ARTICLES
https://doi.org/10.1038/s41591-020-01192-7

nature medicine

An algorithmic approach to reducing unexplained pain disparities in underserved populations

Emma Pierson[1,2], David M. Cutler[3], Jure Leskovec[4], Sendhil Mullainathan[5]✉ and Ziad Obermeyer[6]

Underserved populations experience higher levels of pain. These disparities persist even after controlling for the objective severity of diseases like osteoarthritis, as graded by human physicians using medical images, raising the possibility that underserved patients' pain stems from factors external to the knee, such as stress. Here we use a deep learning approach to measure the severity of osteoarthritis, by using knee X-rays to predict patients' experienced pain. We show that this approach dramatically reduces unexplained racial disparities in pain. Relative to standard measures of severity graded by radiologists, which accounted for only 9% (95% confidence interval (CI), 3–16%) of racial disparities in pain, algorithmic predictions accounted for 43% of disparities, or 4.7× more (95% CI, 3.2–11.8×), with similar results for lower-income and less-educated patients. This suggests that much of underserved patients' pain stems from factors within the knee not reflected in standard radiographic measures of severity. We show that the algorithm's ability to reduce unexplained disparities is rooted in the racial and socioeconomic diversity of the training set. Because algorithmic severity measures better capture underserved patients' pain, and severity measures influence treatment decisions, algorithmic predictions could potentially redress disparities in access to treatments like arthroplasty.

Pain is widespread and unequally distributed in society. Like many other causes of pain, knee osteoarthritis, which affects 10% of men and 13% of women over 60 years of age in the United States[1], disproportionately affects underserved populations; people of color score far higher on knee pain scales than do white individuals[2–6]. Understanding these racial disparities in pain is important for clinical decision making and public policy but also for understanding pain disparities for a variety of other medical problems[7,8].

Two explanations for these disparities have been proposed. First, underserved patients might have more severe osteoarthritis within the knee. Alternatively, underserved patients could have more aggravating factors external to the knee. For example, the same physical ailments in different populations can produce very different experienced pain due to life stress, social isolation or other factors[7–9]. These two explanations have very different treatment implications: psychosocial interventions target causes external to the knee, whereas physical therapy, medication and orthopedic procedures address causes within the knee[10–12].

Research to date has indirectly implicated factors external to the knee. Methodologically, this is demonstrated by defining an objective measure of osteoarthritis severity based on knee X-rays and then measuring the extent of pain disparities that remain after adjusting for severity. Typically, large differences in pain remain even after adjustment[2–4,13]. For example, even though Black patients have more severe osteoarthritis based on standard radiographic measures (Kellgren–Lawrence grade (KLG)), adjusting for KLG only slightly decreased measured Black–white disparities in pain[3,13]. These findings that pain disparities remain even when adjusting for radiographic osteoarthritis severity, however, depend heavily on how severity is measured. The relationship between radiographic severity and pain is debated. Many patients with mild or no disease as measured by radiographic severity suffer pain, and many patients with structural damage on X-ray or even magnetic resonance imaging (MRI) experience no or very little pain[14–16]. Standard radiographic measures such as KLG, developed decades ago in white British populations, might miss physical causes of pain in people of color[17,18]; further, there are known racial and socioeconomic biases in how a patient's pain is perceived by observers[19,20]. If the pain experienced by underserved populations is caused by objective factors missing from current measures, a range of painful, treatable knee ailments would be misattributed to factors external to the knee.

In this paper, we use a machine-learning approach to discriminate between the 'within the knee' and 'external to the knee' hypotheses. We produce a new algorithmic measure of osteoarthritis severity from radiographs alone. We use a dataset of knee radiographs from a diverse sample of 4,172 patients in the United States who had or were at high risk of developing knee osteoarthritis. As part of an NIH-funded study[21], bilateral fixed flexion knee radiographs were obtained and scored by radiologists on summary measures of radiographic severity (for example, KLG) and other objective features (for example, osteophytes and joint space narrowing (JSN)). Patients also reported a knee-specific pain score (Knee injury and Osteoarthritis Outcome Score (KOOS)), derived from a multi-item survey on pain experienced during various activities (for example, fully straightening the knee[22]).

Summary statistics of the 4,172 participants, who generated 36,369 observations (one for each knee at each time point) are provided in Table 1. Black patients had substantially higher pain levels across knees and time points compared with non-Black patients (97% of whom were white). Black patients experienced severe pain 58% of the time (KOOS ≤ 86.1, a standard threshold for severe pain[23]), compared with 38% for patients overall (P value for racial difference, <0.001). The median Black patient had worse pain than 75% of non-Black patients. Black patients had a pain score that was 10.6 KOOS points higher than that of non-Black patients

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Microsoft Research, Cambridge, MA, USA. [3]Department of Economics, Harvard University, Cambridge, MA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA. [6]School of Public Health, University of California at Berkeley, Berkeley, CA, USA.
✉e-mail: sendhil.mullainathan@chicagobooth.edu

Table 1 | Dataset summary statistics

	Training–development	Validation
Sample size		
No. individuals	2,877	1,295
No. observations	25,049	11,320
Demographics		
Black	17%	16%
Lower-income (<$50,000 per year)	38%	39%
Non-college graduates	39%	38%
Female	58%	56%
Mean age, baseline visit (s.d.)	61.1 (9.2)	61.0 (9.1)
Mean BMI, baseline visit (s.d.)	28.7 (4.9)	28.4 (4.6)
Fraction of knees with severe osteoarthritis (KLG ≥ 2)		
All	45%	46%
Black	60%	56%
Lower-income (<$50,000 per year)	52%	49%
Non-college graduates	52%	49%
Fraction of knees with severe pain score (KOOS ≤ 86.1)		
All	37%	38%
Black	53%	58%
Lower-income (<$50,000 per year)	44%	43%
Non-college graduates	46%	45%

(*P* value for racial difference, <0.001); for comparison, the s.d. in the dataset was 16.2 KOOS points. We found similar pain disparities across socioeconomic groups. Across knees and time points, 43% of lower-income patients and 45% of lower-education patients had severe pain (versus 38% overall; both *P* values < 0.001). Extended Data Fig. 3 provides statistics on the overlap between the Black, lower-education and lower-income patient groups; the groups were not independent.

Black patients also had more severe osteoarthritis, with 56% of knees having KLG ≥ 2 versus 46% of knees overall (*P* value for racial difference, <0.001), with similar trends across socioeconomic groups. But despite this higher disease severity, controlling for KLG scores does not fully account for the higher pain levels experienced by Black patients. Table 2 shows that the racial disparity in pain was 10.6 KOOS points, without controlling for any severity measures, compared with 9.7 points when controlling for KLG, meaning that KLG accounted for only 9% of the pain disparity (95% CI, 3–16%). Results were similar for other underserved groups, with KLG accounting for only 16% (95% CI, 5–29%) and 8% (95% CI, −1% to 18%) of the pain disparity by income and education, respectively. These results replicate findings in the literature[3,13] and suggest that objective osteoarthritis severity does not account for a large proportion of the pain disparity between racial and socioeconomic groups. However, this judgment is dependent on the objective measure used (in this case, KLG), which could incorporate a range of inaccuracies. For example, KLG scores were developed decades ago in white British populations, which might not reflect the experience of osteoarthritis in diverse populations[17,18].

To generate an alternative measure, we trained a convolutional neural network to predict the reported pain score for each knee using each X-ray image, using a randomly selected training and development dataset of 25,049 radiographs (2,877 patients). We generated predictions in an independent validation (held-out) set of 11,320 radiographs (1,295 patients, mean age, 61.0 years; 56% female; 16% Black; 39% with income <$50,000; 38% non-college graduates). The following results are shown for the validation set alone, and no patients from the training or development sets were included in the validation set.

The resulting severity measure, denoted by algorithmic pain prediction (ALG-P), summarizes the objective features present in the radiograph that predict pain. As a preliminary check of the network's ability to predict pain, the Pearson correlation, Spearman correlation, root mean square error (RMSE) and mean absolute error of ALG-P for KOOS pain score were estimated; area under the curve (AUC) for predicting severe pain was also calculated (KOOS ≤ 86.1)[23]. As a preliminary check of validity, we found that the network's ability to predict pain was at least as good as that of the KLG measure. The Pearson R^2 value was 0.16 for ALG-P (95% CI, 0.13–0.19) versus 0.10 for KLG (95% CI, 0.08–0.13), representing a relative increase of 61% (95% CI, 38–86%). Further details and performance metrics (AUC for severe pain, etc.) are provided in Extended Data Fig. 4.

We found that disparities in osteoarthritis pain can be better accounted for by differences in this new measure of radiographic disease severity, relative to the standard measure KLG. As shown in Table 2, ALG-P accounted for 43% (95% CI, 33–56%) of the racial pain disparity, 4.7 times more than did KLG (95% CI, 3.2–11.8). It also accounted for 2.0 times more of the disparity by income (32% versus 16%) and 3.6 times more of the disparity by education (30% versus 8%). Importantly, these results were not specific to the KLG scoring system. Racial and socioeconomic disparities in pain persisted when controlling for alternative measures (for example, Osteoarthritis Research Society International (OARSI) joint space narrowing (JSN) grade[24]) or when controlling for the radiologist interpretation of the MRI (as measured by the MRI Osteoarthritis Knee Score (MOAKS)[25]) for the 22% of observations with MRI studies of the knee available (Methods).

Several sensitivity tests were run to determine whether the algorithm's predictive performance was driven by confounding factors or true signal in radiographs (for further details, see Methods). First, when ALG-P was grouped into five bins with sizes equal to those for KLG, the explanatory power of ALG-P was still greater than that of KLG, demonstrating that the algorithm was not simply learning a more granular version of KLG. Consistent with this, regressing ALG-P on KLG and image features that are commonly measured radiologically yielded an R^2 of only 73%. Second, importantly, ALG-P did not simply learn how to reconstruct race or socioeconomic status, and thereby pain, from radiographs, because it remained predictive for pain when controlling for race and socioeconomic status and achieved better predictive performance for pain than did KLG, even within racial and socioeconomic subgroups. Third, there was no evidence that ALG-P was gaining predictive power from image artifacts (or predicting pain only by predicting other features, such as body mass index (BMI)) (Fig. 1), nor that it was learning a radiographic predictor specific to one recruitment site; see Methods for additional robustness checks.

After ruling out these explanations, we attempted to understand how ALG-P predictions reduce pain disparities. We hypothesized that the algorithm's key advantage was learning from a diverse dataset, with nearly 20% Black patients and many lower-income and lower-education patients. Statistics on the overlap between the Black, lower-education and lower-income patient groups are provided in Extended Data Fig. 3. This was tested by retraining the neural network under two experimental conditions: (1) using a non-diverse training set from which all minority patients (for example, all Black patients; we also performed analogous experiments by removing all lower-income patients and all lower-education patients) had been removed and (2) using an equally sized diverse training set from which a subset of non-minority patients were removed. While models trained under both conditions outperformed KLG,

Table 2 | Reducing unexplained racial and socioeconomic disparities in pain

	Pain disparity (KOOS points) after controlling for			Reduction in pain disparity after controlling for		Ratio of reduction
	No severity measures	Radiographic severity (KLG)	Algorithmic severity (ALG-P)	Radiographic severity (KLG)	Algorithmic severity (ALG-P)	ALG-P to KLG
Race	10.6 (8.3, 12.9)	9.7 (7.4, 11.9)	6.1 (3.7, 8.3)	9% (3%, 16%)	43% (33%, 56%)	4.7 (3.2, 11.8)
Income	4.2 (2.8, 5.6)	3.5 (2.3, 4.9)	2.9 (1.6, 4.1)	16% (5%, 29%)	32% (18%, 50%)	2.0 (1.4, 4.4)
Education	5.3 (3.7, 6.7)	4.9 (3.5, 6.2)	3.7 (2.4, 5.0)	8% (−1%, 18%)	30% (18%, 44%)	3.6 (2.1, *)

The first three columns report racial and socioeconomic pain disparities (in KOOS points) without any controls for severity (first column, equivalent to the difference in mean pain scores between groups), when controlling for the clinician's severity measure KLG (second column) and when controlling for algorithmic severity measure ALG-P (third column). Controlling for either severity measure reduces racial and socioeconomic pain disparities, with the algorithmic severity measure achieving a larger reduction. The final three columns quantify the sizes of these reductions. The fourth column reports how much pain disparities are reduced by controlling for KLG, relative to controlling for no severity measures (that is, the reduction in the second column relative to the first). The fifth column reports how much pain disparities are reduced by controlling for ALG-P, relative to controlling for no severity measures (that is, the reduction in the third column relative to the first). The final column reports the ratio of the reductions in disparities by ALG-P versus KLG. In parentheses are 95% CIs computed by cluster bootstrapping at the patient level. The asterisk in the bottom-right entry (*) indicates that the upper limit of the CI was not defined, because the CI for the denominator included zero.

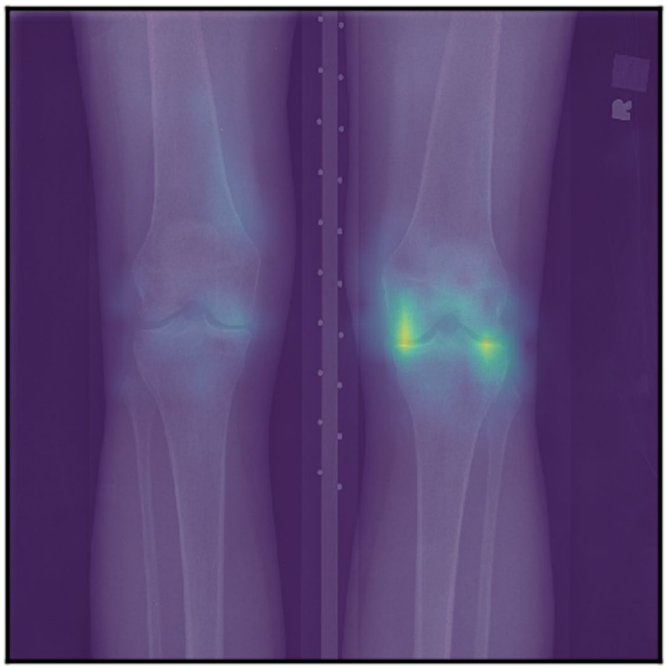


Fig. 1 | Heatmap of a representative X-ray image. The model's prediction target is the pain score in the knee appearing on the right side of the image. Regions that influence the prediction more strongly are shown in brighter colors.

models trained on the diverse training sets achieved better predictive performance for pain and greater reductions in racial and socioeconomic pain disparities than models trained on the non-diverse training sets of the same size (Extended Data Fig. 1). The model trained on a dataset with no Black patients reduced the racial pain disparity by only 2.3× KLG, as opposed to an average of 4.9× for models trained on five randomly sampled diverse training sets of the same size (P value for difference, <0.001 for all five randomly sampled training sets; results when removing all lower-income or all lower-education patients were similar). Thus, training set diversity contributes to the algorithm's ability to reduce disparities.

In addition to raising important questions regarding how we understand potential sources of pain, our results have implications for the determination of who receives arthroplasty for knee pain. While radiographic severity is not part of the formal guideline in allocations for arthroplasty (which only requires evidence of radiographic damage[26]), empirically, patients with higher KLGs are more likely to receive surgery[27]. Consequently, we hypothesize that underserved patients with disabling pain but without severe radiographic disease could be less likely to receive surgical treatments and more likely to be offered non-specific therapies for pain. This approach could lead to overuse of pharmacological remedies, including opioids, for underserved patients and contribute to the well-documented disparities in access to knee arthroplasty[10,27,28].

Our findings are consistent with previous literature reporting that underserved patients are less likely to receive knee surgery[29]. In our data, Black patients have 0.78 lower odds (95% CI, 0.64–0.96) of receiving knee surgery, as do lower-income (0.63; 95% CI, 0.54–0.74) and lower-education patients (0.85; 95% CI, 0.74–0.99). Patients from underserved populations are also more likely to be treated with opioids (odds ratios, 2.17 for Black (95% CI, 1.58–2.99), 1.78 for lower-income (95% CI, 1.34–2.37) and 2.33 for lower-education patients (95% CI, 1.74–3.11).

Disparities in pain, particularly those remaining after adjustment for standard radiographic severity, could contribute to these observations. Patients with greater radiographic severity are empirically more likely to receive arthroplasty[27] (although formal arthroplasty guidelines simply require presence of radiographic damage[26]). Arthroplasty removes tissue objectively affected by degenerative disease and thereby relieves pain (though no trials specifically demonstrated that benefit varies by radiographic appearance[10,28]). As a result, most total knee replacements occur in patients with end-stage knee osteoarthritis[11].

ALG-P identifies a subgroup of patients who have severe pain, based on the radiographic appearance of the knee; however, this appearance is not consistent with severe osteoarthritis as defined by commonly used radiographic grading systems. It is possible that these patients would benefit from arthroplasty, but because radiographic osteoarthritis severity partially determines the decision to offer surgery (along with pain, function and quality of life), these patients may not be offered surgery. Because these patients, with severe pain and high ALG-P but lower osteoarthritis severity (KLG), were more likely to be Black, limitations of standard measures could contribute to disparities in access to arthroplasty. To test this hypothesis, we replicated a procedure previously used in an analysis of arthroplasty allocation, using severe knee pain (KOOS ≤ 86.1) and severe osteoarthritis (KLG ≥ 3) to identify patients in our dataset who were likely under most active consideration for arthroplasty[27]. These patients were then compared with patients identified using our alternative eligibility rule comprising severe pain and our alternate measure of severe osteoarthritis, severe ALG-P, as opposed to severe KLG. Table 3 illustrates the differences between the existing and simulated guidelines for allocation to arthroplasty. The same number of knees were classified as having severe osteoarthritis when using both the KLG and ALG-P severity measures.

Table 3 | Potential eligibility for surgery: comparing KLG and ALG-P

	Knees potentially eligible for surgery (%)		Knees in severe pain and not eligible for surgery (%)	
	Using KLG	Using ALG-P	Using KLG	Using ALG-P
Black	11% (7%, 15%)	22% (17%, 27%)	51% (45%, 57%)	40% (34%, 46%)
Lower-income	10% (8%, 12%)	13% (10%, 15%)	36% (33%, 40%)	34% (31%, 38%)
Lower-education	9% (7%, 11%)	14% (11%, 16%)	38% (35%, 42%)	33% (30%, 37%)

However, measuring severity with ALG-P, rather than with KLG, would double the potential eligibility for arthroplasty for Black patients, increasing it from 11% to 22% of knees ($P<0.001$). Using ALG-P would also decrease the fraction of knees that have severe pain and are ineligible for surgery from 51% to 40% among Black patients ($P<0.001$). Among the population not currently eligible for surgery, patients with the highest ALG-P severity scores were also the patients most likely to be taking analgesics, including opioids (odds ratio, 1.24 for a 1-s.d. worsening in ALG-P; $P=0.008$). As arthroplasty is known to reduce pain, this reallocation of surgery could potentially narrow the racial and socioeconomic disparities in pain as well as reduce the use of opioids for those in severe pain[30].

In summary, we used a machine-learning algorithm to show that standard radiographic measures of severity overlook objective but undiagnosed features that disproportionately affect diagnosis and management of underserved populations with knee pain. As radiographic severity is a key input to management decisions, we propose that our new algorithmic measure ALG-P could potentially enable expanded access to treatments for underserved patients.

This study has limitations. While the Osteoarthritis Initiative (OAI) dataset used for our analysis enrolled a diverse patient group from sites across the USA, our findings need to be validated in independent populations. This would also serve as a check on overfitting, which was minimized by creating a separate validation set before beginning any analysis. The analysis of access to arthroplasty for underserved populations is speculative. We can estimate who might receive surgery, based on pain and radiographic severity, but do not observe the surgical decision-making process. Similarly, it was not possible to assess how using ALG-P as a decision aid would affect patient outcomes in the current study. Finally, a central question we were not able to address is which features of the knee the algorithm is using. Beyond our study, this is generally difficult to determine with neural networks, and fully explaining the signal that algorithms find remains a pressing topic for future work, if algorithms are to be responsibly deployed in medical decision making. Caution is warranted because, while ALG-P accounts for significantly more of the variance in pain than does KLG, the variance accounted for by both methods is low. This low variance does not prevent us from studying disparities between racial or socioeconomic groups, as it is a common feature in studies of disparities in complex, unpredictable traits. The goal in such studies is not to explain all the variance between people, but to understand the group disparities that persist when controlling for relevant contextual variables. Still, one interesting possibility for future work would be to explore whether predictive performance for pain could be improved using deep learning models with different architectures, for example, architectures that accommodate three-dimensional data to make predictions from MRI or combine images from multiple time points to leverage longitudinal datasets[31,32].

One promising option for integrating our algorithm into clinical practice is to use it as a decision aid, rather than as a replacement, for human clinicians (for example, by showing the clinician a heatmap of affected regions within the knee (Fig. 1) alongside the ALG-P score). Such cooperation between humans and algorithms was shown to improve clinical decision making in some settings[33], although this approach is not without challenges, such as physicians potentially placing incorrect levels of weight on algorithmic predictions[34]. More broadly, our results illustrate how algorithms can be used to identify and reduce disparities in healthcare.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-01192-7.

Received: 20 March 2020; Accepted: 24 November 2020;
Published online: 13 January 2021

References

1. Zhang, Y. & Jordan, J. M. Epidemiology of osteoarthritis. *Clin. Geriatr. Med.* **26**, 355–369 (2010).
2. Eberly, L. et al. Psychosocial and demographic factors influencing pain scores of patients with knee osteoarthritis. *PLoS ONE* **13**, e0195075 (2018).
3. Allen, K. D. et al. Racial differences in self-reported pain and function among individuals with radiographic hip and knee osteoarthritis: the Johnston County Osteoarthritis Project. *Osteoarthr. Cartil.* **17**, 1132–1136 (2009).
4. Collins, J. E., Katz, J. N., Dervan, E. E. & Losina, E. Trajectories and risk profiles of pain in persons with radiographic, symptomatic knee osteoarthritis: data from the Osteoarthritis Initiative. *Osteoarthr. Cartil.* **22**, 622–630 (2014).
5. Allen, K. D. et al. Racial differences in osteoarthritis pain and function: potential explanatory factors. *Osteoarthr. Cartil.* **18**, 160–167 (2010).
6. Bolen, J. et al. Differences in the prevalence and impact of arthritis among racial/ethnic groups in the United States, National Health Interview Survey, 2002, 2003, and 2006. *Prev. Chronic Dis.* **7**, A64 (2010).
7. Poleshuck, E. L. & Green, C. R. Socioeconomic disadvantage and pain. *Pain* **136**, 235–238 (2008).
8. Anderson, K. O., Green, C. R. & Payne, R. Racial and ethnic disparities in pain: causes and consequences of unequal care. *J. Pain* **10**, 1187–1204 (2009).
9. Krause, N. et al. Psychosocial job factors associated with back and neck pain in public transit operators. *Scand. J. Work Env. Health* **23**, 179–186 (1997).
10. Deveza, L. A. & Bennell, K. Management of knee osteoarthritis. *UpToDate* https://www.uptodate.com/contents/management-of-knee-osteoarthritis (2019).
11. Losina, E., Thornhill, T. S., Rome, B. N., Wright, J. & Katz, J. N. The dramatic increase in total knee replacement utilization rates in the United States cannot be fully explained by growth in population size and the obesity epidemic. *J. Bone Joint Surg. Am.* **94**, 201–207 (2012).
12. Hochberg, M. C. et al. Effect of intra-articular sprifermin vs placebo on femorotibial joint cartilage thickness in patients with osteoarthritis: the FORWARD randomized clinical trial. *JAMA* **322**, 1360–1370 (2019).
13. Vina, E. R., Ran, D., Ashbeck, E. L. & Kwoh, C. K. Natural history of pain and disability among African–Americans and Whites with or at risk for knee osteoarthritis: a longitudinal study. *Osteoarthr. Cartil.* **26**, 471–479 (2018).
14. Neogi, T. et al. Association between radiographic features of knee osteoarthritis and pain: results from two cohort studies. *BMJ* **339**, b2844 (2009).
15. Bedson, J. & Croft, P. R. The discordance between clinical and radiographic knee osteoarthritis: a systematic search and summary of the literature. *BMC Musculoskelet. Disord.* **9**, 116 (2008).
16. Sayre, E. C. et al. Associations between MRI features versus knee pain severity and progression: data from the Vancouver longitudinal study of early knee osteoarthritis. *PLoS ONE* **12**, e0176833 (2017).
17. Kellgren, J. H. & Lawrence, J. S. Radiological assessment of osteo-arthrosis. *Ann. Rheum. Dis.* **16**, 494–502 (1957).
18. Haug, W., Compton, P. & Courbage, Y. (eds.) *The Demographic Characteristics of Immigrant Populations* Vol. 38 (Council of Europe, 2002).
19. Cheek, N. N. & Shafir, E. The thick skin bias in judgments about people in poverty. *Behav. Public Policy* **4**, 1–26 (2020).

20. Hoffman Kelly, M., Trawalter, S., Axt Jordan, R. & Oliver, M. N. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between Blacks and whites. *Proc. Natl Acad. Sci. USA* **113**, 4296–4301 (2016).
21. Nevitt, M. C., Felson, D. T. & Lester, G. The Osteoarthritis Initiative. https://nda.nih.gov/oai/ (2006).
22. Roos, E. M., Roos, H. P., Lohmander, L. S., Ekdahl, C. & Beynnon, B. D. Knee injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. *J. Orthop. Sports Phys. Ther.* **28**, 88–96 (1998).
23. Englund, M., Roos, E. M. & Lohmander, L. S. Impact of type of meniscal tear on radiographic and symptomatic knee osteoarthritis: a sixteen-year followup of meniscectomy with matched controls. *Arthritis Rheum.* **48**, 2178–2187 (2003).
24. Altman, R. D. & Gold, G. E. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthr. Cartil.* **15**, A1–A56 (2007).
25. Hunter, D. J. et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). *Osteoarthr. Cartil.* **19**, 990–1002 (2011).
26. Rankin, E. A., Alarcon, G. S., Chang, R. W. & Cooney, L. M. Jr NIH Consensus Statement on total knee replacement December 8–10, 2003. *J. Bone Joint Surg. Am.* **86**, 1328–1335 (2004).
27. Losina, E. et al. Lifetime medical costs of knee osteoarthritis management in the United States: impact of extending indications for total knee arthroplasty. *Arthritis Care Res.* **67**, 203–215 (2015).
28. Lingard, E. A. & Riddle, D. L. Impact of psychological distress on pain and function following knee arthroplasty. *J. Bone Joint Surg. Am.* **89**, 1161–1169 (2007).
29. Skinner, J., Weinstein, J. N., Sporer, S. M. & Wennberg, J. E. Racial, ethnic, and geographic disparities in rates of knee arthroplasty among Medicare patients. *N. Engl. J. Med.* **349**, 1350–1359 (2003).
30. Riddle, D. L., Perera, R. A., Jiranek, W. A. & Dumenci, L. Using surgical appropriateness criteria to examine outcomes of total knee arthroplasty in a United States sample. *Arthritis Care Res.* **67**, 349–357 (2015).
31. Xu, Y. et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin. Cancer Res.* **25**, 3266–3275 (2019).
32. Bien, N. et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med.* **15**, e1002699 (2018).
33. Steiner, D. F. et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).
34. Uyumazturk, B. et al. Deep learning for the digital pathologic diagnosis of cholangiocarcinoma and hepatocellular carcinoma: evaluating the impact of a web-based diagnostic assistant. In *Machine Learning for Health ML4H* (NeurIPS, 2019).

Methods

Dataset. Clinical and radiological data were employed from the OAI, a multicenter, longitudinal study of participants aged 45–79 years who had or were at high risk of developing knee osteoarthritis[21]. Study data were anonymized, and this analysis was deemed exempt from review by the Stanford IRB.

Data were analyzed from five time points (baseline visit and 12-, 24-, 36- and 48-month follow-ups). Each observation in the dataset corresponds to one knee for one person at one time point. Observations were removed if they were missing pain scores, KLG, age, race, sex, socioeconomic status or a knee X-ray image that passed the study's quality control. After applying these filters, 4,172 of the original 4,796 patients (87%) were included.

We randomly divided the data at the patient level (not the image level) into a training set, which was used to optimize model weights; a development set, which was used to conduct hyperparameter search and rank models by RMSE for pain score; and a blinded validation (held-out) set (approximately one-third of patients), for which no statistical analyses were performed until the model training procedure, including all hyperparameters, was finalized. (Extended Data Fig. 2 summarizes the analysis pipeline.) We confirmed that all statistics reported in Table 1 were balanced between the training-development and validation (held-out) set (all P values for differences, >0.05). All results were reported on the validation (held-out) set. All CIs and P values were computed by clustering at the patient level to account for repeated observations from each patient. All P values are two-sided.

Radiological images and preprocessing. Bilateral fixed flexion knee X-rays were used in the analysis and preprocessed using standard methods (for example, as in Rajpurkar et al.[35]). Each image was normalized by first dividing pixel values by the maximum pixel value (so that all pixel values were in the range 0–1) and then z scoring (subtracting the mean and dividing by the s.d. across all pixels). Using alternate image normalization methods (z scoring each image individually or z scoring using the mean and s.d. of the ImageNet dataset that the neural network was originally trained on) did not substantially affect performance. Images were downsampled to 1,024 × 1,024 pixels. Images were removed if they did not pass quality control filters, as annotated in the OAI X-ray image metadata.

Study outcomes. As part of the OAI study, images were scored by radiologists on radiographic features of osteoarthritis, including summary measures of severity (for example, KLG) and other features (for example, osteophytes and JSN)[17,24,36].

KLG, a standard measure of osteoarthritis severity, is a five-level categorical variable (0–4), with increasing grades indicating increasing disease severity[17,36]. KLG ≥ 2 is used as a standard threshold for radiological osteoarthritis[21]. Besides KLG, 18 other radiographic features, which quantify osteophytes, JSN, subchondral sclerosis, cysts, chondrocalcinosis and attrition, were also used to train the neural network and to interpret its predictions, as described below. For the scoring of radiographic features, while some images were assessed multiple times by independent teams (referred to as projects 15 and 37), only the assessments from project 15 were used in the analysis, because project 37 assessed only a non-random subset of participants. The OAI only assessed these additional 18 radiographic features (besides JSN, which was assessed in all participants) for participants who developed radiographic osteoarthritis in at least one knee (KLG ≥ 2) at any time point. Therefore, in this analysis, radiographic features were set to zero for other participants; in other words, it was assumed that participants who were never assessed to have osteoarthritis, and thus were not assessed for other radiographic features of osteoarthritis, did not display these features. To ensure that results were not specific to using KLG, a sensitivity analysis was performed using OARSI JSN[24]. Knee MRIs were also collected for a subset of patients and scored using the MOAKS method[25], which we used for another similar sensitivity analysis in this subset.

KOOS pain score was used as a measure of self-reported pain[22]. KOOS is a knee-specific score (0–100, with lower scores indicating greater pain) derived from a multi-item survey on how often patients experience knee pain and pain severity during various activities (for example, 'straightening the knee fully'); as usual, responses to each survey question were aggregated into a single score[22].

Neural network training. A convolutional neural network was trained to predict KOOS pain score for each knee using each X-ray image. The input to the network was an X-ray of both knees, meaning that each X-ray for each person at each time point yielded two separate observations, one for each knee. To ensure that the prediction target was always the KOOS pain score in the knee that appeared on the right side of the image, we flipped the original image horizontally when necessary (that is, when the target knee appeared on the left side of the original image). The network was provided with both knees on the hypothesis that asymmetry between the knees might be predictive for pain; empirically, using both knees slightly improved prediction performance.

To give the network additional information about each image and guide it toward learning medically meaningful features, the network was trained to predict both KOOS pain score (its primary objective) and 19 radiographic features (KLG and the 18 additional radiographic features). For each training example, the network tried to minimize the following loss:

$$\left(Y_{\text{true}} - Y_{\text{predicted}}\right)^2 + \lambda \sum_j \left(C^{(j)}_{\text{true}} - C^{(j)}_{\text{predicted}}\right)^2$$

Y is the KOOS pain score, $C^{(j)}$ is the z-scored jth image feature, and λ is a weight chosen by hyperparameter search. (Because the primary objective was to predict KOOS pain score, the RMSE for predicting KOOS pain score was used as the criterion for selecting model hyperparameters, as described below.) Intuitively, this loss encourages the network to learn to predict the KOOS pain score, its primary objective, but also the radiographic features and thereby learn a representation of the knee X-ray that captures medically relevant information. We emphasize that the additional features were not used as input to the network; the network only used the knee X-ray as input.

The network used a ResNet-18 architecture, with network weights pretrained on ImageNet[37,38]. Deeper layers of the architecture were then fine-tuned on the OAI dataset. The training dataset was augmented by applying random horizontal and vertical translations to each image[39]. Adam[40] was used to optimize network weights, with an initial learning rate that decayed by a factor of 2 each time the loss plateaued. To mitigate overfitting, early stopping was used, and model weights were set at the completion of training to those after the epoch with the lowest RMSE for KOOS pain score on the development set. Random search was used to choose the network hyperparameters, including the batch size, magnitude of the horizontal and vertical translations for dataset augmentation, network architecture and number of layers to fine tune, optimizer to use and optimizer hyperparameters, the number of epochs to train for and the learning rate schedule. After finalizing the network architecture and training procedure, multiple models were trained (initialized with different random seeds), and the top five models (as measured by RMSE for KOOS pain score on the development set) were ensembled[41]. Training was performed on four Nvidia XP GPUs. Analysis was performed using Python 3.5.

Quantifying pain disparities. The main outcome was racial disparities in pain between Black (16% of patients in the validation set) and non-Black patients (84%, of whom 97% were white). Disparities by two socioeconomic measures were also considered: whether the patient had an annual income below $50,000 (39% of patients) and whether they had graduated from college (38% had not). Differences in pain scores across groups were first quantified without controlling for osteoarthritis severity, using mean KOOS pain score between groups (for example, racial pain disparity was defined as the difference in mean pain between Black and non-Black patients). Extended Data Fig. 3 reports the mean KOOS pain score for each race and socioeconomic subgroup.

We then computed the racial and socioeconomic pain disparities that remained when controlling for radiographic osteoarthritis severity. To do so, our approach was to fit a linear regression with KOOS pain score as the dependent variable and two independent variables: binary race or socioeconomic group and a measure of osteoarthritis severity (see below for specifics). The pain disparity was defined as the coefficient on binary race or socioeconomic group; that is, the gap in mean pain between racial or socioeconomic groups when controlling for severity.

We defined two alternative measures of osteoarthritis severity. First, we used the network's predicted pain score ALG-P; this can be thought of as summarizing the radiographic features that are linked to pain, as quantified by the network. Second, we used the radiologist's assessment of severity, as measured by KLG. To ensure fair comparison of explanatory power between ALG-P and KLG, we first rescaled KLG by predicting pain from KLG (in the combined training and development sets) using a regression in which KLG was coded as a categorical variable, with a separate coefficient for each of the five levels; this allowed for maximum flexibility in predicting pain from KLG, in case the relationship between the two was nonlinear. Lasso regression was used as a standard technique to prevent overfitting[42]. Conceptually, the output of this regression model (which was generated in the held-out set) was a rescaled KLG on the same scale as KOOS pain score and thus the same scale as ALG-P.

An alternate procedure would have been to fit a regression controlling for KLG coded as a categorical variable (rather than for rescaled KLG). We favored the procedure used in this paper because it treats the clinical and algorithmic pain predictions consistently; for both predictors, the training–development sets are used to learn a pain predictor, and then that predictor is assessed on the validation set. This avoids potential overfitting to the validation set. However, the two procedures are extremely similar, and we confirmed that the procedure used in this paper yielded estimates of pain disparities that were essentially identical to those produced by the alternate procedure. The income pain disparity estimates differed by 0.2% (3.529 versus 3.524), the racial pain disparity estimates differed by 0.6% (9.664 versus 9.718), and the education pain disparity estimates differed by 0.3% (4.879 versus 4.895).

Because our analysis performs a regression of pain on severity score and binary racial or socioeconomic group, it implicitly fits a model in which the relationship between pain and severity score is the same for both groups. As a robustness check, we performed an additional regression that included an interaction between group and severity score and assessed the significance of the interaction term. In all cases, the interaction term was small (at most, one-quarter

of the main slope effect) and not statistically significant after multiple hypothesis correction (Bonferroni-adjusted $P > 0.05$). This indicates that the relationship between pain and severity score did not differ significantly across groups. As an additional check that our results were not sensitive to the use of linear regression to quantify the pain gap (and the parametric assumption of equal slopes across groups), we performed an alternate computation in which we quantified the pain gap as the sum of gaps between groups at each of the five severity levels (0, 1, 2, 3 and 4), weighting each level by the number of knees at that level. This procedure is a non-parametric means of fully accounting for any differences across racial or socioeconomic groups in the relationship between severity score and pain. Our results remained extremely similar under this alternate definition of the pain gap; our estimation of the pain gap changed by less than 5% in all cases (for both severity scores and all three racial or socioeconomic groups).

Comparing predictive powers of ALG-P and KLG. We found that ALG-P explained 61% (95% CI, 38–86%) more of the variance in pain than did KLG, indicating that the knee X-rays did contain signal for predicting pain that KLG did not capture. The Pearson R^2 for ALG-P was 0.16 (compared to 0.10 for KLG) (Extended Data Fig. 4). When regressing pain on both ALG-P and KLG, the coefficient on ALG-P remained significant ($P < 0.001$), but the coefficient on KLG became insignificant ($P = 0.20$). This indicates that ALG-P captured the signal for pain that was present in KLG, while also capturing signal that KLG did not.

Not only did ALG-P correlate with patients' current pain scores, it also identified patients who went on to have significantly worse future pain trajectories over the follow-up period. When controlling for pain score at baseline, a 1-s.d. worsening in ALG-P corresponded to 1.5× higher odds (95% CI, 1.4–1.7) that patients would be in severe pain at follow-up (combining data across all follow-up visits). Binning ALG-P into five categories of the same size as KLG bins, patients with a binned ALG-P ≥ 2 had 1.7× (95% CI, 1.5–2.0) higher odds of being in severe pain at follow-up when controlling for pain at baseline; patients with a binned ALG-P of 4, the highest grade, had 2.9× (95% CI, 1.9–4.5) higher odds of being in severe pain at follow-up. ALG-P also significantly predicted progression of KLG, even after controlling for KLG at baseline; a 1-s.d. change in ALG-P predicted a 0.07-s.d. worsening in KLG at follow-up (95% CI, 0.06–0.08).

Visualizing image regions that influenced predictions. To compute the degree to which a region of the image influenced the neural network's predicted pain score, the region was 'masked' out, by replacing it with a circle, the value of which was the mean pixel value for the image, using Gaussian smoothing to prevent sharp boundaries[43]. The absolute change in the neural network's predicted pain level (comparing the masked image to the original image) was then computed. This process was repeated for a 32 × 32 grid of regions, evenly tiling the 1,024 × 1,024-pixel image, allowing computation of a heatmap for how much masking each region of the image affected the neural network's prediction (Fig. 1). As an additional robustness check, class activation mapping was used, which similarly indicated that the neural network's prediction was, as expected, primarily influenced by the knee that appeared on the right side of the image, although it was also somewhat influenced by the contralateral knee[44]. (Because the predicted output variable was continuous, for class activation mapping, each filter was upweighted by its weight in the final fully connected layer.)

Allocation of arthroplasty following clinical guidelines. To simulate how arthroplasty would be differentially allocated when using KLG versus ALG-P as a severity measure, we replicated a procedure previously used in an analysis of arthroplasty allocation by identifying patients with severe pain (KOOS ≤ 86.1) and severe osteoarthritis (KLG ≥ 3)[23,27]. A different guideline was then simulated, for which eligibility was driven by severe pain and severe ALG-P, instead of severe pain and severe KLG. To do so, we used the categorical version of ALG-P, on the same scale of 0–4 as for KLG, by dividing the continuous ALG-P into five bins with the same size as KLG bins; arthroplasty was then allocated to knees with severe pain (KOOS ≤ 86.1) and severe osteoarthritis (categorical ALG-P ≥ 3). The same number of knees were classified as having severe osteoarthritis under both severity measures; only the ranking of knees changed. In this analysis, knees were excluded that had already had any knee surgery, and only knees at baseline were considered; neither of these decisions substantially altered results.

Validation of training and image processing pipeline. As a check that the overall training and image preprocessing procedure was able to extract meaningful signal from the image, a model with the same architecture used for the main prediction task was trained to predict KLG (rather than KOOS pain score) from the images. This prediction task was chosen because it was the subject of substantial research, allowing validation of the pipeline used in this analysis in comparison to previous studies[45,46]. Predictive performance on this task was comparable to that of previous studies using models specifically designed to predict KLG (mean square error, 0.35 as compared to 0.48 and 0.50 in previous studies; R^2, 0.87)[45,46]. This indicates that the model was able to extract clinically relevant signal from the image, even on a task it was not originally designed to perform.

Robustness to alternate measures of disease severity. To confirm that results were not specific to the measure of osteoarthritis severity used (KLG), the main analyses were repeated using two alternate measures of osteoarthritis severity. First, OARSI JSN grade was used as a measure of severity, defining a single severity measure by taking the maximum grade over the medial and lateral compartments, which is a standard procedure[24,47]. Similar to the results when comparing to KLG, ALG-P predicted more of the variance in pain (R^2, 0.16) than did JSN (R^2, 0.09), the prediction performance of which was comparable to that of KLG (R^2, 0.10). ALG-P also achieved greater reductions in racial and socioeconomic pain disparities than did JSN: a 3.9× greater reduction in the education pain disparity (30% versus 8%), a 2.1× greater reduction in the income pain disparity (32% versus 16%) and a 7.7× greater reduction in the racial pain disparity (43% versus 6%).

To confirm that results were not specific to radiographic measures of image severity, the main analyses were repeated using MOAKS scores of knee MRIs for the 22% of observations for which they were available[25]. Following a previously used procedure for summarizing MOAKS scores, we extracted MOAKS scores assessing bone marrow lesions, cartilage and meniscus variables; aggregated subscores by taking the maximum within each knee compartment; and applied a threshold to the resulting value to produce a binary variable[48]. This resulted in ten binary variables summarizing the MOAKS scores. On the subset of observations for which MOAKS scores were available, ALG-P predicted more of the variance in pain (R^2, 0.20) than did the MOAKS summary measures, either on their own (R^2, 0.14) or when combined with radiographic features (R^2, 0.16). ALG-P also achieved greater reductions in racial and socioeconomic pain disparities than did the MOAKS summary measures: a greater reduction in the education pain disparity (44% versus 22%), the income pain disparity (52% versus 32%) and the racial pain disparity (52% versus 2%).

Robustness check: ALG-P is not merely a more granular KLG. ALG-P's superior predictive performance could come from the fact that it is a continuous prediction for pain, while KLG is confined to coarser bins (five categories). To test for this, we produced a categorical version of ALG-P, on the same scale of 0–4 as for KLG, by dividing the continuous ALG-P into five bins with the same size as KLG bins. The categorical version of ALG-P still achieved superior predictive power (R^2, 0.15 versus 0.10 for KLG and 0.16 for the continuous ALG-P). It also narrowed racial and socioeconomic pain disparities more than did KLG; it narrowed the racial pain disparity by 4.5× more than did KLG (similar to the original value of 4.7× for the continuous ALG-P), the education pain disparity by 3.4× more than KLG (similar to the 3.6× value for continuous ALG-P) and the income pain disparity by 1.9× more than KLG (similar to the 2.0× value for continuous ALG-P). Of note, the categorical version of ALG-P agreed with KLG only 49% of the time, indicating that ALG-P was actually reranking individuals and not simply learning a more granular version of KLG.

Robustness check: ALG-P is not only reweighting features already known to radiologists. The model could have achieved its predictive performance by simply recovering factors known to radiologists and reweighting them to produce a score different from KLG; for example, placing more weight on osteophytes rather than on sclerosis. To test this, correlations of ALG-P with 19 radiographic features (KLG and an additional 18 radiographic features relevant to osteoarthritis, for example, osteophytes, JSN and sclerosis, as described above) were examined. First, the coefficient of ALG-P in a regression with KOOS pain score as the dependent variable was calculated (0.94; 95% CI, 0.85–1.03, without controlling for radiographic features) and then compared to the coefficient on ALG-P when variables controlling for known radiographic features were added (0.95; 95% CI, 0.80–1.10). The fact that the coefficient did not change indicates that the model's explanatory power for pain was not fully captured by currently measured radiographic features. While ALG-P correlated with a number of radiographic features, with KLG (R^2, 0.57) and all four osteophyte features (R^2, 0.41–0.52) explaining the largest fraction of the variance in ALG-P, ALG-P could not be fully explained by the radiographic features (R^2, 0.73) together.

Robustness check: ALG-P is not simply learning to predict race or socioeconomic status. ALG-P could be narrowing disparities in pain by simply learning how to predict race or socioeconomic status from the knee image. As patients from underserved groups have higher pain, simply learning to predict group membership from the image could produce some signal for predicting pain, without picking up on any independent signal for pain itself. To check that ALG-P's predictive power did not derive merely from predicting race and socioeconomic status, we verified that ALG-P still significantly predicted pain when controlling for our binary variables for race, income and education. In a regression with KOOS pain score as the dependent variable, the coefficient on ALG-P was 0.94 (95% CI, 0.85–1.03) without controlling for binary race or socioeconomic variables and 0.83 (95% CI, 0.74–0.93) when controlling for all three binary race or socioeconomic variables. Thus, the coefficient on ALG-P remained highly statistically significant and similar in magnitude when controlling for race or socioeconomic status. We also verified that ALG-P achieved better predictive performance for pain than did KLG across all six race or socioeconomic groups in our analysis (Black or non-Black, higher- or lower-income and higher- or lower-education).

Robustness check: predictions are not driven merely by image artifacts. The model could be gaining predictive power from image artifacts, for example, related to the study site in which patients were recruited[49]. To check for this, standard visualization techniques were used to assess which regions of the X-rays most influenced the model's predictions. Figure 1 provides a representative example, illustrating that the model's predictions did not appear to be influenced by image artifacts; rather, they were influenced by the expected knee (that is, on the right side of the image) and by regions of the knee (femorotibial joint space and surroundings) that were clinically relevant and consistent with previous work[36,45]. In the heatmap, warmer colors indicate regions of the image that influence the neural network's predictions more strongly.

As an additional check that the model was not merely picking up image artifacts, linear regression was used to assess whether ALG-P still significantly predicted KOOS pain score when controlling for the recruitment site and time point at which imaging was conducted; whether the left or right knee was affected; and the individual's age, sex, marital status, current and maximum BMI, history of knee surgery or injury and smoking or drinking behavior. The coefficient on ALG-P in a regression with KOOS pain score as the dependent variable remained highly statistically significant and similar in magnitude when these controls were included (coefficient, 0.94 (95% CI, 0.85–1.03) without controls; 0.77 (95% CI, 0.67–0.87) with controls), and these controls explained only 32% of the variance in ALG-P.

BMI is an especially plausible source of predictive power, as it is likely detectable from knee radiographs and known to be correlated with pain[50]. Hence, we further confirmed that our predictive power was not only due to predicting BMI by stratifying the dataset by BMI category (18.5–25, 25–30, 30–35 and >35) and confirming that ALG-P still achieved larger R^2 values than did KLG for each BMI group.

In sum, these results indicate that the model was unlikely to be deriving its predictive power merely from image artifacts.

Robustness check: ALG-P generalizes across sites. Previous work has shown that neural network performance on medical data can suffer when networks are tested on data from locations or hospitals that they were not trained on[49]. To assess whether the pain prediction model generalized across the five OAI recruitment sites, we altered the training set such that the model was trained on only four of the five sites; model performance was assessed using the held-out fifth site as a validation set. This experiment was repeated for all five recruitment sites. For all five sites, the algorithmic pain score achieved a higher R^2 value on the held-out site than did KLG and achieved greater reductions in racial and socioeconomic pain disparities. Taking an unweighted average across all five held-out sites, the algorithmic pain predictor achieved an R^2 value of 0.13 (as opposed to 0.10 for KLG and 0.14 for the original ALG-P), a reduction in the racial pain disparity of 31% (as opposed to 7% for KLG), a reduction in the income pain disparity of 27% (as opposed to 13% for KLG) and a reduction in the education disparity of 20% (as opposed to 3% for KLG).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Anonymized imaging and clinical data to reproduce results of this study are available online at https://nda.nih.gov/oai/.

Code availability

Code to reproduce the results of this study is available online at https://github.com/epierson9/pain-disparities.

References

35. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).
36. Kohn, M. D., Sassoon, A. A. & Fernando, N. D. Classifications in brief: Kellgren–Lawrence classification of osteoarthritis. *Clin. Orthop. Relat. Res.* **474**, 1886–1893 (2016).
37. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
38. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
39. Perez, L. & Wang, J. The effectiveness of data augmentation in image classification using deep learning. Preprint at https://arxiv.org/abs/1712.04621 (2017).
40. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* (ICLR, 2014).
41. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
42. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
43. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* **8689**, 818–833 (Springer, 2014).
44. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition* 2921–2929 (IEEE, 2016).
45. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P. & Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci. Rep.* **8**, 1727 (2018).
46. Antony, J., McGuinness, K., O'Connor, N. E. & Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *International Conference on Pattern Recognition* 1195–1200 (IEEE, 2016).
47. Sheehy, L. et al. Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the Multicenter Osteoarthritis Study (MOST). *Osteoarthr. Cartil.* **23**, 1491–1498 (2015).
48. Cutler, D. M., Meara, E. R. & Stewart, S. T. Socioeconomic status and the experience of pain: an example from knees. NBER working paper 27974 (2020); https://www.nber.org/papers/w27974
49. Zech, J. R. et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PLoS Med.* **15**, e1002683 (2019).
50. Rogers, M. W. & Wilder, F. V. The association of BMI and knee pain among persons with radiographic knee osteoarthritis: a cross-sectional study. *BMC Musculoskelet. Disord.* **9**, 163 (2008).

Acknowledgements

We thank K. Blumer, J. Duryea, J. Irvin, P.W. Koh, S. Lamb, G. Lester, S. Li, K. Lin, B. McCann, A. Miller, L. Pierson, C. Olah, M. Raghu, P. Rajpurkar, N. Roth, C. Ruiz, C. Sabatti and participants at several seminars and meetings for helpful comments. We acknowledge financial support from Hertz and NDSEG graduate fellowships and the US Social Security Administration (SSA) grant RDR18000003, funded as part of the Retirement and Disability Research Consortium.

Author contributions

E.P., D.M.C., J.L., S.M. and Z.O. jointly analyzed the results and wrote the paper.

Competing interests

E.P. is employed by Microsoft Research. S.M. and Z.O. have equity interests in LookDeep Health (healthcare services), Dandelion (healthcare services) and Spur Labs (healthcare services). Z.O. has equity interests in Berkeley Data Ventures (consulting).

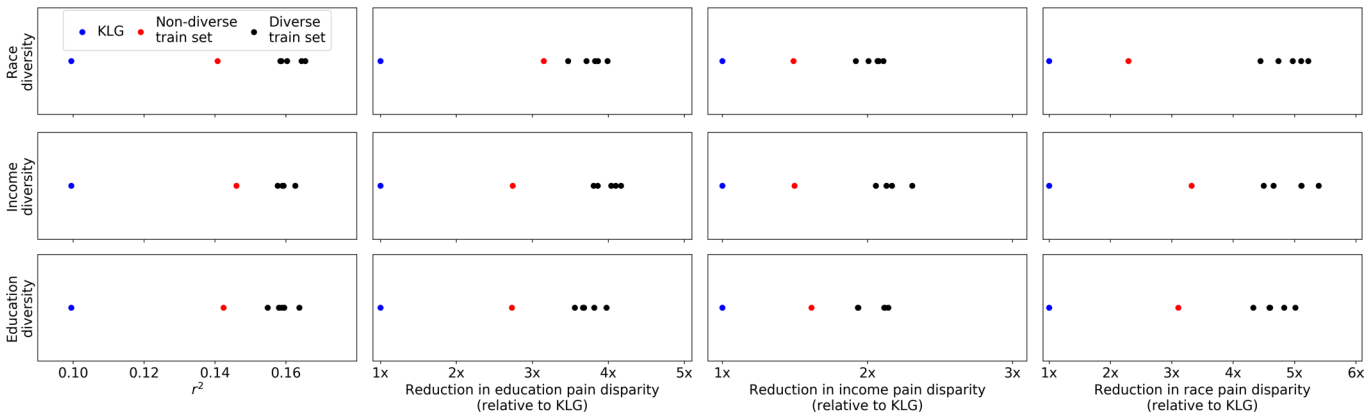
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-020-01192-7.

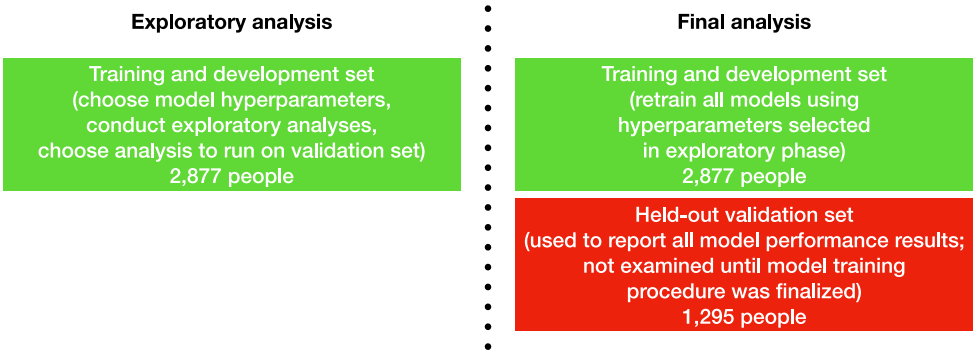
Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-01192-7.

Correspondence and requests for materials should be addressed to S.M.

Peer review information Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



Extended Data Fig. 1 | The effect of dataset diversity on model performance. Each row of plots shows the effect of removing one minority group from the training set: from top, Black, lower-income, and lower-education patients. Each column of plots shows one metric: from left, R^2 in predicting KOOS pain score, and the reductions in the education, income, and racial pain disparities (relative to KLG). In each subplot, the blue dot shows, as a baseline, the performance of KLG. The red dot shows the performance of a neural network trained on a non-diverse training set, with all minority patients removed. The five black dots show the performance of neural networks trained on five diverse training sets of equal size, with five random subsets of non-minority patients removed; in all cases, the diverse training sets yield superior performance to non-diverse training sets of equal size.



Extended Data Fig. 2 | Analysis pipeline. Prior to conducting any analysis, 1,295 patients (red box) were reserved as a held-out validation set to assess final results. In the exploratory phase, the remaining patients were analyzed as follows: a training set was used to optimize model weights, and a development set to select model hyperparameters and conduct early stopping to avoid overfitting. The main analyses to run on the held-out validation set were determined prior to examining it, and the hyperparameters were finalized. In the final analysis, all models were retrained using the hyperparameters chosen in the exploratory phase, and model predictions were assessed on the 1,295 patients in the held-out validation set.

Black	Lower-income	Lower-education	KOOS pain score
No	No	No	89.8 (89.0, 90.6)
No	No	Yes	86.9 (85.1, 88.7)
No	Yes	No	89.2 (87.6, 90.8)
No	Yes	Yes	85.5 (83.7, 87.2)
Yes	No	No	81.5 (77.8, 85.2)
Yes	No	Yes	81.1 (74.9, 87.3)
Yes	Yes	No	80.5 (74.6, 86.4)
Yes	Yes	Yes	74.2 (70.8, 77.5)

Extended Data Fig. 3 | Pain levels among overlapping racial and socioeconomic subgroups. Race and socioeconomic status are correlated: among Black patients, 61% were lower-education and 63% were lower-income, while among non-Black patients, 34% were lower-education and 34% were lower-income.

Predictive performance for pain	KLG	ALG-P
Pearson R^2	0.10 (0.08, 0.13)	0.16 (0.13, 0.19)
Spearman R^2	0.08 (0.07, 0.11)	0.14 (0.11, 0.16)
RMSE	15.4 (14.7, 16.0)	14.9 (14.3, 15.4)
Mean absolute error	11.9 (11.5, 12.2)	11.3 (10.9, 11.7)
AUC (severe pain)	0.64 (0.62, 0.66)	0.69 (0.67, 0.71)

Extended Data Fig. 4 | Predictive performance for pain. 95% CIs are computed by cluster bootstrapping at the patient level.

nature research

Corresponding author(s): Sendhil Mullainathan

Last updated by author(s): Nov 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A; no custom code was used to collect the data in the study.
Data analysis	All data analysis was performed using Python 3.5 with standard libraries. Code to reproduce the results of this study is available online at https://github.com/epierson9/pain-disparities.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Anonymized imaging and clinical data to reproduce results of this study are available online at https://nda.nih.gov/oai/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Longitudinal, quantitative cohort study which studies racial and socioeconomic disparities in pain by predicting pain from knee radiographs.
Research sample	4,172 patients who had or were at high risk of developing knee osteoarthritis aged 45-79; mean age at baseline, 61; 56% female (validation set; 58% female in train set). All data was previously collected by the Osteoarthritis Initiative (OAI) and further details are provided in the Methods section and the OAI protocol. The sample is not representative of the US population because it deliberately over-samples patients with or at risk for osteoarthritis, consistent with the purpose of the OAI and of the present study.
Sampling strategy	The entire OAI cohort was used; we did not conduct any further sampling. Further details about the sample are provided in the Methods section and in the OAI protocol.
Data collection	We did not collect any data; all data collection procedures are described in the OAI protocol and the Methods section. Data was collected by trained medical professionals and OAI scientists using questionnaires and MRI/x-ray equipment.
Timing	We analyze data from the baseline visit (collected 2004-2006), the 12 month followup (2005-2007); the 24-month followup (2006-2008); the 36 month followup (2007-2009) and the 48 month followup (2008-2010).
Data exclusions	To ensure all observations had the requisite data for the present study, observations were removed if they were missing pain scores, Kellgren-Lawrence grade (KLG), age, race, sex, socioeconomic status, or a knee x-ray image which passed the study's quality control. After applying these filters, 4,172 of the original 4,796 patients (87%) were included. All exclusion criteria were determined prior to any analysis of the validation set.
Non-participation	At the final timepoint in our analysis, 89% of participants were available for clinic visit or telephone interview. The reasons for dropout included that participants were contacted but did not provide data (3%); withdrew from the study (4%); could not be contacted (3%); or were deceased (1%). https://nda.nih.gov/oai/study-details/schedule-of-assessments.html provides additional statistics.
Randomization	Not relevant; this is not a randomized controlled trial and participants were not randomized into experimental groups. The purpose of this study is not to assess a treatment, but to evaluate algorithmic performance in reducing unexplained pain disparities.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	See above.

Ethics oversight	Study data were anonymized and this analysis was deemed exempt from review by the Stanford IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.