AI models can be racially biased when trained on unbalanced data sets, researchers find

April 13 2022

Credit: Pixabay/CC0 Public Domain

Using the UK Biobank, a large-scale biomedical database, researchers from the School of Biomedical Engineering & Imaging Sciences have

found that AI models can be racially biased if they are trained on unbalanced databases, meaning where AI models are used, misdiagnoses would occur for under or non-represented races.

In the majority of cardiovascular diseases (CVDs), there are known associations between sex/race and epidemiology, pathophysiology, clinical manifestations, effects of therapy, and outcomes. Although these differences do not have proven causative links with race and gender, their presence remains a potential concern about the performance of AI models in cardiovascular imaging.

The paper, published in *Frontiers in Cardiovascular Medicine*, looked at the performance of AI models based on cardiac MR imaging that are used to derive biomarkers of the heart.

It was shown that if those biomarkers are used for the diagnosis of heart failure, for instance, there would be more misdiagnoses in minority races than there would be for majority races.

The researchers found statistically significant differences in segmentation performance scores between races as well as in absolute/relative errors in volumetric and functional biomarkers, showing that the AI model was biased against minority racial groups, even after correction for possible confounders.

Lead researcher Dr Andrew King, Reader in Medical Image Analysis, School of Biomedical Engineering & Imaging Sciences, said researchers need to consider the training data when they are deploying these models into clinical practice to ensure that there is adequate representation of racial groups.

For years, clinicians have been relying on manual/semi-automatic segmentation approaches to trace the cardiac chamber contours. The

process, however, is tedious, time-consuming and prone to subjective errors.

But researchers have proposed to use AI and machine learning (ML) models for tasks such as automatic cardiac functional quantification. These methods are now starting to move towards broader clinical translation.

In an earlier work, the researchers identified three methods that can use the same data but develop a model which is fairer and has a more equal performance for different racial groups.

These methods take into account that the databases used for training are unbalanced, for instance, the white group accounts for 80 percent of the data and the other racial groups for the remaining 20 percent. The first method aims to modify the training sampling strategy to remove the discrimination. Effectively, the method fools the AI model into thinking that the database is balanced when in reality it is not.

The second method aims to combine the segmentation task with a classification task that will aim to predict the race of the subject based on the images. By trying to combine these two tasks the model learns to segment the heart in a less biased way.

The final strategy aims to train a separate model per race group. The main disadvantage of this strategy is it requires race knowledge to apply the model, and this is not always available in all clinical settings.

Dr. King says that "this is an important time for the future of AI. Techniques are starting to be used in the real world including in high-stakes applications like medicine. If we don't make sure that AI techniques are fair then it may erode public trust in their use. Future research should bear this in mind and ensure that all sectors of society

benefit equally from AI."

More information: Esther Puyol-Antón et al, Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation, *Frontiers in Cardiovascular Medicine* (2022). DOI: 10.3389/fcvm.2022.859310

Provided by King's College London