npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01270-x

A scoping review of reporting gaps in FDA-approved AI medical devices

Check for updates

Vijaytha Muralidharan[1,13] ✉, Boluwatife Adeleye Adewale[2,3,13], Caroline J. Huang[4], Mfon Thelma Nta[5], Peter Oluwaduyilemi Ademiju[6], Pirunthan Pathmarajah[1], Man Kien Hang[7,8], Oluwafolajimi Adesanya[9], Ridwanullah Olamide Abdullateef[10,11], Abdulhammed Opeyemi Babatunde[10], Abdulquddus Ajibade[10], Sonia Onyeka[1], Zhou Ran Cai[1], Roxana Daneshjou[1,12,14] & Tobi Olatunji[6,14]

Machine learning and artificial intelligence (AI/ML) models in healthcare may exacerbate health biases. Regulatory oversight is critical in evaluating the safety and effectiveness of AI/ML devices in clinical settings. We conducted a scoping review on the 692 FDA-approved AI/ML-enabled medical devices approved from 1995-2023 to examine transparency, safety reporting, and sociodemographic representation. Only 3.6% of approvals reported race/ethnicity, 99.1% provided no socioeconomic data. 81.6% did not report the age of study subjects. Only 46.1% provided comprehensive detailed results of performance studies; only 1.9% included a link to a scientific publication with safety and efficacy data. Only 9.0% contained a prospective study for post-market surveillance. Despite the growing number of market-approved medical devices, our data shows that FDA reporting data remains inconsistent. Demographic and socioeconomic characteristics are underreported, exacerbating the risk of algorithmic bias and health disparity.

To date, the FDA has approved 950 medical devices driven by artificial intelligence and machine learning (AI/ML) for potential use in clinical settings[1]. Most recently, the FDA has launched the Medical Device Development Tools (MDDT) program, which aims to "facilitate device development, timely evaluation of medical devices and promote innovation", with the Apple Watch being the first approved device for this regulatory process[2,3]. As AI/ML studies begin to translate to clinical environments, it is crucial that end users can evaluate the applicability of devices to their unique clinical settings and assess sources of bias and risk.

One definition of algorithmic bias in the context of AI/ML health systems is instances when an algorithm amplifies inequities and results in poor healthcare outcomes[4]. Some defined sub-categories of algorithmic bias are listed in Box 1[5,6].

Despite the rise in awareness of algorithmic bias and its potential implications on the generalizability of AI/ML models[7], there is a paucity of standardized data reporting by regulatory bodies including the FDA that provide reliable and consistent information on the development, testing, and training of algorithms for clinical use. This limits accurate analysis and evaluation of algorithmic performance, particularly in the context of under-represented research groups such as ethnic minorities, children, maternal health patients, patients with rare diseases, and those from lower socioeconomic strata. Deploying devices that cannot be transparently evaluated by end users may increase health disparity and is particularly relevant in the context of emerging clinical trials and real-world deployment[8]. To date, there has been limited review of this published data.

Here, we investigate AI-as-medical-device Food and Drug Administration (FDA) approvals from 1995–2023 to examine the contents, consistency, and transparency in FDA reporting of market-approved devices with a focus on bias. We focus our study on the limited published FDA data and associated papers, in the format of a scoping review.

[1] Department of Dermatology, Stanford University School of Medicine, Stanford, CA, 94304, USA. [2] Babcock University Teaching Hospital, Ilishan-Remo, Ogun State, Nigeria. [3] Neuroscience and Ageing Research Unit, Institute for Advanced Medical Research and Training, College of Medicine, University of Ibadan, Ibadan, Nigeria. [4] Stanford University School of Medicine, Stanford, CA, 94304, USA. [5] Department of Medicine and Surgery, Afe Babalola University, Ado-Ekiti, Nigeria. [6] Intron Health, Lagos, Nigeria. [7] Co:Helm, New York, NY, USA. [8] University College London, London, UK. [9] School of Medicine, University of Illinois, Urbana-Champaign, Urbana, IL, 60612, USA. [10] Faculty of Clinical Sciences, College of Medicine, University of Ibadan, Ibadan, Nigeria. [11] College Research and Innovation Hub (CRIH), College of Medicine, University of Ibadan, Ibadan, Nigeria. [12] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, 94304, USA. [13] These authors contributed equally: Vijaytha Muralidharan, Boluwatife Adeleye Adewale. [14] These authors jointly supervised this work: Roxana Daneshjou, Tobi Olatunji. ✉ e-mail: vmurali5@stanford.edu

Box 1 | Categories of algorithmic bias

Representation bias	Data underrepresents or misrepresents subsets of the population measurement bias—data inaccurately reflects the variables.
Omitted variable bias	Omitted data. For example, the “Black-box” nature of many algorithms makes it difficult to figure out which variables were used and omitted.
Aggregation bias	Conclusions are drawn about individuals based on observations about a larger group.
Linking bias	Correlations that AI draws about particular users based on the characteristics of other users that may not be accurate in a heterogeneous population.
Label bias	Bias that arises when the same object is labeled differently based on differing perceptions by the annotator.

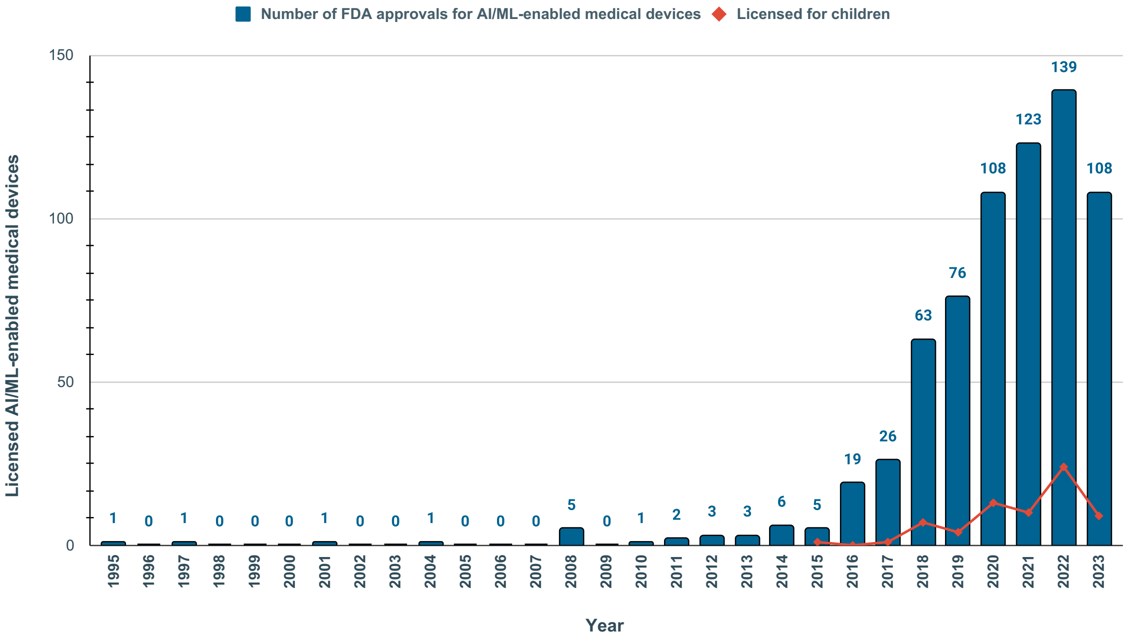

Fig. 1 | Trends in FDA licensing of AI/ML-enabled medical devices. The blue bars show the number of Food and Drug Administration (FDA) approvals for medical devices from 1995 to 2023. The red line plot shows the number of FDA approvals for AI/ML-enabled medical devices that are licensed for children. Total number of FDA approvals for AI/ML-enabled medical devices have increased steeply since 2016. FDA approvals for children-licensed devices have increased steadily since 2018.

Results

Distribution of device approval across clinical specialties

A total of 692 SSEDs of FDA-approved AI-enabled medical devices/software were analyzed. There was a steady increase in annual FDA approvals for AI-enabled medical devices with a mean of 7 between 1995 and 2015, increasing to 139 approvals in 2022 (Fig. 1). The regulatory class of each device included in the study was determined and categorized according to the United States Food and Drug Administration (FDA) classification system. Only 2 (0.3%) of the devices belonged to the regulatory Class III (devices posing the highest risk), while the vast majority (99.7%) of the devices belonged to Class II (devices whose safety and underlying technology are well understood and therefore considered to be lower risk)[9].

Table 1 shows the distribution of 408 approved devices across organ systems. The top three organ systems represented amongst approved medical devices are the circulatory (20.8%), nervous (13.6%), and reproductive (7.2%). The least represented are the urinary (1.2%) and endocrine (0.7%) systems (Table 1). Each device in the FDA database is classified under a particular medical specialty (Fig. 2). The FDA classification shows that the most represented medical specialty is Radiology (532 approvals; 76.9%) with the fewest approvals in Immunology, Orthopedics, Dental Health, Obstetrics, and Gynecology (Fig. 2 and Table 2). A total of 284 (40.1%) approved devices could not be categorized to an organ system either because (1) the clinical indication was not specific to one system or because (2) the function of the device cuts across multiple organ systems (e.g., whole-body imaging system/software). As such, there are some differences between the categories of organ system and medical specialty. For instance, 70 (10.1%) of the devices are classified by the FDA under the cardiovascular field despite 144 (20.8%) approvals specific to the circulatory system (Tables 1 and 2).

Reporting data on statistical parameters and post-market surveillance

Indication for use of the device was reported in most (678; 98.0%) SSEDs (Fig. 3a) and 487 (70.4%) SSEDs contained a pre-approval performance study. However, 435 (62.8%) provided no data on the sample size of the subjects. Although 319 (46.1%) provided comprehensive detailed results of performance studies including statistical analysis, only 13 (1.9%) of them included a link to a scientific publication with further information on the safety and efficacy of the device (Fig. 4). Only 219 (31.6%) SSEDs provided data on the underlying machine-learning technique. Only 62 device documents (9.0%) contained a prospective study for post-market surveillance. Fourteen (2.0%) SSEDs addressed reporting of potential adverse effects of medical devices on users.

Race, ethnicity, and socioeconomic diversity

Patient demographics in algorithmic testing data were only specified in 153 (22.1%) SSEDs, with 539 (77.9%) not providing any demographic data (Fig. 3b). Only 25 (3.6%) provided information on the race and/or ethnicity of tested or intended users. Socioeconomic data on tested or intended users were provided for only 6 (0.9%) of devices (Fig. 3c).

Age diversity

There were 134 (19.4%) SEEDs with available information on the age of the intended subjects. Upon examining age diversity in approved devices, the first FDA approval for a device licensed for children was in 2015. Between 2015 and 2022, the annual FDA approvals for the pediatric age group steadily increased from 1 to 24 in total. Despite this rise, the proportion of pediatric-specific approvals relative to the total approvals (for adults and pediatrics combined) has remained low, fluctuating between 0.0% and 20.0% (Fig. 1 and Table 3). Although 4 (0.6%) devices were exclusively developed for children, we found 65 more devices that have been approved for use in both adult and pediatric populations, thus bringing the total number of approvals for the pediatric population to 69 (10.0%). Testing and validation of devices in children and adults was reported in only 134 (19.4%) SSEDs (Fig. 5a, b). The distribution of devices for children ($n = 69$) across medical specialties falls under just 5 categories, following a similar pattern as earlier observed for the entire population with lead representation in the fields of Radiology (72.5%; $n = 69$), Cardiovascular health (14.4%; $n = 69$) and Neurology (10.1%; $n = 69$) (Fig. 5c). There were only three (0.4%) approved devices that focused exclusively on geriatric health.

Sex diversity

When examining sex reporting transparency, there were a total of 39 (5.6%) approvals exclusively for women's health, 36 of them focusing on the detection of breast pathology. The remaining three were designed to aid cervical cytology; determine the number and sizes of ovarian follicles; and perform fetal/obstetrics ultrasound. Of the 10 (1.5%) devices that were exclusively for men, eight of them were indicated in diagnostic and/or therapeutic procedures involving the prostate, while the remaining two were for seminal fluid analysis.

Discussion

Our study highlights a lack of consistency and data transparency in published FDA AI/ML approval documents which may exacerbate health disparities. In a similar study examining 130 FDA-approved AI medical devices between January 2015 and December 2020, 97% reported only

Table 1 | Distribution of FDA-approved AI-enabled medical devices by organ system

System	Number of approvals (%)
Circulatory	144 (20.8)
Nervous	94 (13.6)
Respiratory	48 (6.9)
Reproductive	47 (6.8)
Musculoskeletal	33 (4.8)
Gastrointestinal	24 (3.5)
Urinary	8 (1.2)
Hematologic	5 (0.7)
Endocrine	5 (0.7)

Table 2 | Primary medical specialty associated with FDA approval

Medical specialty	Number of approval (%)
Radiology	532 (76.9)
Cardiovascular	70 (10.1)
Neurology	20 (2.9)
Hematology	15 (2.2)
Gastroenterology/urology	11 (1.6)
Ophthalmology	9 (1.3)
Anesthesiology	6 (0.9)
Clinical chemistry	6 (0.9)
General and plastic surgery	5 (0.7)
Microbiology	5 (0.7)
General hospital	3 (0.4)
Ear nose & throat	2 (0.3)
Pathology	4 (0.6)
Dental	1 (0.1)
Immunology	1 (0.1)
Obstetrics and gynecology	1 (0.1)
Orthopedic	1 (0.1)

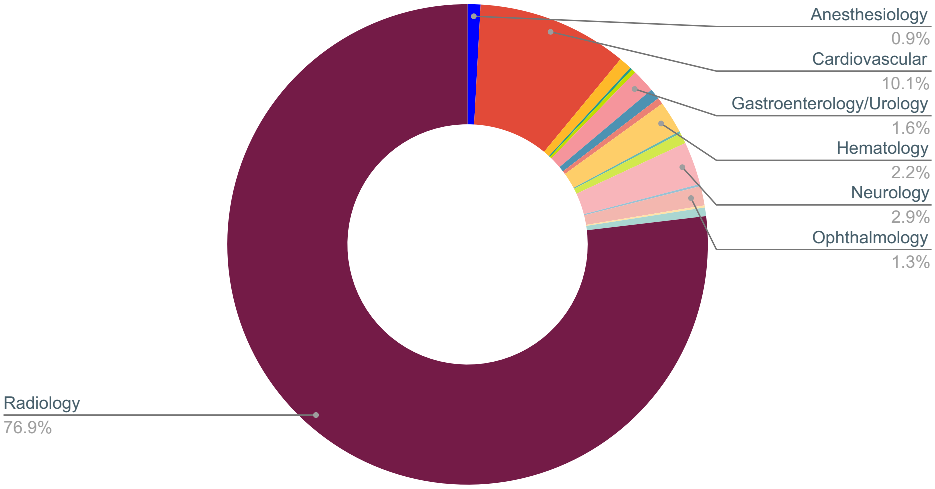

Fig. 2 | FDA-assigned medical specialty. Distribution of medical specialties represented in 692 FDA approvals between 1995 and 2023.

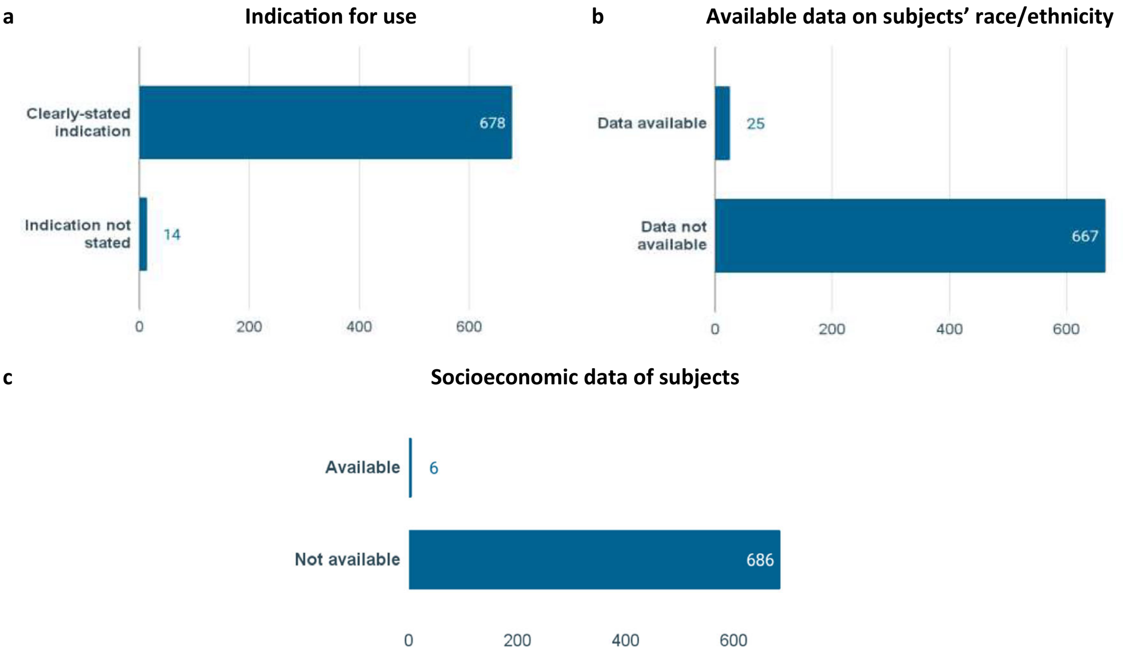

Fig. 3 | Reported information on intended subjects. Number of FDA summary documents providing relevant sociodemographic information about intended subjects of medical device. **a** Only 14 (2.0%) summary documents lacked clearly stated indications for the use of the medical device. **b** The majority (667; 96.4%) of the documents did not provide any information about race/ethnicity of intended subjects and how this may impact the application of the device. **c** The majority (686; 99.1%) of the documents did not provide any information about socioeconomic context of intended subjects.

Fig. 4 | Accessibility to publications supporting safety, efficacy, and transparency. Only few summary documents provide a means of accessing published scientific publication that validates the use of the medical device.

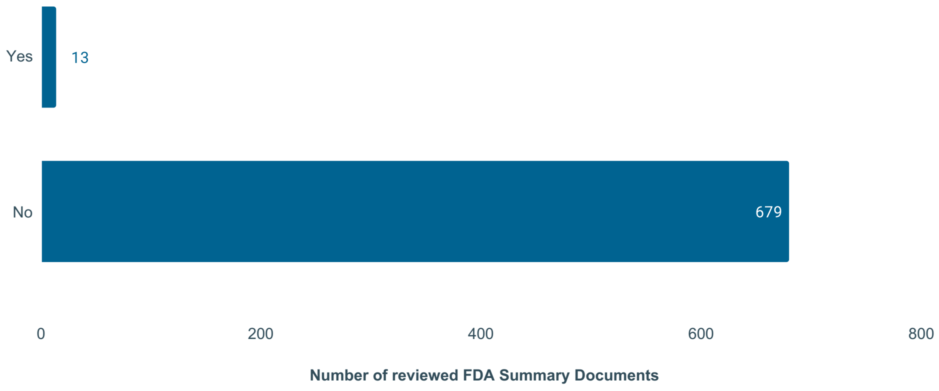

Table 3 | Annual trends in FDA approvals for AI-enabled medical devices and software for children (2015–2023)

Year	Number of FDA approvals for AI/ML-enabled medical devices	Licensed for children (percentage relative to total FDA approvals %)
2015	5	1 (20.0)
2016	19	0 (0.0)
2017	26	1 (3.8)
2018	63	7 (11.1)
2019	76	4 (5.3)
2020	108	13 (12.0)
2021	123	10 (8.1)
2022	139	24 (17.3)
2023	108	9 (8.3)

retrospective evaluations; prospective studies did not evaluate high-risk devices; 72% did not publicly report whether the algorithm was tested on more than one site; and 45% did not report basic descriptive data such as sample size[10,11]. A lack of consistent reporting prevents objective analysis of the fairness, validity, generalizability, and applicability of devices for end users. As our results describe, only 37% of device approval documents contained information on sample size. As the clinical utility of algorithmic data is limited by data quantity and quality[12], a lack of transparency in sample size reporting significantly limits the accurate assessment of the validity of performance studies, and device effectiveness[13].

Only 14.5% of devices provided race or ethnicity data. Recent literature strongly emphasizes the risks of increasing racial health disparity through the propagation of algorithmic bias[14–16]. A lack of racial and ethnic profiling in publicly available regulatory documents risks further exacerbating this important health issue[17,18]. The FDA has recognized the potential for bias in AI/ML-based medical devices and has initiated action plans ("Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical

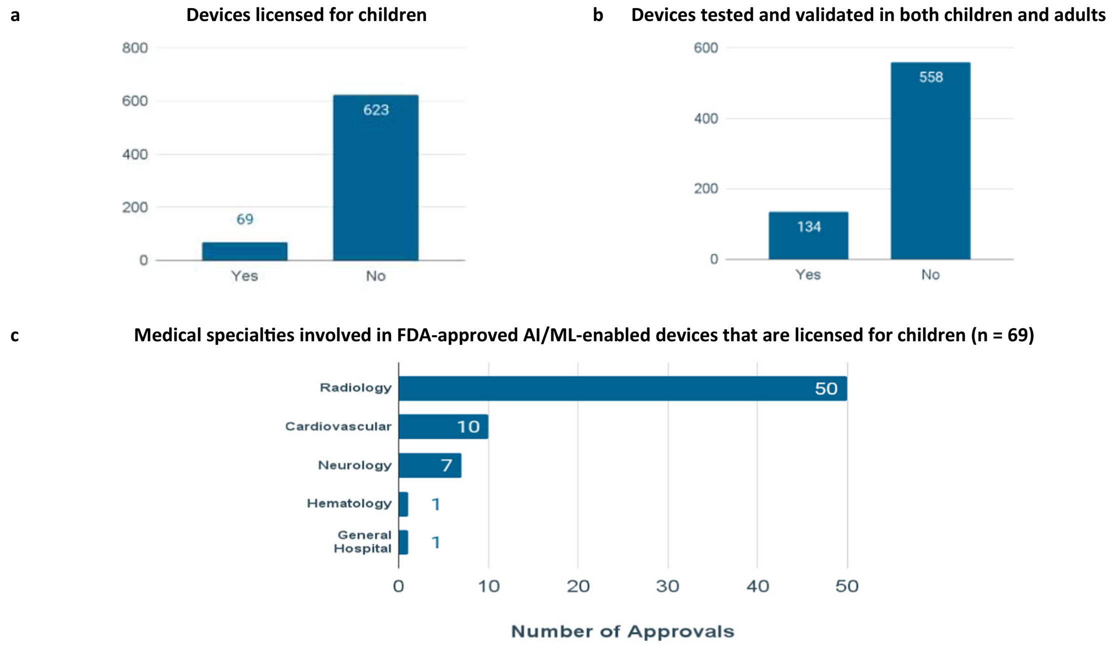

Fig. 5 | FDA approvals for pediatrics-licensed medical devices. The proportion of total FDA approvals **a** licensed for children is 10.0% **b** tested and validated in both children and adults is 19.4%. **c** Radiology has the highest representation among FDA-approved devices for children.

Device (SaMD) Action Plan") in January 2021 to address these concerns[19,20]. However, despite these efforts, our study highlights reporting inconsistencies that may continue to propagate racial health disparities[21]. In light of these results, there is a pressing need for transparent and standardized regulatory frameworks that explicitly consider racial diversity in the evaluation and reporting of AI/ML medical devices[22]. Other strategies to mitigate racial bias may include adopting adversarial training frameworks and implementing post-authorization monitoring to ensure AI/ML devices perform equitably across all patient demographics[23,24].

While AI/ML presents potential opportunities to reduce socioeconomic disparity in health, a lack of representation of target users across varied economic strata risks the propagation of health disparity in higher and lower-income groups[25]. As with other clinical research domains, a lack of representation of lower socioeconomic groups including those in remote and rural areas, risks neglect of those most likely to benefit from improved access to healthcare[26,27]. Our data shows that only 0.6% of approved devices contained specific data detailing the socioeconomic striate of users in testing and/or algorithmic training datasets. This data renders it difficult to predict the potential clinical and financial impacts of approved medical devices on economic population subsets. Furthermore, a lack of socioeconomic data prevents accurate and robust cost-effectiveness analyses that may significantly impact the availability and impact of medical algorithms or devices[28,29]. Studies have underscored disparities rooted in socioeconomic factors, impacting the performance of AI/ML technologies[4,30,31]. Initiatives promoting diversity in data collection and consideration of model performance across socioeconomic groups are paramount and must be incorporated in the assessment of market approval for emerging technologies[32].

With only 19.4% of devices providing information on the age of intended device users, our study suggests that the evaluation and approval process of medical AI devices by the FDA lacks comprehensive data on age diversity. Recent literature across specialties demonstrates differential performances in algorithms trained on adult or pediatric data[33,34]. As an example, a study exploring echocardiogram image analysis suggested that adult images could not be appropriately generalized to pediatric patients and vice versa[35]. A lack of transparent age reporting, therefore, risks propagating age-related algorithmic bias, with potential clinical, ethical, and societal implications on the target population[33,36]. Mitigating age bias requires a concerted effort to ensure that training and testing datasets appropriately match intended users. Further, with only 0.6% of devices approved specifically for the pediatric age group, our findings identify equity gaps in the representation of children in AI/ML market-approved devices[37,38]. Medical devices that have demonstrated their inclusion in pediatric patient populations include the *MEDO ARIA* software, which assists in the image-based assessment of developmental hip dysplasia (DDH) in infants aged 0 to 12 months[39]. The *EarliPoint System* is recommended for developmental disability centers to aid in diagnosing and assessing Autism Spectrum Disorder (ASD) in children aged 16 to 30 months, and the *Cognoa ASD Diagnosis Aid*, for diagnosing ASD in children aged 18 to 72 months[40,41].

With our findings showing that only 0.4% of approved devices cater specifically to geriatric health needs, specific considerations should be considered for the older adult population. Despite having the highest proportion of healthcare utilization, geriatric patients are traditionally underrepresented in clinical research[42]. A recent WHO ethical guidance document outlines the potential societal, clinical, and ethical implications of ageism in medical research, and describes the lack of geriatric representation as a health hazard in light of aging populations[43]. Initiatives such as the National Institutes of Health (NIH) Inclusion Across the Lifespan policy aim to promote the participation of older adults in research studies, which may help equitize the potential impacts of algorithmic development for this population, considering unique ethical and clinical considerations[44,45]. Similar to considerations for children, we propose that regulatory bodies encourage market approval documents to make clear intentions to test and train on a geriatric population and ensure that appropriate validation methods are in place to ensure the appropriate generalization of model outputs to specific geriatric health needs[46–49]. Examples of medical devices representing the geriatric population include *NeuroRPM*, designed to quantify movement disorder symptoms in adults aged 46 to 85 with Parkinson's disease[50]. NeuroRPM's suitability for clinic and home environments facilitates remote visits for patients with limited access to in-person care[50]. Another device, *icobrain*, automates labeling, visualization, and volumetric quantification of brain structures in potential dementia patients[51]. For osteoarthritis, the *Knee OsteoArthritis Labeling Assistant (KOALA)* software measures joint space width and radiographic features on knee X-rays, aiding in risk stratification and highlighting the importance of preventative screening in geriatrics[52].

Box 2 | Recommendations for regulatory reporting of AI/ML devices

Explicit representation standards	The FDA should explicitly mandate that AI/ML medical device submissions provide comprehensive demographic information, including race, ethnicity, age, gender, and socioeconomic status of the studied populations. This ensures that the algorithms are developed and validated on diverse datasets, minimizing the risk of biased outcomes and ensuring equitable representation. These statistics should be reported separately for training, validation, and deployment data.
Inclusive validation criteria	Establish clear guidelines for validation studies to include diverse and representative populations. Manufacturers should be required to demonstrate the performance of their AI/ML algorithms across various demographic groups, ensuring that the technology is effective for minorities, marginalized, and underrepresented communities.
Transparency requirements	Implement stringent transparency standards, mandating clear documentation of the machine learning techniques used, the sources of training data, and the rationale behind algorithmic decisions. Transparency ensures that end-users, healthcare providers, and regulatory authorities have insight into the algorithm's decision-making processes.
Post-market surveillance for equity	Establish a robust post-market surveillance framework specifically focused on monitoring the performance of AI/ML devices in real-world settings. This ongoing evaluation should include an analysis of outcomes across populations representative of all patients for whom the tool is indicated, with a particular emphasis on detecting and addressing any emerging biases affecting minorities and under-represented groups.
Interdisciplinary review panels	Form interdisciplinary review panels within the FDA consisting of experts in data science, healthcare disparities, and ethics. This ensures a holistic evaluation of AI/ML submissions, incorporating perspectives that can identify and address potential biases and disparities in health outcomes.
Incentives for diversity and equity	Establish incentives, such as expedited review processes or regulatory benefits, for manufacturers who proactively address issues of representation, validation rigor, and bias mitigation in their AI/ML device submissions. This encourages industry-wide commitment to fairness and equity.

Our study also examined variations in the representation of different medical specialties among approved medical devices. Specialties most commonly represented include Radiology, Cardiology, and Neurology[1]. Promoting clinical equity requires a more balanced representation of specialties and disease systems in digital innovation. Whilst we appreciate that AI/ML research is limited by data availability and data quality, industry, academia, and clinicians must advocate for equality of innovation amongst specialties, to include a broad range of conditions and patient populations in medical device development and testing that may potentially benefit[53]. As the FDA is a US-based regulator, our review does not examine the representation of specialties or conditions outside the US, in particular in Low- and Middle-Income Countries (LMICs) which contain over 80% of the global burden of disease[54–56]. Many countries do not have the regulatory capacity to release approval documentation, and thus future studies must incorporate international data availability, collaboration, and cohesion[57]. Regulatory bodies both within and outside the USA must attempt to align technological development with key priorities in national and global disease burden to promote global equity[58].

Our results showed that transparency in study enrollment, study design methodology, statistical data, and model performance data were significantly inconsistent amongst approved devices. While 70.4% of studies provided some detail on performance studies before market approval, only 46.1% provided detailed results of the performance studies. In 62.9% of devices, there was no information provided on sample size. Transparency is crucial in addressing the challenges of interpretability and explainability in AI/ML systems, and our current findings suggest that evaluation cannot be comprehensively conducted across approved FDA devices[59]. Models that are transparent in their decision-making process (interpretability) or those that can be elucidated by secondary models (explainability) are essential for validating the clinical relevance of any outcomes and ensuring that devices that may be incorporated in clinical settings and thus enhanced transparency must be incorporated in future approvals[22,60]. Further ethical considerations encompass a range of issues, including patient privacy, consent, fairness, accountability, and algorithmic transparency[61]. Including ethics methods in both study protocols and future regulatory documents may minimize privacy concerns arising from the potential misuse, and increase end-user confidence[62,63].

Only 142 (20.5%) of the reviewed devices provide statements on potential risks to end users. Further, only 13 (1.9%) approval documents included a corresponding published scientific validation study, providing evidence of their safety and effectiveness. Underreporting of safety data in approved devices limits the ability of end users to determine generalizability, effectiveness, cost-effectiveness, and medico-legal complexities that may occur from device incorporation[64]. It is therefore paramount that regulatory bodies such as the FDA advocate for a mandatory release of safety data and considerations of potential adverse events. One example of an approved device reporting adverse effects is the *Brainomix 360 e-ASPECTS*, a computer-aided diagnosis (CADx) software device used to assist the clinician in the characterization of brain tissue abnormalities using CT image data[65]. Its safety report highlights some of the potential risks of incorrect scoring of the algorithm, the potential misuse of the device to analyze images from an unintended patient population, and device failure.

Box 2 details some recommendations that may be adopted by the FDA and similar regulatory bodies internationally to reduce the risk of bias and health disparity in AI/ML.

While the FDA's Artificial Intelligence/Machine Learning (AI/ML) Action Plan outlines steps to advance the development and oversight of AI/ML-based medical devices[20], including initiatives to improve transparency, post-market surveillance, and real-world performance monitoring[66], our study highlights that there remain several clinically relevant inconsistencies in market approval data that may exacerbate algorithmic biases and health disparity.

Poor data transparency in the approval process limits some of the conclusions that can be reliably drawn in this study and limits quality assessment of post-market performance and real-world effectiveness. The paucity of sociodemographic data provided in the SSEDs begets the question of whether applicants were failing to track sociodemographics or if they were simply failing to report them. The FDA SSED template[67] does stipulate disclosure of risks and outcomes that may be impacted by sex, gender, age, race, and ethnicity. Thus, we can only justifiably assume that based on what is available and accessible, the paucity of the subgroup analysis data results from the failure of applicants to track the sociodemographics rather than an overall failure of the application process to capture the relevant information. However, it is clear that devices are approved despite not having all of this information available, making the case also for the potential failure of enforcement of these metrics before approval. Considering that most companies do not publish their post-market outcome data (only 1.9% have published data available) and are currently not mandated to do so, our findings are limited by what is accessible. This further re-emphasizes the argument for rigorous and more transparent regulation by bodies such as the FDA to protect end consumers and enhance post-market evaluation and the assessment of real-world effectiveness.

The authors also acknowledge that as this review focuses only on market-approved devices by the FDA in the United States, the results may not be universally generalizable. However, the authors do believe that the results and concepts highlighted in this scoping review are globally relevant. They hope that this paper can form the basis of further studies that focus on devices in varied settings and that this paper will motivate greater global data transparency to promote further health equity in emerging technologies. The authors also note that in recent months, there have been additional AI/ML market approved devices added to the 510k database that have not been included in this evaluation. While this is a limitation, the authors believe that the rapidly accruing number of devices makes the findings of this paper further relevant, demanding quick regulatory review and action.

The ramifications of inadequate demographic, socioeconomic, and statistical information in the majority of 510(k) submissions to the FDA for AI/ML medical devices approved for clinical use are multifaceted and extend across societal, health, legal, and ethical dimensions[12,33,68]. Addressing these informational gaps is imperative to ensure the responsible and equitable integration of AI/ML technologies into clinical settings and the appropriate evaluation of demographic metrics in clinical trials. Additional focus must be given to under-represented groups who are most vulnerable to health disparities as a consequence of algorithmic bias[33,69].

Methods

Based on the intended use, indications for use, and associated risks, the FDA broadly classifies devices as class I (low-risk), class II (medium-risk), and class III (high-risk). Class I and II devices are often cleared through the 510(k) pathway in which applicants submit a premarket notification to demonstrate safety and effectiveness and/or substantial equivalence of their proposed device to a predicate device. Class III (high-risk) devices are defined as those that support human life, prevent impairment of health, or pose potential unreasonable health risk(s)[69]. Evaluation of this class of devices follows the FDA's most rigorous device approval pathway known as premarket approval (PMA)[70].

The Food, Drug, and Cosmetic Act subparagraph 520(h)(1)(A) requires that a document, known as a Summary of Safety and Effectiveness Data (SSED) be made publicly available following FDA approval. SSEDs are authored by applicants using a publicly accessible template provided by the FDA[67]. The document intends to provide a balanced summary of the evidence for the approval or denial of the application for FDA approval. To be approved, it is required that the probable benefits of using a device outweigh the probable risks. The studies highlighted in the SSED should provide reasonable evidence of safety and effectiveness[67].

Thus, we conducted a scoping review of AI-as-a-medical-device approved by the FDA between 1995 and 2023, using FDA Summary of Safety and Effectiveness Data (SSED). This scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines[71]. A completed PRISMA-ScR checklist is included in Supplementary Table 1. A protocol was not registered.

We included all SSEDs of FDA-approved AI/ML-enabled medical devices between 1995 and 2023 made publicly available via https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.[1] Each SSED was reviewed by an expert in computer science, medicine, or academic clinical research who identified, extracted, and entered relevant variables of interest (Supplementary Table 2). Data was then computed into a Microsoft Excel spreadsheet. Counts and proportions of each variable were generated using Microsoft Excel. The spreadsheet and analysis worksheet on Microsoft Excel have been made available publicly via https://zenodo.org/records/13626179.

Variables of interest were determined per the Consolidated Standards of Reporting Trials - Artificial Intelligence (CONSORT-AI) extension checklist which is a guideline developed by international stakeholders to promote transparency and completeness in reporting AI clinical trials[72]. Equivocal or unclear information identified in each SSED was then evaluated in consensus.

Primary outcome measures included frequency of race/ethnicity reporting, age reporting, and availability of sociodemographic data of the algorithmic testing population provided in each approval document. Secondary outcomes evaluated the representation of various medical specialties, organ systems, and specific patient populations such as pediatric and geriatric in approved devices.

Data availability

Data supporting this study are included within the article and supporting materials. All FDA summary of safety and effectiveness data (SSED) documents are publicly available and accessible via https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. The extracted data and analysis worksheet are published on Zenodo and available via https://zenodo.org/records/13626179.

Received: 18 May 2024; Accepted: 23 September 2024;
Published online: 03 October 2024

References

1. U.S. Food and Drug Administration (FDA). Artificial intelligence and machine learning (AI/ML)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (2024).
2. Center for Devices and Radiological Health. Medical Device Development Tools (MDDT). https://www.fda.gov/medical-devices/medical-device-development-tools-mddt (2024).
3. Ajraoui, S. & Ballester, B. R. Apple Watch AFib history feature makes medical device history. https://www.iqvia.com/blogs/2024/05/apple-watch-afib-history-feature-makes-medical-device-history (2024).
4. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* **9**, 020318 (2019).
5. Chu, C. H. et al. Ageism and artificial intelligence: protocol for a scoping review. *JMIR Res. Protoc.* **11**, e33211 (2022).
6. Jiang, H. & Nachum, O. Identifying and correcting label bias in machine learning. *Proc. Mach. Learn. Res.* **108**, 4621–4630 (2020).

7. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
8. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: a call for open science. *Patterns* **2**, 100347 (2021).
9. The Pew Charitable Trusts. How FDA regulates artificial intelligence in medical products. https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2021/08/how-fda-regulates-artificial-intelligence-in-medical-products (2021).
10. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
11. Wu, E. et al. Toward stronger FDA approval standards for AI medical devices. HAI Policy Brief. 1–6 (2022).
12. Mashar, M. et al. Artificial intelligence algorithms in health care: is the current Food and Drug Administration regulation sufficient? *JMIR AI* **2**, e42940 (2023).
13. Ahmed, M. I. et al. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus* **15**, e46454 (2023).
14. Nazer, L. H. et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit. Health* **2**, e0000278 (2023).
15. Delgado, J. et al. Bias in algorithms of AI systems developed for COVID-19: a scoping review. *J. Bioeth. Inq.* **19**, 407–419 (2022).
16. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
17. Fox-Rawlings, S. R., Gottschalk, L. B., Doamekpor, L. A. & Zuckerman, D. M. Diversity in medical device clinical trials: do we know what works for which patients? *Milbank Q.* **96**, 499–529 (2018).
18. Hammond, A., Jain, B., Celi, L. A. & Stanford, F. C. An extension to the FDA approval process is needed to achieve AI equity. *Nat. Mach. Intell.* **5**, 96–97 (2023).
19. Abernethy, A. et al. The promise of digital health: then, now, and the future. *NAM Perspect.* **2022** https://doi.org/10.31478/202206e (2022).
20. U.S. Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. https://www.fda.gov/media/145022/download?attachment (2021).
21. Mittermaier, M., Raza, M. M. & Kvedar, J. C. Bias in AI-based models for medical applications: challenges and mitigation strategies. *Npj Digital Med.* **6**, 113 (2023).
22. Arora, A. et al. The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **29**, 2929–2938 (2023).
23. Cary, M. P. et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Aff.* **42**, 1359–1368 (2023).
24. Ferrara, E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci* **6**, 3 (2023).
25. d'Elia, A. et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. *Fam. Med. Community Health* **10**, e001670 (2022).
26. Gurevich, E., El Hassan, B. & El Morr, C. Equity within AI systems: what can health leaders expect? *Health Manag. Forum* **36**, 119–124 (2023).
27. Thomasian, N. M., Eickhoff, C. & Adashi, E. Y. Advancing health equity with artificial intelligence. *J. Public Health Policy* **42**, 602–611 (2021).
28. Paik, K. E. et al. Digital determinants of health: health data poverty amplifies existing health disparities—a scoping review. *PLoS Digit Health* **2**, e0000313 (2023).
29. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
30. Topol, E. J. Welcoming new guidelines for AI clinical research. *Nat. Med.* **26**, 1318–1320 (2020).
31. Green, B. L., Murphy, A. & Robinson, E. Accelerating health disparities research with artificial intelligence. *Front. Digit. Health* **6**, 1330160 (2024).
32. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
33. Muralidharan, V., Burgart, A., Daneshjou, R. & Rose, S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *npj Digital Med.* **6**, 166 (2023).
34. Busnatu, Ș. et al. Clinical applications of artificial intelligence—an updated overview. *J. Clin. Med.* **11**, 2265 (2022).
35. 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).
36. Van Kolfschooten, H. The AI cycle of health inequity and digital ageism: mitigating biases through the EU regulatory framework on medical devices. *J. Law Biosci.* **10**, lsad031 (2023).
37. Joshi, G. et al. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. *Electronics* **13**, 498 (2024).
38. Berghea, E. C. et al. Integrating artificial intelligence in pediatric healthcare: parental perceptions and ethical implications. *Children* **11**, 240 (2024).
39. U.S. Food and Drug Administration (FDA). 510(k) summary: MEDO ARIA [Premarket notification submission K200356]. https://www.accessdata.fda.gov/cdrh_docs/pdf20/K200356.pdf (2020).
40. U.S. Food and Drug Administration (FDA). 510(k) summary: EarliPoint System [Premarket notification submission K213882]. https://www.accessdata.fda.gov/cdrh_docs/pdf21/K213882.pdf (2021).
41. U.S. Food and Drug Administration (FDA). De Novo summary (DEN200069): Cognoa ASD Diagnosis Aid. https://www.accessdata.fda.gov/cdrh_docs/pdf20/DEN200069.pdf (2020).
42. Shenoy, P. & Harugeri, A. Elderly patients' participation in clinical trials. *Perspect. Clin. Res.* **6**, 184–187 (2015).
43. Centers for Disease Control and Prevention (CDC), National Center for Health (NIH) Statistics. Older adult health. https://www.cdc.gov/nchs/fastats/older-american-health.htm (2024).
44. World Health Organization (WHO). Global report on ageism. https://iris.who.int/bitstream/handle/10665/340208/9789240016866-eng.pdf?sequence=1 (2021).
45. Rudnicka, E. et al. The World Health Organization (WHO) approach to healthy ageing. *Maturitas* **139**, 6–11 (2020).
46. Choudhury, A., Renjilian, E. & Asan, O. Use of machine learning in geriatric clinical care for chronic diseases: a systematic literature review. *JAMIA Open.* **3**, 459–471 (2020).
47. Bernard, M. A., Clayton, J. A. & Lauer, M. S. Inclusion across the lifespan: NIH policy for clinical research. *JAMA* **320**, 1535–1536 (2018).
48. Lau, S. W. et al. Participation of older adults in clinical trials for new drug applications and biologics license applications from 2010 through 2019. *JAMA Netw. Open.* **5**, e2236149 (2022).
49. Pitkala, K. H. & Strandberg, T. E. Clinical trials in older people. *Age Ageing* **51**, afab282 (2022).
50. U.S. Food and Drug Administration (FDA). 510(k) summary: NeuroRPM [Premarket notification submission K221772]. https://www.accessdata.fda.gov/cdrh_docs/pdf22/K221772.pdf (2022).
51. U.S. Food and Drug Administration (FDA). 510(k) summary: icobrain [Premarket notification submission K192130]. https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192130.pdf (2019).
52. U.S. Food and Drug Administration (FDA). 510(k) summary: Knee OsteoArthritis Labeling Assistant (KOALA) [Premarket notification submission K192109]. https://www.accessdata.fda.gov/cdrh_docs/pdf19/K192109.pdf (2019).
53. De Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Med.* **5**, 2 (2022).
54. Ndubuisi, N. E. Noncommunicable diseases prevention in low- and middle-income countries: an overview of health in all policies (HiAP). *Inquiry* **58**, 46958020927885 (2021).

55. Mathers, C. D. & Loncar, D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* **3**, e442 (2006).
56. *Global Action Plan for the Prevention and Control of Noncommunicable Diseases, 2013–2020* (World Health Organization, 2013).
57. Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S. & Barteit, S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *Npj Digital Med.* **5**, 162 (2022).
58. Tappero, J. W. et al. US Centers for Disease Control and Prevention and its partners' contributions to global health security. *Emerg. Infect. Dis.* **23**, S5–S14 (2017).
59. Farah, L. et al. Assessment of performance, interpretability, and explainability in artificial intelligence-based health technologies: what healthcare stakeholders need to know. *Mayo Clin. Proc. Digit. Health* **1**, 120–138 (2023).
60. Amann, J. et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLoS Digit. Health* **1**, e0000016 (2022).
61. Jeyaraman, M., Balaji, S., Jeyaraman, N. & Yadav, S. Unraveling the ethical enigma: artificial intelligence in healthcare. *Cureus* **15**, e43262 (2023).
62. Sounderajah, V. et al. Ethics methods are required as part of reporting guidelines for artificial intelligence in healthcare. *Nat. Mach. Intell.* **4**, 316–317 (2022).
63. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
64. Zhou, K. & Gattinger, G. The evolving regulatory paradigm of AI in MedTech: a review of perspectives and where we are today. *Ther. Innov. Regul. Sci.* **58**, 456–464 (2024).
65. U.S. Food and Drug Administration (FDA). 510(k) summary: Brainomix 360 e-ASPECTS [Premarket notification submission K221564]. https://www.accessdata.fda.gov/cdrh_docs/pdf22/K221564.pdf (2022).
66. Gilbert, S. et al. Algorithm change protocols in the regulation of adaptive machine learning-based medical devices. *J. Med. Internet Res.* **23**, e30545 (2021).
67. U.S. Food and Drug Administration (FDA). Summary of Safety and Effectiveness (SSED) Template. 3–16. https://www.fda.gov/media/113810/download (2024).
68. Abràmoff, M. D. et al. Considerations for addressing bias in artificial intelligence for health equity. *npj Digital Med.* **6**, 170 (2023).
69. Jain, A. et al. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. *JAMA Health Forum* **4**, e231197 (2023).
70. U.S. Food and Drug Administration (FDA). Premarket approval (PMA). https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-approval-pma (2019).
71. Tricco, A. C. et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann. Intern. Med.* **169**, 467–473 (2018).
72. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).

Author contributions

V.M., B.A.A., R.D. and T.O. were involved in conceptualization, study design and methodology. V.M., B.A.A., C.J.H., M.N., P.A., P.P., M.K.H., O.A., R.O.A., A.O.B., A.A., S.O. and Z.R.C. were responsible for screening data sources, data extraction and entry. B.A.A. and T.O. provided formal data analysis. V.M., B.A.A., C.J.H., R.D. and T.O. ensured data accuracy. V.M., B.A.A., C.J.H., M.N., P.A., P.P., M.K.H. and A.O.B. wrote the original draft of the manuscript. V.M., B.A.A., C.J.H., M.N., P.A. and P.P. revised and wrote the final draft. R.D. and T.O. revised the final draft and provided supervision for the project. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01270-x.

Correspondence and requests for materials should be addressed to Vijaytha Muralidharan.

Reprints and permissions information is available at http://www.nature.com/reprints