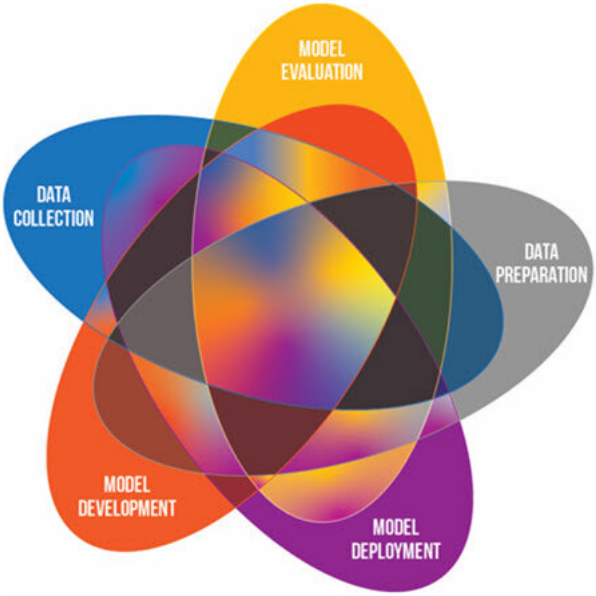
# Roadmap to fair AI: Revealing biases in AI models for medical imaging

April 26 2023




In recent years, artificial intelligence (AI) has been recognized as a powerful tool in the field of medical imaging. However, these models can be subject to several biases, leading to inequities in how they benefit both doctors and patients. Understanding these biases and how to mitigate them is the first step towards a fair and trustworthy AI. Credit: MIDRC, midrc.org/bias-awareness-tool.

Artificial intelligence and machine learning (AI/ML) technologies are constantly finding new applications across several disciplines. Medicine is no exception, with AI/ML being used for the diagnosis, prognosis, risk assessment, and treatment response assessment of various diseases. In

particular, AI/ML models are finding increasing applications in the analysis of medical images. This includes X-ray, computed tomography, and magnetic resonance images. A key requirement for the successful implementation of AI/ML models in medical imaging is ensuring their proper design, training, and usage. In reality, however, it is extremely challenging to develop AI/ML models that work well for all members of a population and can be generalized to all circumstances.

Much like humans, AI/ML models can be biased, and may result in differential treatment of medically similar cases. Notwithstanding the factors associated with the introduction of such biases, it is important to address them and ensure fairness, equity, and trust in AI/ML for medical imaging. This requires identifying the sources of biases that can exist in medical imaging AI/ML and developing strategies to mitigate them. Failing to do so can result in differential benefits for patients, aggravating healthcare access inequities.

As reported in the *Journal of Medical Imaging (JMI)*, a multi-institutional team of experts from the Medical Imaging and Data Resource Center (MIDRC)—including medical physicists, AI/ML researchers, statisticians, physicians, and scientists from regulatory bodies—addressed this concern. In this comprehensive report, they identify 29 sources of potential bias that can occur along the five key steps of developing and implementing medical imaging AI/ML from data collection, data preparation and annotation, model development, model evaluation, and model deployment, with many identified biases potentially occurring in more than one step. Bias mitigation strategies are discussed, and information is also made available on the MIDRC website .

One of the main sources of bias lies in data collection. For example, sourcing images from a single hospital or from a single type of scanner can result in a biased data collection. Data collection bias can also arise

due to differences in how specific social groups are treated, both during research and within the healthcare system as a whole. Moreover, data can become outdated as medical knowledge and practices evolve. This introduces temporal bias in AI/ML models trained on such data.

Other sources of bias lie in data preparation and annotation and are closely related to data collection. In this step, biases can be introduced based on how the data is labeled prior to being fed to the AI/ML model for training. Such biases may stem from personal biases of the annotators or from oversights related to how the data itself is presented to the users tasked with labeling.

Biases can also arise during model development based on how the AI/ML model itself is being reasoned and created. One example is inherited bias, which occurs when the output of a biased AI/ML model is used to train another model. Other examples of biases in model development include biases caused by unequal representation of the target population or originating from historical circumstances, such as societal and institutional biases that lead to discriminatory practices.

Model evaluation can also be a potential source of bias. Testing a model's performance, for instance, can introduce biases either by using already biased datasets for benchmarking or through the use of inappropriate statistical models.

Finally, bias can also creep in during the deployment of the AI/ML model in a real setting, mainly from the system's users. For example, biases are introduced when a model is not used for the intended categories of images or configurations, or when a user becomes over-reliant on automation.

In addition to identifying and thoroughly explaining these sources of potential bias, the team suggests possible ways for their mitigation and

best practices for implementing medical imaging AI/ML models. The article, therefore, provides valuable insights to researchers, clinicians, and the general public on the limitations of AI/ML in medical imaging as well as a roadmap for their redressal in the near future. This, in turn, could facilitate a more equitable and just deployment of medical imaging AI/ML models in the future.

**More information:** Karen Drukker et al, Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment, *Journal of Medical Imaging* (2023). DOI: 10.1117/1.JMI.10.6.061104

Provided by SPIE