Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]

[a]Research Institute for Signals, Systems and Computational Intelligence sinc(i), Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas CONICET, Santa Fe CP3000, Argentina; [b]Instituto de Matemática Aplicada del Litoral, Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas, Santa Fe CP3000, Argentina; and [c]Facultad de Ingeniería, Universidad Nacional de Entre Ríios, Oro Verde CP3100, Argentina

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved April 30, 2020 (received for review October 30, 2019)

Artificial intelligence (AI) systems for computer-aided diagnosis and image-based screening are being adopted worldwide by medical institutions. In such a context, generating fair and unbiased classifiers becomes of paramount importance. The research community of medical image computing is making great efforts in developing more accurate algorithms to assist medical doctors in the difficult task of disease diagnosis. However, little attention is paid to the way databases are collected and how this may influence the performance of AI systems. Our study sheds light on the importance of gender balance in medical imaging datasets used to train AI systems for computer-assisted diagnosis. We provide empirical evidence supported by a large-scale study, based on three deep neural network architectures and two well-known publicly available X-ray image datasets used to diagnose various thoracic diseases under different gender imbalance conditions. We found a consistent decrease in performance for underrepresented genders when a minimum balance is not fulfilled. This raises the alarm for national agencies in charge of regulating and approving computer-assisted diagnosis systems, which should include explicit gender balance and diversity recommendations. We also establish an open problem for the academic medical image computing community which needs to be addressed by novel algorithms endowed with robustness to gender imbalance.

gendered innovations | deep learning | computer-aided diagnosis | medical image analysis | gender bias

Artificial intelligence (AI) influences almost every aspect of our daily life. The media articles we read, the movies we watch, even the driving road map we take are somehow influenced by these systems. In particular, the rise of AI in healthcare during the last few years is changing the way medical doctors diagnose, especially when dealing with medical images. AI systems cannot only augment the information provided by such images with useful annotations (1, 2), but they are also starting to take autonomous decisions by performing computer-aided diagnosis (CAD) (3, 4).

Although the interest in performing fair and unbiased evaluations of AI medical systems has existed since the 1980s (5), the ethical aspects of AI have gained relevance in the last few years. It has been shown that human bias, such as gender and racial bias, may not only be inherited but also amplified by AI systems in multiple contexts (6–9). For example, face recognition systems have been shown to exhibit accuracy disparities depending on gender and ethnicity, with darker-skinned females being the most misclassified group (10). This tendency of AI systems to learn biased models, which reproduce social stereotypes and underperform in minority groups, is especially dangerous in the context of healthcare (11, 12).

In recent years, the research community of gendered innovations has largely contributed to create awareness and integrate sex and gender analyses into all phases of basic and applied research (13). However, such assessment in the context of medical imaging and CAD remains largely unexplored. In this work, we perform a large-scale study that quantifies the influence of gender imbalance in medical imaging datasets used to train AI-based CAD systems. It is worth mentioning that most of the existing work dealing with imbalanced data in the context of deep learning focuses on cases where it is related to the target classes (14, 15). In our study, this would translate to an imbalance in terms of number of patients per pathology. However, note that, in this case, the imbalance is given by a demographic variable different from the target class: gender, which is generally neglected. Our results show that using gender-imbalanced datasets to train deep learning-based CAD systems may affect the performance in pathology classification for minority groups.

Results and Discussion

A model based on deep neural networks, which achieves state-of-the-art results when diagnosing 14 common thoracic diseases using X-ray images (16), was implemented to perform CAD. We employed the area under the receiver operating characteristic curve (AUC) (17) to quantify its performance. Fig. 1 shows the experimental results obtained when training the classifier under different gender imbalance ratios. In Fig. 1*A*, the box plots aggregate the results for 20 experiments using fully imbalanced datasets. The blue boxes represent the performance for models trained only with male images, while orange boxes indicate training with female-only images. Both models are evaluated over male-only (Fig. 1 *A*, *Top*) and female-only (Fig. 1 *A*, *Bottom*) test images. A consistent decrease in performance is observed when using male patients for training and female for testing (and vice-versa). The same tendency was confirmed when evaluating three different deep learning architectures in two X-ray datasets with different pathologies.

We also explored intermediate imbalance scenarios, where both female and male patients were present in the training dataset but considering different proportions (0%/100%,

Author contributions: A.J.L., N.N., V.P., D.H.M., and E.F. designed research; A.J.L., N.N., and E.F. performed research; A.J.L., N.N., V.P., D.H.M., and E.F. analyzed data; and A.J.L., N.N., V.P., D.H.M., and E.F. wrote the paper.

The authors declare no competing interest.

Data deposition: The modified version of source code of the original convolutional neural networks (CNNs) with our auxiliary scripts, the data splits used in our experiments, and the additional results for all of the CNN architectures in both datasets can be accessed in GitHub at https://github.com/N-Nieto/GenderBias_CheXNet.

[1]A.J.L. and N.N. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: eferrante@sinc.unl.edu.ar.

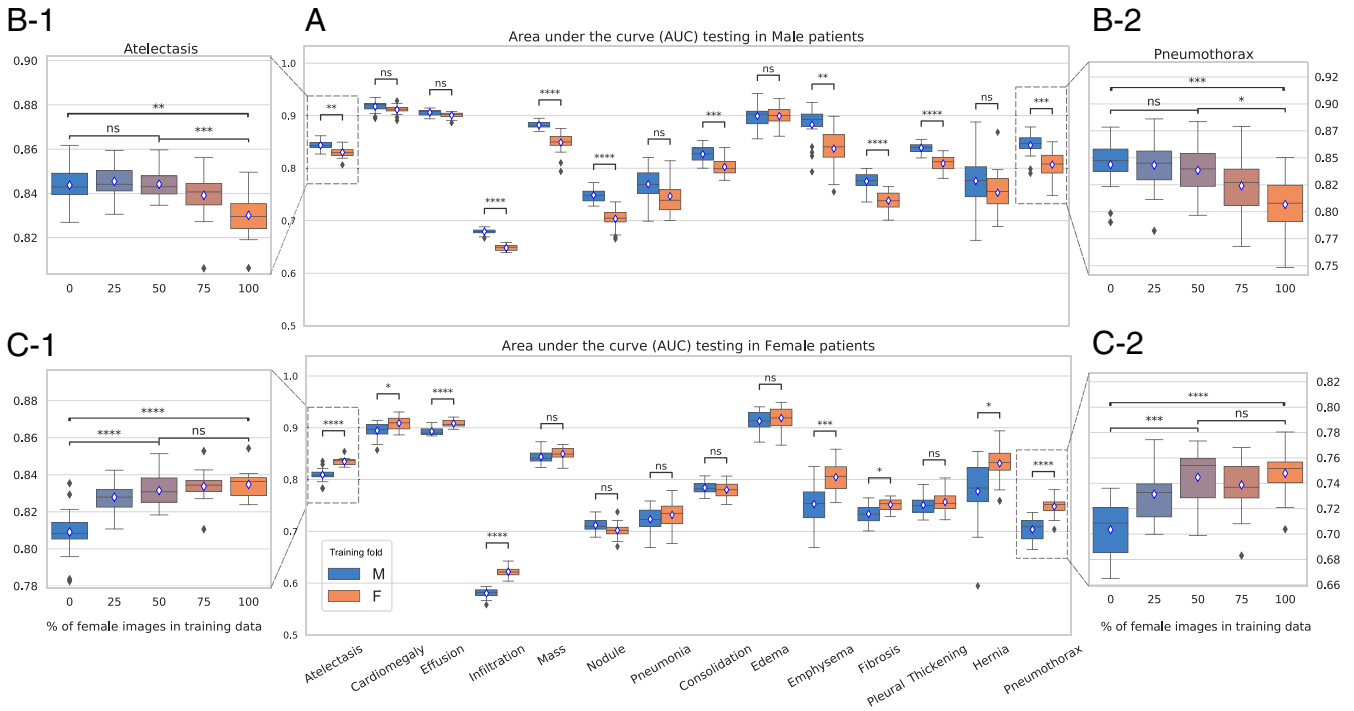

Fig. 1. Experimental results for a DenseNet-121 (18) classifier trained with images from the NIH dataset (16, 19) for 14 thoracic diseases under different gender imbalance ratios. (*A*) The box plots aggregate the results for 20 folds, training with male-only (blue) and female-only (orange) patients. Both models are evaluated given male (*Top*) and female (*Bottom*) test folds. A consistent decrease in performance is observed when using male patients for training and female for testing (and vice versa). (*B* and *C*) AUC achieved for two exemplar diseases under a gradient of gender imbalance ratios, from 0% of female images in training data to 100%, with increments of 25%. In *B*, *1* and *2* show the results when testing on male patients, while, in *C*, *1* and *2* present the results when testing on female patients. Statistical significance according to Mann–Whitney *U* test is denoted by **** ($P \leq 0.00001$), *** ($0.00001 < P \leq 0.0001$), ** ($0.0001 < P \leq 0.001$), * ($0.001 < P \leq' 0.01$), and not significant (ns) ($P > 0.01$).

25%/75%, and 50%/50%). Fig. 1 *B* and *C* shows the average classification performance for two exemplar diseases, Pneumothorax and Atelectasis, under such gradient of gender imbalance ratios (indicated with the percentage of female patients used for training). We found that, even with a 25%/75% imbalance ratio, the average performance across all diseases in the minority class is significantly lower than a model trained with a perfectly balanced dataset. Moreover, we did not find significant differences in performance between models trained with a gender-balanced dataset (50% male and 50% female) and an extremely imbalanced dataset from the same gender. In other words, a CAD system trained with a diverse (and balanced) dataset achieved the best performance for both genders. Altogether, our results indicate that diversity provides additional information and increases the generalization capability of AI systems. Thereafter, it also suggests that diversity should be prioritized when designing databases used to train machine learning-based CAD systems.

Our study shows that gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis based on convolutional neural networks (CNNs), with significantly lower performance in underrepresented groups. We provide experimental evidence in the context of X-ray image classification for such potential bias, aiming to raise the alarm not only within the medical image computing community but also for national agencies in charge of regulating and approving medical systems. As an example, let us take the US Food and Drug Administration. Even though they have released several documents related to the importance of gender/sex issues in the design and evaluation of clinical trials and medical devices (21), when looking at the specific guidelines to obtain the certification to market medical computer-aided systems (22, 23), there is no explicit mention of gender/sex as one of the relevant demographic variables that should describe the sampled population. Similar issues are observed in the medical imaging community. Albeit a few datasets provide this information at the subject level, most public datasets of similar characteristics do not contain gender/sex information at the patient level to date [e.g., the recent MIMIC-CXR (24) x-ray dataset or the Retinal Fundus Glaucoma Challenge (REFUGE) database of ophtalmological images (25), just to name a few]. The same tendency is observed in many of the datasets included in a recent analysis of 150 databases from grand challenges on biomedical image analysis (26), which provides recommendations for database and challenge design, where there is no explicit mention of the importance of sex/gender demographic information.

In general, it is well known that CNNs tend to learn representations useful to solve the task they are being trained for. When we go from male to female images (or vice versa), structural changes in the images appear, leading to a change in data distribution which explains the decrease in performance. Algorithmic solutions to such "domain adaptation" problems (27) should be engineered, especially in cases when it is difficult to obtain gender-balanced datasets [e.g., Autism Brain Imaging Data Exchange (ABIDE) I (28)].

Materials and Methods

Datasets. We use the NIH Chest-XRay14 dataset (16, 19), which includes 112,120 chest X-ray images from 30,805 patients, labeled with 14 common thorax diseases (including hernia, pneumonia, fibrosis, emphysema, edema, cardiomegaly, pleural thickening, consolidation, mass, pneumothorax, nodule, atelectasis, effusion, and infiltration). Labeling was performed according to an automatic natural language processing analysis of the radiology reports. The dataset provides demographic information including the patient's gender: 63,340 (56.5%) images for male and 48,780 (43.5%) images for female patients. Following the demographic variables reported in the

original dataset publication (19), we used the term "gender" to characterize our imbalance study. However, given that some anatomical attributes are reflected in X-ray images, the term *sex* could be more accurate, according to the Sex and Gender Equity in Research guidelines (29). The CheXpert database (30) was also used to confirm that our observations generalize for different datasets. It contains 224,316 chest radiographs of 65,240 patients with diagnostic information ($\sim 60\%$ male and $\sim 40\%$ female). The uncertainty labels included in CheXpert were interpreted as negative following the U-Zeros approach discussed in the original paper (30).

Deep Learning Model. Deep neural networks are machine learning methods with multiple abstraction levels, which compose simple but nonlinear modules transforming representations at one level into a representation at a higher, slightly more abstract level (31). A special type of deep neural network, known as CNNs, was used to implement the CAD system (19, 20). Results shown in Fig. 1 correspond to a Densely Connected CNN (DenseNet) architecture with 14 outputs, one for each disease (18).

We adopted a Keras implementation of the DenseNet-121 which has been shown to achieve state-of-the-art results in X-ray image classification (16). The network has 121 convolutional layers and a final fully connected layer producing a 14-dimensional output, after which we apply an element-wise sigmoid nonlinearity. A model pretrained on ImageNet (32) was used to initialize the network weights. We trained it end to end using Adam optimizer with standard parameters ($\beta 1 = 0.9$ and $\beta 2 = 0.999$), a batch size of 32, and an initial learning rate of 0.001 that was decayed by a factor of 10 each time the validation loss plateaued after an epoch. Additionally, we evaluated two other CNN architectures, the ResNet (33) and the Inception-v3 (34), confirming that our observations generalize for different neural models.

Methodology. Since images can be labeled with multiple diseases, we implemented an automatic method to construct random splits, which guarantees that male and female folds will have the same number of images per pathology. Given a frontal X-ray image, the CAD system predicts the presence or absence of the 14 thoracic diseases. Two models were trained in each experiment, one considering male-only datasets, while the other considered female-only training datasets. Intermediate imbalance scenarios were also analyzed, in which female and male images were presented in the training dataset at different proportions (0%/100%, 25%/75% and 50%/50%). To avoid other sources of bias, care was taken to guarantee, by training data construction, that male and female folds include the same number of pathological cases per class. For the NIH Chest-XRay14, every split included 48,568 images. For the CheXpert dataset, every split included 27,147 images. The same experiment was performed 20 times, using different random splits. In the testing phase, both models were evaluated in male and female patients separately. The classification performance was measured by the well-known AUC (17).

Data Availability. The NIH Chest-XRay14 dataset is publicly available at https://nihcc.app.box.com/v/ChestXray-NIHCC. The CheXpert dataset is publicly available at https://stanfordmlgroup.github.io/competitions/chexpert/. The source code of the original CNNs is publicly available at https://github.com/brucechou1983/CheXNet-Keras. The modified version of this code with our auxiliary scripts, the data splits used in our experiments, and the additional results for all of the CNN architectures in both datasets can be accessed at https://github.com/N-Nieto/GenderBias_CheXNet.

ACKNOWLEDGMENTS. E.F. is a beneficiary of an AXA Research Fund grant. We gratefully acknowledge NVIDIA Corporation for the donation of the graphics processing units used for this research, and the support of Universidad Nacional del Litoral (Grants CAID-PIC-50220140100084LI and 2016-082) and Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (Grants PICT 2014-2627, 2018-3907, and 2018-3384).

1. G. Litjens *et al.*, A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
2. R. Lindsey *et al.*, Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 11591–11596 (2018).
3. A. Esteva *et al.*, Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
4. J. De Fauw *et al.*, Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
5. B. Chandrasekaran, On evaluating artificial intelligence systems for medical diagnosis. *AI Mag.* **4**, 34–34 (1983).
6. J. Zou, L. Schiebinger, AI can be sexist and racist—It's time to make it fair. *Nature* **559**, 324–326 (2018).
7. M. Hutson *et al.*, Even artificial intelligence can acquire biases against race and gender. *Science*, 10.1126/science.aal1053 (2017).
8. T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings" in *Advances In Neural Information Processing Systems*, D. D. Lee, S. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates, 2016), vol. 29, pp. 4349–4357.
9. G. Stanovsky, N. A. Smith, L. Zettlemoyer, Evaluating gender bias in machine translation. arXiv:1906.00591 (3 June 2019).
10. J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proc. Machine Learning Res.* **81**, 77–91 (2018).
11. J. Wiens *et al.*, Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
12. D. S. Char, N. H. Shah, D. Magnus, Implementing machine learning in health care—Addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
13. L. Schiebinger, M. Schraudner, Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering. *Interdiscipl. Sci. Rev.* **36**, 154–167 (2011).
14. G. Haixiang *et al.*, Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **73**, 220–239 (2017).
15. J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).
16. P. Rajpurkar *et al.*, CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225 (14 November 2017).
17. T. Fawcett, An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874 (2006).
18. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks" in *Proceedings Of the IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2017), pp. 4700–4708.
19. X. Wang *et al.*, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2017), pp. 2097–2106.
20. C. Qin, D. Yao, Y. Shi, Z. Song, Computer-aided detection in chest radiography based on artificial intelligence: A survey. *Biomed. Eng. Online* **17**, 1–23 (2018).
21. US Food and Drug Administration, Understanding sex differences at FDA. https://www.fda.gov/science-research/womens-health-research/understanding-sex-differences-fda. Accessed 23 March 2020.
22. US Food and Drug Administration, Clinical performance assessment: Considerations for computer-assisted detection devices applied to radiology images and radiology device data—Premarket approval (PMA) and premarket notification [510(k)] submissions. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology. Accessed 23 March 2020.
23. US Food and Drug Administration, Computer-assisted detection devices applied to radiology images and radiology device data—Premarket notification [510(k)] submissions. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/computer-assisted-detection-devices-applied-radiology-images-and-radiology-device-data-premarket. Accessed 23 March 2020.
24. A. E. Johnson *et al.*, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
25. J. I. Orlando *et al.*, Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med. Image Anal.* **59**, 101570 (2020).
26. L. Maier-Hein *et al.*, Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 1–13 (2018).
27. M. Wang, W. Deng, Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135–153 (2018).
28. A. Di Martino *et al.*, Enhancing studies of the connectome in autism using the Autism Brain Imaging Data Exchange II. *Sci. Data* **4**, 170010 (2017).
29. S. Heidari, T. F. Babor, P. De Castro, S. Tort, M. Curno, Sex and gender equity in research: Rationale for the SAGER guidelines and recommended use. *Res. Integrity Peer Rev.* **1**, 2 (2016).
30. J. Irvin *et al.*, Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
31. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
32. J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2009), pp. 248–255.
33. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings Of the IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2016), pp. 770–778.
34. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision" in *Proceedings of IEEE CVPR* (Institute of Electrical and Electronic Engineers, 2016), pp. 2818–2826.