. Pham HH, Nguyen NH, Tran TT, Nguyen TNM, Nguyen HQ. PediCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *Sci Data*. 2023;10(1):240. doi: 10.1038/s41597-023-02102-5

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Diana Belkis Moreno Benítez, Roberto Antonio Vega Bonilla, Flor M Escalante Freitez.

Writing–review and editing: Diana Belkis Moreno Benítez, Roberto Antonio Vega Bonilla, Flor M Escalante Freitez.