

Pitfalls and Best Practices in Evaluation of AI Algorithmic Biases in Radiology



Paul H. Yi, MD • Preetham Bachina, BS • Beepul Bharti, BS • Sean P. Garin, BS • Adway Kanhere, MSE • Pranav Kulkarni, BS • David Li, MD • Vishwa S. Parekh, PhD • Samantha M. Santomartino, BA • Linda Moy, MD • Jeremias Sulam, PhD

From the Department of Radiology, St Jude Children's Research Hospital, 262 Danny Thomas Pl, Memphis, TN 38105-3678 (P.H.Y.); Johns Hopkins University School of Medicine, Baltimore, Md (P.B.); Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Md (B.B., J.S.); Uniformed Services University of the Health Sciences, Bethesda, Md (S.P.G.); Institute for Health Computing, University of Maryland School of Medicine, Baltimore, Md (A.K., P.K.); Department of Medical Imaging, Western University Schulich School of Medicine & Dentistry, London, Ontario, Canada (D.L.); Department of Diagnostic and Interventional Imaging, McGovern Medical School at The University of Texas Health Science Center at Houston (UTHealth Houston), Houston, Tex (V.S.P.); Drexel University School of Medicine, Philadelphia, Pa (S.M.S.); and Department of Radiology, New York University Grossman School of Medicine, New York, NY (L.M.). Received June 8, 2024; revision requested July 29; final revision received December 12; accepted December 19, 2024. Address correspondence to P.H.Y. (email: Paul.Yi@stjude.org).

P.H.Y. and J.S. supported in part by the National Institutes of Health (R01CA287422).

Conflicts of interest are listed at the end of this article.

See also the editorial by Davis in this issue.

Radiology 2025; 315(2):e241674 • <https://doi.org/10.1148/radiol.241674> • Content codes:  

Despite growing awareness of problems with fairness in artificial intelligence (AI) models in radiology, evaluation of algorithmic biases, or AI biases, remains challenging due to various complexities. These include incomplete reporting of demographic information in medical imaging datasets, variability in definitions of demographic categories, and inconsistent statistical definitions of bias. To guide the appropriate evaluation of AI biases in radiology, this article summarizes the pitfalls in the evaluation and measurement of algorithmic biases. These pitfalls span the spectrum from the technical (eg, how different statistical definitions of bias impact conclusions about whether an AI model is biased) to those associated with social context (eg, how different conventions of race and ethnicity impact identification or masking of biases). Actionable best practices and future directions to avoid these pitfalls are summarized across three key areas: (a) medical imaging datasets, (b) demographic definitions, and (c) statistical evaluations of bias. Although AI bias in radiology has been broadly reviewed in the recent literature, this article focuses specifically on underrecognized potential pitfalls related to the three key areas. By providing awareness of these pitfalls along with actionable practices to avoid them, exciting AI technologies can be used in radiology for the good of all people.

© RSNA, 2025

Artificial intelligence (AI) and deep learning (DL) have generated excitement and optimism in radiology for their potential to transform the field of radiology. Potential uses of AI in radiology include automated diagnoses of complex diseases at medical imaging (1–3), triage of potentially actionable findings in the emergency department (4,5), and automated extraction of patient outcomes information from free-text reports (6). Despite the excitement around AI in radiology, troubling findings of algorithmic biases, or AI biases, among different demographic groups have been reported (7–11), with performance disparities disadvantaging historically underrepresented groups (Fig 1A). Subsequent studies have identified potential factors leading to these biases, such as the lack of demographic diversity in datasets used to train DL models (12,13) and the ability of DL models to predict patient demographics such as biologic sex and self-reported race (14–16) on the basis of images alone. Because these algorithmic biases frequently disadvantage historically underrepresented groups, they risk perpetuating—or exacerbating—pre-existing health inequities at scale.

Despite the growing awareness of problems related to fairness of AI models in radiology, evaluation of algorithmic biases is challenging for reasons that span both clinical and technical domains. From a clinical perspective, identifying biases is often difficult or impossible because medical imaging datasets frequently either do not report demographics or they report limited demographic information (eg, age and sex but not race) (8,12). From a technical perspective, although algorithmic fairness concepts are well defined in the machine learning communities (17–19), they are not always easily translated into clinical concepts important to

the radiology and medical communities because of differences in how bias is conceptualized between the groups. From a sociotechnical perspective, DL models can predict self-reported race from a medical image (14,15), which suggests a possible mechanism of biased DL models in radiology.

To guide the careful consideration and mitigation of biases in AI in radiology, this article summarizes potential pitfalls in the evaluation of algorithmic biases along with best practices to avoid these pitfalls and future directions to mitigate them across three key areas: (a) medical imaging datasets, (b) demographic definitions, and (c) statistical evaluations of bias (Fig 1B). Although recent articles (20–24) have broadly reviewed AI bias in radiology, this article focuses specifically on underrecognized potential pitfalls related to the three key areas. Despite being underrecognized, recognition of these pitfalls is critically important to ensure the safe and trustworthy use of AI in radiology.

Medical Imaging Datasets

Medical imaging datasets are the foundation for the development and evaluation of AI models in radiology (25), including for the evaluation of algorithmic biases. Both radiology societies (25–31) and independent research groups (32–35) have led efforts to curate and publicly release large medical imaging datasets to promote the development of AI models in radiology. Several AI competitions have been held on the basis of these datasets to crowdsource AI solutions for clinically important problems, ranging from diagnosis of pulmonary embolism (26) and identification of traumatic spine injuries (28) on CT scans to segmentation of brain tumors on MRI scans (36–38).

Abbreviations

AI = artificial intelligence, DL = deep learning

Summary

Evaluation of algorithmic biases, or artificial intelligence biases, is challenging in radiology due to incomplete reporting of demographic information in medical imaging datasets, variability in definitions of demographic categories, and inconsistent statistical definitions of bias.

Essentials

- Medical imaging datasets should report demographic variables, such as age, sex, race, and ethnicity, as a standard practice.
- Reporting demographic variables is necessary to provide measures of diversity of the patients represented in these datasets, as well as to facilitate measurements of bias in artificial intelligence (AI) in radiology.
- It is important to be precise and specific in the demographic definitions used to evaluate algorithmic biases, or AI biases, in radiology.
- Precise and specific demographic definitions in radiology ensure robust and valid conclusions regarding the presence and magnitude of any algorithmic biases.
- Statistical definitions of bias used in AI evaluations in radiology should be consistent with standard notions of demographic bias and chosen on the basis of specific clinical use cases and deployment settings.

These datasets have facilitated the development of state-of-the-art AI models for numerous clinical applications—often with accompanying descriptions of their technical designs and open-sourced code (39)—providing benefits for the clinical radiology

and AI research communities. Several datasets have further become the de facto reference standard benchmark datasets for radiology AI, such as the popular National Institutes of Health ChestX-ray14 (40), Stanford CheXpert (32), and MIMIC-CXR (34) chest radiograph datasets (9,41,42).

Pitfalls Related to Demographic Reporting in Medical Imaging Datasets

Recognizing the importance of these datasets for the rapid development of AI in radiology over the past several years, it is necessary to highlight potential fairness-related pitfalls related to the curation and assembly of medical imaging datasets. The first pitfall is whether patient demographics have been reported or collected in the first place. Demographic reporting is central to the evaluation and mitigation of biases in AI because it provides a summary of the diversity of data that might be used to train an AI model and allows for identification of biases through subgroup analyses. Although this consideration may seem obvious, medical imaging datasets frequently do not report demographics. A review of 23 publicly available chest radiograph datasets found that 17% and 26% of datasets, respectively, did not report demographics in any form in aggregate and at the image level (12). Reporting demographics at the image level is particularly relevant for bias evaluation, as aggregate demographic reporting does not allow for subgroup analyses to identify potential biases. A similar evaluation of the popular data science competition platform, Kaggle (43), identified 24 medical imaging datasets used to host data science competitions

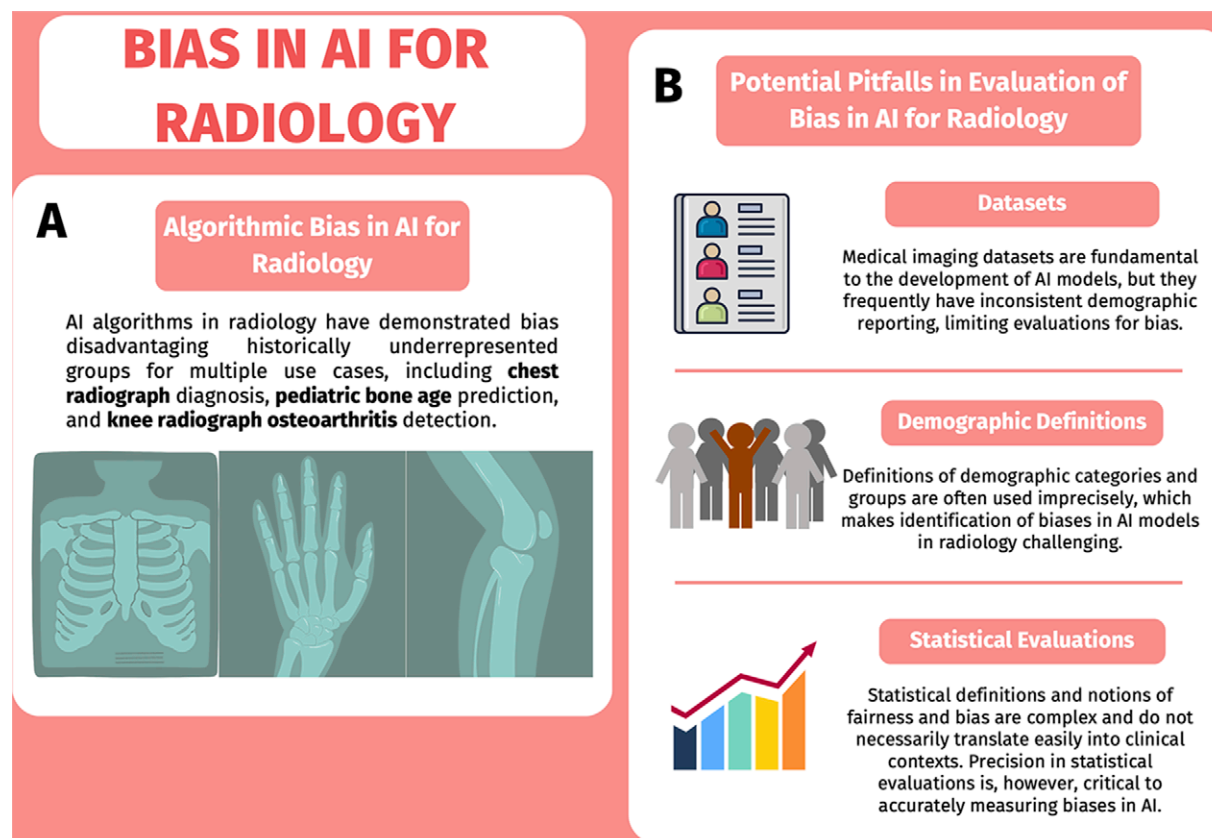


Figure 1: Diagrams of (A) bias and (B) evaluation of bias. (A) Algorithmic bias, or artificial intelligence (AI) bias, in radiology has been demonstrated for multiple use cases. (B) The evaluation of AI bias in radiology has several potential pitfalls related to datasets, demographic definitions, and statistical evaluations. The neural network graphics created by Loxaxs from Wikimedia Commons were modified under Creative Commons license (CCO 1.0).

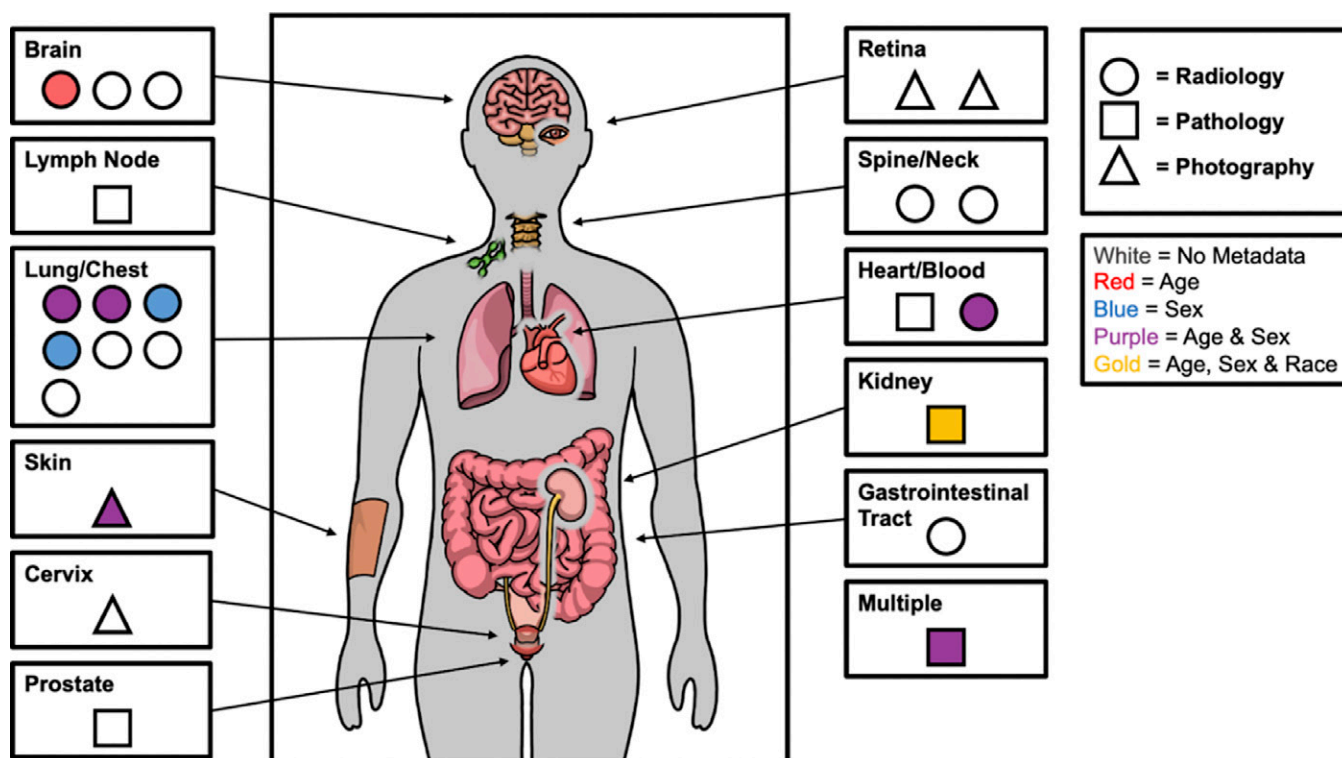


Figure 2: Figure illustrates demographic reporting practices of medical imaging datasets on Kaggle platform (<https://www.kaggle.com>) stratified by body part, imaging type, and types of demographic metadata reported. The majority of datasets did not report any demographics and those that did, reported age and/or sex only with the exception of one that reported age, sex, and race. Data are from Garin et al (43). Image courtesy of Sean P. Garin.

primarily focused on radiologic and pathologic imaging. Of these 24 Kaggle-hosted datasets, only nine reported any demographics (Fig 2).

Although reporting demographics in medical imaging datasets is critical for the proper evaluation of biases in AI models, it is equally important to determine which specific demographics should be reported. Real-world health disparities have been well documented in several diseases across numerous demographic groups, including age, sex, race, ethnicity, and socioeconomic status (44–47), frequently disadvantaging underrepresented minority groups. Similar disparities have been confirmed when using AI in radiology, with minority groups frequently having worse performance compared with majority groups for tasks ranging from chest radiograph disease classification (9,10,42,48) to cardiac MRI anatomy segmentation (49). Because these biases span multiple demographic groups, medical imaging datasets should ideally report a comprehensive set of demographic characteristics. Unfortunately, the status quo for demographic reporting has been suboptimal. A review of 23 public chest radiograph datasets (12) found that the datasets frequently reported age and sex (83% and 78% of datasets, respectively), but far fewer reported race or ethnicity (13% of datasets) or health insurance (4%)—a proxy for socioeconomic status. Similar findings were reported by Garin et al (43) in their review of Kaggle medical imaging datasets, which found that nine of 24 datasets reported demographic information; of these, five reported age and sex, two reported sex only, one reported age only, and one reported age, sex, race, and ethnicity (Fig 2). Although these findings highlight pitfalls in pre-existing imaging datasets, they may reflect the fact that medical journals do not necessarily mandate specific demographic

reporting beyond age and sex (50), although the reporting of other relevant variables, such as race and ethnicity, are encouraged (51). This notion is corroborated by work showing that AI research articles in radiology journals frequently do not report sociodemographic variables, including race or ethnicity (52). An important consideration for answering the question of what demographic variables should be reported is how these variables should be defined, which will be covered in the subsequent section.

Because medical imaging datasets are frequently obtained as a convenience sample (20) (eg, consecutive series of patients presenting to a hospital), they will likely have some degree of imbalance in the representation of demographic groups that reflects the underlying population and/or pre-existing health inequities in access to care (46,47,53). For example, a study evaluating racial bias of a cardiac MRI anatomy segmentation DL tool used a dataset of 5903 cardiac MRI scans from the UK Biobank, which consisted of 81% White patients; this percentage mirrors the representation of White patients in England and Wales from which the UK Biobank data are drawn (54). Ideally, medical imaging datasets would balance demographic representation by design whenever possible. In a similar vein, potential confounding non-demographic variables may be present in medical imaging datasets that reflect underlying differences in health outcomes or access to health care between different demographic groups. These potential confounders can include scanner brand and model (55), department where the image was obtained (eg, inpatient vs outpatient) (55), radiographic views (56), hospital where the imaging was performed (57), and disease prevalence (57). Strikingly, these variables can be predicted by DL models based on medical images alone through uncanny mechanisms. DL models are able

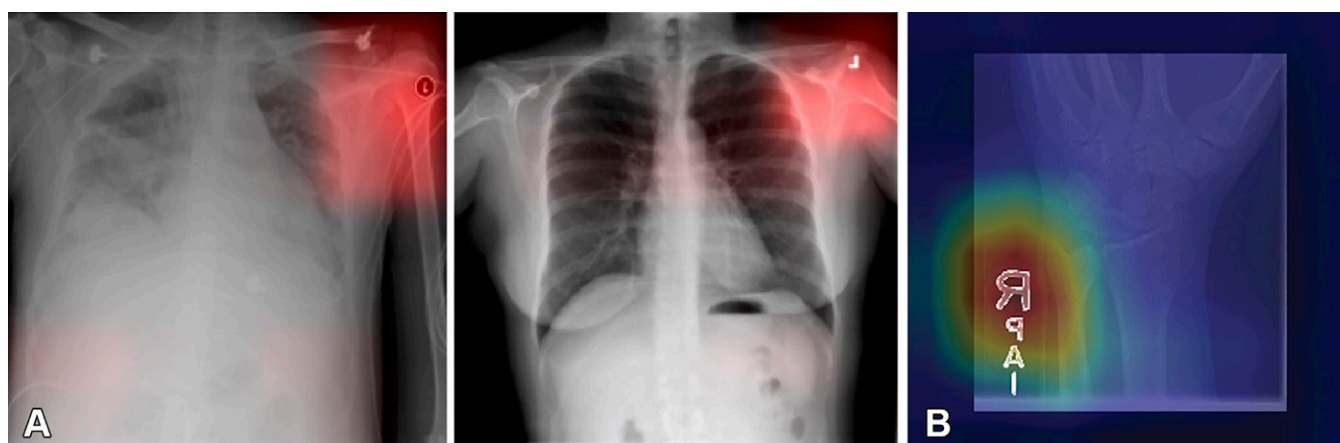


Figure 3: (A) Images from a deep learning (DL) model in radiology that can learn to identify confounding features related to bias and unfair predictions, including laterality markers (image annotations indicate the side of the body being viewed [right vs left]) to identify the hospital at which a chest radiograph was obtained. Images adapted and reprinted from reference 57, an open-source article, published under the Creative Commons license (CC BY 4.0). (B) Image from a DL model that can make a diagnosis of radiographic abnormality on extremity radiographs, also known as shortcut learning. Reprinted, with permission, from reference 59.

to identify the hospital or site at which a radiograph was obtained by using laterality markers (57), image annotations indicating the side of the body being viewed (right vs left). These laterality markers can be used as shortcuts (57–60) to make the right diagnoses for the wrong reasons (ie, learning to make a diagnosis based on spurious correlation [eg, laterality marker], rather than the disease or condition itself) (Fig 3). Evaluations for biases in AI models should account for known confounding factors to ensure that conclusions about fairness of AI models are rigorously established (61).

Technical Solutions to Mitigate or Augment Imperfect Datasets

Although these pitfalls related to datasets stem from clinical realities and challenges, recent technical advances may provide solutions to mitigate their negative impact on the development of biased AI models and to evaluate for AI biases in the absence of demographic information (Fig 4). When demographic variables are not reported, the use of pseudolabels for demographic variables provided by DL models previously trained to predict demographics (14,15,62,64) is a promising approach to provide estimates of potential bias in AI models (Fig 4A). Pseudolabels, or predictive labels, are labels assigned to unlabeled data from the outputs of a trained AI algorithm. These pseudolabels can be applied to statistical methods for providing mathematically provable guarantees on the degree of disparities between demographic groups (62) (Fig 4B). However, even in the best case, where demographics and confounding variables are reported, achieving equal performance of a predictive algorithm for all demographic groups and all possible confounding variables will be generally impossible.

Generative AI may be able to help. Generative AI is a set of machine learning techniques that are able to sample—or generate—new data from a specific distribution, including images, text, and video. Generative AI approaches to image synthesis (65) provide the potential to create synthetic imaging datasets with more balanced representation of these demographic and confounding variables. Recent work by Ktena et al (63) demonstrated that using generative AI diffusion image synthesis models to augment medical imaging datasets with demographic imbalances, disease prevalence,

and other potential confounding features resulted in better downstream DL models. These models were more robust and had smaller demographic-based performance disparities compared with DL models trained on the real images alone (Fig 4B). Although these technical approaches are early in their development, they indicate promise for the use of advanced AI and statistical methods to overcome pitfalls related to measurement of AI biases in the setting of imperfect datasets.

Avoiding Pitfalls Related to Medical Imaging Datasets: Best Practices and Future Directions

Best practices and future directions to avoid the pitfalls related to medical imaging datasets are summarized in Table 1. These best practices include collecting and reporting as many demographic variables and common confounding features as possible and collecting and sharing raw imaging data without institution-specific postprocessing, whenever feasible. Future directions are also presented to guide future research and curation of medical imaging datasets.

Pitfalls Related to Demographic Definitions

Although demographic variables are a standard reporting element for datasets and research studies in radiology (50,66), heterogeneity in how these variables are defined are an important and insidious pitfall for the evaluation of AI biases in radiology (Fig 5). Many demographic categories such as gender and race are not biologic variables (51,67) but are self-identified characteristics informed by many factors, including society and lived experiences. They are nonetheless how biases and health disparities are measured and acted upon, for example, by allocating more health care resources toward underrepresented groups.

Sex and Gender

The terminologies used for demographic concepts in radiology AI research can be imprecise and/or confusing. For example, the terms *sex* (biologic) and *gender* are often used interchangeably, despite representing two different concepts (67) (Fig 5A). *Sex* is a biologic category defined by a person's genetic chromosomal makeup (eg, XX or XY chromosome for male and female,

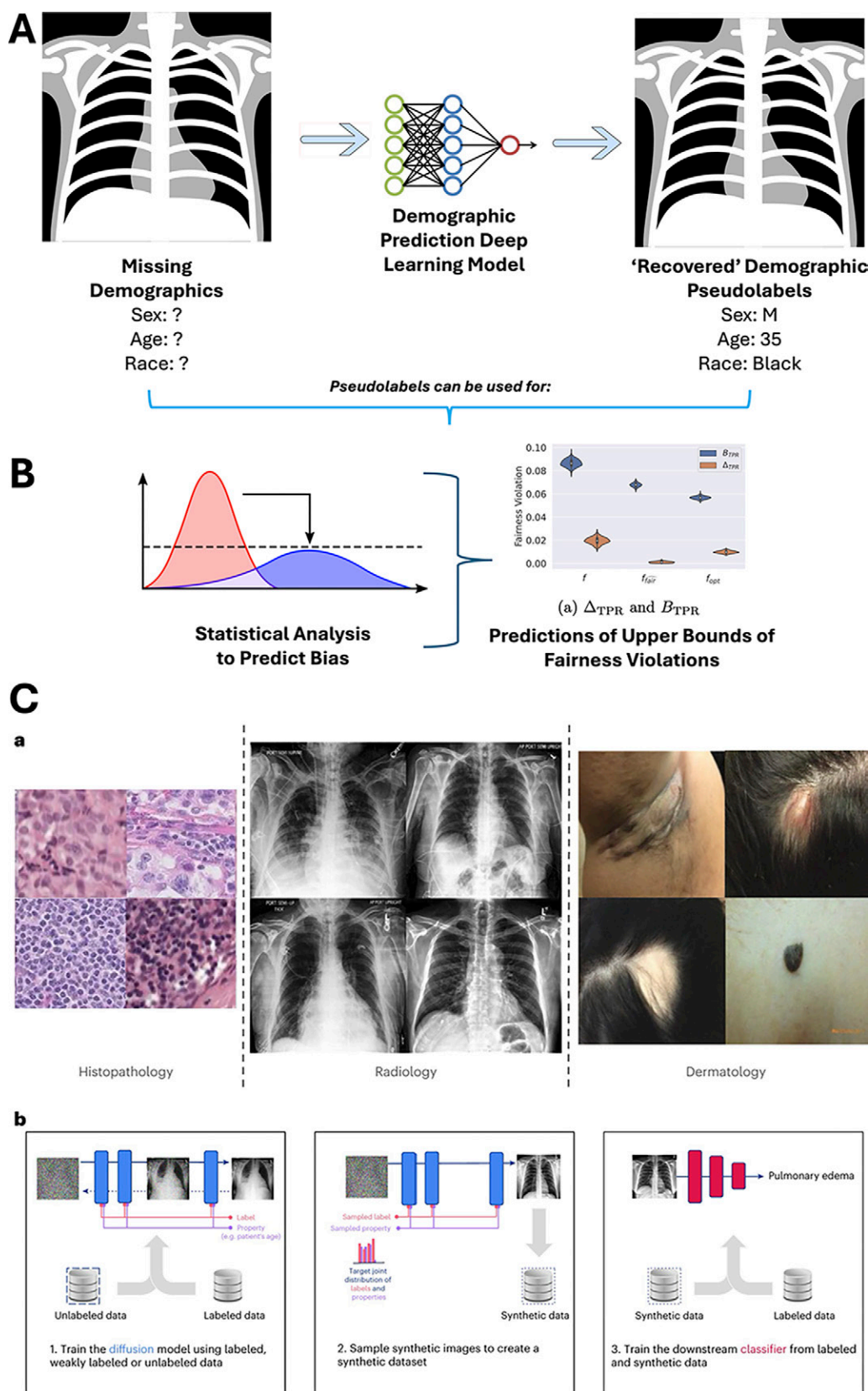


Figure 4: Technical approaches to addressing pitfalls in dataset limitations for the evaluation of artificial intelligence (AI) bias in radiology. **(A)** Deep learning models trained to identify patient-reported demographics on medical images can be used to “recover” pseudolabels for demographics, which allow for estimates of dataset diversity and potential biases. The neural network graphic (middle graphic) was modified under Creative Commons license (CCO 1.0) and the chest radiograph graphics (right and left graphics) were created by Jmarchn from Wikimedia Commons, under Creative Commons license (CC BY-SA 3.0). **(B)** These pseudolabels (labels assigned to unlabeled data from the outputs of AI algorithms trained to predict that label) can be used in conjunction with advanced statistical methods to predict upper bounds for the degree of fairness violations and performance disparities for an AI model tested on a dataset even in the absence of demographic labels. Graph on the left is a free image from Rawpixel, licensed under a Creative Commons license (CCO 1.0) (<https://www.rawpixel.com/>). Chart on the right is adapted and reprinted from reference 62, an open-source article, published under the Creative Commons license (CC BY 4.0). **(C)** Generative AI models (AI trained to generate new data, including images, text, and video) can be used to create synthetic medical images to augment datasets, which can be used to train subsequent disease classification AI models that have decreased fairness disparities. Reprinted from reference 63, an open-access article, published under Creative Commons license (CC BY 4.0).

respectively), while *gender* is an individual's self-identification as male, female, or nonbinary (67); accordingly, an individual's sex can be biologically male but their gender can be female. Nevertheless, the terms *sex* and *gender* have been used interchangeably in AI radiology research. A landmark study by Larrazabal et al (10) used the Stanford CheXpert dataset to conclude that

gender-imbalanced datasets lead to biased DL chest radiograph disease classifiers, despite the fact that CheXpert defines the male and female variable as biologic sex (68). MIMIC, a widely-used public multimodal medical dataset (69), which includes chest radiographs (34), defines its demographic variable *gender* as “the genotypical sex of the patient” (70). Recognizing the importance

Table 1: Best Practices and Future Directions for Avoiding Pitfalls Related to Medical Imaging Datasets for Evaluation of Fairness of AI in Radiology

Best Practice Recommendations	Future Directions (Open Questions)
When assembling medical imaging datasets, intentionally collect and report demographic information as allowed within patient privacy regulations, such as the Health Insurance Portability and Accountability Act, or HIPAA.	What is the best way to standardize reporting of demographics in medical imaging datasets, especially considering that demographic categorizations can differ between societies?
Collect as many demographic variables as possible, with a suggested minimum set including age, sex and/or gender, race, and ethnicity.	How useful and accurate are AI-generated demographic pseudolabels* to provide estimates of diversity in datasets that do not have demographic information reported for use cases beyond chest radiographs (62)?
Collect common confounding features in medical imaging datasets, such as imaging view, hospital site, inpatient versus outpatient imaging, and scanner brand and model.	How useful and reliable are generative AI† tools to augment datasets with synthetic images to balance unavoidable demographic imbalances?
Whenever possible, collect and report raw imaging data, without location and/or institution-specific postprocessing.	How should we identify and account for nondemographic confounding features to prevent shortcut learning and downstream bias in AI models in radiology?

Note.—AI = artificial intelligence.

* Pseudolabels are labels assigned to unlabeled data from the outputs of AI algorithms trained to predict that label.

† Generative AI is AI trained to generate new data, including images, text, and video.

of gender inclusivity for society and for radiology research and practice (71,72), it is critical that AI researchers and radiologists evaluating AI models for biases use the proper terminology to refer to sex and gender.

Race and Ethnicity

Similar to sex and gender, race and ethnicity are frequently conflated as being the same category (9) (Fig 5B), despite their being two separate social constructs that have meant different things at different times, and which continue to evolve (73). Broadly speaking, *race* refers to broad categories generally based on ancestry and physical characteristics, whereas *ethnicity* refers to one's cultural identity based on things such as language, customs, and religion (51,74). It is important to understand that concepts of race and ethnicity do not necessarily translate outside of a specific society (51); unless otherwise stated, the discussions of race and ethnicity in this article will be based on the U.S. concepts of these demographic categories for illustrative purposes.

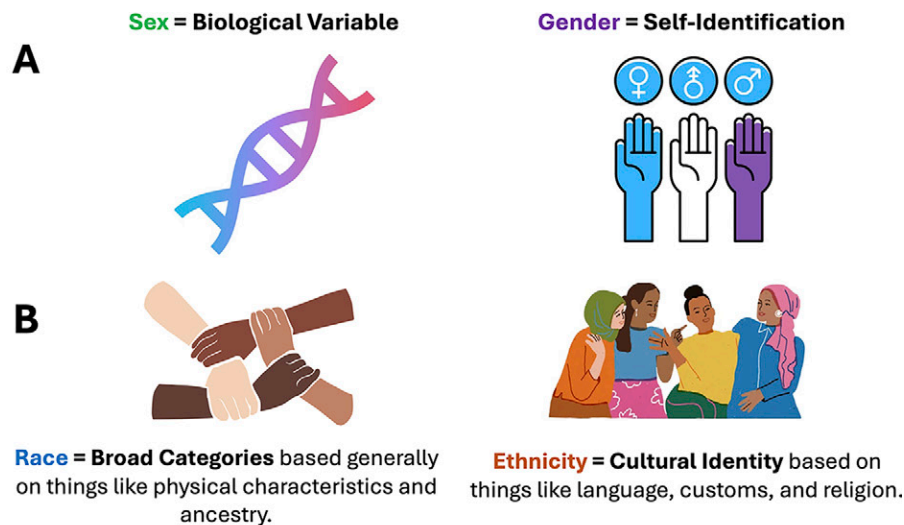
The U.S. Census categorizes race into six groups (75): White, Black, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and Some Other Race (for people who do not identify with a single race), noting that individuals can also select more than one racial group (eg, for people who identify as biracial); and ethnicity into two categories: Hispanic or Latino or Not Hispanic or Latino. Despite this well-defined distinction between race and ethnicity, AI biases have been evaluated in radiology research by combining race and ethnicity into a single group (eg, considering Hispanic or Latino ethnicity as a single race/ethnicity category [7,9,42] or a single ethnicity category [15]). Of note, prior studies have used the provided race/ethnicity labels already present in widely used public datasets (34,76), which underscores the importance of addressing the proper collection of demographic data at the dataset curation stage. In addition to respecting the racial and ethnic self-identities of people, ensuring accurate measurements of race- and/or ethnicity-based biases in AI models is important to allow for apples-to-apples comparison of bias evaluations, and also because these serve as recommendations upon which health policy decisions are made. If conclusions

about the presence and degree of race- or ethnicity-based biases are erroneously made because of variable conceptualizations of these demographics, then the resultant health policies could be made in error, potentially perpetuating pre-existing inequities.

The real-world impact of algorithmic bias evaluation is illustrated in work by Obermeyer et al performed in 2019 (77), which identified racial bias in a health risk prediction algorithm used to allocate health insurance resources. This algorithm systematically disadvantaged Black patients compared with White patients; specifically, for the same algorithmic risk score, Black patients were much sicker than White patients (77). These biases were determined to result from the algorithm predicting health care costs as a proxy for illness (77). These findings underscore the reality that structural biases are frequently hidden in datasets, which are reflective of societal disparities; for example, disparities due to unequal access to imaging centers between demographic groups (53,78,79) or differences in disease prevalence or severity (46,57). This work led to real-world change through the health insurance company whose algorithm was audited as the company worked with the study authors to retrain this algorithm to be less biased (77). Real-world change also resulted from influencing lawmakers (80) and regulators (81) to hold health care insurance companies and hospitals accountable by using commercial health care prediction tools in ways that are safe and equitable.

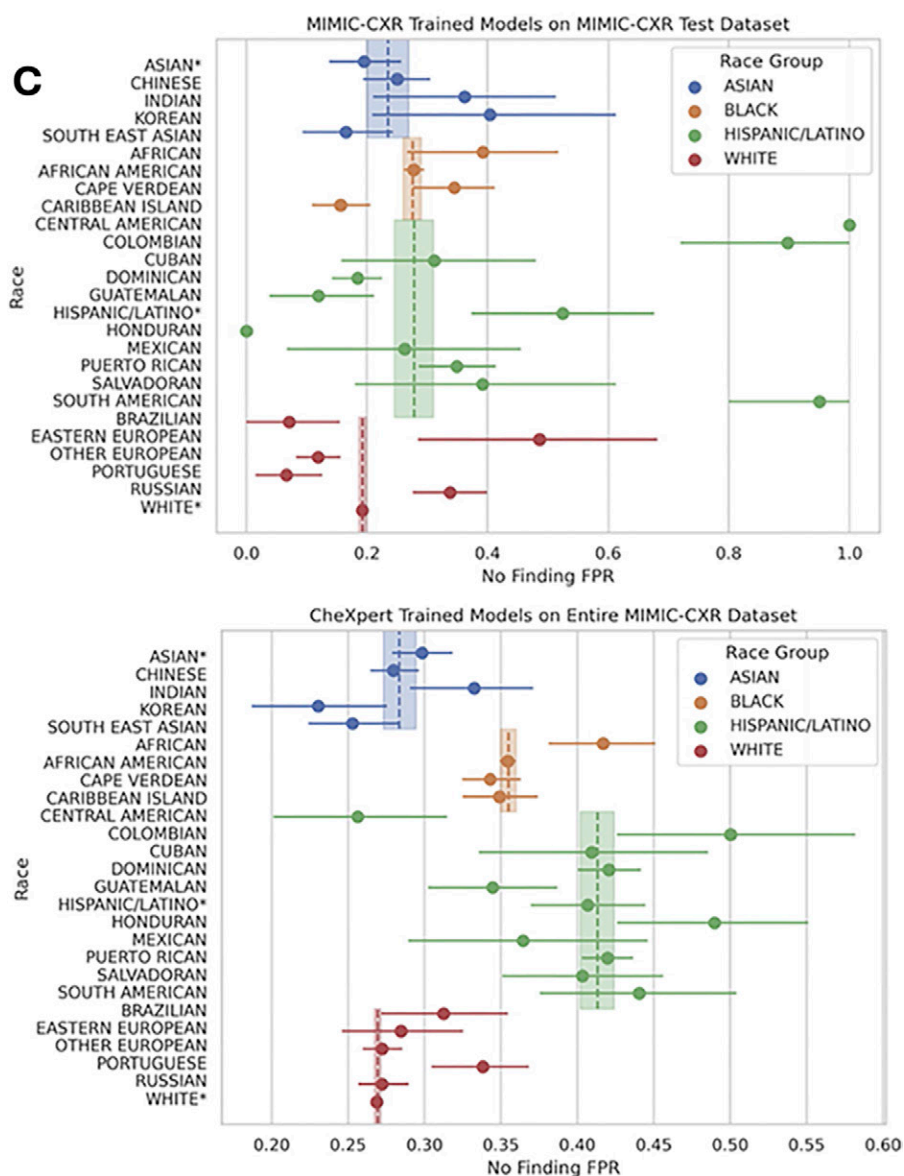
Although they may seem like static concepts, demographic categories are fluid, having had different meanings in different societies and times in history (51), and they continue to evolve. In 2023, motivated by the recognition of the inadequacies of current coarse race categories to represent the diversity in the U.S. population (82), the Biden administration proposed major changes to the demographic conceptions of race and ethnicity with the additions of new race categories to include Middle Eastern or North African—separating this group from the historical White category—and listing Hispanic or Latino as a new race category (73). Moving toward more granular concepts of race and ethnicity is a goal that is not only socially conscious and beneficial, but one that has important health care implications in evaluations of bias in AI. Although race labels such as those used by the U.S. Census are

Figure 5: Demographic definitions are often used imprecisely but have specific semantic distinctions and meanings that are critical for the evaluation of artificial intelligence (AI) bias in radiology. **(A)** Male, female, and other related categories can fall under sex and/or gender, which are two separate categories. **(B)** Similarly, race and ethnicity are often conflated, but represent two distinct concepts and categories. Using granular ethnicity labels (eg, Korean or Indian) can help identify clinically meaningful performance disparities in AI models in radiology that can go hidden when measuring such biases using coarse race labels (eg, Asian). **(C)** Forest plots show granular underdiagnosis rates. In this example, there are several hidden underdiagnosis disparities identified within each coarse racial group when evaluating granular ethnicity labels that often exceeded the variation between coarse racial groups. Granular groups labeled with an asterisk are the patients who only reported a coarse race or ethnicity. FPR = false-positive rate. Reprinted, with permission, from reference 8.



a convenient way to categorize people loosely based on supposed area of genetic ancestry, they are imperfect and coarse as they broadly categorize large swaths of people with considerable diversity. For example, the race group *Asian* encompasses an entire continent with ethnic subgroups, such as Indian, Chinese, and Korean (8,83). Importantly, these coarse racial labels can mask clinically meaningful medical differences, such as diabetes being nearly twice as common in U.S. adults of Indian descent compared with those of Chinese descent (84).

When evaluating for AI biases in radiology, similar masking of clinically meaningful performance disparities can occur if the differences between more granular ethnic groupings are not considered. Bachina et al (8) previously evaluated whether coarse race labels mask underdiagnosis disparities that exist between more granular ethnicity labels for state-of-the-art DL chest radiograph disease classification models. They first confirmed previous findings that DL models trained on two large U.S. datasets had higher underdiagnosis rates in non-White patients compared with White patients (9). They then showed that coarse race labels masked real underdiagnosis disparities between granular ethnic groups (Fig 5C). Moreover, these variations in underdiagnosis bias *within* a coarse racial group frequently exceeded those *between* coarse labels. For example, a DL model trained on the CheXpert dataset (32) had underdiagnosis rates ranging from 23.1% (Korean) to 33.2% (Indian) compared with 23.5% for



the coarse Asian category (8). Similar findings have been confirmed for clinical risk prediction models using tabular data (83). These findings highlight the importance of evaluating for

Table 2: Best Practices and Future Directions for Avoiding Pitfalls Related to Demographic Definitions for Evaluation of Fairness of AI in Radiology

Best Practice Recommendations	Future Directions (Open Questions)
Be specific and precise in the terminologies used for defining demographic groups (eg, sex and gender are not the same thing). Do not conflate separate but related demographic categories (eg, race and ethnicity). Ensure that conventions used are consistent with those appropriate for a specific society, time, and place, with considerations for both individual lived experiences and conventions used to inform health policy. Be aware of changing societal norms for demographic identification and adapt evaluations of bias in AI in radiology with current categorizations.	How do we best measure biases reflective of potential differences in genetic ancestry when using coarse labels of race and more granular, yet still coarse, labels of ethnicity (8)? How do we compare results of studies or evaluations of AI in radiology claiming conclusions about bias in the setting of heterogeneous categorizations used for various demographics? What is most important for assessing bias in AI—individual’s identities, which inform people’s lived experiences, or government-defined categorizations, which may inform health policy and regulation?

Note.—AI = artificial intelligence.

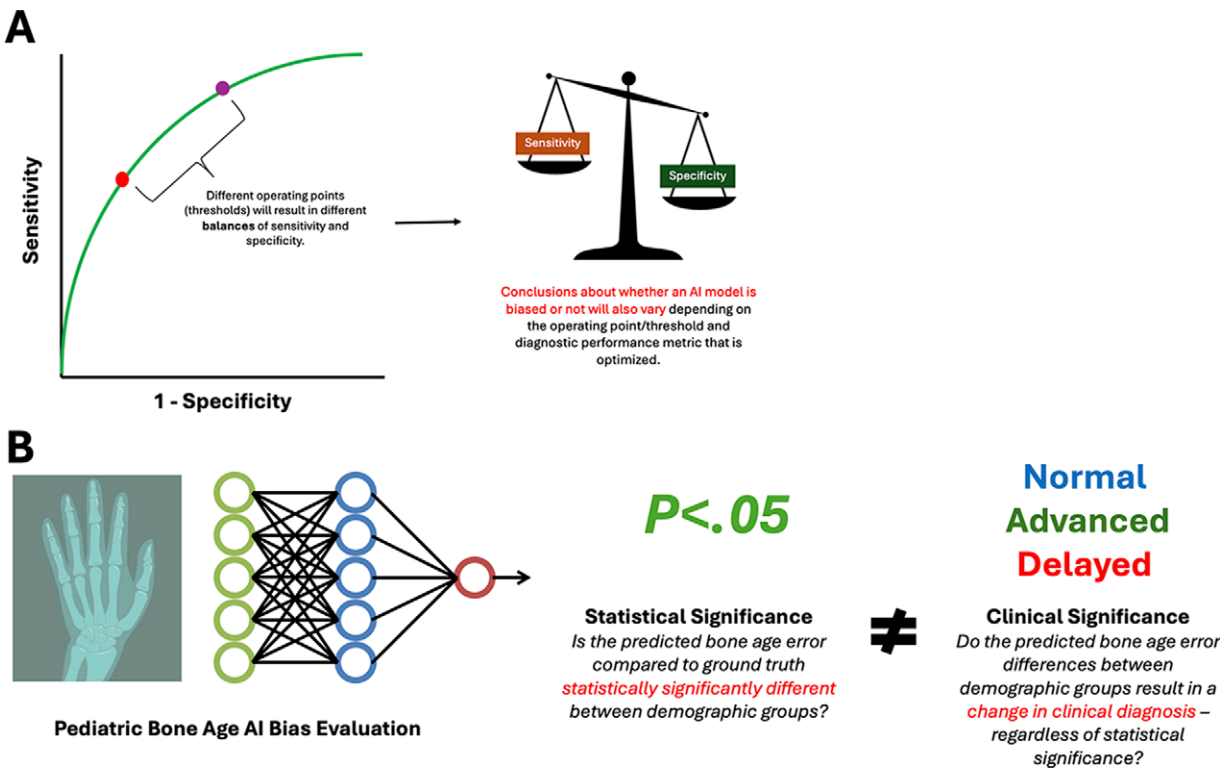


Figure 6: Statistical evaluations of artificial intelligence (AI) bias in radiology have several pitfalls and considerations to ensure clinically relevant conclusions are drawn. These include **(A)** recognizing the paradox of the incompatibility of fairness metrics, where different fairness metrics cannot be fulfilled simultaneously, analogous to how a receiver operating characteristic curve requires choice of threshold points that have trade-offs between sensitivity and specificity, and **(B)** distinguishing between statistical and clinical significance when evaluating measured biases. The neural network graphic created by Loxaxs from Wikimedia Commons was modified under Creative Commons license (CC0 1.0).

biases in AI using demographic categories that best represent the identities of individuals, as well as their potential unique risks for disease that may be related to their underlying genetic ancestry. As demographic concepts evolve in society, radiologists—and regulatory entities—evaluating AI for biases will also need to adapt their evaluations accordingly.

Avoiding Pitfalls Related to Demographic Definitions: Best Practices and Future Directions

Best practices and future directions to avoid these pitfalls related to demographic definitions and conventions are summarized in Table 2. These include being specific with demographic terminologies used, avoiding conflation of separate but related

demographic categories such as race and ethnicity, and applying these categories in accordance with societal norms. Future directions are presented to guide specific and precise evaluations of demographic bias in AI in radiology.

Pitfalls Related to Statistical Evaluations of Bias

Evaluating for demographic bias in AI models in radiology is critically important to promote the safe and equitable use of these technologies in clinical care. At the same time, the difficulty in measuring these biases is increased by an existing gap between technical and/or statistical and clinical domains (Fig 6). Although definitions of bias and fairness have been well-established and accepted for years by the statistics and machine learning

Table 3: Best Practices and Future Directions for Avoiding Pitfalls Related to Statistical Evaluations of Bias and Fairness of AI in Radiology

Best Practice Recommendations	Future Directions (Open Questions)
Use standard and well-accepted notions of demographic bias evaluation based on clinically relevant comparisons of AI model performance between different demographic groups.	How will conclusions reached in research evaluations of bias in AI models impact or be used to inform health policy locally (eg, at a hospital level), nationally (eg, U.S. Food and Drug Administration), and internationally? Accordingly, how should these evaluations be performed, and by whom?
Choose fairness metrics that are specific and most relevant to the clinical use case and deployment setting of interest.	How should standards for statistical evaluation of AI bias in radiology be operationalized in a way that achieves consensus between the technical and clinical communities?
Be mindful that different operating points of a predictive model will result in different performance, and thus potentially different demographic biases. Ensure that such operating points and thresholdings are documented in research and by vendors providing commercial AI products.	Given that metrics of demographic bias are in general incompatible, how can clinicians and computer scientists alike be best guided to choose the right metric for the right clinical context?
Do not conflate statistical significance with clinical significance. Ensure that any conclusions about bias are placed into clinical context (ie, how will the differences or disparities identified translate into clinical impact?).	How do we best define clinically relevant metrics and model downstream clinical impact of biased AI in radiology?

Note.—AI = artificial intelligence.

communities (17–19), these are typically general notions that were not conceptualized with radiology or medicine in mind. Even the terminologies used for statistical concepts that radiologists take for granted can be different. In fact, the term *bias* can have different meanings. In statistics, *bias* refers to a discrepancy between the expected value of an estimated parameter and its true value (20,21,23). In the context of statistical learning, *bias* toward a certain type or class of solutions, or predictors, is necessary for any learning task. However, in this article, the term *bias* is used in the context of demographic fairness, which reflects differences in metrics between different demographic groups (7–9).

Evaluating for biases or disparities in health care outcomes between different demographic groups is not unique to AI but is a standard part of all clinical research, whereby two or more groups are compared for a specific outcome in relation to a single intervention, event, or risk factor of interest. For example, an evaluation of racial and ethnic disparities in COVID-19 disease severity on chest radiographs statistically compared chest radiograph severity scores between Black and White patients (46). Applied to AI models, evaluations for biases are performed via subanalyses according to the demographic group for an AI model of interest (7–9,11,42,85). For instance, an evaluation of race-based biases of a commercial AI product for digital breast tomosynthesis cancer screening compared false-positive rates between different race groups (11). This notion of bias is also well accepted by the statistics and machine learning communities (comparing one AI model on different test sets consisting of demographic groups of interest) (17). Despite this notion of clinical AI bias as evaluations of differences in model performance between demographic groups, a major pitfall is using a different notion of bias.

A seminal article by Larrazabal et al (10) published in 2020 systematically evaluated the impact of male-to-female imbalance in training datasets on the development of biased chest radiograph DL disease classifiers. However, the notion of bias used in this study is different from standard notions of bias. Rather than comparing the performance of DL models trained on datasets with varying degrees of male-to-female imbalance on test sets of men versus women, this study compared the performance of two

DL models (trained with different levels of representation) on one test set at a time (men or women). Conclusions about disparities in the performances of these models do not necessarily indicate bias in the notions defined previously, because they might simply indicate that one model outperforms the other in general (regardless of the sex of the individuals in the test set).

The more precise way to evaluate for bias would be to evaluate differences in performance of a single AI model tested on images from men versus women (9,42,85). This is a standard notion of demographic fairness that allows for the auditing of a model for disparities in performance for different demographic groups. Ensuring that standard notions of bias are used when evaluating AI in radiology is critical because the conclusions reached may be different depending on the specific comparisons performed, with implications for downstream health policy and local hospital AI deployment decisions. Moreover, while the notions of bias used by Larrazabal et al (10) are difficult—if even possible—to mitigate, definitions of demographic bias based on performance metrics of a given model when tested on different demographic groups are easier to correct and alleviate by means of statistical tools that are much better understood (18,19,62).

Pitfalls Related to Statistical Evaluations of Fairness of AI

Once standard notions of bias are established, additional pitfalls exist for the statistical evaluations of AI biases. Chief among these is the concept of the incompatibility of *fairness metrics* (17,19,86). This is the observation that different notions of fairness are frequently incompatible and cannot be satisfied simultaneously; therefore, there is no universal fairness metric that can be applied to all use cases and problems (19). There are several different fairness metrics that emphasize different statistical performance metrics, which have been previously described in detail (20); for example, *demographic parity* (18,87) evaluates whether the predictions made by an AI model are independent of a sensitive (eg, demographic) attribute, whereas *equalized odds* (18,88) evaluates whether true-positive rates and false-positive rates are equal between different groups. Because specific fairness metrics emphasize different characteristics of a diagnostic test, an AI model that

Table 4: Suggested Courses of Action to Mitigate Demographic Biases in AI in Radiology

Course of Action	Problems Addressed
Form a consensus panel to define and continuously update standards for reporting demographics and other patient characteristics that can result in protected demographic groups in different contexts.	Improving dataset reporting of demographics.
Develop a standardized framework to identify and address potential nondemographic confounding features that could contribute to algorithmic biases, such as clinical site (eg, inpatient vs outpatient) and scanner type and/or model.	Improving dataset reporting of potential nondemographic contributors to biases in AI in radiology.
Develop a standard lexicon of terminology for concepts of fairness and AI bias in radiology.	Ensuring standardized terminology for research and discussion about demographic biases in AI in radiology.
Develop standardized statistical evaluation frameworks for evaluation of demographic bias of AI algorithms in radiology grounded in clinical contexts.	Standardizing the statistical evaluation of AI biases in radiology in clinically meaningful ways.
Create checklists for AI research manuscripts' reporting of key elements relevant to evaluation and mitigation of demographic biases in AI in radiology, including reporting of demographic information, statistical definitions of algorithmic fairness and bias employed and their justification, and discussion of subgroup analyses.	Facilitating reproducible, transparent, and rigorous scientific manuscripts using AI in radiology with regards to demographic fairness.

Note.—AI = artificial intelligence.

is fair under one definition of fairness may not be fair simultaneously under another definition of fairness (except under stringent special cases). For example, if an AI model has equal sensitivity between old and young patients and is, therefore, fair and unbiased based on sensitivity, it may not necessarily be fair when comparing specificity (Fig 6A). This is a reality of the well-known tradeoffs of sensitivity and specificity of any diagnostic or predictive tool, which can be illustrated by the ubiquitous area under the receiver operating characteristic curve (89,90). Furthermore, because predictive models must choose an operating point, or threshold, on the receiver operating characteristic curve, different operating points used to emphasize or optimize certain diagnostic measures (eg, optimizing for specificity for a confirmation examination) may result in different conclusions about bias or fairness in the same AI model. Conclusions about whether an AI model is biased or not should also consider the impact of different thresholds.

These statistical considerations must ultimately be made with the clinical context and use case in mind, with the goal of preventing harms while maximizing benefits (17). In addition to considering the statistical performance metrics that are most desirable in an AI tool (eg, sensitivity for a screening examination), radiologists evaluating AI must consider the clinical relevance of specific metrics beyond mere statistical significance in the appropriate clinical and societal context. A special remark is also warranted regarding differences between statistical and clinical significance (91). Indeed, a difference in predictive performance among different groups, even if statistically significant, might not imply a change in the clinical outcome or the patient's treatment (Fig 6B).

For example, for pediatric bone age prediction, a significant error in predicted bone age may not necessarily result in any clinical difference in terms of the diagnosis rendered for normal, advanced, or delayed skeletal maturity (7). The converse is also possible. Beheshtian et al (7) previously evaluated biases in an award-winning bone age DL model and found that although the differences in mean absolute difference of predicted bone age compared with the reference standard between men and women were not statistically significant, there were higher proportions of clinically significant errors (ie, those that would change the

diagnosis of skeletal maturity as normal, advanced, or delayed) in women compared with men. This difference in conclusions about bias highlights the importance of the use of clinically relevant metrics to assess bias in AI models, since the clinical impact is the ultimate outcome of interest for AI models in radiology. Whether considering high-level notions of algorithmic bias, specific statistical definitions and metrics, and clinically relevant metrics for use cases, it is critical that AI models be evaluated with the right metric for the right context.

Avoiding Pitfalls Related to Statistical Evaluation of Biases and Fairness: Best Practices and Future Directions

Best practices and future directions to avoid these pitfalls related to the statistical evaluation of AI biases and fairness in radiology are summarized in Table 3, including the use of standard statistical notions of demographic biases, choosing fairness metrics most relevant to a particular clinical use case, and not conflating statistical significance with clinical significance.

Suggested courses of action to mitigate demographic biases and improve future statistical evaluations of AI bias in radiology are summarized in Table 4.

Conclusion

Despite the growing awareness of fairness problems of artificial intelligence (AI) in radiology, evaluation of algorithmic biases, or AI biases, remains challenging. To guide the evaluation of demographic biases in AI in radiology, this article summarizes best practices to identify and mitigate potential pitfalls in evaluation of algorithmic biases related to medical imaging datasets, demographic definitions, and statistical evaluations of bias. This article also provides future directions with open questions for further research and suggested initial courses of action to mitigate demographic biases. These future directions include establishing standards for dataset demographic reporting and statistical evaluations of AI biases that are clinically relevant, as well as developing and validating technical methods for mitigating algorithmic biases; for example, using generative AI to improve dataset diversity and reconciling tensions that result from complexities that are technical (eg, incompatibility

of fairness metrics), clinical (eg, how do we best define clinically relevant fairness metrics), and social (eg, which demographic categorizations should we use for a given society). Although AI bias in radiology is worrisome, by being aware of these pitfalls and ways to mitigate them, we will be better equipped to use these promising technologies for the benefit of all people.

Deputy Editor: Vicky Goh

Scientific Editor: Sarah Atzen

Disclosures of conflicts of interest: **P.H.Y.** Grants or contracts from National Cancer Institute, National Institutes of Health, American College of Radiology, and RSNA; consulting fees from Bunkerhill Health; associate editor of *Radiology: Artificial Intelligence*; vice chair of Society for Imaging Informatics in Medicine Program Planning Committee; and stock or stock options from Bunkerhill Health. **P.B.** No relevant relationships. **B.B.** No relevant relationships. **S.P.G.** No relevant relationships. **A.K.** No relevant relationships. **P.K.** No relevant relationships. **D.L.** No relevant relationships. **V.S.P.** No relevant relationships. **S.M.S.** No relevant relationships. **L.M.** Editor of *Radiology*; grant from Siemens, Gordon and Betty Moore Foundation, Mary Kay Foundation, and Google; consulting fees from Lunit, iCAD, and Guerbet; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events from iCAD and Guerbet; board member of International Society for Magnetic Resonance in Medicine and Society of Breast Imaging; stock or stock options in Lunit. **J.S.** Grant from National Institutes of Health (R01CA287422).

References

- Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 2017;284(2):574–582.
- Buda M, Wildman-Tobriner B, Hoang JK, et al. Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists. *Radiology* 2019;292(3):695–701.
- Ding Y, Sohn JH, Kawczynski MG, et al. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using ¹⁸F-FDG PET of the Brain. *Radiology* 2019;290(2):456–464.
- Rothenberg SA, Savage CH, Abou Elkassem A, et al. Prospective Evaluation of AI Triage of Pulmonary Emboli on CT Pulmonary Angiograms. *Radiology* 2023;309(1):e230702.
- Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med* 2018;24(9):1337–1341.
- Lehnen NC, Dorn F, Wiest IC, et al. Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis. *Radiology* 2024;311(1):e232741.
- Beheshtian E, Putnam K, Santomartino SM, Parekh VS, Yi PH. Generalizability and Bias in a Deep Learning Pediatric Bone Age Prediction Model Using Hand Radiographs. *Radiology* 2023;306(2):e220505.
- Bachina P, Garin SP, Kulkarni P, et al. Coarse Race and Ethnicity Labels Mask Granular Underdiagnosis Disparities in Deep Learning Models for Chest Radiograph Diagnosis. *Radiology* 2023;309(2):e231693.
- Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021;27(12):2176–2182.
- Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA* 2020;117(23):12592–12594.
- Nguyen DL, Ren Y, Jones TM, Thomas SM, Lo JY, Grimm LJ. Patient Characteristics Impact Performance of AI Algorithm in Interpreting Negative Screening Digital Breast Tomosynthesis Studies. *Radiology* 2024;311(2):e232286.
- Yi PH, Kim TK, Siegel E, Yahyavi-Firouz-Abadi N. Demographic Reporting in Publicly Available Chest Radiograph Data Sets: Opportunities for Mitigating Sex and Racial Disparities in Deep Learning Models. *J Am Coll Radiol* 2022;19(1 Pt B):192–200.
- Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* 2020;324(12):1212–1213.
- Gichoya JW, Banerjee I, Bhimreddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* 2022;4(6):e406–e414.
- Adleberg J, Wardeh A, Doo FX, et al. Predicting Patient Demographics From Chest Radiographs With Deep Learning. *J Am Coll Radiol* 2022;19(10):1151–1161.
- Duffy G, Clarke SL, Christensen M, et al. Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit Med* 2022;5(1):188.
- Machine Learning Glossary. Fairness. Google Dev. <https://developers.google.com/machine-learning/glossary/fairness>. Accessed May 27, 2024.
- Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. arXiv 1610.02413 [preprint] <https://doi.org/10.48550/arXiv.1610.02413>. Posted October 7, 2016. Accessed June 2024.
- Barocas S, Hardt M, Naryanan A. Fairness and machine learning: Limitations and opportunities. MIT Press; 2023. <https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/>. Accessed June 5, 2024.
- Tejani AS, Ng YS, Xi Y, Rayan JC. Understanding and Mitigating Bias in Imaging Artificial Intelligence. *RadioGraphics* 2024;44(5):e230067.
- Tejani AS, Retson TA, Moy L, Cook TS. Detecting Common Sources of AI Bias: Questions to Ask When Procuring an AI Solution. *Radiology* 2023;307(3):e230580.
- Faghani S, Khosravi B, Zhang K, et al. Mitigating Bias in Radiology Machine Learning: 3. Performance Metrics. *Radiol Artif Intell* 2022;4(5):e220061.
- Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. *Radiol Artif Intell* 2022;4(5):e210290.
- Zhang K, Khosravi B, Vahdati S, et al. Mitigating Bias in Radiology Machine Learning: 2. Model Development. *Radiol Artif Intell* 2022;4(5):e220010.
- Kitamura FC, Prevedello LM, Colak E, et al. Lessons Learned in Building Expertly Annotated Multi-Institution Datasets and Hosting the RSNA AI Challenges. *Radiol Artif Intell* 2024;6(3):e230227.
- Colak E, Kitamura FC, Hobbs SB, et al. The RSNA Pulmonary Embolism CT Dataset. *Radiol Artif Intell* 2021;3(2):e200254.
- Flanders AE, Prevedello LM, Shih G, et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiol Artif Intell* 2020;2(3):e190211.
- Lin HM, Colak E, Richards T, et al. The RSNA Cervical Spine Fracture CT Dataset. *Radiol Artif Intell* 2023;5(5):e230034.
- Lakhani P, Mongan J, Singhal C, et al. The 2021 SIIM-FISABIO-RSNA Machine Learning COVID-19 Challenge: Annotation and Standard Exam Classification of COVID-19 Chest Radiographs. *J Digit Imaging* 2023;36(1):365–372.
- Jiang B, Ozkara BB, Zhu G, et al. Assessing the Performance of Artificial Intelligence Models: Insights from the American Society of Functional Neuroradiology Artificial Intelligence Competition. *AJNR Am J Neuroradiol* 2024;45(9):1276–1283.
- Prevedello LM, Halabi SS, Shih G, et al. Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions. *Radiol Artif Intell* 2019;1(1):e180031.
- Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv 1901.07031 [preprint] <https://arxiv.org/abs/1901.07031>. Posted January 21, 2019. Accessed February 7, 2022.
- Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Med Image Anal* 2020;66:101797.
- Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019;6(1):317.
- Calabrese E, Villanueva-Meyer JE, Rudie JD, et al. The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. *Radiol Artif Intell* 2022;4(6):e220058.
- Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv 2107.02314 [preprint] <https://doi.org/10.48550/arXiv.2107.02314>. Posted July 5, 2021. Accessed May 2024.
- Kazerooni AF, Khalili N, Liu X, et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv 2305.17033 [preprint] <https://doi.org/10.48550/arXiv.2305.17033>. Posted May 26, 2023. Accessed May 2024.
- Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* 2015;34(10):1993–2024.
- Pan I, Cadrin-Chênevert A, Cheng PM. Tackling the Radiological Society of North America Pneumonia Detection Challenge. *AJR Am J Roentgenol* 2019;213(3):568–574.
- Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 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 (CVPR). IEEE; 2017; 3462–3471.
- Murphy ZR, Venkatesh K, Sulam J, Yi PH. Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: A Comparison of Performance, Sample Efficiency, and Hidden Stratification. *Radiol Artif Intell* 2022;4(6):e220012.
- Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput* 2021;26:232–243.

43. Garin SP, Parekh VS, Sulam J, Yi PH. Medical imaging data science competitions should report dataset demographics and evaluate for bias. *Nat Med* 2023;29(5):1038–1039.
44. Sun TY, Walk IV OJBD, Chen JL, Nieva HR, Elhadad N. Exploring Gender Disparities in Time to Diagnosis. *arXiv* 2011.06100 [preprint] <https://doi.org/10.48550/arXiv.2011.06100>. Posted November 11, 2020. Accessed May 2024.
45. Spencer CS, Gaskin DJ, Roberts ET. The quality of care delivered to patients within the same hospital varies by insurance type. *Health Aff (Millwood)* 2013;32(10):1731–1739.
46. Joseph NP, Reid NJ, Som A, et al. Racial and Ethnic Disparities in Disease Severity on Admission Chest Radiographs among Patients Admitted with Confirmed Coronavirus Disease 2019: A Retrospective Cohort Study. *Radiology* 2020;297(3):E303–E312.
47. Suleyman G, Fadel RA, Malette KM, et al. Clinical Characteristics and Morbidity Associated With Coronavirus Disease 2019 in a Series of Patients in Metropolitan Detroit. *JAMA Netw Open* 2020;3(6):e2012270.
48. Kulkarni P, Chan A, Navarathna N, Chan S, Yi PH, Parekh VS. Hidden in Plain Sight: Undetectable Adversarial Bias Attacks on Vulnerable Patient Populations. *arXiv* 2402.05713 [preprint] <https://doi.org/10.48550/arXiv.2402.05713>. Posted February 8, 2024. Accessed May 2024.
49. Puyol-Antón E, Ruijsink B, Mariscal Harana J, et al. Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. *Front Cardiovasc Med* 2022;9:859310.
50. Radiology Publication Instructions for Authors. *Radiology*. <https://pubs.rsna.org/page/radiology/author-instructions?doi=10.1148%2Fradiology&publicationCode=radiology>. Accessed May 30, 2024.
51. Flanagan A, Frey T, Christiansen SL; AMA Manual of Style Committee. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA* 2021;326(7):621–627.
52. Driessen R, Bhatia N, Gichoya JW, Safdar NM, Balthazar P. Sociodemographic Variables Reporting in Human Radiology Artificial Intelligence Research. *J Am Coll Radiol* 2023;20(6):554–560.
53. Jose O, Stoeckl EM, Miles RC, et al. The Impact of Extreme Neighborhood Socioeconomic Deprivation on Access to American College of Radiology-accredited Advanced Imaging Facilities. *Radiology* 2023;307(3):e222182.
54. Ethnic group census maps. UK Office for National Statistics. <https://www.ons.gov.uk/census/maps/choropleth/identity/ethnic-group/ethnic-group-tb-20b/asian-asian-british-or-asian-welsh-bangladeshi>. Accessed May 30, 2024.
55. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019;2(1):31.
56. Wu E, Wu K, Zou J. Explaining medical AI performance disparities across sites with confounder Shapley value analysis. *arXiv* 2111.08168 [preprint] <https://arxiv.org/abs/2111.08168>. Posted November 12, 2021. Accessed February 8, 2022.
57. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med* 2018;15(11):e1002683.
58. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut Learning in Deep Neural Networks. *Nat Mach Intell* 2020;2(11):665–673.
59. Yi PH, Malone PS, Lin CT, Filice RW. Deep Learning Algorithms for Interpretation of Upper Extremity Radiographs: Laterality and Technologist Initial Labels as Confounding Factors. *AJR Am J Roentgenol* 2022;218(4):714–715.
60. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell* 2021;3(7):610–619.
61. Mukherjee P, Shen TC, Liu J, Mathai T, Shafaat O, Summers RM. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. *Nat Med* 2022;28(6):1159–1160.
62. Bharti B, Yi P, Sulam J. Estimating and Controlling for Fairness via Sensitive Attribute Predictors. *arXiv* 2207.12497 [preprint] <https://doi.org/10.48550/arXiv.2207.12497>. Posted July 25, 2022. Accessed May 2024.
63. Ktena I, Wiles O, Albuquerque I, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat Med* 2024;30(4):1166–1173.
64. Yi PH, Wei J, Kim TK, et al. Radiology “forensics”: determination of age and sex from chest radiographs using deep learning. *Emerg Radiol* 2021;28(5):949–954.
65. Chambon P, Bluethgen C, Delbrouck JB, et al. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. *arXiv* 2211.12737 [preprint] <https://doi.org/10.48550/arXiv.2211.12737>. Posted November 23, 2022. Accessed May 2024.
66. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* 2020;2(2):e200029.
67. Clayton JA, Tannenbaum C. Reporting Sex, Gender, or Both in Clinical Research? *JAMA* 2016;316(18):1863–1864.
68. Garbin C, Rajpurkar P, Irvin J, Lungren MP, Marques O. Structured dataset documentation: a datasheet for CheXpert. *arXiv* 2105.03020 [preprint] <https://arxiv.org/abs/2105.03020>. Posted May 7, 2021. Accessed February 8, 2022.
69. The Medical Information Mart for Intensive Care. MIMIC. <https://mimic.mit.edu/>. Accessed May 30, 2024.
70. Patients table. MIMIC. <https://mimic.mit.edu/docs/iv/modules/hosp/patients/>. Published 2020. Accessed May 30, 2024.
71. Doo FX, Zavaletta V, Carroll EF, Ellis KL, Rosenkrantz AB. Turning a Page in the Yellow Journal: Figure Legends and Gender-Inclusive Patient Descriptors. *AJR Am J Roentgenol* 2022;219(1):1–2.
72. Goldberg JE, Moy L, Rosenkrantz AB. Assessing Transgender Patient Care and Gender Inclusivity of Breast Imaging Facilities Across the United States. *J Am Coll Radiol* 2018;15(8):1164–1172.
73. Wang HL. New “Latino” and “Middle Eastern or North African” checkboxes proposed for U.S. forms. NPR. <https://www.npr.org/2023/01/26/1151608403/mena-race-categories-us-census-middle-eastern-latino-hispanic>. Published January 26, 2023. Updated April 7, 2023. Accessed May 30, 2024.
74. Yin K. Why We Confuse “Race” and “Ethnicity”: A Lexicographer’s Perspective. *Conscious Style Guide*. <https://consciousstyleguide.com/why-we-confuse-race-ethnicity-lexicographers-perspective/>. Published 2019. Accessed May 30, 2024.
75. United States Census Bureau. Measuring Racial and Ethnic Diversity for the 2020 Census. *Census.gov*. <https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html>. Accessed May 30, 2024.
76. Gertych A, Zhang A, Sayre J, Pospiech-Kurkowska S, Huang HK. Bone age assessment of children using a digital hand atlas. *Comput Med Imaging Graph* 2007;31(4-5):322–331.
77. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447–453.
78. Peña MA, Sudarshan A, Muns CM, et al. Analysis of Geographic Accessibility of Breast, Lung, and Colorectal Cancer Screening Centers Among American Indian and Alaskan Native Tribes. *J Am Coll Radiol* 2023;20(7):642–651.
79. Mango VL, Stoeckl EM, Reid NJ, et al. Impact of High Neighborhood Socioeconomic Deprivation on Access to Accredited Breast Imaging Screening and Diagnostic Facilities. *J Am Coll Radiol* 2023;20(7):634–639.
80. Booker Wyden Health Care Letters. <https://www.scribd.com/document/437954989/Booker-Wyden-Health-Care-Letters#>. Accessed June 7, 2024.
81. Attorney General Bonta Launches Inquiry into Racial and Ethnic Bias in Healthcare Algorithms. State of California Department of Justice, Office of the Attorney General. <https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-inquiry-racial-and-ethnic-bias-healthcare>. Published August 31, 2022. Accessed June 7, 2024.
82. The term “Asian American” doesn’t serve everyone it covers. *Vox*. <https://www.vox.com/identities/22380197/asian-american-pacific-islander-aapi-heritage-anti-asian-hate-attacks>. Accessed May 31, 2024.
83. Movva R, Shanmugam D, Hou K, et al. Coarse race data conceals disparities in clinical risk score performance. *arXiv* 2304.09270 [preprint] <https://doi.org/10.48550/arXiv.2304.09270>. Posted April 18, 2023. Accessed May 2024.
84. Vicks WS, Lo JC, Guo L, et al. Prevalence of prediabetes and diabetes vary by ethnicity among U.S. Asian adults at healthy weight, overweight, and obesity ranges: an electronic health record study. *BMC Public Health* 2022;22(1):1954.
85. Li D, Bharti B, Wei J, Sulam J, Yi PH. Sex Imbalance Produces Biased Deep Learning Models for Knee Osteoarthritis Detection. *Can Assoc Radiol J* 2023;74(1):219–221.
86. Friedler SA, Scheidegger C, Venkatasubramanian S. On the (im)possibility of fairness. *arXiv* 1609.07236 [preprint] <https://arxiv.org/abs/1609.07236>. Posted September 23, 2016. Accessed May 31, 2024.
87. Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC. A clarification of the nuances in the fairness metrics landscape. *Sci Rep* 2022;12(1):4209.
88. Romano Y, Bates S, Candès EJ. Achieving Equalized Odds by Resampling Sensitive Attributes. *arXiv* 2006.04292 [preprint] <https://arxiv.org/abs/2006.04292>. Posted June 8, 2020. Accessed November 5, 2022.
89. Erickson BJ, Kitamura F. Magician’s Corner: 9. Performance Metrics for Machine Learning Models. *Radiol Artif Intell* 2021;3(3):e200126.
90. Krupinski EA. Evaluating AI Clinically—It’s Not Just ROC AUC! *Radiology* 2021;298(1):47–48.
91. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspect Clin Res* 2015;6(3):169–170.