Not So Black and White: Confounders Mediate AI Prediction Of Race On Chest X-Rays

Preetham Bachina[1,2], Sean Garin[1], Pranav Kulkarni[1], Adway Kanhere[1], Daniel Kargilis[1,2], Vishwa S Parekh[1], Paul H Yi[1]
[1] University of Maryland Medical Intelligent Imaging (UM2ii) Center, Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD
[2] Johns Hopkins University School of Medicine, Baltimore, MD

Background.

Excitement over automated diagnosis on chest x-rays (CXRs) using artificial intelligence (AI) has been dampened by reports of deep learning (DL) algorithms demonstrating biased performance against historically disadvantaged demographic groups. One hypothesized source of these algorithmic biases is the ability of AI to identify patient demographics such as age, sex, and race using CXRs, as shown by recent studies (Gichoya et al. Lancet Digital Health 2022). However, because these demographic variables are complex, having components of biology and environment – race, for example, is a social construct, albeit one with some components of genetic ancestry – it is unclear what it means for a DL algorithm to identify race on a CXR. Recently, Duffy et al. (NPJ Digital Medicine 2022), demonstrated that DL predictions of race on echocardiogram data were primarily mediated by confounding features, namely sex and age. Our purpose was to determine 1) whether DL models could be trained to accurately predict patient age, sex or race on CXRs and 2) assess the impact of other confounding demographic variables on these demographic predictions.

Methods.

We trained DL algorithms to identify age, sex, and race on CXRs using datasets from two USA academic medical centers: 1) MIMIC-CXR dataset (Beth Israel Deaconess Medical Center; Boston, MA) and 2) CheXpert (Stanford Medical Center; Palo Alto, CA), comprising 227,835 and 224,316 images, respectively. Each dataset was labeled with the following self-reported demographic conventions: 1) Age (0-20, 21-40, 41-60, 61-80, 80+ years old), 2) Sex (Male vs. Female), and 3) Race (White, Non-White). Each dataset was randomly split into 70/10/20% train/validation/test splits and used to finetune ResNet-34 DL classification models pretrained on ImageNet (transfer learning) for each demographic classification task (e.g., male vs. female); testing was performed on the testing data splits.

To evaluate the impact of confounding demographic variables on these DL models' predictions, we created several test sets adjusted for varying proportions of potential confounders. For example, to evaluate the impact of confounding variables of sex on DL prediction of race, we created separate test sets where there were equal proportions of race categories, but variable sex proportions skewed by x% (ranging from 0 to 100%), where the test set contained x% of White patients as female and x% of non-White patients as male. Similarly, to evaluate the impact of age as a confounder, separate test sets were created where race was skewed in the same fashion using <40 or >40 years old as the two confounding categories. A similar procedure was applied to the DL prediction models for age and sex.

Performance on the test sets with varying proportions of potential confounding demographic variables was characterized with area under the ROC curve (AUC) with a one-vs-all approach.

Results.

Overall, the DL prediction models for all three demographic variables (age, sex, race) performed well on all test sets, with AUCs ranging from approximately 0.8 (age) to >0.9 (race and sex). When predicting binary race classification (White or non-White) confounded by sex or age, the CheXpert race DL model performance consistently decreased as confounding in the MIMIC test set increased. For example, on test sets with a skew of 0% for sex and age, the CheXpert model predicted race with an AUC of ~0.93 for both. However, for every 10% increase in skew, the AUC dropped by ~0.07 and ~0.08 for sex and age, respectively, such that at a skew of 100%, the CheXpert model predicted race with an AUC of ~0.87 and ~0.85 for sex and age, respectively. In contrast, the DL sex prediction models did not demonstrate changes in performance based on confounding in the test sets based on race or age. For example, the CheXpert sex prediction model predicted sex with an AUC of ~0.99 for both and race and age skewed test sets. Finally, for DL age prediction models, we found variable impact of confounders in the test set. For example, the CheXpert age classifying model predicted age on the MIMIC test sets with an AUC of ~0.8 for the race skewed test sets, regardless of the degree of skew. On the other hand, on a test set skewed by 0% with respect to sex, these same models had a performance of ~0.85 which consistently dropped as skew increased, such that on a test set skewed by 100% with respect to sex, the age model has a performance of ~0.79.

Conclusion.

Our findings support the conclusion that DL-based prediction of certain demographic variables from CXRs, namely race, may primarily be mediated by the detection of other confounding features. Given the complexities of demographic identity – including the complex interplays between biology, environment, and culture --understanding what exactly it means for AI to predict "race" and other patient demographics from CXRs is of vital importance. Further study is needed to better understand how this confounding may influence DL algorithms demonstrating bias towards different demographic groups and reevaluate the clinical utility of DL algorithms in a clinical context.