Lost in Translation: International Limitations of Deep Learning Demographic Classifiers Developed on US Imaging Datasets

Preetham Bachina, MSE, Medical Student, Dual-Affiliate, Johns Hopkins University School of Medicine, St. Jude Children's Research Hospital

Sean Garin; Pranav Kulkarni; Aditya Kulkarni, MS; Adway Kanhere, MSE; Vishwa Parekh, PhD; Jeremias Sulam, PhD; Paul Yi, MD

Introduction

Deep learning (DL) models can learn demographic attributes from chest X-rays (CXRs), creating a risk for biased diagnostic outputs. While debiasing techniques aim to mitigate the use of demographic representations, they are predominantly developed and validated on US datasets despite the goal to deploy fair AI globally. To investigate the consistency of demographic embeddings targeted by debiasing techniques, we evaluated the international generalizability of US-trained DL models for sex, age, and race classification across 10 datasets.

Hypothesis

DL demographic classifiers trained on US datasets will perform poorly in non-US populations.

Methods

We curated 10 geographically diverse CXR datasets spanning 4 continents (Figure 1). We trained separate DL models (ResNet-34, ImageNet-pretrained) to classify sex, age or race using three US datasets (CheXpert, MIMIC, NIH) each split into 70/10/20% train/validation/test sets. The US test sets were used for internal and cross-dataset external model testing. The remaining 7 non-US datasets were used solely for external testing. Model performance was assessed using weighted AUROC (wAUC) or predictive accuracy when applicable, with 95% CIs estimated via bootstrap resampling (n=1000).

Results

US-trained sex and age classification models performed well on internal and external US testing (sex wAUC: >0.98; age wAUC: >0.80; Figure 2A). These models generalized well to most non-US datasets (sex wAUC: >0.90; age wAUC: >0.80; Figure 2A). US-trained race classification models achieved high internal and external performance on US datasets (wAUC: >0.82). However, when applied to international datasets, race predictions deviated substantially from census-derived national demographic distributions (Figure 2B).

Conclusion

While US-trained sex and age DL classifiers generalized well across geographic and cultural settings, race classifiers did not. These findings suggest that DL models' ability to predict race from CXRs may not reflect a generalizable imaging biomarker outside of a specific cultural context, which may undermine global efforts to reduce bias by targeting the algorithmic encoding of race.

Figure(s)

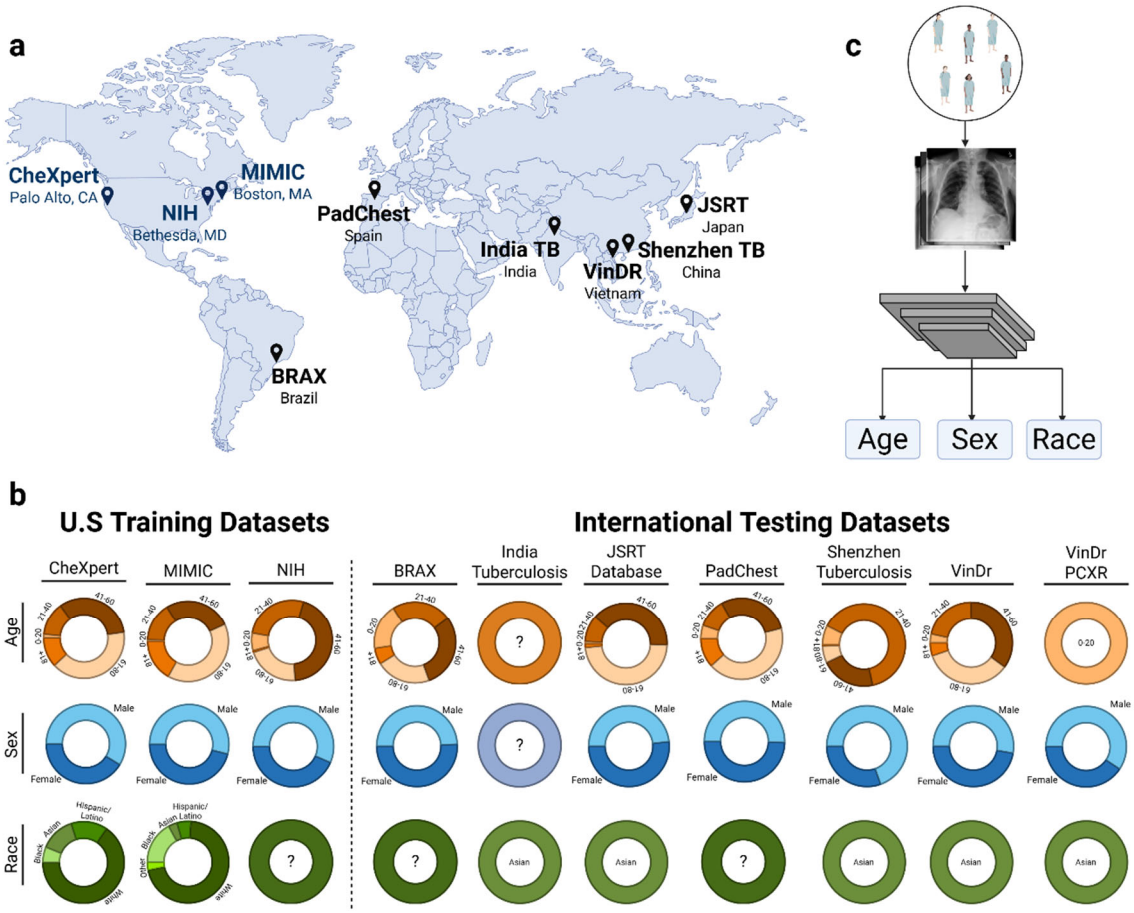

Figure 1. Overview of Study Pipeline. a. Three Chest X-ray datasets (blue) from the United States were used in model training and evaluation: CheXpert (Stanford, CA), MIMIC (Boston, MA), and NIH ChestX-ray14 (Bethesda, MD). Seven international datasets (black) were used for external validation: BRAX (Brazil), India Tuberculosis (India), JSRT Database (Japan), PadChest (Spain), Shenzhen Tuberculosis (China), and VinDr and VinDr-PCXR (Vietnam). b. Demographic distribution for each dataset is shown. For analysis, we focused on standardized subgroups based on available labels and literature: sex (male, female), age (0-20, 21-40, 41-60, 61-80, 81+ years), and race (Asian, Black, White, Hispanic/Latino). c. For each US dataset with sufficient labeled samples for a demographic attribute, a ResNet-34 model (pretrained on ImageNet) was trained to predict the relevant demographic (sex, age, or race).

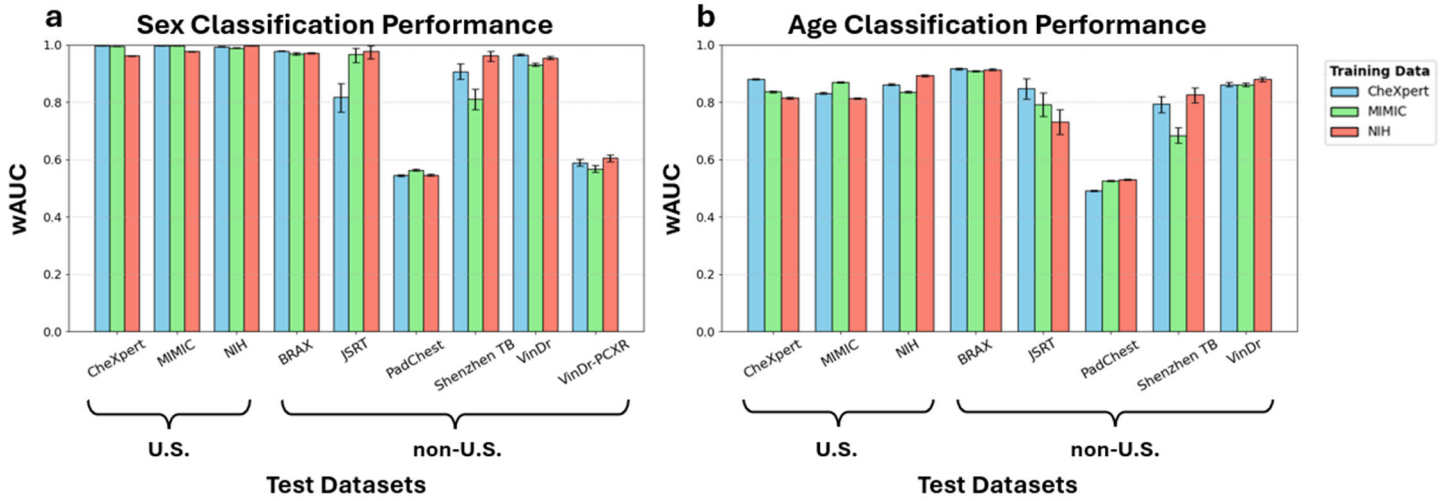

Figure 2A. Performance of sex and age classifying deep learning models trained on US datasets. a. wAUC of sex classifying model for each respective test dataset. b. wAUC of age classifying model for each respective test dataset. Error bars indicate 95% CIs estimated using nonparametric bootstrap sampling (n = 1000).

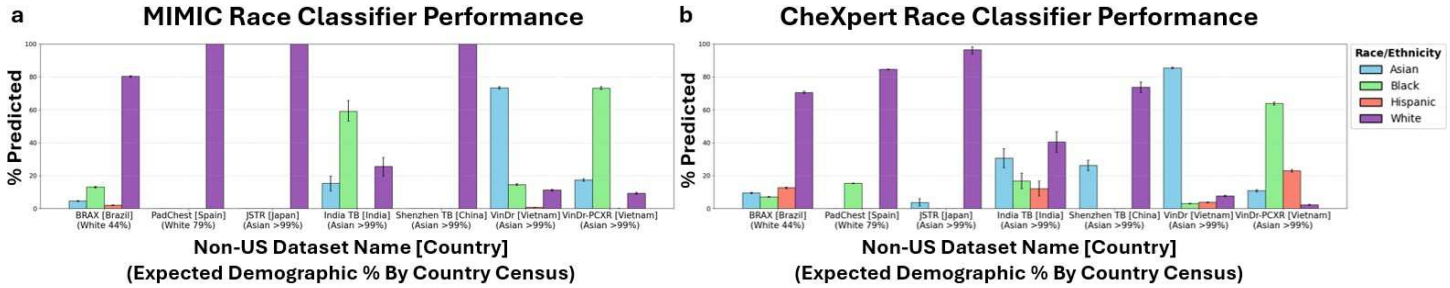

Figure 2B. Performance of US-trained race classifying deep learning models on non-US sourced data. a. MIMIC race classifying model predictions b. CheXpert race classifying model predictions. Error bars indicate 95% CIs estimated using nonparametric bootstrap sampling (n = 1000). For datasets sourced from Asia, India TB [India], JSRT [Japan], Shenzhen TB [China], VinDr [Vietnam], and VinDr-PCXR [Vietnam], census data indicate that over 99% of the population in each country aligns with the US Census racial category "Asian". For BRAX [Brazil] from South America, census data report 43.5% of the population aligns with the US Census racial category "White". For PadChest [Spain] from Europe, 79% of the population aligns with the US Census racial category "White".

Keywords

Applications; Artificial Intelligence; Machine Learning