Beyond Static Bias: Quantifying Fairness Variability in CheXpert

Ines Ayed
Department of Mathematics and Computer Science
University of the Balearic Islands
07122 Palma, Spain
ines.ayed@uib.es

Gabriel Moyà Alcover
Department of Mathematics and Computer Science
University of the Balearic Islands
07122 Palma, Spain
gabriel.moya@uib.es

Fernando Alonso-Fernandez
School of Information Technology
Halmstad University
Halmstad, Sweden
feralo@hh.se

Antoni Jaume-i-Capó
Department of Mathematics and Computer Science
University of the Balearic Islands
07122 Palma, Spain
antoni.jaume@uib.es

Abstract

Fairness in machine learning is typically assessed through static point-estimate metrics that overlook the robustness and reliability of model behavior under biased data. We introduce a statistical framework to analyze the relationship between the variability of dataset bias and the variability of a model's fairness gaps. Using Monte Carlo simulation, we quantify bias in the CheXpert dataset and find that while bias is small, it is consistently stable with near-zero variance across its five most common pathologies. Applying a mixed-effects model, we then examine how this stable bias relates to fairness variability in leaderboard models. We find that model fairness can fluctuate unpredictably even when dataset bias is modest but stable, revealing a hidden robustness failure in fairness evaluations. Our results underscore the need to move beyond static fairness metrics toward evaluation methods that explicitly characterize robustness under subpopulation and distribution shifts, aligning with the broader goals of building reliable machine learning under imperfect data.

1 Introduction

Machine learning (ML) algorithms have reached high accuracy in medical image diagnosis, making them a great tool to democratize medical access and improve the accuracy of disease diagnosis [1]. However, research points to the fact that these algorithms are under-diagnosing certain demographic groups, where the AI algorithm may incorrectly identify a person with a disease as healthy, potentially delaying necessary medical care [2, 3]. This compromises their reliability in practice from a fairness perspective.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Reliable ML from Unreliable Data.

For example, a study by Seyyed et al. [2] showed that AI models produce significant differences in the accuracy of automated chest x-ray diagnosis across racial and other demographic groups such as females or Black patients, even when the models only had access to the chest x-ray itself. Similarly, Larrazabal et al. [3] found a consistent decrease in performance of AI systems for computer-assisted diagnosis for underrepresented genders in training datasets. With the increasing deployment of these algorithms in real-world applications, there is a growing fear of mirroring or amplifying human bias in typically marginalized groups.

One potential source of algorithmic bias is the bias present within the training dataset. Thus, it is important to analyze the demographic bias in datasets used in medical imaging tasks. While public datasets from various regions worldwide are becoming more accessible, there remains a critical lack of standardized metrics to assess and quantify demographic bias in these datasets [4]. Without such metrics, it is difficult to determine the extent to which biases exist and how they may impact model performance.

We hypothesize that dataset bias and fairness are related. To investigate this, we introduce a framework combining Monte Carlo simulations and mixed-effects models to systematically quantify the relationship between dataset bias and model fairness metrics. Our simulations reveal that dataset bias is consistent and stable, suggesting it is an intrinsic feature of the CheXpert dataset rather than a sampling artifact. In contrast, fairness gaps can be highly variable across resampled datasets.

Mixed-effects modeling further shows that the impact of dataset bias varies by metric. Demographic Parity Gap and Equalized Odds Gap are more sensitive to overall dataset bias (as measured by Cramer's V; a measure of association between demographics and pathology labels) than Equal Opportunity Gap, indicating that the relationship between bias and fairness is not uniform. Additionally, model variability contributes less to fairness variability than dataset composition, emphasizing the critical role of data in shaping model outcomes.

Our key contributions are:

- Application of Monte Carlo simulations to quantify variability in fairness metrics.
- Aggregation and analysis at the patient level to assess demographic subgroup effects.
- Use of mixed-effects models to link dataset bias with fairness gaps while accounting for hierarchical data structure.

2 Related Work

Existing radiography datasets provide general demographic statistics but rarely quantify bias systematically [5]. While prior studies have measured fairness in AI for medical imaging, they typically focus on disparities in performance across demographic groups without statistically relating these disparities to underlying dataset bias [2, 3].

Recent studies, such as [6], discuss metrics to quantify demographic bias in facial expression recognition datasets. These metrics define bias as statistical imbalances in group representation. However, such works do not account for variability or uncertainty in the bias estimates, nor do they extend to more complex datasets such as chest X-rays, where multiple pathologies and demographic correlations exist.

Our work adapts these bias quantification approaches to the medical imaging context, incorporating both bias and its uncertainty. By combining Monte Carlo resampling and mixed-effects modeling, we provide a framework to systematically link demographic bias with model fairness, capturing both the magnitude and reliability of fairness metrics across subgroups.

3 Methods

3.1 Dataset and Preprocessing

All demographic attributes in this study were derived from the publicly available CheXpert dataset [7], which contains de-identified chest radiographs and metadata collected at Stanford Hospital between 2002 and 2017. CheXpert is released for research under a data use agreement, and all data in this study were used in compliance with its terms.

3.1.1 Demographic Data

The CheXpert dataset records gender as Male or Female. In our analysis, we retained these two categories; a single entry with unknown gender was excluded.

Patient age is provided as a continuous variable. For analysis, we stratified age into categorical intervals [0–18, 18–40, 40–60, 60–80, 80+], following the grouping reported in [8].

Race labels are self-reported. Our categorization synthesizes prior approaches in medical imaging fairness research. Following [8], we adopt the subgroups White, Black, and Asian. In line with the work of Seyyed et al. [2], we treat Hispanic as a distinct subgroup rather than aggregating it into Other, motivated by both its substantial representation (n=1,455) and its clinical relevance for health equity research. The final race categories used in this study are Asian, Black, White, Hispanic, Other, and Unknown (for missing or unrecorded entries).

3.1.2 Uncertainty Handling

The CheXpert dataset provides target labels with four possible values: positive (1), negative (0), uncertain (-1), and unmentioned (blank). For binary classification, we treated unmentioned labels as negative and adopted the U-zeros approach, mapping all uncertain labels to 0 as in [2]. This provides a conservative strategy, as in clinical applications the cost of a false negative (missed diagnosis) is typically considered higher than that of a false positive. Other uncertainty handling strategies, such as U-ones (mapping uncertain labels to 1) or **U-ignore** (excluding uncertain labels), could be investigated in future work to assess the robustness of fairness analyses across different label assumptions.

3.1.3 Target Labels

Facial Expression Recognition (FER) datasets typically follow a single-label classification paradigm, where each image is assigned exactly one emotion. In contrast, chest X-ray datasets, such as CheXpert, involve multi-label classification, as a single scan can exhibit multiple co-occurring conditions, such as pneumonia, edema, and pleural effusion. CheXpert contains 14 target labels, reflecting a wide range of thoracic conditions. Existing approaches for multi-label classification fall into two main categories [9]. Binary label approaches treat each disease as a separate, disjoint classification task, ignoring potential correlations between labels. Label correlation approaches leverage dependencies or co-occurrence among diseases to improve predictive performance, reflecting clinical practice where co-occurring conditions inform diagnosis.

For this study, we focus on five target labels, Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion, chosen based on their clinical importance, prevalence, and use as competition tasks in CheXpert [7], and adopt a disjoint binary classification setup. This approach simplifies modeling and facilitates fairness evaluation, particularly when examining the variability of model predictions across repeated patient observations and under standard statistical assumptions of independence.

3.2 Quantifying Bias

Bias and fairness are closely related but distinct concepts. We differentiate between dataset bias metrics, which capture statistical imbalances in the representation of demographic groups, and model fairness metrics, which quantify performance gaps across those groups. We hypothesize that the variability of dataset bias metrics is related to the variability of model fairness gaps.

3.2.1 Dataset Bias Metrics

We selected three demographic bias metrics recommended by Dominguez et al. [6] for use in facial expression recognition datasets, and adapted them to the CheXpert setting: Shannon Evenness Index (SEI), Simpson's Diversity Index (1–D), and Cramer's V (ϕ_C).

Shannon Evenness Index (SEI) [10] measures the homogeneity of group representation. Values range from 0 (highly imbalanced) to 1 (perfectly balanced), providing a normalized measure of diversity.

Simpson's Diversity Index (1–D) [11] captures the probability that two randomly selected individuals belong to different demographic groups. Like SEI, it characterizes representation bias by quantifying the degree of demographic diversity independent of target labels.

Together, SEI and Simpson's Index summarize representation bias, defined as unequal demographic group distributions in a dataset. These metrics can be applied to any dataset since they do not require information about clinical or target labels.

Cramer's V (ϕ_C) [12] quantifies associations between a demographic variable and a target label, capturing stereotypical bias. It is derived from the chi-squared statistic, which assumes independence of observations and requires sufficiently large expected frequencies (≥ 5 per cell) for validity.

Applying Cramer's V in medical imaging introduces challenges not present in FER data. In CheXpert, the same patient may contribute multiple radiographs (different views or time points), violating independence assumptions and potentially inflating chi-squared statistics. Moreover, demographic features remain constant across repeated observations, further compounding dependence. To mitigate these issues, we restrict our analysis to major demographic categories (gender, age, race) and ensure adequate subgroup sample sizes to preserve statistical validity.

3.2.2 Model Fairness Gaps

To assess how dataset bias metrics relate to model behavior, we evaluated fairness gaps using three widely studied criteria. These metrics are standard in the algorithmic fairness literature and capture complementary notions of fairness [13]:

- **Equalized Odds**: Requires both false positive and false negative rates to be similar across demographic groups, ensuring that errors are not disproportionately concentrated in one group.
- **Equal Opportunity**: A relaxed version of Equalized Odds, requiring only that true positive rates are equal across groups, thus focusing specifically on underdiagnosis risks.
- **Demographic Parity**: Requires predicted label distributions to be independent of demographic attributes, regardless of ground-truth labels.

We applied these definitions to a subset of top-performing models from the CheXpert competition leaderboard [14]. Fairness gaps were computed by measuring the absolute differences in metric values across demographic subgroups (gender, age, and race) for each of the five clinically important classification tasks described earlier. To capture variability and reduce sensitivity to sample splits, we applied Monte Carlo resampling to the test dataset of CheXpert, producing a distribution of fairness gap estimates for each model–subgroup combination.

3.3 Experimental Design

3.3.1 Patient-level Aggregation

To account for multiple images per patient in the dataset, we aggregate image-level labels to a single patient-level label using a maximum rule: a patient is assigned a positive label (1) for a given pathology if at least one of their associated images is positive. This aggregation, applied to the training set, ensures a one-to-one mapping between patients and data points, which is critical for satisfying the assumption of record independence in subsequent statistical analyses.

3.3.2 Monte Carlo Simulation

To assess the robustness and statistical variability of both dataset bias and model fairness metrics, we performed patient-level Monte Carlo simulations with 5,000 bootstrap iterations, a standard practice consistent with recommendations in the statistical literature [15].

For dataset bias, the simulation was conducted on the training set using the aggregated patient-level labels. In each iteration, patients were sampled with replacement, and dataset bias metrics such as Simpson's Index and Cramer's V were computed across demographic groups.

For model fairness, the simulation was conducted on the test set, where all images corresponding to each bootstrapped patient were retained. This allowed computation of fairness gaps (Equalized Odds, Equal Opportunity, and Demographic Parity) across models, pathologies, and demographic subgroups.

This two-stage approach generates empirical distributions for each metric, from which means, standard deviations, and 95% confidence intervals can be derived. By explicitly accounting for both sampling variability in the data and variability across model predictions, this procedure provides a rigorous assessment of metric reliability under potentially unstable or unrepresentative data.

3.3.3 Mixed-Effects Model

To investigate the relationship between dataset bias metrics and model fairness gaps, we employed a mixed-effects modeling framework [16]. The response variable is the fairness gap for a given metric (Equalized Odds, Equal Opportunity, or Demographic Parity), and the independent variable is the corresponding dataset bias metric (Cramer's V).

The hierarchical nature of the data motivates the use of mixed-effects models. Each leaderboard model provides multiple fairness measurements across pathologies and demographic groups, and each patient contributes multiple images. Mixed-effects models allow us to account for this nested structure by including both fixed effects (capturing systematic differences across demographic groups and pathologies) and random effects (capturing correlations among repeated measurements within the same model and variability across models). The model formula is:

$$\text{Fairness gap} \sim \text{Cramer's V} \times \text{Demographic group} + \text{Pathology} + (1 \mid \text{Model})$$

where $(1 \mid \text{Model})$ denotes a random intercept per model. Interaction terms between Cramer's V and demographic group allow the model to capture group-specific sensitivities to dataset bias.

Prior to modeling, measurements were aggregated at the patient level to prevent skewing by patients with multiple correlated images. We also restricted the analysis to major demographic groups (gender, age, and race) to avoid multicollinearity from small or intersectional subgroups.

This mixed-effects approach provides robust estimates of how variability in dataset bias translates to variability in fairness gaps while properly accounting for the hierarchical structure and repeated measurements. Model convergence and fit were checked for all metrics to ensure reliable inference.

Limitations Aggregating images to a patient-level dataframe, particularly when using the maximum value of labels, can lead to loss of detailed information. This simplification may obscure finer-grained associations between labels and demographic attributes and prevents analysis of temporal dynamics across multiple observations for the same patient. As a result, some nuances of bias or fairness variability at the image level could be underestimated.

4 Results

4.1 Dataset Bias Analysis

We first examined the distribution of image counts per patient in the training dataset. Most patients (72.9%) had between one and three images, whereas a small subset contributed substantially more, resulting in a long-tailed distribution (See Figure 1). This motivates the use of patient-level aggregation to prevent over-representation of patients with many images in subsequent analyses.

4.1.1 Representational Bias

We assessed dataset representativeness and subgroup balance using two complementary metrics, Shannon evenness and Simpson diversity indices. Results indicate an even gender distribution (mean evenness = 0.991, 95% CI [0.990, 0.992]; Simpson diversity = 0.494, 95% CI [0.493, 0.495]), consistent with near-equal representation across the two gender categories. Although the Simpson diversity value appears low, this is expected given only two categories, and it actually reflects a near-maximal diversity for gender. For age groups, balance was lower but still relatively high (evenness = 0.824, 95% CI [0.823, 0.826]; diversity = 0.719, 95% CI [0.717, 0.720]). Considering that the maximum possible Simpson diversity for five age categories is 0.8, this indicates that the dataset includes a broad mix of age groups, with only slight overrepresentation of certain brackets. In contrast, race distribution showed the largest imbalance, with lower evenness (0.690, 95% CI [0.686,

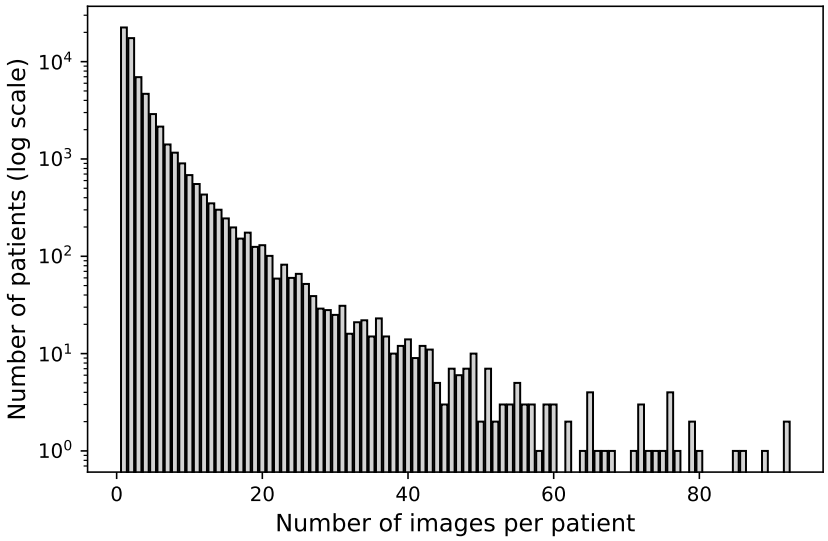

Figure 1: Distribution of image counts per patient in the training dataset.

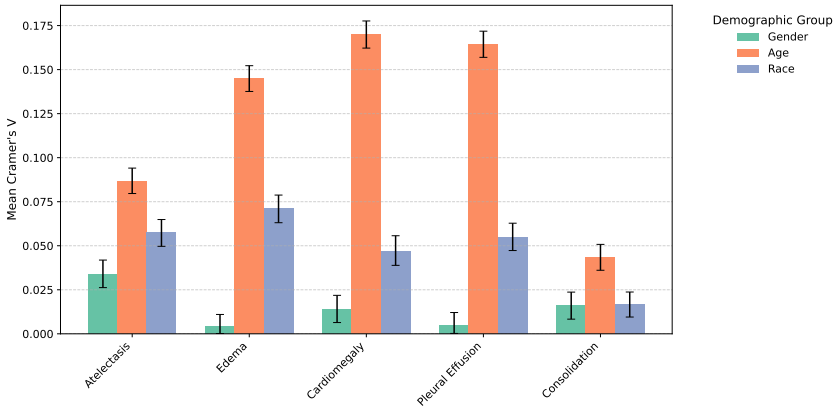

Figure 2: Cramer's V by pathology and demographic group.

0.693]) and diversity (0.640, 95% CI [0.637, 0.643]), reflecting underrepresentation of several racial subgroups.

Standard deviations were consistently small (<0.002) and confidence intervals tight, indicating that the bootstrap estimates are stable and precise. Collectively, these results suggest that while the dataset is well-balanced by gender, age coverage is reasonably broad, and race remains a source of representational bias, which could have implications for fairness evaluations.

4.1.2 Stereotypical Bias

Analysis of Cramer's V across pathologies and demographic groups revealed that associations between age and certain conditions were modest but consistent (see Figure 2). For instance, age grouping showed higher Cramer's V for Edema (0.145, 95% CI [0.138, 0.152]) and Cardiomegaly (0.170, 95% CI [0.162, 0.178]), indicating a systematic relationship between patient age and these pathologies in the dataset. Biases across gender and race were generally smaller, with mean Cramer's V values below 0.08. Even modest Cramer's V values, such as 0.14, are meaningful in a dataset of this scale, reflecting subtle but systematic demographic imbalances. Standard deviations were consistently low (<0.004) and confidence intervals very narrow, confirming that these associations are stable and not due to random variation.

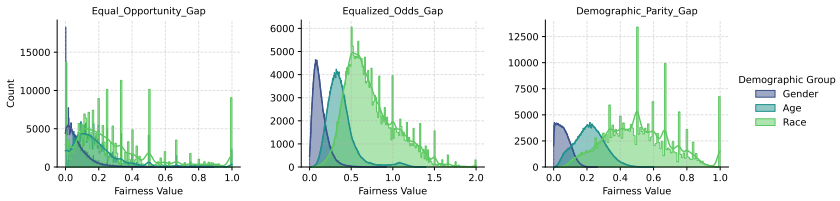

Figure 3: Distribution of fairness metrics by demographic group.

4.2 Fairness Gaps Variability

We assessed model fairness across demographic groups using three metrics: Demographic Parity Gap, Equal Opportunity Gap, and Equalized Odds Gap. Results indicate that fairness gaps vary substantially by subgroup. Race exhibits the largest gaps (e.g., mean Equalized Odds Gap = 0.733, std = 0.303), while gender shows the smallest disparities (mean = 0.127, std = 0.086). Age subgroups fall in between (mean = 0.375, std = 0.176). Standard deviations indicate notable variability in fairness outcomes across subgroups, particularly for race. The distributions of these metrics are visualized in Figure 3, which shows histograms for each metric by demographic group. The figure illustrates both the central tendencies and the spread of fairness gaps, confirming that while some disparities are small (e.g., gender), others remain substantial and variable, particularly for race.

4.3 Mixed-Effects Model Results

We fitted mixed-effects models to examine the relationship between dataset bias, as measured by Cramer's V, and fairness gaps across three metrics: Equal Opportunity Gap, Equalized Odds Gap, and Demographic Parity Gap. Models included random intercepts for leaderboard models to account for repeated measurements, fixed effects for demographic groups and pathologies, and interactions between Cramer's V and demographics.

For the Equal Opportunity Gap, Cramer's V was negatively associated with fairness for most groups (Coef = -0.659, SE = 0.008, $p < 0.001$), with the interaction term indicating that gender had a stronger sensitivity (Coef = 1.695, SE = 0.038, $p < 0.001$). Pathology effects were modest but significant, with Edema (Coef = 0.062) and Consolidation (Coef = 0.066) showing positive associations. Random effect variance per model was small (0.004), indicating that model-to-model variability contributed little relative to dataset bias.

For the Equalized Odds Gap, associations were weaker but still significant (Cramer's V Coef = -0.057, SE = 0.009, $p < 0.001$), and gender again showed heightened sensitivity (interaction Coef = 0.472, SE = 0.041, $p < 0.001$). Pathology effects varied, with Pleural Effusion showing a notable negative effect (Coef = -0.071).

For the Demographic Parity Gap, Cramer's V exhibited a strong positive effect (Coef = 0.997, SE = 0.004, $p < 0.001$), while the interaction with gender was strongly negative (Coef = -2.903, SE = 0.019, $p < 0.001$). Pathology-specific effects were significant but small (e.g., Pleural Effusion Coef = -0.080). Random effect variance was essentially zero, indicating that the hierarchical model structure did not substantially affect the estimates.

These results indicate that systematic dataset biases, quantified via Cramer's V, are significantly associated with fairness gaps, with certain demographic groups, particularly gender, showing stronger sensitivity. Variability across leaderboard models was minimal, confirming that observed effects are robust to model differences.

5 Discussion

Our findings reveal systematic differences in demographic balance across subgroups. Gender is nearly balanced, with a Simpson diversity value close to the theoretical maximum for two groups. In contrast, age distribution skews toward older patients, with younger adults and children underrepresented. Racial diversity is moderate, but the dataset is dominated by White patients, with minority groups (e.g., Native, Hispanic, Black) present in very small numbers. While gender representation is strong,

age and racial composition may introduce fairness concerns, particularly for underrepresented groups. The narrow confidence intervals indicate these imbalances are stable and not due to sampling noise.

Associations between age and certain pathologies, such as Edema and Cardiomegaly, likely reflect both genuine prevalence patterns and dataset-specific collection biases. While such relationships are expected clinically, their persistence in a machine learning dataset raises concerns for reliability; models may inadvertently exploit demographic correlates rather than underlying disease features. Gender and race showed weaker associations, yet even subtle imbalances can propagate through model predictions, especially in high-capacity models. These results underscore the importance of auditing demographic–pathology relationships when building reliable models from clinical data.

Results of fairness gaps variability highlight that even when models are trained on the same dataset, fairness outcomes can vary substantially across demographic groups. The large disparities observed for race suggest persistent underrepresentation or structural imbalances in the data that propagate through model predictions. In contrast, the smaller gaps for gender indicate relatively balanced representation in this subgroup. The high standard deviations for some metrics, particularly for race, indicate that fairness is sensitive to the sampling of patients, emphasizing the importance of robust evaluation methods.

The mixed-effects models demonstrate that even modest dataset biases can propagate into measurable fairness gaps. Gender consistently shows heightened sensitivity to Cramer's V across all metrics, suggesting that models trained on the dataset may systematically disadvantage one gender if biases are present. Pathology-specific effects, while smaller, indicate that dataset composition interacts with disease prevalence to influence fairness outcomes. Random effect variance was minimal across leaderboard models, implying that differences in model architecture or training approach contribute less to fairness variability than underlying dataset biases. This emphasizes the importance of auditing dataset representativeness when developing and deploying models.

Taken together, these findings highlight that subtle but systematic demographic imbalances, even when small in magnitude, can significantly affect fairness and reliability. Careful dataset auditing and bias mitigation are therefore essential for medical ML. In particular, fairness metrics proved unstable in small subgroups, where a single misclassification could dramatically shift results. This volatility poses a fundamental challenge; fairness evaluation is not only about bias but also about the reliability of subgroup estimates.

Finally, our study treated each scan independently, ignoring longitudinal dependencies from repeated measurements. Incorporating temporal structure would enable more robust fairness assessments over time. Future work should also move beyond static evaluations. We envision *a Fairness Volatility Index* to systematically quantify subgroup stability across datasets and over time, aligning fairness evaluation with the broader goals of reliable machine learning under real-world data limitations.

6 Conclusion

Our study shows that fairness metrics in medical imaging are not only influenced by dataset bias but also by the reliability of subgroup estimates. Even when dataset bias, as measured by Cramer's V, is stable, fairness gaps can fluctuate substantially, especially in small or underrepresented subgroups where a single misclassification can drastically change the metric. Mixed-effects modeling revealed that fairness gaps are significantly associated with dataset bias, with demographic groups such as gender showing greater sensitivity. However, random effects across models were small, indicating that instability arises more from data distribution than from model variation.

These findings underscore the need for fairness evaluations that account for subgroup size and reliability, paving the way for methods that measure not only bias but also its stability.

Acknowledgments and Disclosure of Funding

The authors acknowledge Gemini for assisting with drafting and revising text and refining methodological details, including exploring statistical approaches such as patient-level bootstrapping and mixed-effects modeling. All scientific content and conclusions are the responsibility of the authors.

This work is funded by Project PID2023-149079OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

References

[1] Mengfang Li, Yuanyuan Jiang, Yanzhou Zhang, and Haisheng Zhu. Medical image analysis using deep learning algorithms. *Frontiers in public health*, 11:1273253, 2023.

[2] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[3] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[4] H Yi Paul, Tae Kyung Kim, Eliot Siegel, and Noushin Yahyavi-Firouz-Abadi. Demographic reporting in publicly available chest radiograph data sets: opportunities for mitigating sex and racial disparities in deep learning models. *Journal of the American College of Radiology*, 19(1):192–200, 2022.

[5] Satvik Tripathi, Kyla Gabriel, Suhani Dheer, Aastha Parajuli, Alisha Isabelle Augustin, Ameena Elahi, Omar Awan, and Farouk Dako. Understanding biases and disparities in radiology ai datasets: a review. *Journal of the American College of Radiology*, 20(9):836–841, 2023.

[6] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5209–5226, 2024.

[7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[8] Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J Wang, Dushyant Sahani, and Shwetak Patel. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305, 2025.

[9] Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and David Zhang. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE journal of biomedical and health informatics*, 24(8):2292–2302, 2020.

[10] Claude E Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois press, 1998.

[11] Edward H Simpson. Measurement of diversity. *nature*, 163(4148):688–688, 1949.

[12] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

[13] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

[14] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexternal: Generalization of deep learning models for chest x-ray interpretation to photos of chest x-rays and external clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 125–132, 2021.

[15] Rand R Wilcox. *Fundamentals of modern statistical methods: Substantially improving power and accuracy*, volume 249. Springer, 2001.

[16] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the use of Monte Carlo simulations and mixed-effects models to study how dataset bias affects fairness metrics in medical imaging, matching the paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are addressed in the last paragraph of Methods section and throughout the Discussion, including small subgroup volatility, independent scan assumption, and lack of temporal modeling.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is primarily empirical and methodological. It does not present formal theorems or proofs, but instead uses statistical models and simulations to analyze dataset bias and fairness variability.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The dataset (CheXpert) is public, subgroup definitions are clearly described, and all fairness metrics and modeling procedures (Monte Carlo simulations, mixed-effects models) are detailed in the Methods section. Together, these provide sufficient information to reproduce the main results and validate the claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The CheXpert dataset is publicly available and methods are described in detail, but code is not released at this stage.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: While no new training was performed, the paper clearly specifies the data source (CheXpert), subgroup definitions, simulation setup, and mixed-effects modeling framework, which together provide the necessary details to interpret and understand the reported results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports confidence intervals for diversity and fairness metrics, and uses mixed-effects models to quantify variability and test statistical significance. This provides appropriate measures of uncertainty and supports the robustness of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not perform model training or other compute-intensive experiments. All analyses are conducted on pre-existing datasets using standard statistical methods and mixed-effects models, which require minimal compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics. No human subjects were directly involved, and all dataset usage follows ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper analyzes demographic bias in medical imaging datasets and proposes methods for more reliable fairness auditing, which can help identify and mitigate unfair treatment of underrepresented groups.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The study analyzes publicly available medical imaging datasets and does not release any high-risk models or sensitive data, so no additional safeguards are required.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (e.g., CheXpert) and code libraries used are publicly available and appropriately cited, with licenses and terms of use respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new datasets or code libraries are released; the paper focuses on analysis and methodology.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The study uses pre-existing medical datasets; no new human subjects or crowdsourcing experiments were conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The study only analyzes de-identified public medical datasets; no new human subjects were involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: An LLM (Gemini) was used only to assist with drafting and revising text, and refining methodological explanations. All scientific content, analyses, and conclusions are solely the responsibility of the authors.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.