Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen and Marzyeh Ghassemi

Artificial intelligence (AI) systems have increasingly achieved expert-level performance in medical imaging applications. However, there is growing concern that such AI systems may reflect and amplify human bias, and reduce the quality of their performance in historically under-served populations such as female patients, Black patients, or patients of low socioeconomic status. Such biases are especially troubling in the context of underdiagnosis, whereby the AI algorithm would inaccurately label an individual with a disease as healthy, potentially delaying access to care. Here, we examine algorithmic underdiagnosis in chest X-ray pathology classification across three large chest X-ray datasets, as well as one multi-source dataset. We find that classifiers produced using state-of-the-art computer vision techniques consistently and selectively underdiagnosed under-served patient populations and that the underdiagnosis rate was higher for intersectional under-served subpopulations, for example, Hispanic female patients. Deployment of AI systems using medical imaging for disease diagnosis with such biases risks exacerbation of existing care biases and can potentially lead to unequal access to medical treatment, thereby raising ethical concerns for the use of these models in the clinic.
b. We measure the underdiagnosis rate in distinct chest X-ray diagnosis models trained in four dataset settings: MIMIC-CXR (CXR, 371,858 images from 65,679 patients)\(^2\), CheXpert (CXP, 223,648 images from 64,740 patients)\(^3\), ChestX-ray14 (NIH, 112,120 images from 30,805 patients)\(^4\), and a multi-source combination of all three (ALL, 707,626 images from 129,819 patients)\(^2\) on shared labels. The CXR, CXP and NIH datasets have relatively equal rates of male and female patients, and most patients are between 40 and 80 years old. Note that the CXP and NIH datasets report only patient sex and age, whereas the CXR dataset additionally reports patient race/ethnicity and insurance type for a large subset of images. In the CXR dataset we note that both race/ethnicity and insurance type are highly skewed. We use the term ‘sex’ to match the reported terminology in the underlying data. Gender presentation plays a large role in societal biases but these data are not routinely collected\(^5\)–\(^7\). More detailed summary statistics for the datasets are listed in Table 1. The full data collection description per dataset is available in the Methods.

Underdiagnosis in under-served patient subpopulations. We find that the underdiagnosis rate for all datasets differs in all considered subpopulations. In Fig. 2a we show the subgroup-specific underdiagnosis for CXR dataset on race/ethnicity, sex, age and insurance type. We observed that female patients, patients under 20 years old, Black patients, Hispanic patients and patients with Medicaid insurance receive higher rates of algorithmic underdiagnosis than other groups. In other words, these groups are at a higher risk of being falsely flagged as healthy, and of receiving no clinical treatment. We summarize a similar analysis of the other datasets (CXP, NIH and ALL) in Table 2 and Extended Data Figs 1–3. Additional data for image counts on the test set per subgroup are given in Supplementary Tables 1–3.

We find that the patterns of bias are consistent across the CXR (Fig. 2a), ALL (Extended Data Fig. 1a) and CXP (Extended Data Fig. 2a) datasets—that is, female and younger patients have the largest underdiagnosis rates. However, in the NIH dataset (Extended Data Fig. 3a), male patients and patients aged >80 years have the largest underdiagnosis rate. This may be partially due to the small subset sizes, given that the test set for patients aged >80 years has only 37 samples with the no finding label with which to estimate FPR. The NIH dataset is also different from the CXP and CXR datasets in several key ways: it contains frontal images only, rather than frontal and lateral images; it does not use the CheXpert labelers\(^8\) to create diagnostic labels; and it has only seven of the shared disease labels instead of 14, meaning that the no finding label denotes the absence of different diseases. Moreover, the NIH dataset originates from a hospital that “...does not routinely provide standard diagnostic and treatment services. Admission is selective: patients are chosen by Institute physicians solely because they have an illness being studied by those Institutes.” (from https://clinicalcenter.nih.gov/about/welcome/faq.html). Thus, the NIH dataset may have less diverse samples than the CXP and CXR datasets, which originate from clinical hospitals (see Methods for more detail).

Underdiagnosis in intersectional groups. We investigate intersectional groups, here defined as patients who belong to two subpopulations, for example, Black female patients. Similar to prior work in facial detection\(^9\), we find that intersectional subgroups (Fig. 2b) often have compounded biases in algorithmic underdiagnosis. For instance, in the CXR dataset, Hispanic female patients have a higher underdiagnosis rate—that is, a no finding FPR—than white female patients (Fig. 2b(i)). Also, the intersectional subgroups of patients who are aged 0–20 years and female, aged 0–20 years and Black, and aged 0–20 years with Medicaid insurance have the largest underdiagnosis rates (Fig. 2b(ii)). The underdiagnosis rate for the intersection of Black patients with another subgroup of age, sex and insurance type (Fig. 2b(iii)) and that for patients with Medicaid insurance with another subgroup of sex, age and race/ethnicity (Fig. 2b(iv)) is also shown in Fig. 2b. We observe that patients who belong to two under-served subgroups have a larger underdiagnosis rate. In other words, not all female patients are misdiagnosed at the same rate (for example, Hispanic female patients are misdiagnosed more than white female patients) (Fig. 2b(i)). The intersectional underdiagnosis rate for the ALL, CXP and NIH datasets is shown in Extended Data Figs 1c, 2c and 3c, respectively, where the intersectional identities are often underdiagnosed even more heavily than the group in aggregate. The most underdiagnosed age groups for female patients are listed under the Female–Age attribute in Table 2.
Likelihood of underdiagnosis in specific diseases. The distribution of disease prevalence in the underdiagnosed patient population is significantly different to that in the general patient population. We compare the disease prevalence in the unhealthy population and the underdiagnosed population for the intersections of race/ethnicity and sex in Supplementary Table 4. For example, underdiagnosed populations are proportionally more likely to have a positive label for lung lesion and less likely to have a positive label for pleural effusion. This suggests that the task of disease detection is more difficult for some diseases than others.

Fairness definitions in a healthcare context. Our study considers underdiagnosis as the main fairness concern, due to its potentially harmful impact on patients, such as causing a delay in receiving treatment (for example, assigning lower priority to the under-diagnosed population in a triage use case). We acknowledge that depending on the use case of the algorithm there are many other fairness definitions one may consider. One such definition is predictive parity, which implies equal positive predictive value, or, equivalently, false discovery rate (FDR) between the groups.

In Supplementary Table 6 we report the additional data for FDR of a no disease diagnosis (that is, the likelihood that the patient is ill given that the classifier predicts no finding). We observe that, similar to FPR and FNR, significant gaps exist across many protected attributes. In particular, these disparities tend to follow a different pattern of that seen for FPR, favoring, for example, female people over male people and younger people over older people. The underlying cause is the difference in prevalence between groups—that is, given that there are far fewer sick people in the 0–20 year age group (Supplementary Tables 1–3), we will have relatively fewer false positives and true negatives, which, keeping all else constant, will cause a decrease in the FDR.

Discussion

We have shown consistent underdiagnosis in three large, public datasets in the chest X-ray domain. The algorithms trained on all settings exhibit systematic underdiagnosis biases in under-served subpopulations, such as female patients, Black patients, Hispanic patients, younger patients and patients of lower socioeconomic status (with Medicaid insurance). We found that these effects persist for intersectional subgroups (for example, Black female patients) but are not consistently worse in the smallest intersectional groups. The specific subpopulations most affected vary in the NIH dataset, specifically male patients and patients aged >80 years, which should be explored further. Beyond these immediate take-aways, there are several topics for further discussion and investigation.

First, we highlight that automatic labeling from notes should be carefully audited. We note that in chest X-ray datasets, there has been a general shift in machine learning from manual image labeling to automatic labeling, with natural language processing (NLP)-based methods used to generate the labels in radiology reports. This has resulted in large annotated chest X-ray datasets that are widely used for training deep learning models and for providing AI solutions. Although automatic labelers have been validated for labeling quality and adapted as reliable ground truth, the

| Table 1 | Summary statistics for all datasets |
|----------|-----------------------------------|
| Subgroup | Attribute | CXR | CXP | NIH | ALL |
|----------|----------|-----|-----|-----|-----|
| Sex (%)  |          |     |     |     |     |
| Male     |           | 52.17 | 59.36 | 56.49 | 55.13 |
| Female   |           | 47.83 | 40.64 | 43.51 | 44.87 |
| Age (%)  |          |     |     |     |     |
| 0–20 years |       | 2.20 | 0.87 | 6.09 | 2.40 |
| 20–40 years |     | 19.51 | 13.18 | 25.96 | 18.53 |
| 40–60 years |     | 37.20 | 31.00 | 43.83 | 36.29 |
| 60–80 years |     | 34.12 | 38.94 | 23.11 | 33.90 |
| >80 years  |           | 6.96 | 16.01 | 1.01 | 8.88 |
| Race/Ethnicity (%) | |     |     |     |     |
| Asian     |           | 3.24 | – | – | – |
| Black     |           | 18.59 | – | – | – |
| Hispanic  |           | 6.41 | – | – | – |
| Native    |           | 0.29 | – | – | – |
| White     |           | 67.64 | – | – | – |
| Other     |           | 3.83 | – | – | – |
| Insurance (%) |        |     |     |     |     |
| Medicare  |           | 46.07 | – | – | – |
| Medicaid  |           | 8.98 | – | – | – |
| Other     |           | 44.95 | – | – | – |
| AUC ± 95% CI |       | 0.834 ± 0.001 | 0.805 ± 0.001 | 0.835 ± 0.002 | 0.859 ± 0.001 |

The datasets studied are MIMIC-CXR (CXR)\(^{1–3}\), CheXpert (CXP)\(^{27}\), ChestX-ray14 (NIH)\(^{28}\) and a multi-source dataset (ALL) composed of aggregated data from the CXr, CXP and NIH datasets using the shared labels (disease labels and the no finding label) in all three datasets. The deep learning model is trained on each of the CXR, CXP, NIH and ALL datasets. The model’s AUCs are then estimated for each of the labels in the CXr (14 labels), CXP (14 labels), NIH (15 labels) and ALL (8 labels) datasets, and are averaged over all of the labels for each dataset. The reported AUC ± 95% confidence interval (CI) for each dataset is then the average of the AUCs for the five trained models with different random seeds using the same train-validation-test split.
Fig. 2 | Analysis of underdiagnosis across subgroups of sex, age, race/ethnicity and insurance type in the MIMIC-CXR (CXR) dataset. 

a. The underdiagnosis rate, as measured by the no finding FPr, in the indicated patient subpopulations. 

b. Intersectional underdiagnosis rates for female patients (b(i)), patients aged 0–20 years (b(ii)), Black patients (b(iii)), and patients with Medicaid (b(iv)). c.d. The overdiagnosis rate, as measured by the no finding FNr in the same patient subpopulations as in a and b. The results are averaged over five trained models with different random seeds on the same train-validation-test splits. 95% confidence intervals are shown. Subgroups with too few members to be studied reliably (≤15) are labeled in gray text and the results for these subgroups are omitted. Data for the Medicare subgroup are also omitted, given that data for this subgroup are highly confounded by patient age.
Articles

...Therefore, we would expect our results to be less biased in clinical care itself, in which under-served subpopulations have been sources of bias amplification concerns36. Moreover, missing data, small sample size and the consistently suboptimal care delivered to some patients with low socioeconomic status may have fewer interactions with the healthcare system, or they may be more likely to visit a teaching or research clinic where clinical reasoning or treatment plans may be different33. Our results may not be replicable in health settings in which the dynamics of sex or racial identity are different, or in which the health insurance system operates differently.

Third, although there are possible post-hoc technical solutions for imposing fairness, it comes with deep flaws. One simple post-processing method for achieving equal FNR and FPR across subgroups is the selection of different thresholds for different groups corresponding to the intersection of their receiver operating characteristic (ROC) curves34,35. However, there are many flaws involved in using a different threshold for each group. For example, for intersectional subgroups with small populations, an accurate approximation of the threshold might be difficult to obtain because of the large degree of uncertainty. The number of thresholds required to be computed also grows exponentially with the number of protected attributes, which makes it largely infeasible for interceptions of three or more protected attributes. Additionally, race and ethnicity are partially social constructs, with unclear boundaries. As a result, self-reported race and ethnicity may be inconsistent and may vary based on individual factors such as age, socioeconomic level or the level of acculturation to society36. This heterogeneity in self-identification may result in lower model performance for patients of groups in which self-identification criteria are more complex. Finally, this solution is ideal only in cases in which the per-group ROC curves have intersections. In cases in which the ROC curves do not intersect, or we desire an FNR–FPR combination not corresponding to an intersection between curves, achieving equal FNR and FPR would require randomization—that is, systematic worsening of the model performance in particular subgroups37. It is unclear whether worsening the overall model performance for one subgroup to achieve equality is ethically desirable. This is especially relevant in the medical context, in which we do not expect that all subgroups will have similar areas under the ROC curve (AUCs), given that the difficulty of the problem often varies with the protected group, for example, with age. We do note that equal FPR alone is easily achievable through threshold adjustments if the underdiagnosis is the main fairness concern. However, such a solution could still induce large overdagnosis (FN) disparities, in addition to requiring knowledge of the patients’ group membership.

Fourth, despite the fact that we do not have the same disease prevalence between subgroups based on real data38,39, and our choice of fairness metrics does not directly involve prevalence between subgroups, we stress that equal underdiagnosis rates between subgroups of age, sex and race/ethnicity are still desired. If a classifier deployed in a clinical pipeline mistakenly underdiagnosed a certain subgroup (for example, Black patients) more than others due to the lower prevalence of the disease, this still leads to disadvantage for members of that group and could lead to serious ethical concerns34.

Fifth, we note that fairness definitions must be chosen carefully in a healthcare context, given that many definitions are not concurrently satisfiable as shown through fairness impossibility theorems38–40. For example, given that the base rates of the two groups are different, it is impossible for them to have equal FNR, FPR and FDR, unless the classifier predicts all samples perfectly40.

Last, regulatory and policy decision-makers must consider underdiagnosis. Our work demonstrates the importance of detailed evaluation of medical algorithms, even those that are built with seemingly robust model pipelines. Given that medical algorithms are increasingly widespread, practitioners should assess key metrics such as differences in underdiagnosis rates and other health disparities during the model development process and again after deployment. Furthermore, the clinical application and historical context of each medical algorithm and the potential biases in data gathering should guide the type and frequency of introspection. Moving AI-based decision-making models from paper to practice without considering the biases that we have shown, as well as the ability of AI-based models to detect attributes such as the race of the patients from X-rays41, may harm under-served patients. We therefore suggest fairness checks, for underdiagnosis to be merged into the regulatory approval of medical decision-making algorithms before deployment, particularly in the case of triage, where underdiagnosis delays access to care. Moreover, developers, practitioners and the clinical staff need to take into account biases such as the underdiagnosis of under-served populations in the AI-based medical decision-making algorithms and its harmful effect on patients. Additionally, given that different fairness metrics are not concurrently satisfiable, a thorough use-based study to analyze the advantages and disadvantages of different fairness metrics is essential. Such studies guide policymakers to standardize the fairness checks of AI-based diagnostic algorithms prior to deployment. Finally, it is important to acknowledge that a rapidly changing research landscape can yield iterative modifications to regulations as we continue to better understand how algorithmic bias can permeate medical algorithms.

### Table 2 | Age and sex subgroups with the most underdiagnosis and least overdiagnosis for all four datasets

| Subpopulation       | CXR  | CXP  | NIH  | ALL  |
|---------------------|------|------|------|------|
| **Most underdiagnosed group** |
| Sex                 | Female | Female | Male | Female |
| Age (years)         | 0–20 | 20–40 | >80  | 0–20 |
| Female–Age (years)  | 0–20 | 20–40 | 0–20 | 0–20 |
| **Least overdiagnosed group** |
| Sex                 | Female | Female | Male | Female |
| Age (years)         | 0–20 | 20–40 | 0–20 | 0–20 |
| Female–Age (years)  | 0–20 | 20–40 | 0–20 | 0–20 |
In conclusion, we demonstrate evidence of AI-based underdiagnosis against under-served subpopulations in diagnostic algorithms trained on chest X-rays. Clinically, underdiagnosis is of key importance because undiagnosed patients incorrect receive no treatment. We observe, across three large-scale datasets and a combined multi-source dataset, which under-served subpopulations are consistently at significant risk of algorithmic underdiagnosis. Additionally, patients in intersectional subgroups (for example, Black female patients) are particularly susceptible to algorithmic underdiagnosis. Our findings demonstrate a concrete way that deployed algorithms (for example, https://models.acrdsi.org/) could escalate existing systemic health inequities if there is not a robust audit of performance disparities across subpopulations. As algorithms move from the laboratory to the real world, we must consider the ethical concerns regarding the accessibility of medical treatment for under-served subpopulations and the effective and ethical deployment of these models.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-021-01595-0.

Received: 20 January 2021; Accepted: 28 October 2021; Published online: 10 December 2021

References

1. Raghavan, M., Barocas, S., Kleinberg, J. & Levy, K. Mitigating bias in algorithmic hiring: evaluating claims and practices. In FAT’10: Proceedings of the 2010 Conference on Fairness, Accountability, and Transparency 469–481 (Association for Computing Machinery, 2020).

2. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25, 1337–1340 (2019).

3. Char, D. S., Eisenstein, L. G. & Jones, D. S. Implementing machine learning in health care: addressing ethical challenges. N. Engl. J. Med. 378, 981–983 (2018).

4. Chen, I. Y., Joshi, S. & Ghassemi, M. Treating health disparities with artificial intelligence. Nat. Med. 26, 16–17 (2020).

5. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453 (2019).

6. Larrazabal, A. J. et al. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl Acad. Sci. USA 117, 12592–12594 (2020).

7. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: fairness gaps in deep chest X-ray classifiers. In Pacific Symposium on Biocomputing 2021 (eds Altman, R. B. et al.) 232–243 (World Scientific Publishing, 2021).

8. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight: reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med. 383, 874–882 (2020).

9. Mamary, A. J. et al. Race and gender disparities are evident in COPD underdiagnoses across all severities of measured airflow obstruction. Chronic Obstruct. Pulm. Dis. 5, 177–184 (2018).

10. Sun, T. Y., Bear Don’t Walk, O. J. IV, Chen, J. L., Reyes Nieva, H. & Elhadad, N. The causes of medical malpractice suits against radiologists in the United States. Radiology 266, 548–554 (2013).

11. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C. & Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 11236–11245 (IEEE, 2019).

12. Wang, X. et al. Learning image labels on-the-fly for training robust classification models. Preprint at https://arxiv.org/abs/2009.10335v2 (2020).

13. Cohen, J. P., Hashir, M., Brooks, R. & Bertrand, H. On the limits of cross-domain generalization in automated X-ray prediction. Proc. Mach. Learn. Res. 121, 136–155 (2020).

14. Iliouzou, I. & Ben Ahmed, M. A novel approach for multi-label chest X-ray classification of common thorax diseases. IEEE Access 7, 64279–64288 (2019).

15. Rimmer, A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ 359, j4683 (2017).

16. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).

17. James, J. T. A new, evidence-based estimate of patient harms associated with hospital care. J. Patient Saf. 9, 122–128 (2013).

18. Whang, J. S., Baker, S. R., Patel, R., Luk, L. & Castro, A. III The causes of medical malpractice suits against radiologists in the United States. Radiology 266, 548–554 (2013).

19. Akbarian, S., Seyyed-Kalantari, L., Khalvati, F. & Dolatabadi, E. Evaluating transfer knowledge in neural networks for medical images. Preprint at https://arxiv.org/abs/2008.13574 (2020).

20. Wang, X. et al. Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology 294, 199–209 (2020).

21. Bhatt, M. L. B., Kent, S. & Bhaskar, R. Pulmonary tuberculosis as differential diagnosis of lung cancer. South Asian J. Cancer 1, 36–42 (2012).

22. Verma, S. & Rubin, J. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) 1–7 (IEEE, 2018).

23. Zhang, H. et al. An empirical framework for domain generalization in clinical settings. In CHIL 21: Proceedings of the Conference on Health, Inference, and Learning 279–290 (Association for Computing Machinery, 2021).

24. Zhang, H., Lu, A. X., Abdalla, M., McDermott, M. & Ghassemi, M. Hurtle words: quantifying biases in clinical contextual word embeddings. In CHIL ’20: Proceedings of the ACM Conference on Health, Inference, and Learning 110–120 (Association for Computing Machinery, 2020).

25. Aranga, S. et al. FIA: Fairness, Accountability, and Transparency in Multilingual Text-based Medical Imaging. In FAT ’19: Proceedings of the Conference on Fairness, Accountability, and Transparency 120–128 (Association for Computing Machinery, 2019).

26. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Re, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In CHIL ’20: Proceedings of the ACM Conference in Health, Inference, and Learning 151–159 (Association for Computing Machinery, 2020).

27. Zhao, J., Wang, T., Yatskar, M., Sordone, V. & Chang, K.-W. Men also like shopping: reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing 2979–2989 (Association for Computational Linguistics, 2017).

28. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern. Med. 178, 1544–1547 (2018).

29. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (NIPS 2016) (eds Lee, D. et al.) 3315–3323.

30. Barocas, S., Hardt, M. & Narayanan, A. Fairness and Machine Learning (FairBook.org, 2019).

31. Morning, A. The racial self-identification of South Asians in the United States. J. Ethn. Migr. Stud. 27, 61–79 (2001).

32. Barocio, M., Cisneros, P. & Loubes, J.-M. Review of mathematical methods for fairness in machine learning. Preprint at https://arxiv.org/abs/2005.13754 (2020).

33. Banerjee, I. et al. Reading race: AI recognises patient’s racial identity in medical images. Preprint at https://arxiv.org/abs/2107.10356 (2021).
Methods

Dataset. We have utilized three large public chest X-ray datasets in this study: MIMIC-CXR (CXR)\(^{3,4}\), CheXpert (CXP)\(^{27}\) and ChestX-ray14 (NIH)\(^{22}\). The CXR dataset was collected from Beth Israel Deaconess Medical Center (Boston, MA, United States) between 2011 and 2016, the CXP dataset was collected from Stanford Hospital (Stanford, CA, United States) between October 2002 and July 2017, and the NIH dataset was collected from the NIH Clinical Center (Bethesda, MD, United States) between 1992 and 2015. The CXR and CXP datasets contain 14 diagnosis labels and the NIH dataset contains 15 diagnosis labels, and all contain one extra label indicating no predicted diagnosis of the other disease labels (‘no finding’). We focus on the no finding label for our underdiagnosis analysis. Disease labels are consistent in CXR and CXP, while only eight labels of the NIH dataset are matched with them. In the multi-source ALL dataset we aggregate the three aforementioned datasets on the eight shared labels.

Dataset collection and inclusion criteria. Because of the size of these large datasets and the fact that no exclusion criteria are mentioned in the dataset descriptions, we do not anticipate any issues with selection bias and assume that the collected datasets are representative of patients at these hospitals over the specified years. Only the CheXpert14 dataset is gathered from the NIH clinical research dedicated hospital, where patients are treated without charge and are selected based on whether the illness is being studied by the Institutes.

The NIH dataset has only frontal view images, whereas the other datasets have both frontal and lateral view images. We include all of the images of each dataset, regardless of the view, in the model training and evaluation. The race/ethnicity and sex data are collected in the MIMIC-CXR dataset and the age is reported at the patient’s first admission. In the CheXpert dataset, sex is assigned by clinicians and the age is reported at the time of the examination. In the CheXpert14 dataset, the sex is self-identified and the age corresponds to the time of the examination. In the MIMIC-CXR dataset, the race/ethnicity and insurance type data were collected only if the patient was admitted to an intensive care unit, therefore there are around 100,000 X-rays for which we do not have these data (these are X-rays done for patients who were admitted only to the emergency department. The reported race/ethnicity in the MIMIC-CXR dataset are white, other, Hispanic, Latino, Black/African American, and American Indian/Alaska Native, and in this study we have used the shorter terminology white, other, Hispanic, Black, and Native for each group, respectively.

Definition and quantification of the fairness metrics. Commonly used fairness definitions such as equality of odds and equality of opportunity\(^{37}\) rely on equalized prediction metrics across subgroups. We evaluate the fairness of models in binarized fairness metrics because binarized prediction is most often required for clinical decision-making at the individual level. To assess model decision biases in underdiagnosed patients we compare underdiagnosis rates across subpopulations. We define the underdiagnosis rate as the FPR of the binarized model prediction for the no finding label at the level of the subgroup (s), that is, FPR\(_s\), for example, female patients) and the intersection (s\(_i\)) identities, that is, FPR\(_{s_i}\), (for example, Black female patients), as given by:

\[
FPR_s = P [\hat{Y} = 1 | s, Y = 0]
\]

(1)

\[
FPR_{s_i} = P [\hat{Y} = 1 | s_i, Y = 0]
\]

(2)

where i, j denote subgroups with distinct attributes, Y is the true label and \(\hat{Y}\) is the predicted label. We then compare these underdiagnosis rates across subpopulations including age and sex in all four datasets, as well as race/ethnicity and insurance type in the CXR dataset specifically.

Additionally, we evaluate the FNR for the no finding label across all subgroups (the definitions are similar to equation (1) and equation (2), but with \(\hat{Y} = 0\) and \(Y = 1\) with the patients belonging to s, or s\(_i\)). This measure is useful to help differentiate between overall model noise (for example, when predictions are flipped at random in either direction), which would result in approximately correlated FPR and FNR rates across subgroups, and selective model noise (for example, when predictions are selectively biased towards a prediction of no finding), which would result in un- or anti-correlated FPR and FNR rates. Although both kinds of noise are problematic, the latter is a form of technical bias amplification because it would show the known bias of the model in the training of the model on the NIH dataset as well as all the other datasets. The train—validation—test set sizes for the ALL dataset are 575,381–67,177–65,068, for the CXR dataset they are 298,137–37,300–36,421, for the CXP dataset they are 178,352–23,022–22,274 and for the NIH dataset they are 98,892–6,855–6,373, respectively. The splits are random, and no patient is shared across splits. We use the same split as in ref.\(^{7}\). The ALL dataset aggregates the original splits of the CXR and CXP and NIH datasets. Therefore, patients in the test set of each individual dataset stay in the test set of the ALL split. We applied center crop and random horizontal flip data augmentation. Similar to ref.\(^{7}\), for the NIH dataset we applied a 10° rotation, and for the other datasets we applied a 15° rotation. Data augmentation for model training.

Model training. The trained models used in this study are identical to that of ref.\(^{7}\) for all datasets, except for the NIH dataset. We train a 121-layer DenseNet\(^{46}\), with weights initialized using ImageNet\(^{42}\). Given that we need the no finding label, we include this label in the training of the model on the NIH dataset as well as all the other datasets. The train—validation—test set sizes for the ALL dataset are 575,381–67,177–65,068, for the CXR dataset they are 298,137–37,300–36,421, for the CXP dataset they are 178,352–23,022–22,274 and for the NIH dataset they are 98,892–6,855–6,373, respectively. The splits are random, and no patient is shared across splits. We use the same split as in ref.\(^7\). The ALL dataset aggregates the original splits of the CXR and CXP and NIH datasets. Therefore, patients in the test set of each individual dataset stay in the test set of the ALL split. We applied center crop and random horizontal flip data augmentation. Similar to ref.\(^7\), for the NIH dataset we applied a 10° rotation, and for the other datasets we applied a 15° rotation. Data augmentation for model training.

Data availability

All three datasets used for this work are public under data use agreements. We have followed the protocols associated with the data sources. All experiments are conducted on observational, retrospective data. All datasets are referenced in the paper: the MIMIC-CXR\(^{3,4}\) dataset is available at https://physionet.org/content/mimic-cxr/2.0.0/; the CheXpert\(^{27}\) dataset is available at https://stanfordmlgroup.github.io/competitions/chexpert/ and the ChestX-ray14\(^{22}\) dataset is available at https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community. Access to all three datasets requires user registration and the signing of a data use agreement, after which access is provided in a timely manner. Only the MIMIC-CXR dataset requires the completion of an additional credentialing process. After following these procedures, the MIMIC-CXR data are available through PhysioNet\(^{42}\). The MIMIC-CXR project page on PhysioNet describes the data access procedure\(^{42}\). The race/ethnicity and insurance type for the patients are not provided directly with the download of the MIMIC-CXR dataset. However, these data are available by merging the patient IDs in MIMIC-CXR with subject IDs in MIMIC-IV\(^{47}\) using the patient and admissions tables. Access to MIMIC-IV requires a similar procedure as MIMIC-CXR and the same credentialing process is applicable for both datasets.
Code availability
The code for training the models on the MIMIC-CXR (CXR)\(^26\), CheXpert (CXP)\(^27\) and ALL datasets is identical to that in https://github.com/LalehSeyyed/CheXclusion. The code for training the ChestX-ray14 (NIH)\(^28\) dataset on 15 labels as well as the code for all of the analyses in this paper is presented in https://github.com/LalehSeyyed/Underdiagnosis_NatMed. We have provided the Conda environment in the same repository for the purpose of reproducibility. We are not able to share the trained model and the true labels and predicted labels CSV files of the test set due to the data-sharing agreement. However, we have provided the patient ID per test splits, random seed and the code. The true label and predicted label CSV files and trained models can then be generated by users who have downloaded the data from the original source following the procedure described in the Data Availability section.

References
42. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252 (2015).
43. Iandola, F. et al. Densenet: implementing efficient ConvNet descriptor pyramids. Preprint at https://arxiv.org/abs/1404.1869v1 (2014).
44. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101, e215–e220 (2000).
45. Johnson, A., Pollard, T., Mark, R., Berkowitz, S. & Horng, S. MIMIC-CXR database. PhysioNet https://doi.org/10.13026/C2JT1Q (2019).
46. Johnson, A. et al. MIMIC-IV (version 0.4). PhysioNet https://doi.org/10.13026/a3wn-hq65 (2020).

Acknowledgements
The authors thank M. Haider for helpful discussions and acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC, grant PDF-516984 to L.-S.-K.), Microsoft Research (M.G.), Canadian Institute for Advanced Research (CIFAR) (M.G.) and an NSERC Discovery Grant (to M.G.). The authors also thank Vector Institute for providing high-performance computing platforms.

Author contributions
L.S.-K., H.Z., M.B.A.M., I.Y.C. and M.G. have substantially contributed to the underlying research and drafting of the paper.

Competing interests
The authors declare no competing interests.

Additional information
Extended data are available for this paper at https://doi.org/10.1038/s41591-021-01595-0.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-021-01595-0.
Correspondence and requests for materials should be addressed to Laleh Seyyed-Kalantari.

Peer review information Nature Medicine thanks Luke Oakden-Rayner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Analyzing underdiagnoses over subgroups of sex, age, within ALL dataset (combined CXR, CXP and NIH dataset on shared labels). Fig. S1. Analyzing underdiagnoses over subgroups of sex, age, within ALL dataset (combined CXR, CXP and NIH dataset on shared labels). The results are averaged over 5 trained model with different random seed ± 95% confidence interval (CI). A. The underdiagnosis rate (measured by 'No Finding' FPR). B. The overdiagnosis rate ('No Finding' False Negative Rate (FNR)) over subgroups of sex, age. C. The intersectional underdiagnosis rates within only female patients. D. Examining the overdiagnosis rate for the intersectional identities. The number of images with actual 0 or 1 'No Finding' label in the age - sex intersections in the test dataset is presented in Supplementary Table 1.
Extended Data Fig. 2 | Analyzing underdiagnoses over subgroups of sex, age, within CheXpert (CXP) dataset. Fig. S2. Analyzing underdiagnoses over subgroups of sex, age, within CheXpert (CXP) dataset. The results are averaged over 5 trained model with different random seed ± 95% CI. A. The underdiagnosis rate is FPR in ‘No Finding’. B. Examining the overdiagnosis rate (‘No Finding’ FNR) over sex and age subgroups, C. The intersectional underdiagnosis rates within only female patients, and D. measure the overdiagnosis rate for the intersectional identities. The subgroups labeled in gray text, with results omitted, indicate the subgroup has too few members (<= 15) to be used reliably. The number of images with actual 0 or 1 ‘No Finding’ label in the age - sex intersections in the test dataset is presented in Supplementary Table 1.
Extended Data Fig. 3 | Analyzing underdiagnoses over subgroups of sex, age, within ChestX-ray14 (NIH) dataset. Fig. S3. Analyzing underdiagnoses over subgroups of sex, age, within ChestX-ray14 (NIH) dataset. The results are averaged over 5 trained model with different random seed ± 95% confidence interval (CI). A. The underdiagnosis rate (‘No Finding’ FPR). B. The over diagnosis rate (‘No Finding’ FNR) over subgroups of sex and age. C. The intersectional underdiagnosis rates within only female patients. D. The over diagnosis rate for the intersectional identities. The subgroups labeled in gray text, with results omitted, indicate the subgroup has too few members (<= 15) to be used reliably. The number of images with actual 0 or 1 ‘No Finding’ label in the age - sex intersections in the test dataset is presented in Supplementary Table 1.
Reporting Summary

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

Statistics

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

n/a  Confirmed

☐  The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐  A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐  The statistical test(s) used AND whether they are one- or two-sided

☐  *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*

☐  A description of all covariates tested

☐  A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐  A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐  For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.

☐  For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐  For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐  Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

| Data collection | Data was publicly available. We have not collected the data and no software was used in our side to collect data. We have provided a full reference on data availability session. |
|-----------------|----------------------------------------------------------------------------------|

| Data analysis   | The code for training the models on MIMIC-CXR (CXR) [26], CheXpert (CXP) [27], and ALL datasets is identical in [https://github.com/LaiehSeyyed/CheXclusion]. The code for training ChestX-ray14 (NIH) [28] datasets on 15 labels as well as the code for all the analyses in this paper is presented in [https://github.com/LaiehSeyyed/Underdiagnosis_NatMed]. We have provided the.yml Conda environment in the same repository for reproducibility purposes. We are not able to share the trained model and the true label and predicted label CSV files of the test set due to the data-sharing agreement. But we have provided the patient ID per test splits, random seed, and the code. Then the true label and predicted label CSV files and trained models can be generated by users who have downloaded the data from the original source following the procedure that is described in 'Data availability' session. |

channels:
- pytorch
- conda-forge
- anaconda
- defaults

dependencies:
- blas=1.0=mkl
- ca-certificates=2019.5.15=0
- certifi=2019.3.9=py36_0
- cffi=1.12.3=py36h2e261b9_0
- cudatoolkit=10.0.130=0
- cycler=0.10.0=py_1
Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All three datasets that have been used for this work are public and under data use agreements. We have followed the data use agreements, and the experiments are based on observational, retroactive data. The datasets are all well-referenced in the paper. Here is the link to each of the datasets:

- MIMIC-CXR (25) dataset is available at: https://physionet.org/content/mimic-cxr/2.0.0/
- ChexNet (27) dataset is available at: https://stanfordmlgroup.github.io/competitions/chexpert/
- ChestX-ray14 (28) dataset is available at: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

Access to all three datasets requires user registration and the signing of a data use agreement. Then access is provided in a timely manner. Only the MIMIC-CXR dataset also requires the completion of a credentialing process, that takes a few hours to be completed. After following this procedure the MIMIC-CXR data is available through PhysioNet (42). The MIMIC-CXR project page on PhysioNet describes the data access procedure (43). The race/ethnicities and insurance type of the patients are not provided naturally with the download of the MIMIC-CXR dataset. However, this data is available through merging the patient IDs in MIMIC-CXR with subject IDs in MIMIC-IV (44) using the patient and admissions tables. Access to MIMIC-IV requires a similar procedure as MIMIC-CXR and the same credentialing process is applicable for access to both datasets.

Field-specific reporting

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

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

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

Life sciences study design

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

| Sample size | In total, we use over 98,000 chest x-ray images for model training, with specific ablations for subgroups and dataset source. While there is no definitive threshold set for convergence of deep neural network model training, the machine learning literature generally suggests that over 50,000 samples is appropriate for convolutional neural network convergence in fine-tuning of natural and medical images. We have randomly samples 10%, 10%, and 80% of the patients of each whole dataset for train, validation and test set, such that each patient medical images belongs to only one of the train, test and validation sets. We have not done any other sampling form the whole dataset in our model development. The detail on the number of images per dataset (sample size) and where they have been collected are presented in Human research participant session in this document (next page.) |
| Data exclusions | No data was excluded. |
| Replication | To ensure reproducibility, we save all model random seeds, and have released the source code for model training upon acceptance. |
| Randomization | As no human subject evaluation was performed, we did not require randomization groups. |
| Blinding | As this study contained no human evaluation or intervention, and only profiles computational models on a fixed dataset, no blinding was necessary |

Reporting for specific materials, systems and methods

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

Policy information about studies involving human research participants

| Population characteristics |  |
|----------------------------|---|
| We have used already existing public data on human chest X-rays and we have not collect them. The distribution of data over sex and age is provided in table 2. One of the datasets has the race and insurance type of the patients where we have reported in the Table 2. |
| # images                   |  |
| CXR: 371,858               |  |
| CXP: 223,648               |  |
| NIH: 112,120               |  |
| Sex                        |  |
| CXR - Male: 52.17%         |  |
| CXR - Female: 47.83%       |  |
| CXP - Male: 59.36%         |  |
| CXP - Female: 40.64%       |  |
| NIH - Male: 56.49%         |  |
| NIH - Female: 43.51%       |  |
| Age - CXR                  |  |
| 0-20: 2.20%                |  |
| 20-40: 19.51%              |  |
| 40-60: 37.20%              |  |
| 60-80: 34.12%              |  |
| 80+: 6.96%                 |  |
| Age - CXP                  |  |
| 0-20: 0.87%                |  |
| 20-40: 13.18%              |  |
| 40-60: 31.00%              |  |
| 60-80: 38.94%              |  |
| 80+: 16.01%                |  |
| Age - NIH                  |  |
| 0-20: 6.09%                |  |
| 20-40: 25.90%              |  |
| 40-60: 43.83%              |  |
| 60-80: 38.94%              |  |
| 80+: 1.01%                 |  |
| CXR: Race/ethnicity        |  |
| Asian: 3.24%               |  |
| Black: 18.59%              |  |
| Hispanic: 6.41%            |  |
| Native: 0.29%              |  |
| White: 67.64%              |  |
| Other: 3.83%               |  |
| ---                         |  |
| CXR - Insurance            |  |
| Medicare: 46.07%           |  |
| Medicaid: 8.38%            |  |
| Other: 44.95%              |  |

Recruitment

Because of the size of these large datasets, and the fact that no exclusionary criteria are mentioned in the dataset descriptions, we do not anticipate any issues with selection bias and assume that the collected datasets are representative of patient's at these hospitals over the specified years. Only the ChestX-ray14 dataset is gathered from the NIH clinical research dedicated hospital where patients are treated without charge, unlike most hospitals, the Clinical Center does not routinely
provide standard diagnostic and treatment services. Admission is selective: patients are chosen by Institute physicians solely because they have an illness being studied by those Institutes’, as mentioned in their website [https://clinicalcenter.nih.gov/about/welcome/faq.html]

NIH dataset has only frontal view images where the other datasets have both frontal and lateral view images. We include all the images of each dataset regardless of their view in the model training and evaluation.

The race/ethnicity and sex data are self-reported in the MIMIC-CXR dataset and age is reported at a patient’s first admission. In the CheXpert dataset, sex is assigned by clinicians and the age is at the time of the examination. In the ChestX-ray14 dataset, the sex is self-identified and the age corresponds to the time of the examination. In the MIMIC-CXR dataset, we only have the race/ethnicity and insurance type data of a patient if the patient was admitted to an ICU, so there are around ~100,000 x rays where we do not have this data [these are x rays done for patients who were only admitted to the emergency department. The reported races/ethnicities in MIMIC-CXR dataset are WHITE, OTHER, HISPANIC/LATINO, BLACK/AFRICAN AMERICAN, AMERICAN INDIAN/ALASKA NATIVE, where in this study we have used shorter terminology White, Other, Hispanic, Black, and Native for each group, respectively.

Ethics oversight

Since we have worked on public, anonymized, retrospectively collected data, and we have not collect any human data ourselves we did not get any organizational/IRB approval. These public datasets are largely and commonly used in the machine learning medical imaging literature.

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