Medical imaging algorithms exacerbate biases in underdiagnosis

Laleh Seyyed-Kalantari[1,2]*, Guanxiong Liu[1,2], Matthew McDermott[3], Irene Y. Chen[3], Marzyeh Ghassemi[1,2]

[1]University of Toronto.

[2]Vector Institute.

[3]Massachusetts Institute of Technology.

*Correspondence to: laleh@cs.toronto.edu.

Artificial intelligence (AI) systems have increasingly achieved expert-level performance, particularly in medical imaging *(1)*. However, there is growing concern that AI systems will reflect and amplify human bias against under-served subpopulations *(2-7)*. Such biases are especially troubling in the context of underdiagnosis: if AI systems falsely predict that patients are healthy, patients would be denied care when they need it most. This use case is particularly relevant in the context of existing health disparities where high underdiagnosis rates for under-served subgroups are well documented *(8-11)*. Although bias in underdiagnosis can potentially delay access to medical treatment unequally, underdiagnosis due of AI has been relatively unexplored. In this work we examine algorithmic underdiagnosis in chest X-ray pathology classifiers and find that classifiers consistently and selectively underdiagnose under-served patients, actively amplifying the existing biases in clinical care. These effects are worse on intersectional subpopulations, e.g., Black females, and persist across three large and a multi-source chest X-ray dataset. Our work demonstrates that deploying AI systems risks exacerbating biases present in current care practices. Developers, clinical staff, and regulators must address the serious ethical concerns of -- and barriers to -- effective deployment of these models in the clinic.

Introduction

As artificial intelligence (AI) algorithms increasingly affect decision making in society *(12)*, researchers have raised concerns about algorithms creating or amplifying biases in calculators documented *(2-8)*. While AI algorithms in specific circumstances can potentially reduce bias *(13)*, direct application of AI has also been shown to systemize bias in a range of settings *(2-7, 15)*. This tension is particularly pressing in healthcare, where AI systems could improve patient health *(4)*, but can also exhibit biases *(2-7)*. Motivated by demonstrations that AI algorithms can match specialist performance in particularly in medical imaging *(1)* and the global radiologist shortage *(16)*, AI-based diagnostic tools looks a clear case for deployment.

While there is much work in algorithmic bias *(14)* and bias in health *(2-7, 8-11)*, the topic of AI-driven underdiagnosis has been relatively unexplored. Crucially, underdiagnosis -- defined as falsely claiming the patient is healthy -- leads to no clinical treatment when a patient needs it most. The existing clinical landscape demonstrates biases in underdiagnosis against under-served subpopulations. For example, Black patients compared to non-Hispanic White with chronic obstructive pulmonary disease are more under-diagnosed *(9)*, in clinical applications the risk score threshold is adjusted racially, which is potentially harmful for Black patients *(8)*, or the quality of care delivered to the patients within the same hospital varies by the patient insurance type *(11)*. Such biases can manifest algorithmically, e.g. females receiving a longer time to diagnosis than males with the same medical conditions *(10)*. In medical imaging specifically, algorithmic bias

Subgroup	Attribute	CXR	CXP	NIH	ALL
	#images	371,858	223,648	112,120	707626
Sex	Male	52.17%	59.36%	56.49%	55.13%
	Female	47.83%	40.64%	43.51%	44.87%
Age	0-20	2.20%	0.87%	6.09%	2.40%
	20-40	19.51%	13.18%	25.96%	18.53%
	40-60	37.20%	31.00%	43.83%	36.29%
	60-80	34.12%	38.94%	23.11%	33.90%
	80+	6.96%	16.01%	1.01%	8.88%
Race	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	0.834 ± 0.001	0.805±0.001	0.835 ± 0.002	0.859±0.001

Table 1. The summary statistics across datasets. The description of MIMIC-CXR (CXR) *(17)*, CheXpert (CXP) *(18)*, and Chest-Xray8 (NIH) *(19)* dataset as well as the multi-source dataset (ALL) composed of the aggregation of CXR, CXP, and NIH on shared labels. The reported AUCs are the averages on 14, 14, 15, and 8 labels of the CXR, CXP, NIH, and ALL dataset.

from chest X-rays has demonstrated asymmetric diagnosis for a range of diseases *(6)* in males patients and female, but underdiagnosis has not been investigated.

In this work, we perform a systematic study of underdiagnosis bias in a chest X-ray prediction model across three large public radiology datasets, MIMIC-CXR (CXR) *(17)*, CheXpert (CXP) *(18)*, and Chest-Xray8 (NIH) *(19)* , as well as a multi-source dataset combining all three on shared diseases. Motivated by known differences in disease manifestation in patients by sex *(6)*, age *(20)*, and race *(8)*, and the effect of insurance type in quality of received care *(11)*, we report results across all of these factors. Also we use insurance type as an imperfect proxy of socioeconomic status – e.g. patients with Medicaid insurance are often low income.

We find that algorithms trained on all settings exhibit systemic underdiagnosis biases in under-served subpopulations, including females, Black and Hispanic, younger patients, and patients of lower socioeconomic status (with Medicaid insurance). Further, we show that our observations are not consistent with an increase in overall noise on these subgroups, but instead are reflective of a specific increase in underdiagnosis alone. We find these effects persist for intersectional subgroups (e.g., Black female), and are not consistently worse in the smallest intersectional groups.

Methodology

We train distinct chest X-ray diagnosis models in four settings: the MIMIC-CXR dataset (CXR, 371,858 images from 65,079 patients) *(17)*, CheXpert (CXP, 223,648 images from 64,740

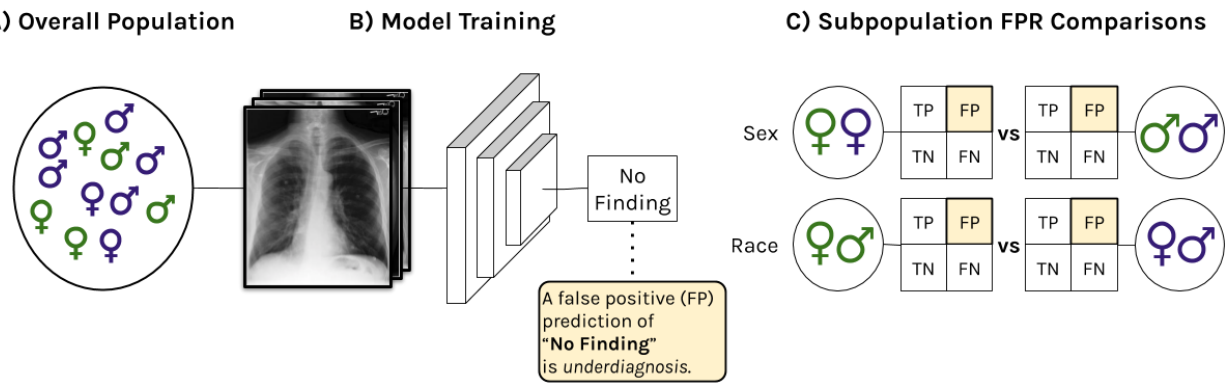

Fig. 1. The model pipeline. We (**A**) examine chest radiographs across several datasets with diverse populations, (**B**) learn a DenseNet from these data (training across all patients simultaneously), predicting the presence of the "No Finding" label -- indicates the algorithm detect no disease for image, and (**C**) compare false positive rate (FPR) of this model on different subpopulations (including sex, race, age, and insurance type) to examine the algorithm's underdiagnosis rate. If the model has a higher underdiagnosis rates on certain subpopulations, such as female or Medicaid patients, this would induce significant health disparities that lead to higher rate of no clinical treatment for certain subpopulations, were the model deployed.

patients) *(18)*, Chest-Xray8 (NIH, 112,120 images from 30,805 patients) *(19)*, and a multi-source combination of all three (ALL, 707626 images from 129,819 patients). 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 only report patient sex[1] and age, whereas the CXR additionally reports patient race and insurance type for a large subset of images. The race and insurance type attributes are highly skewed in the dataset, where White are the majority and patients with Medicaid insurance are the minorities within the dataset. The NIH dataset has only frontal view images where other datasets have also lateral view images. For more detailed summary statistics across datasets, see Table 1. For our medical imaging predictive model and training scheme, we follow best practices *(7)* and train a 121-layer DenseNet *(21)*, with weights initialized using ImageNet *(22)*. Additional details of model training and construction can be found in Appendix Section Model Training Details.

To assess and compare underdiagnosis rates across subpopulations, we compute the false positive rate (FPR) of model prediction for the "No Finding" label, which indicates no disease diagnosed, though patient suffer from at least a diagnostic disease condition. We then compare these FPRs across subpopulations including age and sex on all four datasets, as well as race and insurance type on the CXR dataset specifically. For a visual illustration of our model pipeline, see Fig. 1.

Additionally, we measure the False Negative Rate (FNR) for the "No Finding" label across all subgroups. This measure is useful to help us differentiate between overall model noise (e.g., predictions are flipped at random in either direction), which would result in roughly correlated FPR and FNR rates across subgroups, and selective model noise (e.g., predictions are selectively biased towards a prediction of "No Finding"), which would result in un- or anti-correlated FPR

[1] We use sex as reported in underlying data. Gender presentation plays a large role in societal biases, but this data is not routinely collected *(17, 18, 19)*.

and FNR rates. While both kinds of noise are problematic, the latter is a form of technical *bias amplification*, as it would show the known bias of clinical underdiagnosis is being selectively amplified by the algorithm--i.e., the model is not only failing to diagnose those patients clinicians, but is also failing to diagnose other patients as well.

Results

Underdiagnosis occurs in under-served patient subpopulations. We find the underdiagnosis rate for all datasets differs dramatically based on all measured subpopulations. In Fig. 2A we show the subgroup-specific underdiagnosis for CXR dataset on race, sex, age, and insurance type. We observed female, patients under 20 years old, Black, Hispanic, and patients with Medicaid insurance receive higher rates of algorithmic underdiagnosis than other groups. In other words, these groups are at higher risk of being falsely flagged as "healthy", and receiving no clinical treatment. Results on other datasets follow a similar trend and they are shown in the Appendix.

Underdiagnosis occurs in intersectional groups. We investigate intersectional groups -- here defined as patients who belong to two subpopulations, e.g., Black female. Similar to prior work in face detection *(15)*, we find that intersectional subgroups (see Fig. 2B) often have compounded biases in algorithmic underdiagnosis. For instance, in CXR dataset Hispanic females have a higher "No Finding" FPR than White females (see Fig. 2B-1). Also, patients less than 20 years, female, Black, or patients with Medicaid insurance who are often low income has the largest underdiagnosis rates (see Fig. 2B-2). The underdiagnosis rate for the intersection of Black patients with another subgroup of age, sex, and insurance type (see Fig. 2B-3) and patients with Medicaid insurance with another subgroup of sex, age, and race (see Fig. 2B-4) are also depicted in Fig. 2B. It is observable that the patients who are the member of two under-served subgroups are experiencing larger underdiagnosis rate. In another word, though female in subpopulation study of underdiagnosis rate (Fig. 2A) have shown to have a larger underdiagnosis rate, not all females are misdiagnosed at the same rate (see Fig. 2B-1). The intersection underdiagnosis for other datasets is shown in the Appendix where they also follow a similar pattern.

Underdiagnosis is not a result of subgroup-specific overall noise. As illustrated in Fig. 2C (FNR for 'No Finding'), FPR and FNR show inverse relationships across different under-served subgroups on the CXR dataset (though this finding is consistent across all datasets, see Appendix), for both overall and intersectional subgroups (see Fig. 2D). In other word . For emphasis, we restate that this finding is not consistent with a simple increase in overall noise for specific subgroups, but instead indicates that under-served subpopulations are being aggressively flagged erroneously as healthy by the algorithm, without a corresponding increase in false negatives.

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 incorrectly receive no treatment. We observe, across three large-scale datasets and a combined multi-source dataset, under-served subpopulations are consistently at significant risk of algorithmic underdiagnosis. Additionally, patients in intersectional subgroups (e.g., Black female) are particularly susceptible to algorithmic underdiagnosis.

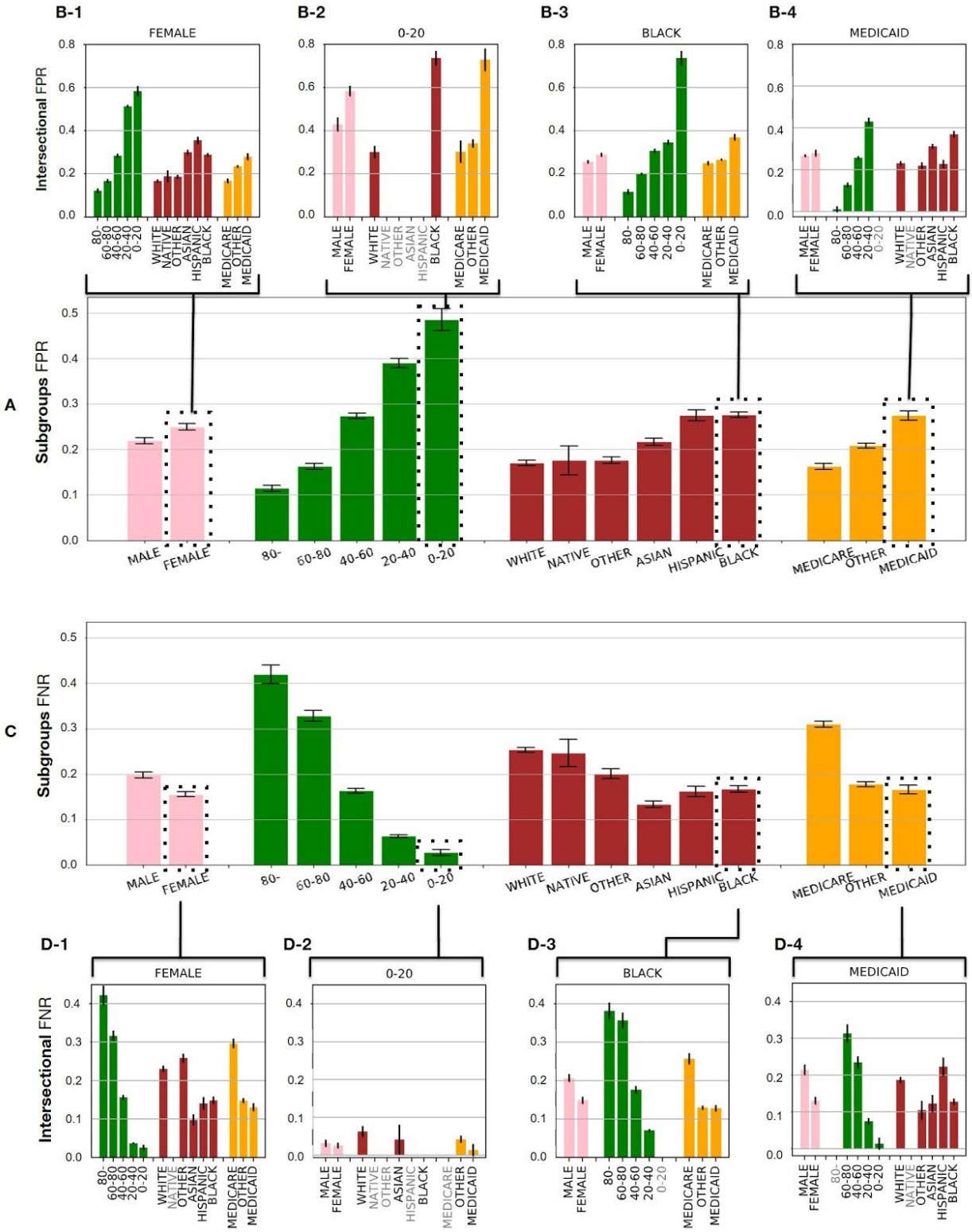

Fig. 2. Analyzing underdiagnoses over subgroups of sex, age, race & insurance type within the MIMIC-CXR (CXR) dataset. The results are averaged over 5 run with different random seed ± 95% confidence interval. **A**. The underdiagnosis rate (measured by "No Finding" FPR). Female, 0-20, Black and/or low-income patients under Medicaid insurance have the largest underdiagnosis

rate, indicating the greatest disparity. **B**. The intersectional underdiagnosis rates within only female (**B1**), ages 0-20 (**B2**), Black (**B3**), or Medicaid (**B4**) patients. In these plots, we see that intersectional identities are often underdiagnosed even more heavily than the group in aggregate (e.g., Medicaid female patients are underdiagnosed more than Medicare female patients). **C, D** We compute the same analyses on the overall subgroups (**C**), and the intersectional subgroups (**D1-4**) but now examining the "No Finding" FNR. If we observed a commensurate increase in FNR alongside the increase in FPR observed in **A**, **B**, this would indicate these results are tracking an increase in overall noise. Instead, we typically observe an inverse correlation between FPR and FNR, indicating the model is selectively underdiagnosing these vulnerable subpopulations. Throughout, subgroups labeled in gray text, with results omitted, indicate the subgroup has too few members (<= 15) to be used reliably.

This discrepancy is especially interesting in the context of known biases in clinical care itself, in which under-served subpopulations are often underdiagnosed by doctors without a simultaneous increase in privileged group overdiagnosis *(9)*. Our prediction labels are provided by these same doctors, and are therefore not an unbiased ground truth -- in other words, our labels should already suffer from this same bias that our model is then additionally exhibiting. This is a form of bias amplification, when a model's predicted outputs amplify a known source of error in the data generative process *(23)* or data distribution *(24)*. This is an especially dangerous outcome for machine learning models in healthcare, as it indicates that the existing biases in health practice risk being magnified, rather than ameliorated, by algorithmic decisions based on large (707,626 images), multi-source datasets. While this evaluation is for chest x-ray diagnostic imaging, this issue is likely widespread across data sources, and prediction tasks.

Our findings demonstrate a concrete way that algorithms escalate existing systemic health inequities. As algorithms move from the lab to the real world, we must consider the ethical concerns about the accessibility of medical treatment for under-served subpopulations and effective and ethical deployment of these models.

References:

1. Pranav, R., et al., CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning, in (CVPR, 2017).
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**(11), 981–983 (2018).
4. Chen, I. Y., Joshi, S. & Ghassemi, M., Treating health disparities with artificial intelligence, *Nat Med.* **26**(1)**,** 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. *PNAS* **117**(23), 12592–12594 (2020).

7. Seyyed-Kalantari, L., et al., CheXclusion: Fairness gaps in deep chest X-ray classifiers, in (PSB, 2021), p. 232–243.

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**(9), 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 Obstructive Pulmonary Diseases: Journal of the COPD Foundation* **5**(3), 177-184 (2018).

10. Sun, T. Y., et al., Exploring gender disparities in time to diagnosis; TY Sun, OJ Walk IV, JL Chen, HR Nieva, N Elhadad – in (ML4H at NeurIPS, 2020).

11. Spencer, C. S., Gaskin, D. J. & Roberts, E. T., The quality of care delivered to the patients within the same hospital varies by insurance type, *Health Aff (Milwood).*, **32**(10)**,** 1731-9 (2013).

12. Raghavan, M., Barocas S. & Kleinberg J., Mitigating bias in algorithmic hiring: Evaluating claims and practices (FAT*, 2020), p. 469–481.

13. Cowgill, B., "Bias and productivity in humans and machines, Upjohn Institute Working Paper", (W. E. Upjohn Institute for Employment Research, Kalamazoo, MI, 2018).

14. Dwork, C., et al., Fairness through awareness, in (ITCS, 2012), p. 214–226.

15. Buolamwini, J. & Gebru, T., Gender shades: Intersectional accuracy disparities in commercial gender classification, *PMLR* **81**, 77–91 (2018).

16. .Rimmer, A., Radiologist shortage leaves patient care at risk, *BMJ*: **359** (2017).

17. Johnson, A. E. W., *et al*., MIMIC-CXR: A large publicly available database of labeled chest radiographs, *Science Data* **6,** 317 (2019).

18. Irvin, J., *et al*., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in (CVPR, 2019).

19. Wang, X., *et al*., ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, (CVPR, 2017), p. 2097-2106.

20. Bhatt, M. L. B., Kant, S. & Bhaskar R., Pulmonary tuberculosis as differential diagnosis of lung cancer, *South Asian J. of cancer* **1**(1), 36-42 (2012).

21. Iandola, F., et al. Densenet: Implementing efficient ConvNet descriptor pyramids, in (CVPR, 2014).

22. Russakovsky, O., et al., Imagenet large scale visual recognition challenge, IJCV, **115**(3) 211-252, (2015).

23. Oakden, L., Dunnmon, J., Carneiro, G. & Re, C., Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, in (CHIL, 2020), p. 151-159.

24. Zhao, J., et al., Men also like shopping: Reducing gender bias amplification using corpus-level constraints, (EMNLP, 2017), p. 2979-2989.

Acknowledgments: we acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) - funding number PDF-516984, Microsoft Research, Canadian Institute for Advanced Research (CIFAR,) NSERC Discovery Grant; **Author contributions**: Each named author has substantially contributed to conducting the underlying research and drafting the manuscript.; **Competing interests:** Authors declare no competing interests.; and **Data and materials availability:** All 3 datasets that we have used for this work are public under data use agreements. MIMIC-CXR dataset available at: https://physionet.org/content/mimic-cxr/2.0.0/ CheXpert dataset is available at: https://stanfordmlgroup.github.io/competitions/chexpert/ ChestX-ray8 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. **Code availability** All code used in the analysis will be available in a public repository for purposes of reproducing or extending the analysis. The link to the code will be added to the text of the paper for the camera ready version.

Supplementary Materials:

Model Training Details
Additional Results
Figs. S1 to S3