# An Algorithmic Approach to Explaining Why the Underserved Feel More Pain

**Emma Pierson**, PhD; Department of Computer Science, Stanford University and Microsoft Research
**David M. Cutler**†, PhD; Department of Economics, Harvard University
**Jure Leskovec**†, PhD; Department of Computer Science, Stanford University
**Sendhil Mullainathan**†*, PhD; Booth School of Business, University of Chicago
**Ziad Obermeyer**†, MD; School of Public Health, University of California at Berkeley

†: **Alphabetical order**
**Corresponding author**: Sendhil Mullainathan, PhD, The University of Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637 (sendhil.mullainathan@chicagobooth.edu)

## Abstract

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 experts using medical images, raising the possibility that underserved patients' pain stems from factors *external to the knee* (e.g., stress). Here we use deep learning to create an alternative measure of 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 explain only 9% (95% CI, 3%–16%) of racial disparities in pain, algorithmic predictions explain 43%, or 4.7x more (95% CI, 3.2x–11.8x), with similar results for lower-income and less-educated patients. These results suggest that much of underserved patients' pain stems from factors *within the knee*, but ones not reflected in standard radiographic measures of severity. We show that the algorithm's ability to explain disparities is rooted in racial and socioeconomic diversity of the training set, and that it does not simply reconstruct race or known radiographic features. Since algorithmic predictions better capture underserved patients' pain, algorithmic predictions could potentially redress disparities in access to treatments: Because patients with severe osteoarthritis are empirically more likely to receive arthroplasty, access to surgery would double for Black patients (22% vs. 11%; $p<0.001$) if an algorithmic severity measure were used instead of standard measures.


.

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 in the United States[1,2] —disproportionately impacts underserved populations: non-whites score nearly twice as high on knee pain scales as do whites.[3–10] Understanding these 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.[11–13]

Two kinds of explanations for these disparities have been proposed. First, underserved patients may simply have more severe osteoarthritis *within the knee*. Alternatively, underserved patients may have more aggravating factors *external to the knee*[*]: for example, the same physical ailments can produce very different experienced pain due to life stress, social isolation, or other factors.[11,12,14,15] These 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.[16–18]

Research to date has indirectly implicated explanations external to the knee. Methodologically, this is demonstrated by defining an objective measure of osteoarthritis severity based on knee x-rays, then measuring the extent of pain disparities that remain after adjusting for severity. Typically large differences remain even after adjustment.[4,8,10,19] For example, even though Black patients have more severe osteoarthritis based on standard radiographic measures (Kellgren-

[*] We use this 'internal' vs. 'external' dichotomy descriptively, but recognize its limitations (e.g., certain local nociceptive processes do not fit cleanly into either category). More precisely, we consider structural factors visible on radiographs to be internal, and all other factors to be external.

Lawrence grade, KLG), adjusting for this only slightly decreases measured Black-white disparities in pain.[8,19]

These results, however, depend heavily on how radiographic osteoarthritis severity is measured. The relationship between radiographic severity and pain is debated: many patients with mild or no disease suffer pain, and many patients with structural damage on x-ray or even MRI have no pain.[20–24] Standard radiographic measures like KLG, developed decades ago in white British populations, may miss physical causes of pain in non-white populations.[25,26] If the pain experienced by underserved populations is caused by objective factors missing from current measures, we would misattribute a range of painful, treatable knee ailments to factors external to the knee.

In this paper, we use machine learning to help 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,[27] bilateral fixed flexion knee radiographs were obtained and scored by radiologists on summary measures of radiographic severity (e.g., Kellgren-Lawrence grade: KLG) and other objective features (e.g., osteophytes and joint space narrowing). 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 (e.g., fully straightening the knee).[28]

*Pain and osteoarthritis severity*

**Table 1** shows summary statistics on the 4,172 participants, who generated 36,369 observations (one for each knee at each time point). Black patients have substantially higher pain levels: across knees and timepoints, Black patients experienced severe pain 58% of the time (thresholding at KOOS≤86.1, a standard threshold for severe pain),[29,30] vs. 38% for patients overall (*p*-value for racial difference < 0.001); the median Black patient has worse pain than 75% of non-Black patients. Black patients have 10.6 KOOS points higher pain than non-Black patients (*p*-value for racial difference < 0.001); for comparison, the standard deviation in the dataset was 16.2 KOOS points. We find similar pain disparities across socioeconomic groups. Across knees and timepoints, 43% of lower-income patients, and 45% of lower-education patients, have severe pain (vs. 38% overall; both *p*-values < 0.001).

Black patients also have more severe osteoarthritis, with 56% of knees receiving KLG≥2 vs. 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 does not fully explain Black patients' higher pain levels. **Table 2** shows the racial disparity in pain is 10.6 KOOS points without controlling for any severity measures vs. 9.7 points controlling for KLG, meaning KLG accounts for only 9% of the pain disparity (95% CI, 3%–16%). Results are similar for other underserved groups: KLG explains 16% (95% CI, 5%–29%) and 8% (95% CI, -1%–18%) of the pain disparity by income and education, respectively.

*Developing an algorithmic severity measure*

So far, our results replicate findings in the literature to date,[8,19] and suggest that objective osteoarthritis severity does not explain 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: it was developed decades ago and in a very specific setting that is unlikely to reflect the experience of osteoarthritis in diverse populations.[25,26]

To generate an alternative measure, we train a convolutional neural network to predict the reported pain score for each knee using each x-ray image, using a randomly selected training/development dataset of 25,049 radiographs (2,877 patients). We generate predictions in an independent validation (hold-out) set of 11,320 radiographs (1,295 patients: mean age 61.0; 56% female; 16% Black; 39% income <$50,000; 38% non-college graduates). All results below are shown for the validation set alone, and no patients appear in both training/development and validation sets.

The resulting severity measure, which we denote ALG-P (algorithmic pain prediction), 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, RMSE, and mean absolute error of ALG-P for KOOS pain score were estimated; AUC for predicting severe pain was also calculated (KOOS≤86.1).[29,30] As a preliminary check of validity, we find that the network's ability to predict pain is at least as good as KLG's: Pearson $R^2$ was 0.16 for ALG-P (95% CI, 0.13–0.19) vs. 0.10 for KLG (95% CI, 0.08–0.13), a relative increase of 61% (95% CI, 38%–86%). Further details and performance metrics (AUC for severe pain, etc.) are in **Table S1**.

We find that disparities in osteoarthritis pain can be better explained by differences in this new measure of radiographic disease severity, relative to the standard measure, KLG. As shown in **Table 2**, ALG-P accounts for 43% (95% CI, 33%–56%) of the racial pain disparity—4.7 times more than KLG (95% CI, 3.2–11.8). It also explains 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 are not specific to the KLG scoring system: racial and socioeconomic disparities in pain persist when controlling for alternative measures (e.g., OARSI joint space narrowing),[31] or when controlling for the radiologist interpretation of the MRI (as measured by the MOAKS score,[32] for the 22% of observations with MRI studies of the knee available). Further details are in **SI Section 1.2**.

*Investigating algorithmic performance*

Several tests were run to determine whether the algorithm's predictive performance was driven by confounding factors versus true signal in radiographs. The tests are briefly enumerated here. First, the algorithm is not simply learning a more granular version of KLG. When ALG-P is grouped into 5 bins with sizes equal to those for KLG, the explanatory power of ALG-P is still greater than that of KLG (**SI Section 1.3**). Consistent with this, regressing ALG-P on KLG and image features that are commonly measured radiologically yields an $R^2$ of only 73% (**SI Section 1.4**). Second, importantly, ALG-P is not simply learning how to reconstruct race or socioeconomic status, and thereby pain, from radiographs, because it remains predictive for pain when controlling for race and socioeconomic status and achieves better predictive performance for pain than does KLG even within racial and socioeconomic subgroups (**SI Section 1.5**). Third, there is no evidence that ALG-P is gaining predictive power from image artifacts (or predicting

pain only by predicting other features like body mass index) (**SI Section 1.6, Figure 1**), nor that it is learning a radiographic predictor specific to one recruitment site (**SI Section 1.7**).

After ruling out these explanations, we attempted to explain how algorithmic predictions reduce pain disparities. We hypothesized that the algorithm's key advantage was learning from a diverse dataset—with nearly 20% Black, and many lower-income and lower-education patients. This was tested by retraining the neural network under two experimental conditions: 1) using a "non-diverse" training set from which all *minority* patients (e.g., all Black patients) had been removed, and 2) using an equally sized "diverse" training set from which a subset of *non-minority* patients had been removed. While models trained under both conditions outperform KLG, models trained on the diverse training sets achieve 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 Figure 2**). A model trained on no Black patients reduces the racial pain disparity by only 2.3x KLG, as opposed to an average of 4.9x for models trained on 5 randomly sampled diverse training sets of the same size ($p$-value for difference $< 0.001$ for all 5 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.

*Implications for osteoarthritis management*

In addition to raising important questions regarding how we understand the causes of pain, these results have concrete implications for who receives arthroplasty for knee pain. While radiographic severity is not part of the formal guideline in allocations for arthroplasty (which simply require evidence of radiographic damage[33]) empirically patients with higher KL grades

are more likely to receive surgery.[34] Consequently, underserved patients in disabling pain but without severe radiographic disease are less likely to receive surgical treatments; conversely, they may be more likely to be offered non-specific therapies for pain. This may lead to overuse of pharmacological remedies, including opioids, for underserved patients, and contribute to well-documented disparities in access to knee arthroplasty.[16,34,35]

Consistent with previous literature, underserved patients are less likely to receive knee surgery in our data: Black patients have 0.78 lower odds (95% CI, 0.64–0.96), 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).

Pain disparities, particularly those remaining after adjustment for standard radiographic severity, may contribute to these trends. Patients with greater radiographic severity are empirically more likely to receive arthroplasty[34] (although formal arthroplasty guidelines simply require *presence* of radiographic damage).[33] Arthroplasty removes tissue objectively affected by degenerative disease, and thereby relieves pain (though no trials have specifically demonstrated that benefit varies by radiographic appearance).[16,36] As a result, “the majority of total knee replacements are performed in patients with end-stage knee osteoarthritis.”[17]

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 since radiographic osteoarthritis severity partially determines the decision to offer surgery (along with pain, function and quality of life) they may not be offered it. Since these patients—with severe pain and high ALG-P, but lower osteoarthritis severity (KLG)—were more likely to be Black, limitations of standard measures may 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 a pool of patients who were likely under most active consideration for arthroplasty.[34] This group of patients was then compared to patients based on an alternative eligibility rule: severe pain and severe ALG-P. The latter was defined to include the same number of patients as had KLG≥3 (Methods).

**Table 3** illustrates the differences between the existing and simulated guideline. Measuring severity with ALG-P rather than KLG would double potential eligibility for arthroplasty for Black patients, increasing it from 11% to 22% of knees ($p < 0.001$); it would also decrease the fraction of knees in severe pain and not eligible 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-standard-deviation worsening in ALG-P*;* $p = 0.008$). Since arthroplasty is known to reduce pain, this reallocation of surgery could potentially narrow

the racial and socioeconomic pain disparities as well as reduce the use of opioids for those in severe pain.[37]

*Conclusion*

This study has limitations. While it enrolled a diverse patient group from sites across the country, it should be validated in independent populations. This would also serve as a check on overfitting, which was minimized by creating a separate validation set prior to beginning any analysis (**Extended Data Figure 1**). 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 impossible to assess how using algorithmic pain predictions as a decision aid would affect patient outcomes. Finally, a central question we were not able to address is what features of the knee the algorithm is using. Beyond our specific study, this is generally difficult to determine with neural networks, and fully explaining the signal that algorithms are finding remains a pressing topic for future work if algorithms are to be responsibly deployed in medical decision-making.[38,39] Caution is warranted because, while ALG-P explains significantly more of the variance in pain than does KLG, the variance explained by both methods is low. The low variance explained does not prevent us from studying disparities between racial or socioeconomic groups, since this 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 which accommodate three dimensional data to make predictions from

MRI, or which combine images from multiple timepoints to leverage the longitudinal nature of the dataset.[40,41]

In summary, we used a machine learning algorithm to show that standard radiographic measures of severity overlook objective but undiagnosed features that disproportionately affect underserved populations. Since radiographic severity is a key input to management decisions, we estimate that new algorithmic measures could expand access to treatments for underserved patients. One promising option for integrating our algorithm into clinical practice is to use it as a decision aid, rather than as a replacement, for a human clinician: for example, by showing the clinician a heatmap **(Figure 1)** and algorithmic severity score. Such cooperation between humans and algorithms has been shown to improve clinical decision-making in some settings,[42] although it also presents challenges: for example, humans do not always place appropriate levels of weight on algorithmic predictions.[43,44] More broadly, our results illustrate how algorithms can be used to explain and reduce disparities in healthcare.

## Tables

**Table 1:** Dataset summary statistics.

| | Training/Development | Validation |
|---|---|---|
| **Sample size** | | |
| # individuals | 2,877 | 1,295 |
| # observations | 25,049 | 11,320 |
| **Demographics** | | |
| Black | 17% | 16% |
| Lower-income (<$50,000/year) | 38% | 39% |
| Non-college graduates | 39% | 38% |
| Female | 58% | 56% |
| Mean age, baseline visit (sd) | 61.1 (9.2) | 61.0 (9.1) |
| Mean BMI, baseline visit (sd) | 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/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/year) | 44% | 43% |
| Non-college graduates | 46% | 45% |

**Table 2:** Explaining 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,*[**]) |

**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%) |

[†]Racial and socioeconomic differences in KOOS pain score. "No severity measures" indicates the difference in mean pain scores without controlling for any severity measures; "KLG" column reports differences in pain score after controlling for KLG; and "ALG-P" column reports differences after controlling for the algorithmic pain score ALG-P rather than KLG. In parentheses are 95% confidence intervals computed by cluster bootstrapping at the patient level.

[‡]The "KLG" column reports how much differences in pain score are reduced by controlling for KLG; the "ALG-P" column reports the same reduction, but after controlling for algorithmic pain score (ALG-P) rather than KLG.

[§]Ratio of reductions in disparities by ALG-P vs. KLG.

[**]The upper limit of the confidence interval is not defined because the confidence interval for the denominator includes zero.

## Figures

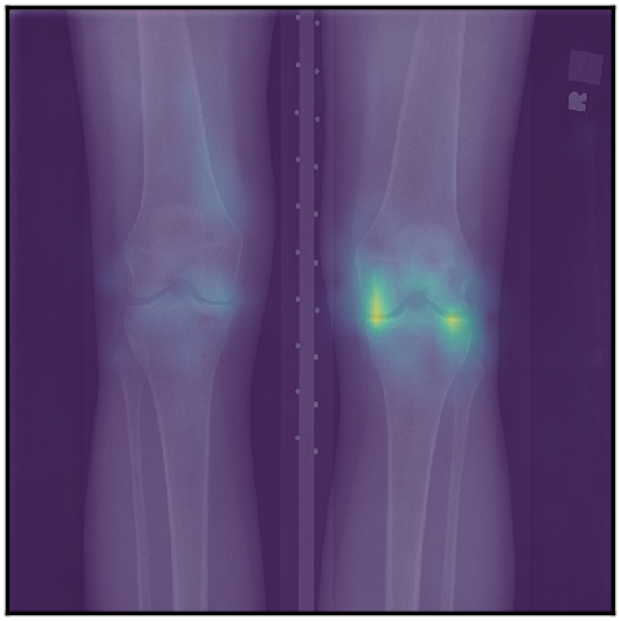


**Figure 1: Heatmap illustrating which regions of a representative x-ray image most strongly influenced the algorithmic pain prediction.** Regions which more strongly influence the prediction are shown in brighter colors. The model's predictions correlate most strongly to the femorotibial joint space and surroundings, consistent with clinical findings and with models previously trained to predict KLG.[45] Consistent with the fact that the model's prediction target is the pain score in the knee appearing on the right, that region of the image most strongly influences the prediction, though it is worth noting that there is some signal in the contralateral knee as well.

## Methods

*Dataset*

Clinical and radiological data were employed from the Osteoarthritis Initiative (OAI), a multicenter, longitudinal study of participants aged 45–79 who had, or were at high risk of developing, knee osteoarthritis.[27] 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, 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.

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 root mean square error (RMSE) for pain score; and a blinded validation (hold-out) set (approximately ⅓ of patients), in which no statistical analyses were performed until the model training procedure, including all hyperparameters, had been finalized. (**Figure S1** summarizes the analysis pipeline.) We confirmed that all statistics reported in **Table 1** were balanced between the train/development and validation (hold-out) set (all *p*-values for differences > 0.05). All results are reported on the validation (hold-out) set. All

confidence intervals and *p*-values are computed clustering at the patient level, to account for repeated observations from each patient. All *p*-values are two-sided.

*Radiological images and pre-processing*

Bilateral fixed flexion knee x-rays were used in the analysis, and pre-processed using standard methods (e.g., as in Rajpurkar et al.).[46] Each image was normalized by first dividing pixel values by the maximum pixel value (so all pixel values were in the range 0–1) and then z-scoring (subtracting the mean and dividing by the standard deviation across all pixels). Using alternate image normalization methods (z-scoring each image individually or z-scoring using the mean and standard deviation of the ImageNet dataset the neural network was originally trained on) did not substantially affect performance. Images were downsampled to 1024 x 1024 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: summary measures of severity (e.g., KLG), and other features (e.g., osteophytes and joint space narrowing).[25,31,47]

KLG, a standard measure of osteoarthritis severity, is a 5-level categorical variable (0 to 4), with increasing grades indicating increasing disease severity.[25,47] KLG≥2 is used as a standard threshold for radiological osteoarthritis.[27] Besides KLG, 18 other radiographic features—which quantify osteophytes, joint space narrowing, 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 analysis because Project 37 assessed only a non-random subset of participants. The OAI only assessed these additional 18 radiographic features (besides joint space narrowing, which was assessed in all participants) for participants who developed radiographic osteoarthritis in at least one knee (KLG $\geq 2$) 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 those features. To ensure that results were not specific to using KLG, sensitivity analysis was performed using OARSI joint space narrowing.[31] Knee MRIs were also collected for a subset of patients and scored using the MOAKS method,[32] which we used for another, similar sensitivity analysis in this subset.

KOOS pain score was used as a measure of self-reported pain.[28] 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 (e.g., "straightening the knee fully"); as usual, responses to each survey question were aggregated into a single score.[28]

*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 which appeared on

the right of the image, we flipped the original image horizontally when necessary (that is, when the target knee appeared on the left 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.

In order to give the network additional information about each image, and guide it towards 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:

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

where $Y$ is the KOOS pain score, $C^{(j)}$ is the z-scored $j^{\text{th}}$ image feature, and $\lambda$ is a weight chosen by hyperparameter search. (Because the primary objective was to predict KOOS pain score, 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 which 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 pre-trained on ImageNet.[48,49] Deeper layers of the architecture were then fine-tuned on the OAI dataset.[50] The training dataset

was augmented by applying random horizontal and vertical translations to each image.[51] Adam was used to optimize network weights, with an initial learning rate that decayed by a factor of 2 each time the loss plateaued.[52] 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.[53] 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.[54] 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.[55] Training was performed on four Nvidia XP GPUs using PyTorch.[56]

*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 not graduated from college (38%). Differences in pain scores across groups were first quantified without controlling for osteoarthritis severity, using mean KOOS pain score between groups (e.g., racial pain disparity was defined as the difference in mean pain between Black and non-Black patients). **Supplementary Table 2** 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/socioeconomic group, and a measure of osteoarthritis severity (see below for specifics). The pain disparity was defined as the coefficient on binary race/socioeconomic group: that is, the gap in mean pain between racial/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 (algorithmic pain prediction): 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: we predicted 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.[57,58] Conceptually, the output of this regression model (which was generated in the hold-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 the 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 the paper yields estimates of pain disparities which are essentially identical to those produced by the alternate procedure. The income pain disparity estimates differed by 0.2% (3.529 vs. 3.524); the race pain disparity estimates differed by 0.6% (9.664 vs. 9.718); and the education pain disparity estimates differed by 0.3% (4.879 vs. 4.895).

Because our analysis performs a regression of pain on severity score and binary racial/socioeconomic group, it implicitly fits a model where 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 where 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 way of accounting fully for any differences across racial/socioeconomic groups in the relationship between severity score and pain. Our results remain extremely similar under

this alternate definition of the pain gap: our estimation of the pain gap changes by less than 5% in all cases (for both severity scores and all three racial/socioeconomic groups).

*Comparing predictive power of ALG-P to 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 which KLG did not capture. The Pearson $R^2$ for ALG-P was 0.16 (compared to 0.10 for KLG) (**Table S1**). 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 non-significant ($p = 0.20$). 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-standard-deviation worsening in ALG-P corresponded to 1.5x 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 of ≥ 2 had 1.7x (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 2.9x (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-standard-deviation change in ALG-P predicted a 0.07-standard-deviation worsening in KLG at follow-up (95% CI, 0.06–0.08).

*Visualizing image regions that influenced predictions*

To compute how much a region of the image influenced the neural network's predicted pain score, the region was "masked" out, by replacing it with a circle whose value was the mean pixel value for the image, using Gaussian smoothing to prevent sharp boundaries.[59] 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 x 32 grid of regions evenly tiling the 1024 x 1024 image, allowing computation of a heat map of how much masking each region of the image affected the neural network's prediction (**Figure 1)**. As an additional robustness check, Class Activation Mapping (CAM) was used and similarly indicated that the neural network's prediction was, as expected, primarily influenced by the knee which appeared on the right of the image, although it was also somewhat influenced by the contralateral knee.[60] (Because the predicted output variable was continuous, for CAM 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 $\leq$ 86.1) and severe osteoarthritis (KLG $\geq$ 3).[29,30,34] A different guideline was then simulated, where 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 0-4 scale as 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 $\leq$ 86.1) and severe osteoarthritis (categorical ALG-P $\geq$ 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 which had already had any knee surgery, and only knees at baseline were considered; neither of these decisions substantially altered results.

## References


1. Zhang Y, Jordan JM. Epidemiology of Osteoarthritis. *Clin Geriatr Med*. 2010;26(3):355-369.
2. Felson DT, Zhang Y. An update on the epidemiology of knee and hip osteoarthritis with a view to prevention. *Arthritis Rheum*. 1998;41(8):1343-1355.
3. Allen KD, Oddone EZ, Coffman CJ, Keefe FJ, Lindquist JH, Bosworth HB. Racial differences in osteoarthritis pain and function: potential explanatory factors. *Osteoarthritis Cartilage*. 2010;18(2):160-167.
4. Eberly L, Richter D, Comerci G, et al. Psychosocial and demographic factors influencing pain scores of patients with knee osteoarthritis. *PloS One*. 2018;13(4):e0195075.
5. Golightly YM, Dominick KL. Racial variations in self-reported osteoarthritis symptom severity among veterans. *Aging Clin Exp Res*. 2005;17(4):264-269.
6. Bolen J, Schieb L, Hootman JM, 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*. 2010;7(3):A64.
7. Centers for Disease Control and Prevention (CDC). Racial/ethnic differences in the prevalence and impact of doctor-diagnosed arthritis--United States, 2002. *MMWR Morb Mortal Wkly Rep*. 2005;54(5):119-123.
8. Allen KD, Helmick CG, Schwartz TA, DeVellis RF, Renner JB, Jordan JM. Racial differences in self-reported pain and function among individuals with radiographic hip and knee osteoarthritis: the Johnston County Osteoarthritis Project. *Osteoarthritis Cartilage*. 2009;17(9):1132-1136.
9. Lachance L, Sowers M, Jamadar D, Jannausch M, Hochberg M, Crutchfield M. The experience of pain and emergent osteoarthritis of the knee. *Osteoarthritis Cartilage*. 2001;9(6):527-532.
10. Collins JE, Katz JN, Dervan EE, Losina E. Trajectories and risk profiles of pain in persons with radiographic, symptomatic knee osteoarthritis: data from the osteoarthritis initiative. *Osteoarthritis Cartilage*. 2014;22(5):622-630.
11. Poleshuck EL, Green CR. Socioeconomic disadvantage and pain. *Pain*. 2008;136(3):235-238.
12. Anderson KO, Green CR, Payne R. Racial and ethnic disparities in pain: causes and consequences of unequal care. *J Pain Off J Am Pain Soc*. 2009;10(12):1187-1204.
13. Cerrada CJ, Tai C, Kumar S, et al. African American Participants Experience Greater Pain Severity and Pain Interference Compared to Non-Hispanic Whites in a Large Scale Virtual Study on Chronic Pain. *Med Decis Making*. 2020;40(1):E317.
14. Krause N, Ragland DR, Greiner BA, et al. Psychosocial job factors associated with back and neck pain in public transit operators. *Scand J Work Env Health*. 1997;23(3):179-186.
15. Gatchel RJ, Polatin PB, Mayer TG. The dominant role of psychosocial risk factors in the development of chronic low back pain disability. *Spine*. 1995;20(24):2702-2709.
16. Deveza LA, Bennell K. Management of knee osteoarthritis. UpToDate. Published 2019. https://www.uptodate.com/contents/management-of-knee-osteoarthritis
17. Losina E, Thornhill TS, Rome BN, Wright J, Katz JN. 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*. 2012;94(3):201-207.
18. Hochberg MC, Guermazi A, Guehring H, et al. Effect of Intra-Articular Sprifermin vs Placebo on Femorotibial Joint Cartilage Thickness in Patients With Osteoarthritis: The FORWARD Randomized Clinical Trial. *JAMA*. 2019;322(14):1360-1370.
19. Vina ER, Ran D, Ashbeck EL, Kwoh CK. Natural History of Pain and Disability among African-Americans and Whites With or At Risk For Knee Osteoarthritis: A Longitudinal Study. *Osteoarthritis Cartilage*.

2018;26(4):471-479.
20. Neogi T, Felson D, Niu J, et al. Association between radiographic features of knee osteoarthritis and pain: results from two cohort studies. *BMJ*. 2009;339(1):b2844-b2844.
21. Bedson J, Croft PR. The discordance between clinical and radiographic knee osteoarthritis: A systematic search and summary of the literature. *BMC Musculoskelet Disord*. 2008;9(1):116.
22. Sayre EC, Guermazi A, Esdaile JM, et al. Associations between MRI features versus knee pain severity and progression: Data from the Vancouver Longitudinal Study of Early Knee Osteoarthritis. Lammi MJ, ed. *PLOS ONE*. 2017;12(5):e0176833.
23. Link TM. Correlations between joint morphology and pain and between magnetic resonance imaging, histology, and micro-computed tomography. *J Bone Joint Surg Am*. 2009;91 Suppl 1:30-32.
24. Culvenor AG, Øiestad BE, Hart HF, Stefanik JJ, Guermazi A, Crossley KM. Prevalence of knee osteoarthritis features on magnetic resonance imaging in asymptomatic uninjured adults: a systematic review and meta-analysis. *Br J Sports Med*. 2019;53(20):1268-1278.
25. Kellgren JH, Lawrence JS. Radiological Assessment of Osteo-Arthrosis. *Ann Rheum Dis*. 1957;16(4):494-502.
26. Haug W, Compton P, Courbage Y. *The Demographic Characteristics of Immigrant Populations*. Vol 38. Council of Europe; 2002.
27. Nevitt MC, Felson DT, Lester G. The osteoarthritis initiative. Published online 2006.
28. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. *J Orthop Sports Phys Ther*. 1998;28(2):88-96.
29. Englund M, Roos EM, Lohmander LS. Impact of type of meniscal tear on radiographic and symptomatic knee osteoarthritis: A sixteen-year followup of meniscectomy with matched controls. *Arthritis Rheum*. 2003;48(8):2178-2187.
30. Wasserstein D, Huston LJ, Nwosu S, Spindler KP. KOOS pain as a marker for significant knee pain two and six years after primary ACL reconstruction: A Multicenter Orthopaedic Outcomes Network (MOON) prospective longitudinal cohort study. *Osteoarthritis Cartilage*. 2015;23(10):1674-1684.
31. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis Cartilage*. 2007;15 Suppl A:A1-56.
32. Hunter DJ, Guermazi A, Lo GH, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). *Osteoarthritis Cartilage*. 2011;19(8):990-1002.
33. Rankin EA, Alarcon GS, Chang RW, Cooney Jr LM. NIH Consensus Statement on total knee replacement December 8-10, 2003. *J Bone Jt Surg*. 2004;86(6):1328.
34. Losina E, Paltiel AD, Weinstein AM, et al. Lifetime medical costs of knee osteoarthritis management in the United States: impact of extending indications for total knee arthroplasty. *Arthritis Care Res*. 2015;67(2):203-215.
35. Skinner J, Weinstein JN, Sporer SM, Wennberg JE. Racial, Ethnic, and Geographic Disparities in Rates of Knee Arthroplasty among Medicare Patients. *N Engl J Med*. 2003;349(14):1350-1359.
36. Lingard EA, Riddle DL. Impact of Psychological Distress on Pain and Function Following Knee Arthroplasty. *J Bone Jt Surg*. 2007;89(6):1161-1169.
37. Riddle DL, Perera RA, Jiranek WA, Dumenci L. Using Surgical Appropriateness Criteria to Examine Outcomes of Total Knee Arthroplasty in a United States Sample. *Arthritis Care Res*. 2015;67(3):349-357.
38. Olah C, Satyanarayan A, Johnson I, et al. The Building Blocks of Interpretability. *Distill*. Published online 2018.
39. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9):1337-1340.
40. Xu Y, Hosny A, Zeleznik R, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin Cancer Res*. 2019;25(11):3266-3275. doi:10.1158/1078-0432.CCR-18-2495
41. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. Saria S, ed. *PLOS Med*. 2018;15(11):e1002699. doi:10.1371/journal.pmed.1002699
42. Steiner DF, MacDonald R, Liu Y, et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer: *Am J Surg Pathol*. 2018;42(12):1636-1646.
43. Yin M, Wortman Vaughan J, Wallach H. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press; 2019:1-12.
44. Uyumazturk B, Kiani A, Rajpurkar P, et al. Deep Learning for the Digital Pathologic Diagnosis of Cholangiocarcinoma and Hepatocellular Carcinoma: Evaluating the Impact of a Web-based Diagnostic

Assistant. *Mach Learn Health ML4H NeurIPS 2019 - Ext Abstr*.
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*. 2018;8(1):1-10.
46. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv*. Published online 2017:1711.052253-9.
47. Kohn MD, Sassoon AA, Fernando ND. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. *Clin Orthop Relat Res*. 2016;474(8):1886-1893.
48. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *Proc IEEE Conf Comput Vis Pattern Recognit*. Published online 2016:770-778.
49. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. *Proc IEEE Conf Comput Vis Pattern Recognit*. Published online 2009:248-255.
50. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Trans Med Imaging*. 2016;35(5):1299-1312.
51. Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv*. Published online 2017:1712.04621.
52. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *Int Conf Learn Represent*. Published online 2014.
53. Caruana R, Lawrence S, Giles CL. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. *Conf Neural Inf Process Syst*. 2001;13:402-408.
54. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. *J Mach Learn Res*. 2013;13(1):281-306.
55. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316(22):2402-2410.
56. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. *Conf Neural Inf Process Syst*. Published online 2017.
57. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288.
58. McNeish DM. Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivar Behav Res*. 2015;50(5):471-484.
59. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. *Eur Conf Comput Vis*. 2014;8689:818-833.
60. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. *IEEE Conf Comput Vis Pattern Recognit*. Published online 2016:2921-2929.

## Acknowledgments

The authors thank Katy Blumer, Jeremy Irvin, Pang Wei Koh, Shengwu Li, Katie Lin, Bryan McCann, Andrew Miller, Leah Pierson, Chris Olah, Maithra Raghu, Pranav Rajpurkar, Nat Roth, Camilo Ruiz, Chiara Sabatti, and participants at several seminars and meetings for helpful comments. E.P. was supported by Hertz and NDSEG graduate fellowships.

## Author contributions

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

## Additional information

Correspondence should be addressed to S.M. (sendhil.mullainathan@chicagobooth.edu).
Supplementary information is available for this paper.

## Code availability

Code will be made publicly available on GitHub upon publication of this manuscript.

## Data availability

Data are available online at https://nda.nih.gov/oai/.

**Supplement to**

# An Algorithmic Approach to Explaining Why the Disadvantaged Feel More Pain

**Emma Pierson**, PhD; Department of Computer Science, Stanford University and Microsoft Research
**David M. Cutler***, PhD; Department of Economics, Harvard University
**Jure Leskovec***, PhD; Department of Computer Science, Stanford University
**Sendhil Mullainathan***, PhD; Booth School of Business, University of Chicago
**Ziad Obermeyer***, MD; School of Public Health, University of California at Berkeley

*alphabetical order

# 1 Supplementary methods

## 1.1 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 has been the subject of substantial research, allowing validation of the pipeline used in this analysis in comparison to previous work.[1,2] Predictive performance on this task was comparable to previous work using models specifically designed to predict KLG (MSE: 0.35 as compared to 0.48 and 0.50 in previous work; $R^2$: 0.87).[1,2] This indicates that the model is able to extract clinically relevant signal from the image even on a task it had not been originally designed to perform.

## 1.2 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 joint space narrowing (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, a standard procedure.[3,4] 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), whose prediction performance was comparable to KLG's ($R^2$: 0.10). ALG-P also achieved greater reductions in racial and socioeconomic pain disparities than did JSN: a 3.9x greater reduction in the education pain

disparity (30% versus 8%), a 2.1x greater reduction in the income pain disparity (32% versus 16%), and a 7.7x 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.[5] 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.[6] This resulted in 10 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 the 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%).

### 1.3 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 0-4 scale as 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 4.5x more than did KLG (similar to the original 4.7x for the continuous ALG-P), the education pain disparity by 3.4x more than KLG (similar to the 3.6x for continuous ALG-P), and the income pain disparity by 1.9x more than KLG (similar to the 2.0x 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.

### 1.4 ALG-P is not just 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 sclerosis. To test this, correlations of ALG-P with 19 radiographic features (KLG and an additional 18 radiographic features relevant to osteoarthritis, e.g., osteophytes, joint space narrowing, and sclerosis, as described in the main Methods) 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), then compared to the coefficient on ALG-P when we added variables controlling for known radiographic features (0.95, 95% CI 0.80–1.10). The fact that the coefficient does 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.

### 1.5 ALG-P is not simply learning to predict race/socioeconomic status

ALG-P could be narrowing disparities in pain by simply learning how to predict race or socioeconomic status from the knee image. Since 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/socioeconomic variables, and 0.83 (95% CI, 0.74–0.93) when controlling for all three binary race/socioeconomic variables. Thus, the coefficient on ALG-P remained highly statistically significant, and similar in magnitude, when controlling for race/socioeconomic status. We also verified that ALG-P achieved better predictive performance for pain than did KLG across all six race/socioeconomic groups in our analysis (Black/non-Black, higher/lower income, and higher/lower education).

### 1.6 Predictions are not driven merely by image artifacts

The model could be gaining predictive power from image artifacts, e.g., related to the study site in which patients were recruited.[7] 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 (i.e., on the right side of the image), and by regions of the knee which were clinically relevant and consistent with previous work.[1,8] In the heatmap, warmer colors indicate regions of the image which influence the neural

network's predictions more strongly, as described in the main Methods.

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 affected knee was left or right; 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—likely detectable from knee radiographs, and known to be correlated with pain.[9] Hence, we further confirmed that our predictive power was not just 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 $R^2$ than KLG on each BMI group.

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

### 1.7 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 they have not been trained on.[7] 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 $R^2$ 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 $R^2$ 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).

## 2 Supplementary tables

**Supplementary Table 1:** Predictive performance for pain.

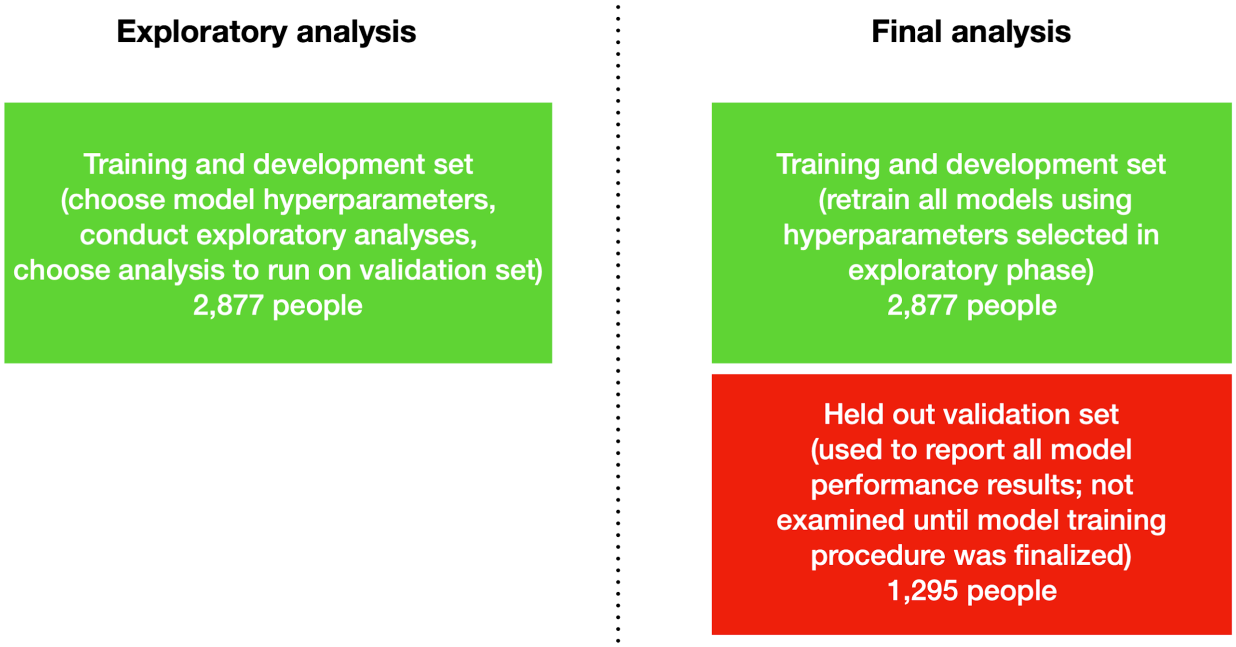
| Predictive performance for pain | KLG | ALG-P |
|---|---|---|
| *Pearson $R^2$* | 0.10 (0.08, 0.13)†† | 0.16 (0.13, 0.19) |
| *Spearman $R^2$* | 0.08 (0.07, 0.11) | 0.14 (0.11, 0.16) |
| *RMSE* | 15.4 (14.7, 16.0) | 14.9 (14.3, 15.4) |
| *Mean absolute error* | 11.9 (11.5, 12.2) | 11.3 (10.9, 11.7)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

**Supplementary Table 2:** 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.

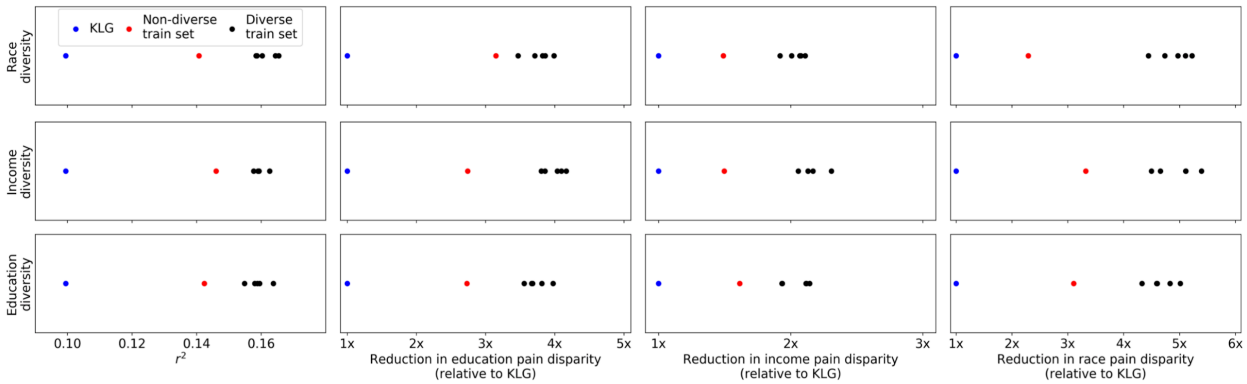
| Black | Lower-income | Lower-education | KOOS pain score |
|---|---|---|---|
| No | No | No | 89.8 (89.0, 90.6) |
| No | No | Yes | 86.9 (85.1, 88.7) |
| No | Yes | No | 89.2 (87.6, 90.8) |
| No | Yes | Yes | 85.5 (83.7, 87.2) |
| Yes | No | No | 81.5 (77.8, 85.2) |
| Yes | No | Yes | 81.1 (74.9, 87.3)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

††95% CIs are computed by cluster bootstrapping at the patient level.

## 3 Extended data figures




**Extended Data Figure 1: The pipeline for analysis, which was conducted in two phases**. Prior to conducting any analysis, 1,295 patients (red box) were reserved as a hold-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 second phase, 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.




**Extended Data Figure 2: 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.

## 4 Supplementary references


1. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Sci Rep*. 2018;8(1):1-10.
2. Antony J, McGuinness K, O'Connor NE, Moran K. Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks. *Int Conf Pattern Recognit*. Published online 2016:1195-1200.
3. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis Cartilage*. 2007;15 Suppl A:A1-56.
4. Sheehy L, Culham E, McLean 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). *Osteoarthritis Cartilage*. 2015;23(9):1491-1498.
5. Hunter DJ, Guermazi A, Lo GH, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). *Osteoarthritis Cartilage*. 2011;19(8):990-1002.
6. Cutler DM, Meara ER, Stewart ST. Socioeconomic Status and Perceptions of Pain. *Mimeo*. Published online 2019.
7. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. *PLOS Med*. 2019;15(11):e1002683.
8. Kohn MD, Sassoon AA, Fernando ND. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. *Clin Orthop Relat Res*. 2016;474(8):1886-1893.
9. Rogers MW, Wilder FV. The association of BMI and knee pain among persons with radiographic knee osteoarthritis: A cross-sectional study. *BMC Musculoskelet Disord*. 2008;9(1):163.