Beyond the *AJR*: "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations"

Judy Wawira Gichoya, MBChB, MS[1], **Leo Anthony Celi, MD, MS**[2,3,4]

[1]Department of Radiology and Imaging Sciences, Emory University, 1364 Clifton Rd NE, Ste AG08, Atlanta, GA 30322.

[2]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA.

[3]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA.

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA.

Summary of the Investigation

Pierson et al. [1] used deep learning techniques on the NIH Osteoarthritis Initiative dataset to predict patients' experienced pain. This dataset is composed of bilateral fixed-flexion knee radiographs from 4172 patients and the corresponding patient-reported knee-specific pain score (Knee Injury and Osteoarthritis Outcome Score [KOOS]) and Kellgren-Lawrence grade (KLG). Their work expands on prior studies that attempted to reduce pain disparities among patients with osteoarthritis. The authors sought to address whether machine learning (ML) techniques discriminate between "within the knee" and "external to the knee" factors that account for osteoarthritis pain disparities.

At baseline, high pain scores were disproportionately observed among patients with lower income and lower education, as well as among Black patients, who reported severe pain scores 58% of the time. Fifty-six percent of Black patients, compared with 46% of patients overall, had a KLG of 2 or greater ($p < .001$). Using a convolutional neural network, the authors developed an algorithmic pain prediction (ALG-P) score using features present in the radiographs. They evaluated their algorithm's ability to reduce the KOOS 10.6-point pain disparity by comparing Black patients with patients of other races in terms of the KOOS without controlling for radiographic severity and the KOOS based on the ALG-P score after controlling for OA severity with the KLG. The ALG-P score explained 43% of the overall pain disparity between Black patients and patients of other races (4.7 times improved disparity explanation compared with disparity explained by the KLG). It also improved the pain disparities between lower- and higher-income patients 2 times better and between lower- and higher-education patients 3.6 times better than the KLG. For 22% of the observations, racial and socioeconomic disparities persisted after controlling for the radiologist's interpretation of the available MRI examination.

Address correspondence to J. W. Gichoya (judywawira@emory.edu).

The study investigated confounding factors by binning the ALG-P score into five groups that more closely matched the KLG with no significant effect on the findings. Further, the algorithm did not overfit on features such as artifacts, body mass index, or site-specific markers. Using simulated arthroplasty referrals to evaluate eligibility, the adoption of ALG-P would double the number of patients with severe pain who are eligible for surgery.

Critical Analysis

This study highlights the potential for ML to improve health equity and illustrates how to assess fairness in algorithms. Illness severity scoring systems commonly used in clinical practice (Acute Physiology and Chronic Health Evaluation [APACHE] IV, Sequential Organ Failure Assessment [SOFA], Oxford Acute Severity of Illness Score [OASIS]) [2] have been found to exhibit racial bias. This bias is partly explained by the limited diversity of patients included in clinical research. Much medical knowledge is based on the understanding of health and disease among White, usually male, patients [3]. The use of fairness metrics can help guide the update of scoring systems. Attention should be given to ensure that the datasets used to train algorithms are diverse.

Several articles describe how ML "outperforms radiologists," yet this article describes how it can significantly augment the work of, rather than replace, radiologists [4]. The prediction of pain has been recently found to improve radiologist performance during spine MRI interpretation [5]. However, before the ALG-P score is deployed, it must be tested in a dataset that does not come from a clinical trial. Its performance must be evaluated on real-world knee radiographs with all the noise that they contain, real-world incomplete health records in which pain is inconsistently documented, and longitudinal images linked to clinical data such as interventions and medications to establish the value added.

Takeaway Point

ML holds promise for improving population health through the development of new metrics to evaluate health disparities, including disparities that may arise from the interpretation of radiologic images.

Acknowledgments

References

1. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021; 27:136–140 [PubMed: 33442014]

2. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health 2021; 3:e241–e249 [PubMed: 33766288]

3. Czerniewicz L It's time to redraw the world's very unequal knowledge map. The Conversation website. theconversation.com/its-time-to-redraw-the-worlds-very-unequal-knowledge-map-44206. Published July 8, 2015. Accessed April 10, 2021

4. Purkayastha S, Trivedi H, Gichoya JW. Failures hiding in success for artificial intelligence in radiology. J Am Coll Radiol 2021; 18(3 pt B):517–519 [PubMed: 33663765]

5. Balza R, Mercaldo SF, Chang CY, et al. Impact of patient-reported symptom information on agreement in the MRI diagnosis of presumptive lumbar spine pain generator. AJR 2021; 217:947–956] [PubMed: 33438459]