nature health

Analysis

https://doi.org/10.1038/s44360-026-00111-3

Underrepresentation of children in public medical imaging datasets

Received: 25 September 2025

Accepted: 17 March 2026

Published online: 24 April 2026

Check for updates

Stanley Bryan Zamora Hua[1,2,3], Nicholas Heller[4], Ping He[1], Alexander J. Towbin[5,6], Irene Y. Chen[2,7], Alex X. Lu[8,11] & Lauren Erdman[9,10,11]

Artificial intelligence (AI) has the potential to transform healthcare for all patients. Yet, there are disproportionately fewer paediatric AI studies and US Food and Drug Administration approvals relative to adults, indicating less effort focused on AI for children. Here, given that innovations in medical AI are accelerated by community-driven research on public datasets, we hypothesized that the disparity in AI for paediatrics is tied to the lack of public paediatric medical imaging data to support their development and evaluation. To that end, we systematically identified and reviewed 203 datasets, revealing that 33% of datasets lacked metadata on patient ages, and when available, children represented less than 2% of patients. To illustrate how a lack of paediatric data can lead to harmful algorithmic bias, we trained models to classify adult cardiomegaly and evaluated them on healthy children. We found a consistent pattern of age-related bias, reproducible across four large-scale public chest X-ray datasets. These findings suggest that the lack of public paediatric data hinders the development of safe AI for children, producing a landscape of adult-first and adult-only AI models with unknown patterns of bias in children.

Artificial intelligence (AI) is revolutionizing patient care with AI systems that can alert doctors when a patient's health will rapidly deteriorate[1,2], detect breast cancer from mammograms[3] and assist in the management of chronic diseases[4]. In children, AI can enable earlier detection and better characterization of disorders[5,6] during childhood, with earlier interventions promising better outcomes[7] and lesser disease burden later in life. While a few paediatric applications of AI have shown promising results[8], the development of AI for paediatric medicine remains limited, with paediatrics accounting for as little as 4% of AI-related studies on PubMed[9].

Two studies in 2024 investigated the approvals of AI-enabled medical devices by the US Food and Drug Administration. Overall, their findings suggest a lack of validation of age-related differences when using these AI systems. The first study analysed approvals from 1995 to 2023 (692 devices total) and found that 81.6% fail to specify age ranges for their intended user population[10]. Similarly, the second study looked at approvals from 1995 to March 2024 (876 devices total), and they identified that 73% omit patient age information in their validation studies[11]. Among the 149 devices authorized for paediatric use, only 28 (18.8%) are known to be validated on children, while 22 (14.8%) are validated only on adults[11]. The absence of rigorous paediatric testing of AI mirrors historical oversights in drug[12] and device development[13,14], where children have been harmed due to lack of age-specific validation and tailored interventions[15–17].

A growing concern is the presence of algorithmic biases embedded in deployed AI systems, with the potential to harm patients on the basis

[1]Center for Computational Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada. [2]Computational Precision Health, University of California, San Francisco, San Francisco, CA, USA. [3]Computational Precision Health, College of Data Science and Society, University of California, Berkeley, Berkeley, CA, USA. [4]Department of Urology, Cleveland Clinic, Cleveland, OH, USA. [5]Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [6]Department of Radiology, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [7]Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [8]Microsoft Research, Boston, MA, USA. [9]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [10]James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [11]These authors contributed equally: Alex X. Lu, Lauren Erdman. e-mail: lauren.erdman@cchmc.org

https://doi.org/10.1038/s44360-026-00111-3

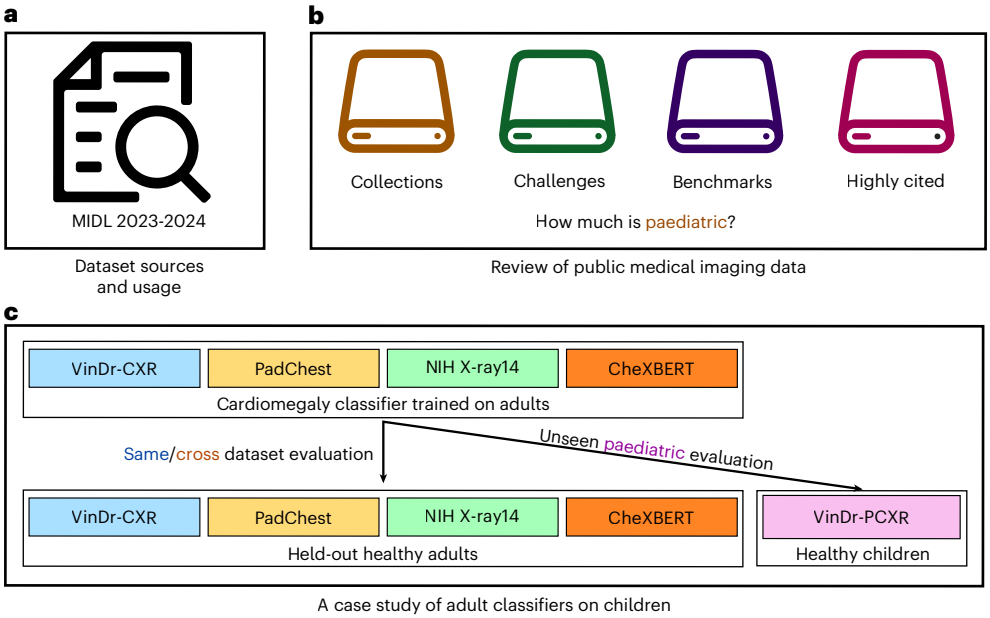

Fig. 1 | Overview of methodology. a, Analysing data use in recent AI for medical imaging papers. MIDL papers are reviewed to scope public dataset sources and usage and presence of paediatric applications of AI for medical imaging. **b**, Review of public medical imaging data. Public medical imaging datasets are taken from multiple sources and reviewed for paediatric representation. **c**, A case study of adult cardiomegaly classifiers on children. Cardiomegaly classifiers are trained on adult chest radiograph datasets and evaluated on held-out healthy adults and healthy children. Icons from OnlineWebFonts (**a**) and SVG (**b**) under a Creative Commons license CC BY 4.0.

of their sociodemographic identities[18]. Recent studies indicate that age is equally as important as sex and race in auditing for bias: AI trained on adult medical data can fail to generalize to children[19–22]. Underlying the challenging transfer of adult AI to children are fundamental differences in anatomy and physiology[23–26], in addition to dissimilar paediatric disease risk, presentation and management[27–34], which may vary substantially across different stages of development. Recent evaluations of AI systems for image interpretation support that models trained on adult data can perform worse in younger children than in older children[19,20,35,36]. Thus, the validation of AI for paediatrics necessitates rigorous testing on children at different ages and stages of development.

The public release of medical datasets of de-identified private health information advances the development of medical AI in many different ways. Benchmark datasets often focus on solving key tasks critical to medical applications (for example, multi-organ segmentation) by enabling the comparison of methods; for this reason, efforts often converge on a select few benchmarks[37,38]. Other datasets are structured as challenges, designed to attract generalist machine learning (ML) practitioners to work on novel problems[39]. Finally, while datasets are distributed with particular use cases in mind, researchers often reuse datasets for research needs beyond their original intent. Within the ML community, repurposing datasets is commonplace: datasets are often pooled together to create larger datasets or enriched with additional annotations for a particular task[40].

Despite the ecosystem of public datasets that help drive AI development, paediatric representation in these crucial resources remains largely unknown. A study in 2022 identified a gap in public paediatric X-ray datasets, hindering the development of AI for paediatric X-ray interpretation. Here, we investigate the paediatric data gap in a much wider scope, quantifying paediatric representation in public medical imaging datasets intended for AI development and evaluation. To understand how this data scarcity shapes downstream AI development, we examine how datasets are released, what age metadata they provide and how the ML community repurposes them. Lastly, we show that excluding children from public datasets can systematically lead to algorithmic bias, revealed only when models are evaluated on children stratified by age.

Results

MIDL studies frequently use public data with limited paediatric application

ML practitioners help drive advancements in AI for medical imaging and often publish these results in conferences such as *Medical Imaging with Deep Learning* (MIDL)[41], focused on cutting-edge deep learning methods for medical image analysis. To understand patterns of dataset use among ML practitioners, we surveyed papers accepted to MIDL in 2023 and 2024. Papers are filtered for those that use structural non-invasive internal imaging data, further defined in the dataset inclusion criteria. For each paper, we recorded the dataset(s) used, and then for all datasets, we identified how the dataset was made available. From this, we identified that datasets belong to either a dataset collection, ML challenge or ML benchmark, or serve as an independent dataset. More details are provided in 'Dataset selection' in the Methods.

We found that a majority (37/46; 80%) of these studies use public datasets in training or evaluating AI models. Among these studies, only one (2.2%) directly applied its work to paediatrics. In addition, 85% (39/46) of studies and 35% (22/62) of cited public datasets failed to report patient ages altogether—a fundamental metadata gap that obscures representational analysis. Only one of the 46 studies used a dataset containing primarily paediatric data, and while 9 additional studies used datasets containing paediatric data, less than 5% of the patients in those datasets were children. Based on these findings, we hypothesized that the limited paediatric applications may result from a lack of public paediatric datasets, prompting a broader review of the representation of paediatric patients in publicly available medical imaging data.

Children are underrepresented in public medical imaging data

To assess for paediatric representation among public medical imaging datasets, we conducted a systematic review of public datasets by collecting, annotating and analysing datasets from each data source (Fig. 1a,b). Our comprehensive search yielded a total of 203 datasets, of which 68 (33.5%) were excluded from further analysis: 2 datasets lacked patient counts, 60 datasets did not provide age information

https://doi.org/10.1038/s44360-026-00111-3

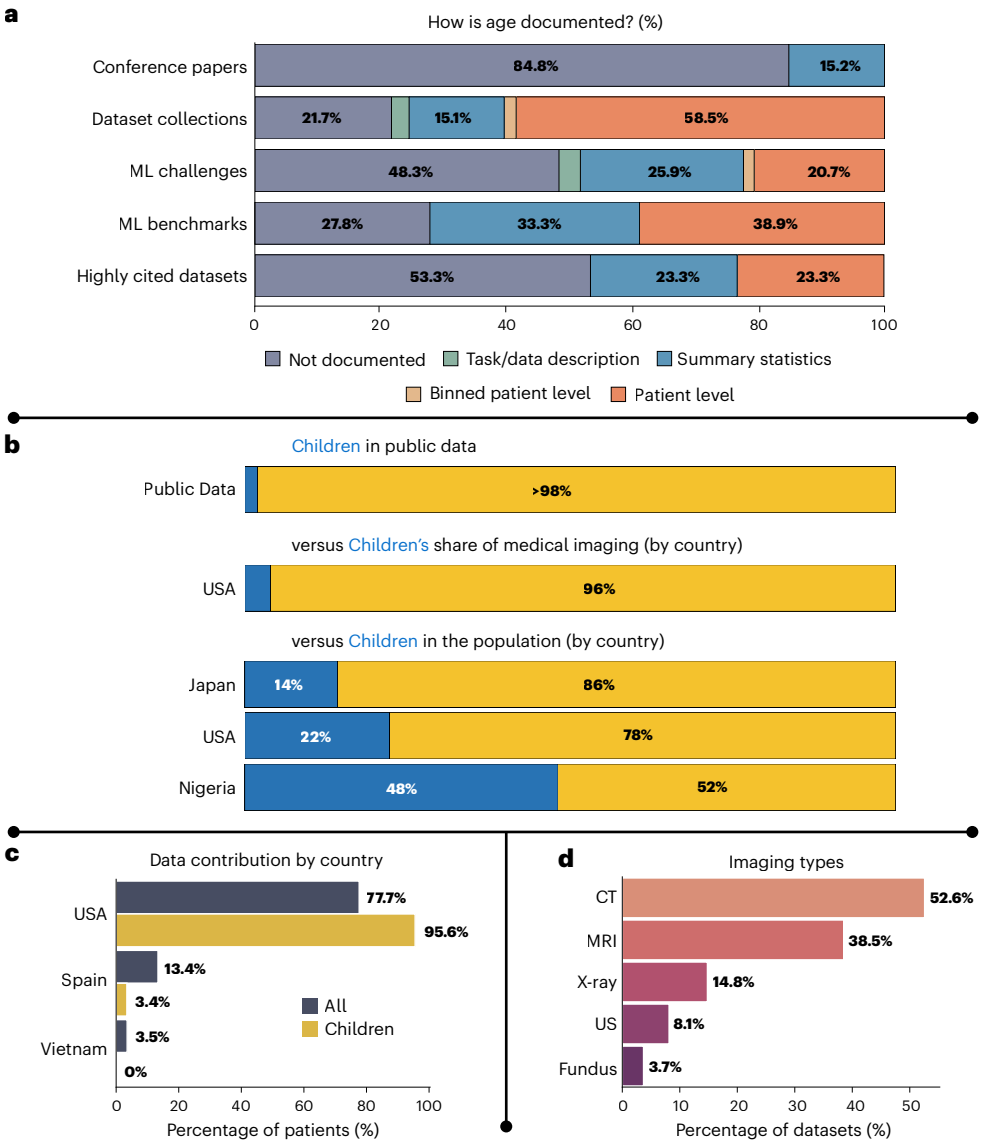

Fig. 2 | Patient age reporting and paediatric representation in public medical imaging datasets. a, Patient ages are rarely provided in public datasets and the papers that cite them. While not always complete, metadata for patient ages are provided in 58.5% of datasets from collections, 20.7% of challenges, 38.9% of benchmark datasets and 23.3% of highly cited datasets. **b**, Children make up less than 2% of patients in public medical imaging datasets. This is contrasted with the percentage of medical imaging examinations that were paediatric in 2016 performed in the USA[42], and population percentages provided by the United Nations[50] in three countries selected from our dataset review. **c**, The USA contributes the majority of patient data among datasets we reviewed: 95.6% of paediatric patients are from the USA. **d**, CT and MRI imaging are overrepresented among released datasets. US, ultrasound.

and 6 datasets were missing both patient count and age metadata. Among datasets that document patient ages, we found that age reporting is not well standardized: 66.7% provide patient-level ages, 28.5% report only summary statistics, and the remainder present age in bins or describe it implicitly within the task description (Fig. 2a). The missing or inconsistently reported patient counts and ages represent fundamental metadata gaps that hinder further analysis.

Analysis of the remaining 135 datasets reveals a paediatric data gap: children represent under 2% (6,026 of 512,608) of patients in public medical imaging datasets (Table 1 and Fig. 2b). Most public patient data were collected from the USA, totalling 76% of all patients and 96% of paediatric patients (Fig. 2c). Yet, public datasets substantially underrepresent paediatric imaging relative to clinical practice. By contrast, children accounted for 4% of all imaging examinations performed across seven healthcare systems in the USA[42]. Among the 135 datasets, 5 (3.7%) were explicitly paediatric-only, while an additional 29 (21.3%) contained paediatric data within largely adult datasets. Children represented between 1% and 2% of patients in The Medical Imaging and Data Resource Center (MIDRC)[43], The Cancer Imaging Archive (TCIA)[44] and Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI)[45] datasets, while in all other dataset sources, children represented less than 1% of patients. Even the established medical segmentation decathlon benchmark[46–48] had zero evidence of paediatric data. Paediatrics-focused ML challenges are rare, with only 2 among the 58 challenges between 2019 and 2024. Given the potential bias of adult overrepresentation in cancer-related datasets, we conducted a sensitivity test by excluding the 70 datasets with patients with cancer, and our findings remain consistent: children represent under 2% of patients (5,883 of 463,180).

Stratifying by modality and task reveal many applications for which virtually no paediatric data exist

Grouping adult and paediatric images across datasets, we found little to no paediatric representation in many modality-specific tasks relative

https://doi.org/10.1038/s44360-026-00111-3

Table 1 | Summary statistics of reviewed public medical imaging datasets, decomposed by data source

	Countries	Institutions	Datasets		Patients		
			All	Filtered	All	Children	%
Dataset collections	11	65	106	80	231,077	4,043	1.75
TCIA	11	61	69	51	42,145	478	1.14
Stanford AIMI	1	2	13	10	111,368	2,096	1.88
MIDRC	1	3	24	19	77,565	1,468	1.89
ML challenges	28	100	58	29	101,181	212	0.21
Grand Challenge	16	53	44	24	74,072	212	0.29
Kaggle	18	51	13	5	27,110	0	0.0
ML benchmarks	14	39	18	15	142,721	83	0.06
MSD	11	30	10	7	2,013	0	0.0
MedFAIR	4	10	7	7	140,108	76	0.05
AMOS	1	2	1	1	600	6	1.1
Total	34	181	203	135	512,608	6,026	1.18
Without cancer	28	105	100	66	463,180	5,883	1.27

Datasets are filtered for those where both the age and number of patients can be inferred. The number of patients was estimated from the 135 filtered datasets. MSD, Medical Segmentation Decathlon; AMOS, Abdominal Multi-Organ Segmentation.

Table 2 | Stratifying the number of medical images by imaging modality (rows) and ML tasks (columns) reveals paediatric data scarcity for many applications

	All		Condition classification		Anatomy segmentation		Lesion segmentation		Image enhancement	
	Adult	Peds	Adult	Peds	Adult	Peds	Adult	Peds	Adult	Peds
X-ray	959,482	7,520	891,720	6136	0	0	0	0	0	0
US	23,767	4,467	192	0	10,830	4,467	751	0	518	0
CT	289,581	914	262,069	124	4,677	373	7,092	309	760	40
MRI	52,390	176	18,905	5	1,903	17	22,174	134	1,283	4
Fundus	117,900	48	117,860	20	40	28	0	0	0	0

Peds, paediatric.

to adults (Table 2). When comparing by imaging modality, the data gap varies greatly with the smallest in ultrasound: an estimated 6 adult ultrasound images exist for every paediatric ultrasound image, while larger gaps are observed for X-ray (128:1), magnetic resonance imaging (MRI) (298:1), computed tomography (CT) (317:1) and fundus (2,457:1). Compared against true adult-to-paediatric imaging rates in the USA in 2016[42], a chi-squared test ($\chi^2(2, N = 371{,}295) = 17{,}368.83$, $P < 0.0001$) indicates that the observed paediatric-to-adult data representation differs from expected: MRI (11:1), ultrasound (15:1) and CT (26:1). The gaps in public adult-to-paediatric data for MRI and CT are more than ten times larger than expected given the amount of paediatric MRI and CTs collected in the USA. We observe the opposite with ultrasound: there are more paediatric ultrasound data relative to adults than expected. However, more than half of the paediatric data we analysed for ultrasound are contributed by one dataset, and this is true for X-ray and fundus as well. EchoNet-Peds[20] is the only dataset our search identified with paediatric ultrasound data. Similarly, 72% of paediatric X-ray data (5,404 images) and 57% of paediatric fundus data (28 images) come from the NIH X-ray14[49] dataset and CHASE DB 1[50], respectively.

Patient ages are often discarded and overlooked in data repackaging efforts

Efforts to increase data available for AI development often repackage existing public datasets by combining datasets or adding annotations on existing datasets. We identified 16 datasets from our search that repackage other datasets. While half of these included age metadata in the original datasets, patient ages were discarded in all of the resulting repackaged datasets. Likewise, we observe this in the development of MedSAM[51], a foundation model for segmenting medical images. To develop and evaluate MedSAM, the authors aggregated 25 CT and 25 MRI public datasets but did not disclose patient ages. Revisiting age information in these datasets, we found that children contributed 9 of the 6,567 CT scans and none of the 7,144 MRI scans. The inclusion of paediatric data appears to be a more recent and intentional development.

BraTS[52] is a well-known recurring competition for brain tumour segmentation, focusing on adults exclusively from 2012 to 2022. However, BraTS[22] 2023 was the first iteration to include paediatric gliomas and required a large multi-institutional data collection effort. While this suggests a shift towards recognizing the importance of paediatrics, newer paediatric datasets currently represent a small fraction over a legacy of adult-focused datasets. These observations suggest that data repackaging efforts, if not intentional about age, may lead to age-imbalanced datasets given the current data landscape.

Stratifying performance by paediatric age in years reveals systematic age biases in adult-trained AI

In the absence of tailored paediatric solutions, physicians often use medical devices and drugs designed for adults on children[15,16]. However, this carries the risk that the device or drug may be less effective or even harmful to children[17,53]. Similar risks exist for AI-enabled medical devices, as models may produce image interpretations that are systematically less accurate on children. Several studies have shown

https://doi.org/10.1038/s44360-026-00111-3

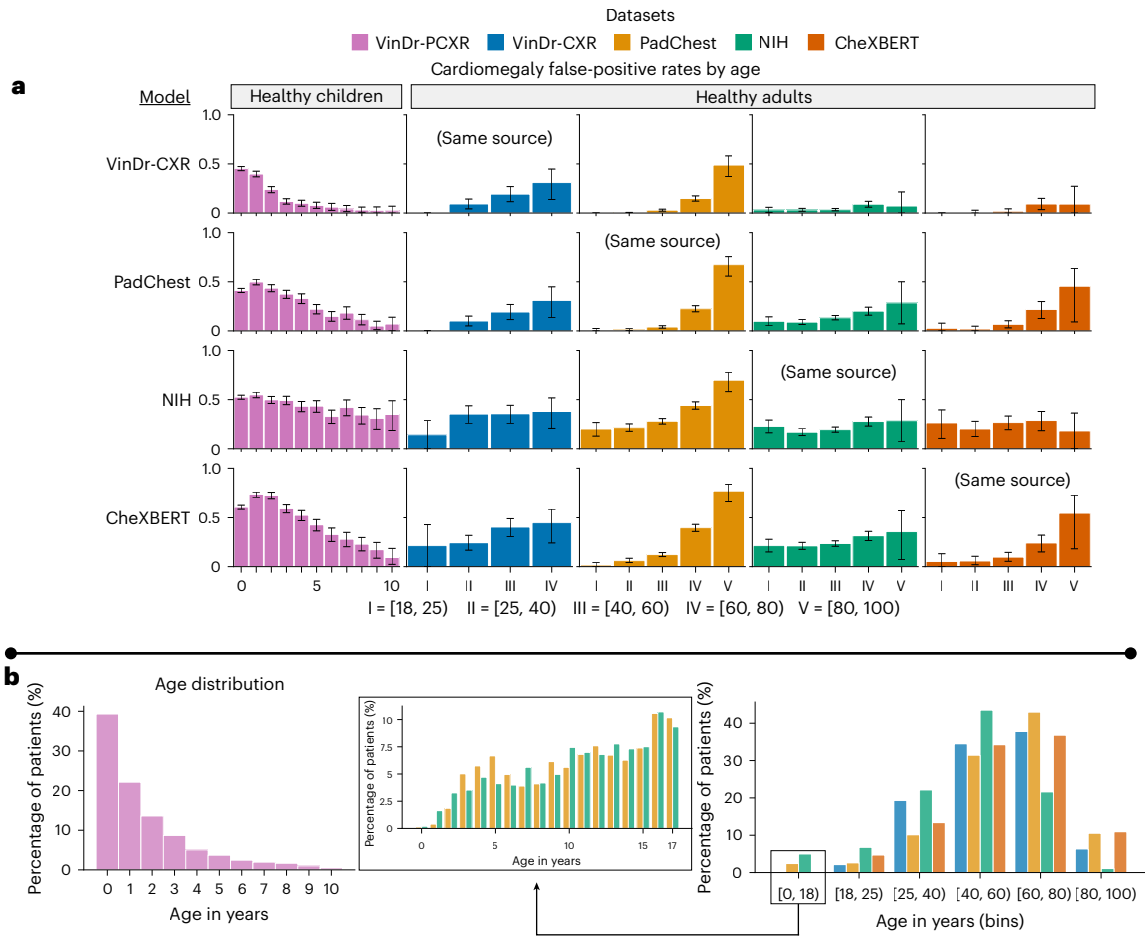

Fig. 3 | Testing for age-based algorithmic bias in models trained on public datasets collected from adult populations. a, Cardiomegaly classifiers trained on adult data increasingly mispredict cardiomegaly on younger children. Points show the false-positive rate computed from the full dataset. Error bars denote 95% confidence intervals obtained using bias-corrected and accelerated bootstrap resampling (12,000 replicates). **b**, Distribution of patient ages in VinDr-PCXR dataset (left) and in the adult-majority datasets (right). For adult-majority datasets with paediatric data (that is, PadChest and NIH), the paediatric age distribution is skewed towards older children with almost no infants at all (middle).

that AI models trained on adult chest radiographs and echocardiograms can perform worse on children[19–21,35,36]. In particular, one study evaluated a commercial AI tool trained to identify eight common findings in chest radiographs and found exceptionally poor performance among children 2 years old and younger[36]. In addition, the authors found that cardiomegaly prediction had a much higher false-positive rate relative to the other seven thoracic findings. For the remaining findings, the authors analysed patient ages for those with incorrect predictions. They found that 81.5% of the 238 misclassified patients were 2 years old and younger, skewed towards infants (<1 year: 45.8%, 1 year: 20.6%, 2 years: 15.1%, 3 years: 2.5%).

From these studies however, it is unclear whether these patterns of bias arise systematically from the lack of paediatric data in model development or are simply an artefact of the specific training set and methods. To understand this phenomenon, we trained AI models to detect cardiomegaly from adult chest radiographs, then evaluated them on healthy children (Fig. 1c). We repeated this process, training models on four adult public chest radiograph datasets and evaluated them on VinDr-PCXR[54], the only large-scale paediatric chest radiograph dataset. We chose cardiomegaly detection because it is a finding clearly affected by a well-established anatomical difference: infants are born with larger hearts relative to their chest[55,56], which is why clinical criteria for diagnosing cardiomegaly differ in this population. We evaluated false positives in healthy children to isolate bias arising from differences in heart presentation between healthy adults and paediatric patients. Nevertheless, differences in clinical determination and disease prevalence impact the utility and performance of AI models in practice and may contribute to additional sources of algorithmic bias that necessitates future study.

Adult cardiomegaly classifiers displayed the highest false positives among patients under 2 years old, and the rate of false positives decreased with age in years (Pearson $r = -0.928$, $P < 0.0001$). This pattern of age-related bias persisted across four public adult chest radiograph datasets (Fig. 3a). In comparison, cross-dataset evaluations of healthy adults showed no consistent patterns of bias. Revisiting paediatric data we extracted from two of the four primarily adult datasets, we observed that older children are more represented with almost no data from infants, despite being the most affected (Fig. 3b). Lastly, we tested whether the lower contrast in paediatric radiographs could explain age-related bias, but we found that histogram matching failed to rescue high false positives in younger children (Supplementary Fig. 2).

Discussion

Our study reveals a critical data gap: children account for less than 2% of patients in public medical imaging datasets supporting AI development and validation. Paediatric data are not only scarce but also fragmented, with each modality dominated by one to two datasets and some tasks showing virtually no paediatric representation. Paediatrics is the focus of only 1 of the 30 (3.3%) most cited medical imaging datasets, 2 of 58 (3.4%) medical imaging challenges between 2017 and 2024, and 1 of 46

(2.1%) MIDL studies in 2023–2024. Beyond paediatric representation, we identify the lack of standardized age reporting and lack of patient identifiers as critical metadata gaps that prevent age-stratified training and evaluation[57]. In addition, we find that current efforts to scale data for AI development through repackaging public datasets frequently omit age metadata, meaning that transparency about patient age representation may degrade as datasets are reused. Together, the absence of both public paediatric data and fine-grained age information limits the ability of researchers to develop novel methods to address paediatric challenges and audit existing models for age-related bias.

The gaps in public paediatric data appear to reflect more than simply the relative amounts of adult to paediatric imaging done in practice. Data from the surveyed public datasets contain fewer paediatric CT and MRI scans than adult-to-paediatric imaging rates in the USA would predict[52], yet contain more paediatric ultrasound data than expected. Data initiatives that prioritize ultrasound as a safer modality for children may partly explain the surplus of paediatric ultrasound data[58,59]. Yet, MRI—a radiation-free modality like ultrasound—shows a larger gap in paediatric data than X-ray, suggesting that imaging safety alone cannot account for how paediatric data are chosen for release. Barriers to paediatric data curation and release are likely contributors. Many such challenges are unique to paediatrics: greater standards for consent, safety and efficacy[49,60–62], all while facing smaller sample sizes, poor patient recruitment and high development costs[62,63]. Although changes in public policy have accelerated paediatric drug and device development, similar incentives have yet to reach medical AI to support paediatric data collection[64–66]. An encouraging sign is the growing number of initiatives that promote paediatric AI equity[9,67–69].

A key challenge in addressing the paediatric data gap is establishing a clear end goal. While increasing paediatric representation is a necessary step, the true end goal is to create safe and effective AI tools for paediatrics. As AI integrates into medical devices, clinicians may unknowingly use AI models trained only on adults. The literature has provided several instances where adult AI models underperforms on children, particularly younger children[19–22,70]. However, these studies do not demonstrate that the source of bias comes from the lack of paediatric data during training. To strengthen the link between paediatric data during training and its impacts on age-related bias, we show how an age-associated bias emerges consistently across models developed from four adult public radiograph datasets, highlighting the necessity of paediatric data among public datasets for evaluation. As our findings suggest, developing safe AI for children requires representation at different ages, and this is simply not available in the vast majority of public medical imaging datasets at present.

A solution to the paediatric data gap will not arise naturally. We urge the broader ML for health community to take action. For healthcare institutions and researchers, we encourage greater collaboration and initiatives to collect, prepare and release AI-ready paediatric data to the public, to support the development of paediatric AI applications. We emphasize the necessity of age representation from infancy to adulthood and the reporting of ages at the patient level. Greater age representation and transparency will help to identify age ranges where AI models are most effective, leading to improved safety beyond paediatric cases. Building on public paediatric data releases, we encourage the creation of challenges and benchmarks to accelerate paediatric AI innovations leveraging community-driven development. We urge policymakers to design policies that encourage paediatric data collection and sharing. Lastly, we argue for greater scrutiny in the evaluation of AI applications for paediatrics, focusing on representation at different stages of childhood and importantly in narrowly defining appropriate age ranges for intended users.

We acknowledge three limitations of our study. First, there may exist paediatric datasets that were not captured in our systematic dataset search, particularly in data platforms without centralized curation such as Zenodo and Harvard Dataverse, or in ML challenges on Grand Challenge that did not grant us timely access to the datasets. Our analysis, however, captures key pathways in which datasets may be found by ML practitioners. Another contributing factor is stricter access controls for paediatric datasets, which ultimately led to the exclusion of most paediatric cancer datasets in TCIA. Second, poor age reporting led to the exclusion of many datasets in our analysis. Although paediatric data may exist among the excluded datasets, our findings suggest they may represent only a small portion. Lastly, although our case study identifies one pattern of age-related bias, we do not expect all adult AI models to share the exact same pattern of bias. Rather, the case study motivates the need to evaluate paediatric AI on children in general and stratified by age.

Despite forming a significant portion of the population, children are almost invisible in the datasets accelerating advancements in AI for medical imaging, constituting less than 2% of public medical imaging datasets. The underrepresentation of paediatric patients and the poor reporting of patient ages in public medical datasets are critical oversights that limit the development of safe AI for paediatric use. This work not only quantifies the extent of this problem but demonstrates the potential of systematic failure on children when using AI models trained only on adults. Lastly, we provide recommendations to encourage the community to take action in addressing the paediatric data gap.

MIDRC
The imaging and associated clinical data downloaded from MIDRC and used for research in this publication were made possible by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under contract 75N92020D00021 and through The Advanced Research Projects Agency for Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Stanford AIMI
This research used data provided by the Stanford AIMI. AIMI curated a publicly available imaging data repository containing clinical imaging and data from Stanford Health Care, the Stanford Children's Hospital, the University Healthcare Alliance and Packard Children's Health Alliance clinics provisioned for research use by the Stanford Medicine Research Data Repository (STARR).

Methods
Dataset inclusion criteria
To be eligible for our systematic review, a dataset must meet the following conditions:

(1) Structural non-invasive internal imaging. The dataset must contain data for medical imaging modalities that capture internal anatomical structures of organs and tissues and do not require breaking the skin or physically entering the body. This strictly includes fundus imaging, ultrasound, MRI, X-ray and CT. To avoid biasing our analyses, we exclude adult-only modalities such as mammography, and because little to no public datasets were found exclusively with fluoroscopy and nuclear imaging, these modalities were additionally excluded. Finally, to support analyses relating paediatric representation in public data to paediatric imaging in practice, we exclude datasets obtained outside of clinical settings.
(2) Accessibility. The dataset must be publicly discoverable and downloadable from the internet with no access restrictions based on the country or institution. If associated only with a publication, the dataset must be made available via an online platform/portal.

To avoid redundancy in the following analyses, we excluded datasets that simply repurpose datasets already annotated. For datasets

that are updated annually such as the Brain Tumor Segmentation challenge (BraTS)[52], only the latest versions (as of 1 January 2025) containing complete and original data are included. If a dataset contains both data that fall within the inclusion criteria and data that fall outside the inclusion criteria, only the subset of data that fall within the inclusion criteria is annotated and included in subsequent analyses.

Dataset annotation

For datasets meeting the inclusion criteria, we attempted to extract the number of patients within each dataset and their ages. To determine the number of patients within each dataset, we used metadata files with patient identifiers as ground truth. When this was not available, we estimated the number of patients in CT, MRI and fundus, assuming that one CT scan, one MRI scan or two fundus images corresponded to one unique patient. For our analyses, we assume that patients are independent across datasets, and while this assumption may not hold, it is impossible to validate given independent anonymization across datasets. Next, we categorized how patient ages are reported and attempted to identify the proportion of paediatric patients: patients aged 17 or younger in each dataset. We identified the following patient age reporting strategies: patient-level, binned patient-level, summary statistics, task/data description and not documented. In the best case, metadata files include the ages of each study participant, allowing for the exact proportion of paediatric patients in a dataset to be determined. In binned patient-level reporting, patient ages are provided in age bins (for example, 20–30 years old). If we could not infer the presence and proportion of paediatric data through available metadata files, we scanned associated publications for summary statistics related to patient ages. If none of the above methods was successful, we attempted to determine if there was no paediatric presence by scanning task or data descriptions for any clear mention or indicators of 'adult-only' data in the associated publication, dataset's official website or portal. Otherwise, a dataset was labelled as not documenting age.

Dataset selection

To curate datasets for our systematic review, we collected and organized them by their source.

Dataset collections. Data collections refer to repositories curated for diverse research and practical applications, including medical imaging. For this study, we focused on clinical data repositories containing medical imaging datasets, either derived from clinical studies or explicitly released for developing ML algorithms. Our analysis included the following dataset collections: (a) TCIA[44], (b) Stanford AIMI[71] and (c) MIDRC[43]. For TCIA and AIMI, we focused on complete datasets that have patient age annotations. For MIDRC, we extracted metadata for 84,016 cases from the explorer (portal v5.39.2, submission v2025.09), then grouped cases by the source dataset and by imaging modality.

Challenges. Challenges play a pivotal role in advancing state-of-the-art ML, often targeting novel or domain-specific applications. ML challenges for medical imaging are commonly hosted on platforms such as the Grand Challenge[71] and Kaggle[72], often in conjunction with conferences. For this study, we annotated datasets from challenges hosted on the listed platforms, launched in the past 5 years (2019–2024).

Benchmarks. Benchmarks consist of one or more pre-existing or novel datasets designed to evaluate and compare algorithm performance on specific and often well-established tasks. While numerous datasets can serve as benchmarks, we focused on three explicitly labelled as such: two benchmarks for multi-organ segmentation: Medical Segmentation Decathlon (MSD)[46] and Abdominal Multi-Organ Segmentation (AMOS)[73], and one benchmark for assessing fairness in medical imaging ML: MedFAIR[74].

Highly cited datasets. Highly cited datasets serve as a means to identify datasets that do not belong to a distinct dataset collection, challenge or benchmark, yet have become foundational resources in the academic community, serving as the basis for numerous significant contributions over time. To identify these datasets, we leveraged Meta's PapersWithCode[75], which ranks imaging datasets by citation count. We selected and annotated the top 30 datasets meeting our inclusion criteria. When aggregating datasets across dataset sources, we remove any duplicate datasets found via PapersWithCode and the other dataset sources. In the analysis restricted to highly cited datasets, all 30 identified datasets were included.

Data breakdown by modality and task. To investigate whether paediatric representation differs by application, we first filtered for datasets where the proportion of paediatric data can be estimated, then categorized datasets by imaging modalities and ML tasks. Specifically, ML tasks are classified under: (a) condition classification (for example, pneumonia classification), (b) anatomy segmentation (for example, organ or tissue segmentation), (c) lesion segmentation (for example, tumour segmentation), (d) disease risk segmentation (for example, fractured rib segmentation and organs-at-risk segmentation) and (e) image enhancement (for example, MRI image reconstruction and synthetic image generation). To perform our analysis, we aggregated the number of medical images across datasets by imaging modality and task. As any one dataset can contain data from multiple imaging modalities, we attempted to identify what proportion of the dataset constitutes each modality, and if not possible, we assumed the images are split evenly across modalities. To assess whether modality-specific data gaps reflect real-world usage, we performed a chi-squared test on the observed counts of paediatric and adult MRI, CT and ultrasound scans, compared with the expected counts based on imaging usage in 2016 from hospitals in the USA[42].

Dataset repackaging. In our review, we observed cases in which a released dataset includes data previously released in a pre-existing public dataset. We term this phenomenon 'dataset repackaging'. We identified all datasets that perform dataset repackaging. On those datasets, we annotated the original datasets they reference and extracted age-related metadata from the original dataset. As many foundation models are often developed on public datasets, we supplemented our analysis by considering patient ages in the CT and MRI datasets used by MedSAM, a medical foundation model.

A case study of paediatric algorithmic bias for cardiomegaly detection

Data. For our experiments, we trained four cardiomegaly classifiers on adult chest radiograph images from large publicly available chest X-ray datasets: (a) VinDr-CXR[76], (b) NIH X-ray14[77], (c) PadChest[78] and (d) CheXBERT[79,80]. Each classifier is evaluated on healthy children with age annotations in the VinDr-PCXR[54] dataset, a dataset of posteroanterior-view chest X-rays from children aged 0–10 years imaged at a major paediatric hospital in Vietnam. The remaining chest X-ray datasets consist primarily of adult data and are used for training, calibration and held-out evaluation on healthy adults. The VinDr-CXR data are from two hospitals in Hanoi, Vietnam, while PadChest data are from a hospital in Barcelona, Spain. Meanwhile, the NIH X-ray14 and CheXBERT are from hospitals in the USA associated with NIH and Stanford, respectively. NIH X-ray14 is the extended version of the original NIH ChestX-ray8 dataset, while CheXBERT is the CheXpert dataset with CheXBERT-assigned labels.

Dataset standardization. To standardize the adult datasets, the following steps are taken:

(1) Exclude paediatric patients aged 17 and under, if any.
(2) Exclude data points without findings annotated or with particularly invalid age annotation.

(3) Keep only posteroanterior-view images.
(4) Sample one image per patient, for patients with data from multiple visits, in priority of: (a) having valid age annotated, and (b) having cardiomegaly.
(5) Exclude patients with findings other than cardiomegaly and patients with uncertain cardiomegaly findings.

Each of the four adult-only datasets are then split independently into a training set, calibration set and held-out set of healthy adults. To create the held-out set, 10% of healthy adults are sampled from each age bin in [18, 25, 40, 60, 80, 100], except for VinDr-CXR, where 20% were sampled due to a smaller number of adults with valid age annotations. The calibration set is curated from 10% of the chest X-rays sampled from patients with cardiomegaly, patients without any findings and patients with a finding other than cardiomegaly. Next, data from patients with findings other than cardiomegaly are removed. Of the remaining data, 75% was used for training, and the remaining 25% for validation (Supplementary Table 1). All images are loaded in RGB channels and resized to (224, 224). For each of the four adult datasets, we computed dataset-specific image normalization parameters (mean, standard deviation) beforehand. These parameters are then used by the respective models to standardize all images for both training and inference.

Model. A ConvNeXt-B (89 M parameters) is chosen as our base neural network architecture[81]. For each of the datasets, a network is trained for at most 50 epochs using the AdamW[82] optimizer with a learning rate of 0.0001 to optimize the binary cross-entropy loss. During each training step, 32 random X-ray images are sampled equally from both classes, and MixUp is used for data augmentation[83]. Parameters from the epoch with the best validation set loss are kept.

Evaluation. The probability threshold (or operating point) for each classification network is adjusted to maximize Youden's *J* statistic (defined as sensitivity + specificity − 1) on their corresponding calibration set. Each adjusted model is then used to predict cardiomegaly on the healthy children in the VinDr-PCXR dataset and healthy adults in each of the adult datasets. The proportion of false positives is reported at the dataset level and stratified by age groups. To determine statistically significant differences, bias-corrected and accelerated bootstrap is used to construct 95% confidence intervals around the false-positive rate[84]. To identify patterns between age in years and false-positive rate, we normalized the false-positive rates by each model and computed a Pearson *r* correlation.

Perturbation analysis. To minimize harmful exposure when acquiring chest radiographs, children typically receive lower radiation dosages compared with adults, resulting in images with possibly lower contrast[59,85]. Neural networks may latch onto this low-level difference in paediatric images. In addition to our evaluation, we explored how accounting for these differences between datasets can impact the bias on paediatric patients. Histogram matching is a technique that allows us to adjust the pixel distribution of one image to match the pixel distribution of one or more images. We utilized histogram matching to adjust the pixel distributions of imaged healthy children (aged 0–1 year old) in the VinDr-PCXR dataset to match the pixel distributions of images in each of the adult-centric datasets, separately. We varied the strength of the histogram matching across different blending ratios. In addition, we plotted pixel histograms for each dataset using random samples of 1,000 images (Supplementary Fig. 1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The annotations of public datasets produced in this study are provided in Dataset 1 (Source Data file 1). To aid users in identifying paediatric data from our dataset search, we also provide additional pointers to these datasets in Dataset 2 (Source Data file 2) disaggregated by imaging modality. These supplementary files are available via GitHub at https://github.com/stan-hua/ped_vs_adults-cxr and via Zenodo at https://doi.org/10.5281/zenodo.18895394 (ref. 86). These annotations may be shared without restriction. The datasets used in this cardiomegaly case study are publicly available and subject to their data use agreements. The MIMIC-CXR, VinDr-CXR and VinDr-PCXR datasets were downloaded from PhysioNet[87]. Meanwhile, PadChest, NIH and CheXBERT were downloaded from their respective sources.

Code availability

Code to reproduce the dataset analyses and cardiomegaly study is available via GitHub at https://github.com/stan-hua/ped_vs_adults-cxr. There, we provide detailed instructions to prepare the chest X-ray datasets for our analyses.

References

1. Verma, A. A. et al. Clinical evaluation of a machine learning-based early warning system for patient deterioration. *CMAJ* **196**, E1027–E1037 (2024).
2. Gallo, R. J. et al. Effectiveness of an artificial intelligence-enabled intervention for detecting clinical deterioration. *JAMA Intern. Med.* **184**, 557–562 (2024).
3. Yala, A. et al. Multi-institutional validation of a mammography-based breast cancer risk model. *J. Clin. Oncol.* **40**, 1732–1740 (2022).
4. Mosquera-Lopez, C. et al. Enabling fully automated insulin delivery through meal detection and size estimation using artificial intelligence. *npj Digit. Med.* **6**, 39 (2023).
5. Airaksinen, M. et al. Assessing infant gross motor performance with an at-home wearable. *Pediatrics* **155**, e2024068647 (2025).
6. Chang, Z. et al. Computational methods to measure patterns of gaze in toddlers with autism spectrum disorder. *JAMA Pediatr.* **175**, 827–836 (2021).
7. Guthrie, W. et al. The earlier the better: an RCT of treatment timing effects for toddlers on the autism spectrum. *Autism* **27**, 13623613231159153 (2023).
8. Ganatra, H. A. Machine learning in pediatric healthcare: current trends, challenges, and future directions. *J. Clin. Med.* **14**, 807 (2025).
9. Richter, F. et al. Toward governance of artificial intelligence in pediatric healthcare. *npj Digit. Med.* **8**, 636 (2025).
10. Muralidharan, V. et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *npj Digit. Med.* **7**, 273 (2024).
11. Brewster, R. C. L., Nagy, M., Wunnava, S. & Bourgeois, F. T. US FDA approval of pediatric artificial intelligence and machine learning-enabled medical devices. *JAMA Pediatr.* https://doi.org/10.1001/jamapediatrics.2024.5437 (2024).
12. Klassen, T. P., Hartling, L., Craig, J. C. & Offringa, M. Children are not just small adults: the urgent need for high-quality trial evidence in children. *PLoS Med.* **5**, e172 (2008).
13. Espinoza, J. C. The scarcity of approved pediatric high-risk medical devices. *JAMA Netw. Open* **4**, e2112760 (2021).
14. Espinoza, J., Shah, P., Nagendra, G., Bar-Cohen, Y. & Richmond, F. Pediatric medical device development and regulation: current state, barriers, and opportunities. *Pediatrics* **149**, e2021053390 (2022).
15. Frattarelli, D. A. et al. Off-label use of drugs in children. *Pediatrics* **133**, 563–567 (2014).
16. Section on Cardiology and Cardiac Surgery & Section on Orthopaedics. Off-label use of medical devices in children. *Pediatrics* **139**, e20163439 (2017).

17. Choonara, I. & Conroy, S. Unlicensed and off-label drug use in children: implications for safety. *Drug Saf.* **25**, 1–5 (2002).
18. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
19. Shin, H. J. et al. Optimizing adult-oriented artificial intelligence for pediatric chest radiographs by adjusting operating points. *Sci. Rep.* **14**, 31329 (2024).
20. Reddy, C. D., Lopez, L., Ouyang, D., Zou, J. Y. & He, B. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *J. Am. Soc. Echocardiogr.* **36**, 482–489 (2023).
21. Zuercher, M. et al. Retraining an artificial intelligence algorithm to calculate left ventricular ejection fraction in pediatrics. *J. Cardiothorac. Vasc. Anesth.* **36**, 3610–3616 (2022).
22. Kazerooni, A. F. et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). Preprint at https://arxiv.org/abs/2404.15009 (2024).
23. Rajagopalan, V. & Mariappan, R. in *Fundamentals of Pediatric Neuroanesthesia* (ed. Rath, G. P.) 15–50 (Springer, 2021).
24. Chang, H. P., Kim, S. J., Wu, D., Shah, K. & Shah, D. K. Age-related changes in pediatric physiology: quantitative analysis of organ weights and blood flows: age-related changes in pediatric physiology. *AAPS J.* **23**, 50 (2021).
25. Fraz, M. M. et al. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.* **59**, 2538–2548 (2012).
26. Kocher, M. & Noonan, K. *Pediatric Musculoskeletal Physical Diagnosis: A Video-Enhanced Guide* (Lippincott Williams & Wilkins, 2020).
27. Ruel, J., Ruane, D., Mehandru, S., Gower-Rousseau, C. & Colombel, J.-F. IBD across the age spectrum: is it the same disease? *Nat. Rev. Gastroenterol. Hepatol.* **11**, 88–98 (2014).
28. Hijiya, N., Schultz, K. R., Metzler, M., Millot, F. & Suttorp, M. Pediatric chronic myeloid leukemia is a unique disease that requires a different approach. *Blood* **127**, 392–399 (2016).
29. Scalapino, K. et al. Childhood onset systemic sclerosis: classification, clinical and serologic features, and survival in comparison with adult onset disease. *J. Rheumatol.* **33**, 1004–1013 (2006).
30. Tarr, T. et al. Similarities and differences between pediatric and adult patients with systemic lupus erythematosus. *Lupus* **24**, 796–803 (2015).
31. Yeh, M. M. & Brunt, E. M. Pathological features of fatty liver disease. *Gastroenterology* **147**, 754–764 (2014).
32. Fialkowski, A. et al. Insight into the pediatric and adult dichotomy of COVID-19: age-related differences in the immune response to SARS-CoV-2 infection. *Pediatr. Pulmonol.* **55**, 2556–2564 (2020).
33. Wheeler, D. S., Wong, H. R. & Zingarelli, B. Pediatric Sepsis – Part I: 'Children are not small adults!'. *Open Inflamm. J.* **4**, 4–15 (2011).
34. Alsubie, H. S. & BaHammam, A. S. Obstructive sleep apnoea: children are not little adults. *Paediatr. Respir. Rev.* **21**, 72–79 (2017).
35. Agarwal, P. et al. Deep learning for pediatric chest X-ray diagnosis: repurposing a commercial tool developed for adults. *PLoS ONE* **20**, e0328295 (2025).
36. Shin, H. J., Son, N.-H., Kim, M. J. & Kim, E.-K. Diagnostic performance of artificial intelligence approved for adults for the interpretation of pediatric chest radiographs. *Sci. Rep.* **12**, 10215 (2022).
37. Sourget, T. et al. [Citation needed] Data usage and citation practices in medical imaging conferences. In *Proc. 7nd International Conference on Medical Imaging with Deep Learning* 1475–496 (PLMR, 2024).
38. Koch, B., Denton, E., Hanna, A. & Foster, J. G. Reduced, reused and recycled: the life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track 2* (NeurIPS, 2021).
39. Reinke, A., Tizabi, M. D., Eisenmann, M. & Maier-Hein, L. Common pitfalls and recommendations for grand challenges in medical artificial intelligence. *Eur. Urol. Focus* **7**, 710–712 (2021).
40. Jiménez-Sánchez, A. et al. Copycats: the many lives of a publicly available medical imaging dataset. In *Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (NeurIPS, 2024).
41. Medical Imaging with Deep Learning. In *Proc. of Machine Learning Research* (eds Oguz, I. et al.) 250 (PMLR, 2024); https://proceedings.mlr.press/v250/
42. Smith-Bindman, R. et al. Trends in use of medical imaging in US health care systems and in Ontario, Canada, 2000–2016. *JAMA* **322**, 843–856 (2019).
43. *MIDRC* (Univ. Chicago, accessed 10 September 2025); https://www.midrc.org/
44. Clark, K. et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
45. Shared Datasets. *Center for Artificial Intelligence in Medicine & Imaging* (Standford Univ., accessed 1 July, 2025); https://aimi.stanford.edu/shared-datasets
46. Antonelli, M. et al. The Medical Segmentation Decathlon. *Nat. Commun.* **13**, 4128 (2022).
47. Halabi, S. S. et al. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* **290**, 498–503 (2019).
48. Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A. & van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**, 501–509 (2004).
49. Hunter, R. B. et al. Navigating the challenges of initiating pediatric device trials – a case study. *J. Clin. Transl. Sci.* **8**, e97 (2024).
50. *World Population Prospects 2024* (United Nations Department for Economic and Social Affairs, 2025).
51. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
52. Bakas, S. et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. Preprint at https://arxiv.org/abs/1811.02629 (2018).
53. van der Zanden, T. M. et al. Benefit–risk assessment of off-label drug use in children: the BRAvO framework. *Clin. Pharmacol. Ther.* **110**, 952–965 (2021).
54. Pham, H. H., Nguyen, N. H., Tran, T. T., Nguyen, T. N. M. & Nguyen, H. Q. PediCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *Sci. Data* **10**, 240 (2023).
55. Edwards, D. K., Higgins, C. B. & Gilpin, E. A. The cardiothoracic ratio in newborn infants. *Am. J. Roentgenol.* **136**, 907–913 (1981).
56. Chia, M. D. et al. Radiographic assessment of the cardiothoracic ratio in the pediatric age group at a tertiary institution in North-Central Nigeria. *Calabar J. Health Sci.* **6**, 65–71 (2022).
57. Garin, S. P., Parekh, V. S., Sulam, J. & Yi, P. H. Medical imaging data science competitions should report dataset demographics and evaluate for bias. *Nat. Med.* **29**, 1038–1039 (2023).
58. Yeung, A. W. K. The 'As Low As Reasonably Achievable' (ALARA) principle: a brief historical overview and a bibliometric analysis of the most cited publications. *Radioprotection* **54**, 103–109 (2019).
59. Seibert, J. A. Tradeoffs between image quality and dose. *Pediatr. Radiol.* **34**, S183–S195 (2004). discussion S234–41.
60. Li, J. S., Cohen-Wolkowiez, M. & Pasquali, S. K. Pediatric cardiovascular drug trials, lessons learned. *J. Cardiovasc. Pharmacol.* **58**, 4–8 (2011).
61. Macrae, D. Conducting clinical trials in pediatrics. *Crit. Care Med.* **37**, S136–S139 (2009).

62. Lagler, F. B., Hirschfeld, S. & Kindblom, J. M. Challenges in clinical trials for children and young people. *Arch. Dis. Child.* **106**, 321–325 (2021).
63. de Wildt, S. N. & Wong, I. C. K. Innovative methodologies in paediatric drug development: a conect4children (c4c) special issue. *Br. J. Clin. Pharmacol.* **88**, 4962–4964 (2022).
64. Pediatric research equity act of 2003. *Public Law* **108**, 155 (2003).
65. Ver Date, J. Best pharmaceuticals for children act. *Public Law* **107**, 109 (2002).
66. *Public Law 115 - 52 - FDA Reauthorization Act of 2017* (Office of the Federal Register & National Archives, 2017).
67. Sammer, M. B. K. et al. Use of artificial intelligence in radiology: Impact on pediatric patients, a white paper from the ACR pediatric AI Workgroup. *J. Am. Coll. Radiol.* **20**, 730–737 (2023).
68. Muralidharan, V., Burgart, A., Daneshjou, R. & Rose, S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *npj Digit. Med.* **6**, 166 (2023).
69. Johnson, K. B., Simonian, M., Adams, L. L. & Schneider, J. H. Toward trustworthy pediatric AI: a call to action from the national academy of medicine. *Pediatrics* **156**, (2025).
70. Jordan, P. et al. Pediatric chest–abdomen–pelvis and abdomen–pelvis CT images with expert organ contours. *Med. Phys.* **49**, 3523–3528 (2022).
71. Grand Challenge. https://grand-challenge.org/ (2024).
72. Kaggle: Your Machine Learning and Data Science Community. *Kaggle* https://www.kaggle.com/ (2024).
73. Yuanfeng, J. et al. AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Proc. of the 36th International Conference on Neural Information Processing Systems (NIPS '22)* 36722–36732 (Curran Assoc., 2022).
74. Zong, Y., Yang, Y. & Hospedales, T. MEDFAIR: benchmarking fairness for medical imaging. Preprint at https://arxiv.org/pdf/2210.01725 (2022).
75. Machine learning datasets. *Papers with Code* https://paperswithcode.com/datasets?mod=medical&page=1 (2024).
76. Nguyen, H. Q. et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data* **9**, 429 (2022).
77. Wang, X. et al. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* 3462–3471 (IEEE, 2017); https://doi.org/10.1109/cvpr.2017.369.
78. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. PadChest: a large chest X-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **66**, 101797 (2020).
79. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. Conf. AAAI Artif. Intell.* **33**, 590–597 (2019).
80. Smit, A. et al. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. Preprint at https://arxiv.org/abs/2004.09167 (2020).
81. Liu, Z. et al. A ConvNet for the 2020s. Preprint at https://arxiv.org/abs/2201.03545 (2022).
82. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2017).
83. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: beyond empirical risk minimization. Preprint at https://arxiv.org/abs/1710.09412 (2017).
84. Efron, B. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **82**, 171 (1987).
85. Martin, L. et al. Paediatric X-ray radiation dose reduction and image quality analysis. *J. Radiol. Prot.* **33**, 621–633 (2013).
86. Hua, S. B. Dataset annotations from: Underrepresentation of children in public medical imaging data. *Zenodo* https://doi.org/10.5281/zenodo.18895394 (2026).
87. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).

Acknowledgements

Data processing, data analysis was performed at the High-Performance Computing Facility, Centre for Computational Medicine, The Hospital for Sick Children, Toronto, Canada. This research was enabled in part by support provided by Compute Ontario (computeontario.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

Author contributions

Study conceptualized by L.E. and A.X.L.; methodology developed by L.E., A.X.L., S.B.Z.H., N.H., P.H., A.J.T. and I.Y.C.; formal analysis performed by S.B.Z.H. and P.H.; data curation performed by S.B.Z.H., L.E., A.X.L., N.H. and P.H.; original draft written by S.B.Z.H., L.E. and A.X.L.; reviewing and editing performed by L.E., A.X.L., S.B.Z.H., N.H., P.H., A.J.T. and I.Y.C.; visualizations prepared by S.B.Z.H.; supervision by L.E. and A.X.L.; financial support for project by L.E.

Competing interests

A.X.L. is an employee of and holds equity in Microsoft. I.Y.C. is an advisor for Vega Health and the Massachusetts Attorney General Office. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44360-026-00111-3.

Correspondence and requests for materials should be addressed to Lauren Erdman.

Peer review information *Nature Health* thanks Sebastian Tschauner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Health* team. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Lauren Erdman

Last updated by author(s): Mar 6, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection
Data analysis	Code to reproduce our analyses is provided in our GitHub repository: https://github.com/stan-hua/ped_vs_adults-cxr We use Python 3.9.19 for all of our analyses. These are a list of conda packages and their versions: cuda-version = "12.6" configobj = "==5.0.9" comet-ml = "3.51" fire = "==0.6.0" jinja2 = "==3.1.4" matplotlib = "==3.9.4" seaborn = "==0.13.2" py-opencv = "==4.12.0" imageio = "==2.37.0" scikit-image = "==0.24.0" scikit-learn = "==1.6.1" openpyxl = "==3.1.5" pytorch-gpu = "==2.5.1" numpy = "==1.26.4" torchmetrics = "1.8.1" lightning = "==2.3.3" pytorch-lightning = "==2.4.0"

```
torchvision = "==0.20.1"
timm = "==1.0.19"
pydicom = "==2.4.4"
pylibjpeg = "==2.0.1"
pylibjpeg-openjpeg = "==2.5.0"
pylibjpeg-libjpeg = "==2.3.0"
```

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All dataset annotations are made available in our GitHub repository: https://github.com/stan-hua/ped_vs_adults-cxr

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We did not report on sex/gender-differences in pediatric representation among public datasets, but we provide sex/gender summary statistics for each dataset when available.
Reporting on race, ethnicity, or other socially relevant groupings	We report on estimates for the representation of children (aged 17 and younger) among patients in public medical imaging datasets. In our cardiomegaly case study, we evaluate adult-trained models on children, stratified by age in years.
Population characteristics	Age statistics are provided for the annotated datasets individually, and summary statistics across datasets are shown in Figure 2, Table 1 and Table 2. The distribution of ages in the chest X-ray studies are plotted in Figure 3.
Recruitment	N/A. Metadata for each dataset was collected from their respective online repositories.
Ethics oversight	N/A. We compiled publicly available data to report on distributions and the frequency of available metadata. No REB/IRB required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is quantitative, performing a systematic review of public medical imaging datasets for AI development. This is followed by a case study to illustrate how absence of pediatric data in public X-ray datasets systematically leads to algorithmic bias.
Research sample	Based on studies from an AI for medical imaging conference, we identified common sources of datasets for AI practitioners: dataset collections, machine learning challenges, benchmarks and highly-cited stand-alone datasets. We organized our dataset search around these sources. For dataset collections, we considered TCIA, Stanford AIMI and MIDRC, which yielded 69, 13 and 24 datasets respectively. For ML challenges, we found 58 datasets across Kaggle, Grand Challenge and RSNA, while for benchmarks, we selected three well-known benchmarks: MSD, AMOS and MedFAIR, containing 10, 1 and 7 sub-datasets each. Lastly, we identify 30 of the most-cited datasets based on PapersWithCode. For our cardiomegaly case study, we used X-ray datasets from PhysioNet (MIMIC-CXR, VinDr-CXR and VinDr-PCXR), Stanford AIMI (CheXBERT), NIH (X-ray18) and BIMCV (PadChest).
Sampling strategy	We comprehensively annotate all included datasets among dataset collections and ML challenges, while we chose the number of highly-cited datasets (30) and benchmarks (3) for convenience. In Table 1, we provide summaries on the number of estimated patients and children from our final list of datasets, broken down by dataset source.

Data collection	Data was collected from online resources. SBZH, LE, AXL, NH, PH manually annotated all datasets, extracting the number of patients within each dataset and their ages from heterogeneous metadata. We provide complete information on the annotation process under Methods / Dataset Annotation.
Timing	For datasets in TCIA and Stanford AIMI, we annotated datasets released before 2025. For MIDRC, we extracted patient metadata from the explorer (portal v5.39.2, submission v2025.09). On challenge platforms Kaggle, Grand Challenge and RSNA, we filtered for challenges within 2019 to 2024. While for benchmarks (MSD, AMOS and MedFAIR) and the top-30 cited datasets analyzed, we did not restrict by time of release.
Data exclusions	We exclude adult-specific imaging modalities and data collections, leading to exclusion of 1 mammography dataset and the exclusion of the UK BioBank data collection. Additional modalities fluoroscope and nuclear imaging were excluded because little to no datasets were found for these modalities. To support analyses comparing pediatric representation in public data to pediatric imaging in practice, we exclude datasets obtained outside of clinical settings, leading to the exclusion of datasets from the OpenNeuro collection.
Non-participation	We use medical imaging datasets made publicly available and rely on the authors of these datasets for data removal based on non-participation.
Randomization	No randomization was utilized in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*