In recent years, there has been rapid growth and expansion in the use of machine learning and other artificial intelligence approaches applied to increasingly rich and accessible health data sets to develop algorithms that guide and support health care. As they make their way into practice, such algorithms have the potential to fundamentally transform how health care decisions are made and, therefore, how patients are diagnosed and treated. While such approaches hold great promise for enabling more precise, accurate, timely, and even fair decision-making when properly developed and applied, there is also growing evidence that systematic biases can lead to unintended and even severe consequences. Mirroring disparities and inequities inherent in our society and health system, such biases can be inherent in not only the underlying data used to develop algorithms but also how algorithmic interventions are deployed.

Elsewhere in JAMA Network Open, Park and colleagues present findings from a study evaluating different approaches to the debiasing of health care algorithms developed to predict postpartum depression (PPD) among a cohort of pregnant women with Medicaid coverage. The researchers, from IBM Research, leveraged the IBM MarketScan Medicaid Database, a deidentified, individual-level claim records data set with approximately 7 million Medicaid enrollees across multiple states, to derive their algorithms. They started by developing 2 sets of machine learning models trained to predict 2 outcomes: (1) diagnosis or treatment for PPD and (2) postpartum mental health service utilization. Their initial, risk-adjusted generalized linear models for each outcome demonstrated a notable difference in the cohort with binarized race, with White patients having twice the predicted likelihood of being diagnosed with PPD compared with Black patients and a significantly higher likelihood of utilizing mental health services. However, as the authors point out, current evidence indicates that while PPD incidence is higher in women with low socioeconomic status, such as the Medicaid enrollees studied, PPD rates are similar across racial and ethnic groups. Therefore, the algorithmically predicted differences derived from the real-world data set are likely the result of underlying disparities, including lower rates of diagnosis and treatment among Black patients.

The authors next examined the relative effectiveness of 3 different algorithmic debiasing methods. These included a preprocessing method called reweighing, an in-processing method called prejudice remover for logistic regression, and an approach to retraining the models without the race variable included called fairness through unawareness. They measured the differences using 2 metrics of fairness: disparate impact (DI), a ratio of estimated favorable outcomes between privileged and underprivileged groups, and equal opportunity difference (EOD), a comparison of true-positive rates between groups. While algorithmic prediction and debiasing performance differed across the models and methods tested, a key finding is that all of the debiasing approaches demonstrated some improvement toward fairness relative to the nondebiased models. Nevertheless, it was also notable that at least in this particular study, the approach of simply disregarding race in the model was not as effective as the other debiasing approaches, likely due to the presence of correlated variables.

While more research is certainly needed to validate and determine which debiasing methods are most appropriately applied to particular settings and conditions, this study suggests that debiasing can address underlying disparities represented in data used to develop and operationalize...
predictive algorithms. Moreover, while the authors set out to test debiasing approaches, they did so precisely because they anticipated that there would likely be inherent biases in algorithms based on even large, well-characterized data sets like the one used. Therefore, this demonstrates another important but perhaps less obvious point. Whether inherent biases are anticipated by the developers or not, evaluation and monitoring of health care algorithms for effectiveness and equity is necessary and, indeed, an ethical imperative.

As algorithmic health care continues to advance, there are myriad reasons why we must concern ourselves with the systematic evaluation and ongoing monitoring of algorithms that drive care. In that sense, this study also represents another important step toward advancing a sorely needed set of approaches and tools for the systematic evaluation and monitoring of health care algorithms.

Just as we would not think of deploying a new pharmaceutical or device into practice without first ensuring its efficacy and safety in the populations to which it will be applied, so too must we recognize the reality that algorithms have the potential for both great benefit and harm and, therefore, require study. Indeed, the processes in place for evaluating and validating therapeutics can offer a useful analogy for the systematic evaluation of artificial intelligence–enabled health care interventions. Building on that analogy, an approach akin to pharmacovigilance is called for. Algorithmovigilance can be defined as the scientific methods and activities relating to the evaluation, monitoring, understanding, and prevention of adverse effects of algorithms in health care. While such activities are certainly relevant during algorithmic development and prior to initial deployment, they must not be limited to that phase of algorithmic use in health care.

Perhaps to an even greater degree than among pharmaceuticals, ongoing evaluation of algorithms used in practice is critical. The nature of how algorithms are developed and deployed in practice can change their anticipated impacts and potentially lead to unintended adverse effects. This has to do with the variability with which algorithms make their way into practice. Beyond concerns shared with pharmaceuticals and devices (eg, their differential effects given individual variation among people), algorithms are also subject to different applications based on factors that often do not apply to pharmaceuticals or devices. These include how they are deployed, who interacts with them, and when in the workflow this takes place.

Beyond the technical considerations related to initial algorithmic performance and validation, it is often the case that algorithmic performance changes as it is deployed against different data, in different settings, and at different times. Moreover, how algorithms are used involves human-computer interactions that add another level of variation and complexity that can change the algorithm’s performance, including how algorithmic outputs are interpreted by different users, the lack of trust in a black-box algorithm, and concerns about becoming overly reliant on the output such that automation is used without critical thought. These and other systemic factors, many difficult to anticipate, can alter the performance characteristics of algorithms and lead to adverse effects and harm to patients and populations. Given the rapidly accelerating pace of change, demand, and promise for care that is guided by adaptive and computable algorithms, the inherent and systemic inequities that exist in our health care system, and the potential for unintended harm, it is imperative that we continue to develop, test, and disseminate tools and capabilities that enable systematic surveillance and vigilance in the development and application of algorithms in health care.
REFERENCES


