

1 Deep learning based phenotyping of medical images improves power 2 for gene discovery of complex disease

3
4 Brianna I. Flynn¹⁺, Emily M. Javan¹⁺, Eugenia Lin², Zoe Trutner², Karl Koenig², Kenoma O.
5 Anighoro², Eucharist Kun¹, Alaukik Gupta^{1,6}, Tarjinder Singh^{3,4,5}, Prakash Jayakumar², and
6 Vagheesh M. Narasimhan^{1,7}

7
8 ¹Department of Integrative Biology, The University of Texas at Austin

9 ²Department of Surgery and Perioperative Care, Dell Medical School

10 ³The Department of Psychiatry at Columbia University Irving Medical Center

11 ⁴The New York Genome Center

12 ⁵The Mortimer B. Zuckerman Mind Brain Behavior Institute

13 ⁶Department of Biomedical Engineering, The University of Texas at Austin

14 ⁷Department of Statistics and Data Science, The University of Texas at Austin

15 ⁺ Co-first authors

16 Abstract

17 Electronic health records (EHRs) are often incomplete and inaccurate, reducing the power of
18 genome-wide association studies (GWAS). Moreover, the variables within these records are
19 often represented in binary codes, masking variation in disease severity among individuals. For
20 some diseases, such as knee osteoarthritis (OA), radiographic assessment is the primary means of
21 diagnosis and can be performed directly from medical images. In this work, we trained a deep
22 learning model (DL-binary) to ascertain knee OA cases from anteroposterior (AP) dual-energy
23 absorptiometry (DXA) scans and achieved clinician level performance. Applying this model
24 across 29,257 individuals from the UK Biobank (UKB), we identified 2,603 (240%) more cases
25 than currently diagnosed in the ICD-10 record. Individuals diagnosed as cases by DL-binary had
26 higher rates of self-reported knee pain, knee pain for longer durations and with increased severity
27 compared to control individuals. We trained another deep learning model to measure the
28 minimum knee joint space width (mJSW), a quantitative phenotype linked to knee OA severity.
29 Despite the DL-binary phenotype and mJSW being highly genetically correlated (92%), the
30 heritability of mJSW was an order of magnitude greater than the ICD-10 code M17 or DL-binary
31 phenotypes. In a GWAS run on mJSW, we identified 18 genome-wide significant loci, as
32 opposed to 1 and 6 at the same sample size using either case-control (DL-binary and ICD-10
33 code M17) phenotype. This improved power also translated to better polygenic risk score (PRS)
34 prediction for knee OA diagnosis in a holdout dataset of 371,686 individuals. We also show that
35 reduced mJSW, but neither case-control phenotype is associated with increased risk of adult
36 fractures, a leading cause of injury-related death in older individuals. For diseases with
37 radiographic diagnosis, our results demonstrate the enormous potential for using deep learning to
38 phenotype at biobank scale, both for improving power for gene discovery and for
39 epidemiological analysis.

40 Introduction

41 For most complex disease traits, clinical endpoints are usually binary (case-control) in
42 nature. In particular, data on disease outcomes from population scale biobanks are only available
43 through recorded ICD-10 billing codes or self-reported diagnosis¹⁻³. While these datasets have
44 provided invaluable insights into the genetic basis of disease, case ascertainment based solely on
45 information available in the EHR or from self-reports can be biased by a multitude of factors
46 including differences in how patients were billed⁴, differential diagnosis due to assessment by
47 clinicians (non-specialist vs specialist)⁵, or differences in classification or diagnosis based on
48 disease severity⁶.

49
50 An alternate approach to ascertaining disease status might be to directly perform clinical-
51 grade assessment from a patient's medical images using a consistent diagnosis protocol.
52 However, this is difficult to achieve at biobank scale where sample sizes can range from
53 hundreds of thousands if not millions of individuals¹. Importantly, for musculoskeletal diseases
54 such as knee OA, radiography is the routine course of diagnosis in the clinic as well as to assess
55 important markers associated with disease progression such as sclerosis, osteophytosis (bone
56 spurs) and narrowing of the space between the femur and tibia, also known as the knee joint
57 space⁷. For such radiographically diagnosed diseases, computer vision approaches for automated
58 phenotyping based on training data from clinicians offer the potential to ascertain both case
59 status and disease severity at scale. Such approaches have already been used for determining
60 pneumonia and SARS-CoV-2 cases from chest X-ray images, with reported accuracy even
61 higher than expert radiologists based on ground truth from molecular information^{8,9}.

62
63 Taking advantage of these developments in computer vision, recent genetic studies have
64 successfully applied these methods to generate image derived phenotypes (IDPs) of distribution
65 of body fat, heart structure, liver fat percentage, and brain morphology, and have linked these
66 novel traits with genome-wide significant loci¹⁰⁻¹⁴. While some recent studies on
67 musculoskeletal disease employ these novel phenotyping approaches¹⁵⁻¹⁷, neither these nor the
68 studies on other traits have specifically investigated how generating quantitative IDPs that
69 underlie binary disease status could be used to improve power for gene discovery at biobank
70 scale.

71
72 Quantitative measurements which provide information about variation in the severity of
73 progression of the disease are already routinely utilized in predicting an individual's risk for
74 complex disease in the clinic. For example, LDL cholesterol levels are a quantitative biomarker
75 measured in blood samples, and are used as a primary biomarker to assess risk for myocardial
76 infarction, among the leading causes of death worldwide¹⁸. Multiple lines of functional evidence
77 suggest that LDL cholesterol levels are also causally linked to heart disease and lowering LDL
78 levels over an entire lifetime through the use of statins is the most widely used long term
79 prescription medication¹⁹. In theoretical work, it has been demonstrated that with equal sample
80 size and when the proportion of cases in a case-control design is equivalent to the prevalence of
81 the disease in the population, the power of a case-control association study is considerably lower
82 than that of a quantitative association study. This is in part because key information about
83 variation in the trait in the sample population is lost when transforming a continuous trait into a
84 binary one²⁰.

85 Building on these foundational ideas, in this work we first trained a binary classification
86 model to identify knee OA cases at clinical level performance and deployed this at biobank scale
87 to compare our radiographically obtained results to the ICD-10 record. Second, we trained an
88 image segmentation algorithm to obtain a quantitative measurement highly correlated with knee
89 OA severity, mJSW, to examine differences in power between GWAS carried out using
90 quantitative approaches versus a case-control design. Third, we generate PRS for each phenotype
91 to evaluate if improvements in statistical power to find novel loci translate to better prediction of
92 ICD-10 record knee OA (M17) in a hold-out dataset of over 300,000 individuals. Finally, we
93 examined epidemiological associations to link our IDPs with an outcome of major clinical
94 relevance.

95 Results

96 Dataset, and quality control of DXA imaging and genetic data

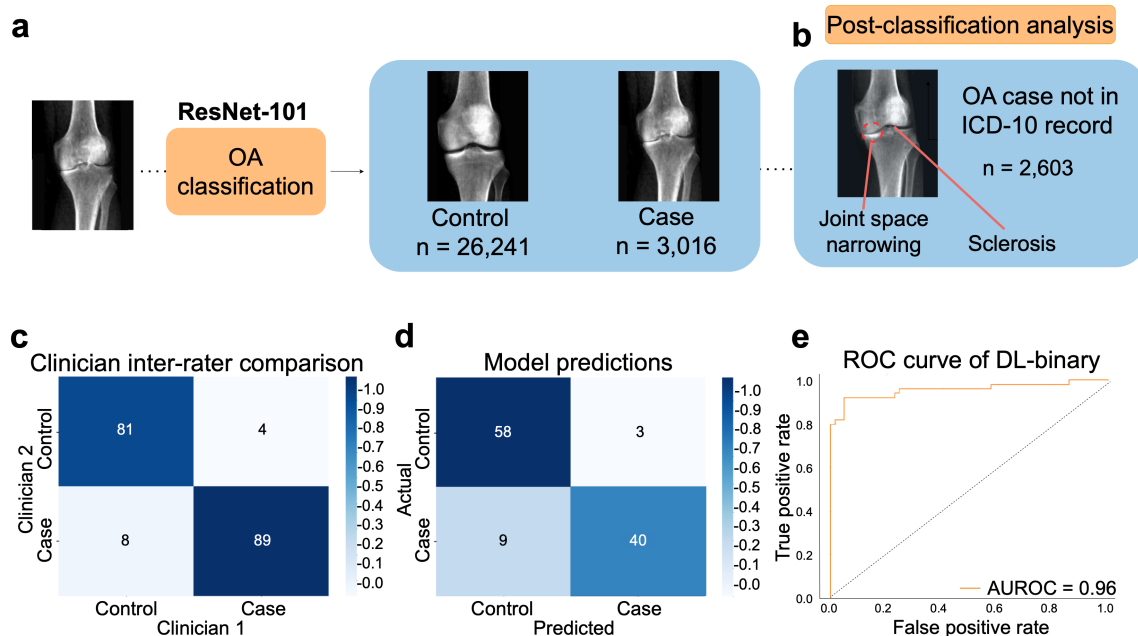
97 To study the genetic basis of knee phenotypes, we jointly analyzed paired DXA imaging
98 and imputed genome sequence data of 42,284 individuals in the UKB. We first restricted the
99 dataset to individuals of white British ancestry, applied standard variant and sample QC and
100 analyzed 12.1 million common bi-allelic SNPs with minor allele frequency $> 0.1\%$ ¹ (**Methods:**
101 Genetic QC). Next, as the bulk imaging data from the UKB comprised of DXA images that
102 reflect scans of different body parts, we used a deep learning approach¹⁵ to subset the imaging
103 dataset to only AP view knee scans. We then removed individuals that had outlier image
104 resolutions or poor quality DXA scans, and padded images to a standard size for processing (see
105 **Methods:** Image segmentation, phenotype measurement and quality control). Post quality-
106 control, we were left with combined imaging and genetic data for a total of 29,257 individuals
107 aged between 46 to 81 with a median age of 64, and a sex ratio of 0.99, consistent with the
108 overall distribution in the UKB (**Methods:** UKB participants and dataset).

109 Automated phenotyping of knee OA achieves clinician level performance

110 To perform automated phenotyping for knee OA based on radiography, we used a binary
111 classification approach based on the Kellgren-Lawrence (KL) grading system²¹ (usually graded
112 0-4, where a 4 is considered the most severe case of radiographic OA) to determine case or
113 control status for each individual reflecting different levels of joint space narrowing, subchondral
114 sclerosis, and the presence of osteophytes. Cases were considered individuals with a KL grade of
115 3 or higher - severe enough that annotating clinicians would consider a candidate for joint
116 replacement surgery in the clinic. Controls were considered individuals who would not be
117 candidates for joint replacement - a grade 2 or lower (**Methods:** Binary classification: DXA scan
118 annotation procedure). To train the deep learning model, we obtained case-control assessment on
119 546 images based on the annotations of three board-certified orthopedic surgeons who
120 independently assessed each image. We then split the dataset so that 80% (436 images) of the
121 data was used for training and 20% (110 images) was used for validation. We next trained a
122 binary classifier (which we refer to as DL-binary) using transfer learning with the ResNet-101
123 architecture²² (**Methods:** Binary classification: Network architecture and model training). The
124 sensitivity and specificity of our model on validation data (that is not used as part of the training

125 process) was within the range of the sensitivity and specificity obtained between two clinicians
 126 grading the same set of images (Clinician sensitivity: 0.92 ± 0.05 , DL-binary sensitivity: $0.82 \pm$
 127 0.07 Clinician specificity: 0.95 ± 0.05 , DL-binary specificity: 0.95 ± 0.06 .) (Fig. 1c, d).

128 **Fig. 1: A deep learning process for automated phenotyping of radiographic knee OA.**



129 **a** ResNet-101 based classifier for binary classification of knee OA, showing an example of a
 130 typical individual diagnosed as a case compared to a control individual. **b** Post-classification
 131 analysis using highlighting regions of the knee that are discriminatory for knee OA. We
 132 confirmed joint space narrowing and sclerosis of the bone (important features for case
 133 classification) are present in cases not reported in the ICD-10 record but identified by the model.
 134 **c** Inter-rater comparison of two clinicians grading a total of 200 AP view knee DXA scans split
 135 roughly equally between cases and controls. **d** Confusion matrix showing performance of the
 136 DL-binary model on validation data. **e** Receiver operating characteristic (ROC) curve for DL-
 137 binary, showing performance of the model under different classification thresholds.

139 **Image based phenotyping reveals twofold more cases compared with ICD-10**
 140 **records**

141 We next deployed our trained model on the remaining 28,725 images of knee DXA scans
 142 from the dataset. We considered an individual a 'case' if our model predicted the individual to
 143 have knee OA on either the left or the right knee, and a control otherwise in line with the ICD-10
 144 code M17 for knee OA. We then assessed how many cases were determined by our deep
 145 learning based binary classification of radiographic OA as compared to what already exists in the
 146 ICD-10 code M17. We found that after deploying the DL-binary classifier, we determined 2,603
 147 more cases compared with the ICD-10 code for knee OA (ICD-10 code M17 1,085 cases, DL-
 148 binary 3,016 cases) (Fig. 1a). To provide additional support for cases reported by DL-binary that
 149 were not already reported in the ICD-10 code M17, our clinical team examined 100 individuals

150 manually and confirmed the presence of osteophytes, reduced joint space and in some cases
151 subchondral sclerosis (**Fig. 1b**). As these alone may not be diagnostic, we also investigated
152 associations with three self-reported measures of knee pain in the UKB: knee pain experienced in
153 the past month (binary), knee pain for 3+ months (binary, and reflecting knee pain experienced
154 over a long duration) and rating of knee pain in the past three months (scale from 0 - 10). We
155 found that in individuals who were newly identified as cases, the rate of self-reported knee pain
156 was significantly higher compared to control individuals (individuals not diagnosed by ICD-10
157 code M17 or by DL-binary) across all three measures we examined (recent pain reported as knee
158 pain in the past month: 49.4% in cases and 27.2% in controls, chi-square statistic = 536.6, $p <$
159 2.2×10^{-16} , chronic pain as determined by knee pain lasting 3 or more months: 80.4% in cases
160 and 70.6% in controls, chi-square statistic = 28.29, $p = 1.5 \times 10^{-7}$ and severity of pain reported in
161 the last 3 months: mean rating of 3.33 in cases and 2.58 in controls, t-test $p = 1.45 \times 10^{-15}$). These
162 results suggest that knee OA is likely underdiagnosed in the ICD-10 record and that our
163 approach is capable of identifying additional true cases not already present in the EHR.

164 Image segmentation to measure joint space width

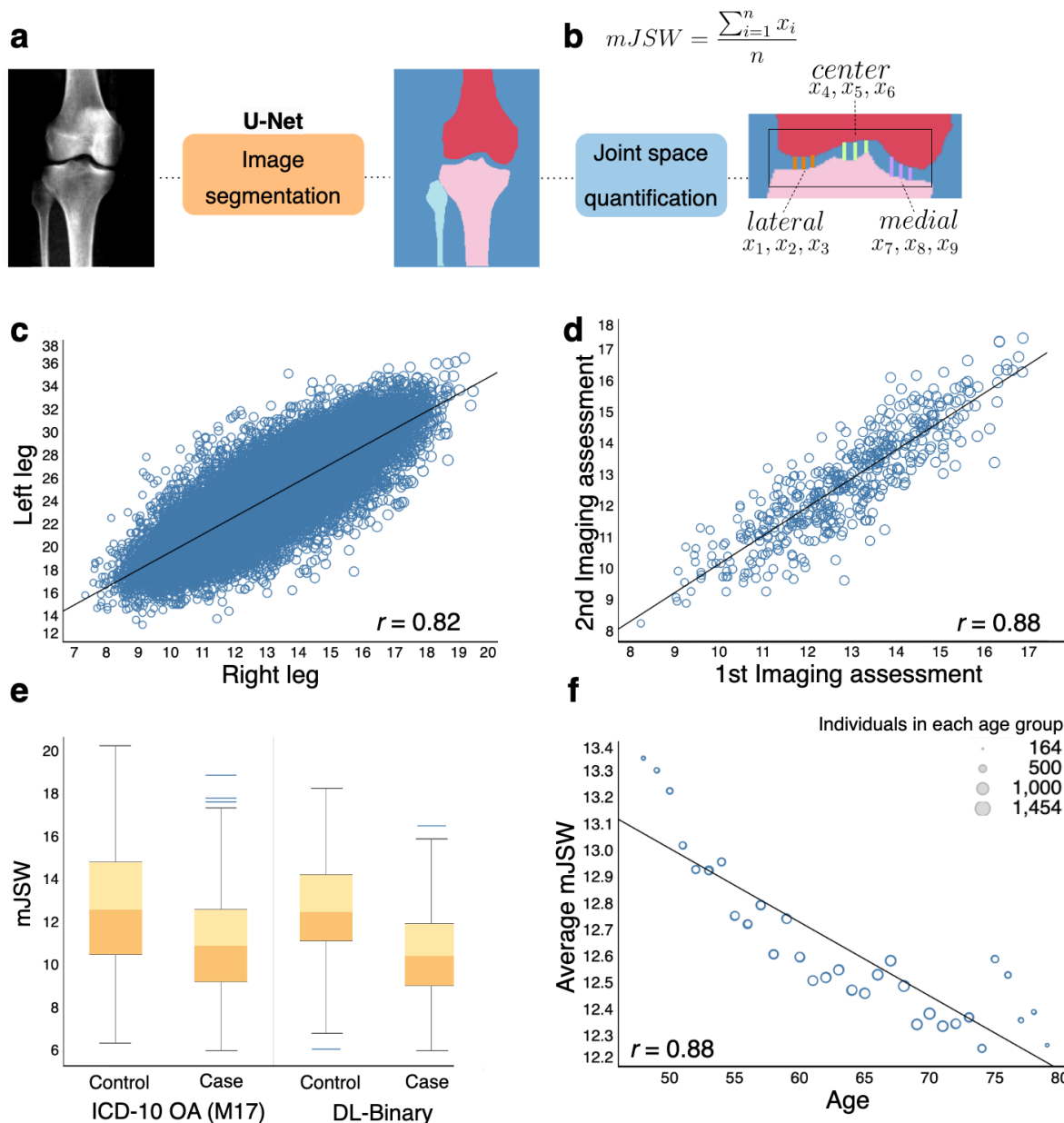
165 To examine knee OA severity beyond simple case-control assessment, we developed a
166 method to obtain a quantitative measurement from knee DXA scans known to be highly
167 associated with the disease: the minimum inter-bone joint space between the femur and tibia,
168 which we refer to as mJSW. To perform automated measurement on the UKB dataset, we first
169 collected training data for 63 DXA scan derived images of the knee (40 training, 23 validation).
170 On each of these images we labeled the positions of the femur, tibia and fibula at pixel level,
171 which were then validated by a team of clinicians. We then trained a deep learning model based
172 on the U-Net architecture²³ with a 34-layer ResNet encoder²² to perform semantic segmentation
173 of the femur, tibia and fibula in each DXA image at pixel-level resolution (**Fig. 2a**). After quality
174 control and image normalization (**Methods: DXA scan image quality control and**
175 **standardization**), we computed mJSW by measuring the distance between the femur and tibia
176 along multiple positions on the medial, lateral and center axes of the joint. We then computed the
177 average of these distances for each leg (**Fig. 2b**). The mJSW measurement is defined as the
178 smallest of the two averages for either leg, returning one phenotype measurement per individual.
179 If the individual only had a right or left leg DXA scan, this was used as the mJSW measurement
180 for that individual. To standardize mJSW measurements across image resolutions, we regressed
181 each of the joint space lengths on the overall height of the individual (**Methods: Image**
182 **segmentation: Measurement and quality control**).

183
184 We evaluated the performance of the segmentation model in several ways. First, the set
185 accuracy, the correspondence between labeled data and annotation of the trained model on
186 validation data, was 0.99. Second, the correlation between measurements taken between the right
187 and left leg was 0.82 (**Fig. 2c**). Third, the correlation between images taken of the same person
188 across two imaging visits was 0.88, despite changes in image resolution, scanner, technician and
189 imaging position, demonstrating that our mJSW measurement process is fairly consistent across
190 biological replicates (**Fig. 2d**). We do not expect to see 100% concordance across these
191 replicates as joint space width often can change in a period of more than 2 years particularly in
192 older individuals, in part due to possible knee joint cartilage degeneration. Fourth, we examined
193 the relationship between mJSW and OA status, both using DL-binary and the ICD-10 code M17

194 case-control data (**Fig. 2e**). As expected, mJSW was significantly lower in cases compared to
195 controls regardless of which case annotation we used (t-test $p < 2.2 \times 10^{-16}$, and $p < 2.2 \times 10^{-16}$ for
196 DL-binary and ICD-10 code M17 respectively). Finally, we examined the relationships between
197 mJSW and age - which is known to be highly associated with knee degeneration (**Fig. 2f**). Again,
198 as expected, we found that the mJSW decreased significantly with age (linear regression, Beta =
199 -0.028, $p < 2.2 \times 10^{-16}$).

200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239

240 **Fig. 2: Deep learning based image segmentation for minimum joint space width (mJSW).**



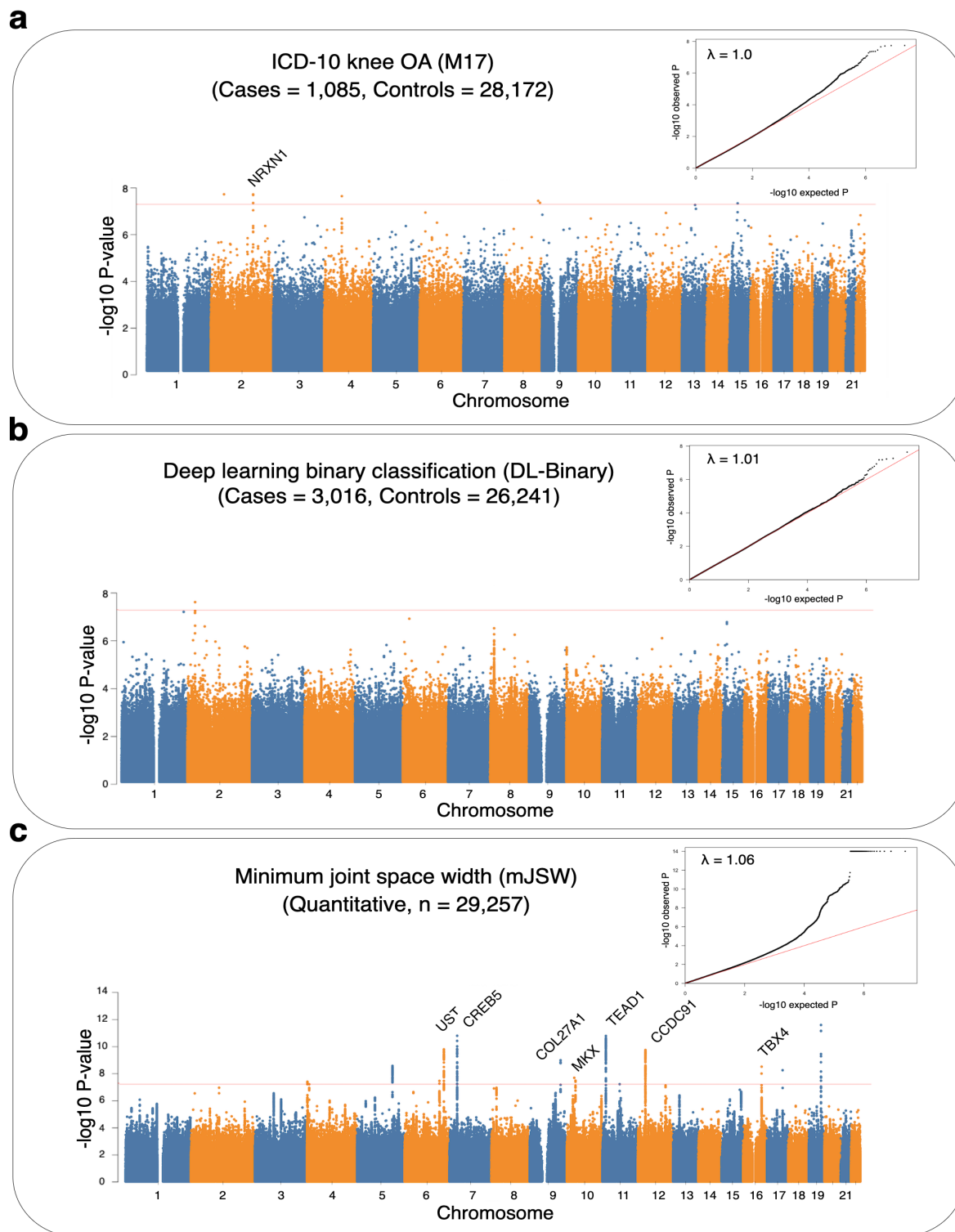
241
 242
 243 **a** Deep learning based image segmentation labeling the femur, tibia, fibula, and background
 244 based on a U-Net architecture. **b** Measurement of mJSW between the tibia and the femur taken
 245 by the average of 3 points each in the lateral, center and medial portions of the knee joint. **c**
 246 Correlation of mJSW between the right and left leg of the same individual ($n=29,257$). **d**
 247 Correlation in calculated mJSW between the first and second imaging visit for the same
 248 individual ($n=461$). **e** mJSW is narrower in cases compared to controls, using both ICD-10 code
 249 M17 and DL-binary case identification. **f** Average mJSW decreases significantly with increasing
 250 age (ages 48 - 79, $r = 0.88$, p -value < 0.0001). Circle size corresponds to the number of
 251 individuals within each age group, with larger diameters equating to higher sample size relative
 252 to smaller circles.

253 Genetic associations using image derived phenotypes

254 Having obtained IDPs related to knee OA, we performed GWAS to link these phenotypes
255 to their genetic basis. After generating summary statistics for each genetic association (**Fig. 3**),
256 we estimated SNP heritability using LD Score regression²⁴ for the three phenotypes: (1) Knee
257 OA as determined by the ICD-10 code M17 data from UKB, (2) Knee OA as determined using
258 DL-binary and (3) mJSW, the quantitative phenotype highly correlated with severity of knee
259 OA. The heritability of both binary phenotypes was low (ICD-10 code M17: 0.02 ± 0.02 and
260 DL-binary: 0.04 ± 0.02). In contrast the heritability of the quantitative phenotype mJSW was
261 0.24 ± 0.02 . Genomic inflation for the three phenotypes also confirmed this trend, with lambda
262 for ICD-10 code M17: 1.0, DL-binary: 1.01, and mJSW: 1.06. Deviations from expectation
263 across the genome are visualized in the qqplots inserts on **Fig. 3**. We found 18 independent loci
264 that reached genome-wide significance in the mJSW GWAS, including one that was also
265 significant in a previously reported GWAS for knee OA with 62,497 cases and 333,557
266 controls²⁵ (**Fig. 3**). We found one locus and six genome-wide significant loci with either binary
267 phenotype respectively (DL-binary and ICD-10 code M17), though mJSW and DL-binary had a
268 genetic correlation of -0.92 ± 0.25 (**Methods: Heritability and genetic correlation**). This suggests
269 substantial improvements in power from using a continuous, quantitative measure associated
270 with knee OA disease severity.

271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298

299 **Fig. 3** Manhattan plots for GWAS performed using three knee OA phenotyping methods.



300
301
302
303
304

a ICD-10 code M17 defined knee OA case and control status. **b** The deep learning based automated case-control phenotype, DL-binary. **c** The deep learning based quantitative endophenotype, mJSW. Loci over the genome-wide significance threshold ($p = 5 \times 10^{-8}$) that are

305 in close proximity to only a single gene are annotated. *Inset:* Quantile-quantile (qq) plot of
 306 deviation of the observed p-value from the theoretical distribution, along with the λ value
 307 quantifying genomic inflation.

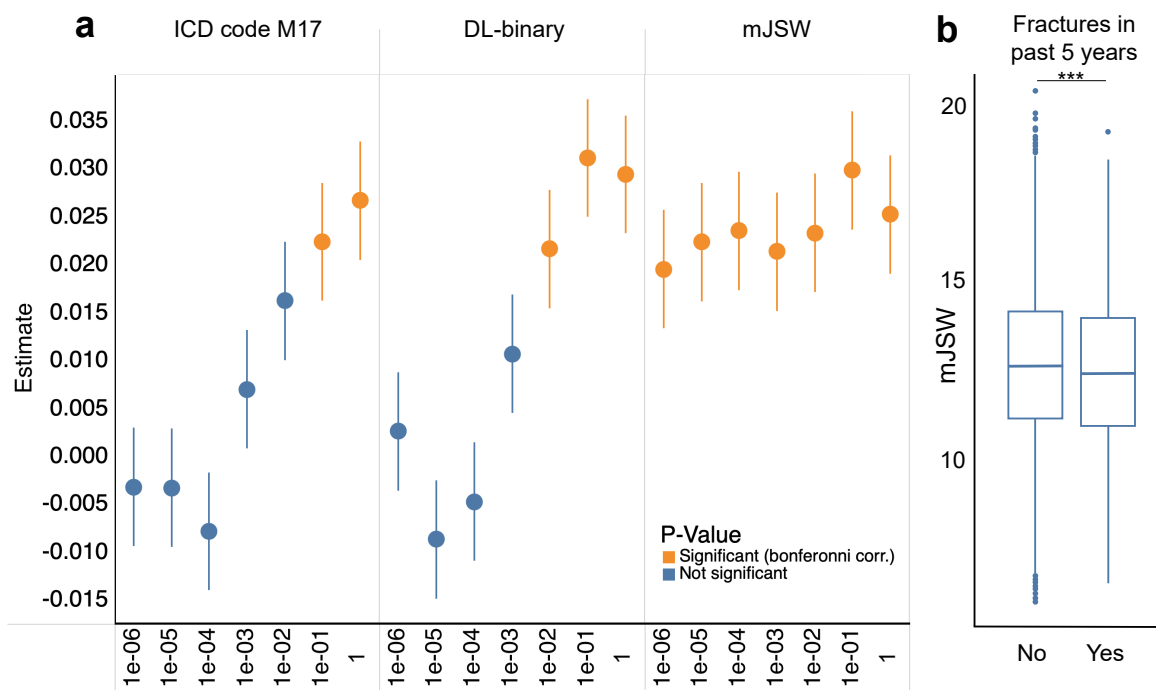
308 Polygenic risk scores for knee joint space are highly predictive of knee OA

309 As our GWAS for mJSW identified many more loci of genome-wide significance
 310 compared to DL-binary or ICD-10 code M17, we wanted to assess if this translated to improved
 311 power to predict knee OA in individuals outside of our DXA imaged sample. We computed PRS
 312 using clumping and thresholding (selecting variants below different p-value thresholds ranging
 313 from 1 to 1×10^{-6}) from the GWAS of ICD-10 code M17, DL-binary and mJSW, and deployed
 314 these scores on 371,686 individuals in the population who were not included in the GWAS
 315 (**Methods:** Polygenic Risk Scoring). We carried out logistic regression with binary presence or
 316 absence of ICD-10 code M17 diagnosed knee OA as the outcome, using z-scores generated from
 317 each of the PRSs as the predictor variable, and the first 20 PCs, age, sex, height and BMI as
 318 covariates. After controlling for these variables and performing multiple hypothesis testing
 319 correction at the level of the total number of associations performed, the mJSW PRS remained
 320 independently associated with knee OA diagnosis regardless of p-value threshold, but the DL-
 321 binary PRS or the ICD-10 code M17 PRS were only significantly associated with knee OA at
 322 certain thresholds, again reflecting differences in power between the various GWASs (**Fig. 4a**).

323

324

Fig 4. Genetic and epidemiological analysis of image derived phenotypes



325 **a** Results of performing logistic regression analysis using the PRS generated from each GWAS
 326 to predict ICD-10 M17 diagnosis on a hold-out dataset of 371,686 UKB individuals, showing the
 327 regression estimate obtained at each p-value threshold (and 1 standard error), colored by whether
 328 the test was significant after Bonferroni correction. **b** Boxplots showing the distribution of
 329 mJSW in patients who experienced a fracture in the past 5 years (n=24,147). mJSW was
 330

331 significantly predictive of fractures in logistic regression analysis, asterisks correspond to p-
332 value significance ($p < 1.10 \times 10^{-3}$).

333 Quantitative phenotyping allows for novel epidemiological associations

334 In addition to improving power in genetic analysis, we wanted to examine if we could use
335 the mJSW phenotype to improve statistical power to detect an important epidemiological
336 outcome in the health record, fractures within the last five years. After controlling for height,
337 sex, age and body fat percentage, mJSW was significantly predictive of fracture in the last five
338 years ($p = 1.10 \times 10^{-3}$) in logistic regression analysis, but not with DL-binary (fractures: $p =$
339 0.79) or with ICD-10 code M17 ($p = 0.171$) (**Fig. 4b**). While previous work on a much smaller
340 sample size of ~ 2000 individuals has shown that knee OA is associated with falls²⁶, our results
341 specifically implicate joint space narrowing with an independent increased risk of fractures, a
342 known cause of death in individuals 65 and over²⁷. These results emerge only upon examining
343 our quantitative phenotype mJSW which captures an element of disease severity, revealing knee
344 OA as an important risk factor for potentially fatal complications from fractures in older adults.

345 Discussion

346 In this study, we demonstrate a deep learning method to directly phenotype OA cases and
347 controls (DL-binary), as well as joint space narrowing (mJSW), from DXA scan derived AP
348 view knee radiographs of the UKB. We compared this image derived phenotyping approach with
349 case-control status already available in the ICD-10 record code M17 on the same set of
350 individuals, to determine if image-derived phenotyping approaches have an effect on statistical
351 power in GWAS.

352
353 We find that the case-control phenotyping using the DL-binary classification method
354 enables us to raise the case count by greater than two fold and circumvents some issues with
355 sourcing cases from the EHR such as variation in specific definitions of OA or differences in a
356 clinician's perception of the disease²⁸. While previous work has shown that the ICD-10 record
357 can have issues identifying individuals with disease for a variety of reasons, our study carrying
358 out image based diagnosis at large scale provides evidence of the extent to which the record can
359 be incomplete.

360
361 Additionally, both case-control methods lack information about disease severity, which
362 may explain why they are underpowered compared to the quantitative measurement mJSW in the
363 genetic and epidemiological analyses we investigate. The high genetic correlation between
364 mJSW and DL-binary (92%) suggests that while the binary case-control phenotype of knee OA
365 is underpowered compared with mJSW, the genetic relationships found between the two
366 phenotypes are consistent with one another.

367
368 While computer vision approaches to extract and analyze DXA scan derived phenotypes
369 are not themselves novel^{16,17,29,30}, this work is amongst the first to use this approach on a disease
370 for which diagnosis is primarily radiographic, to demonstrate that having a quantitative
371 endophenotype that captures additional information about variation in disease severity improves
372 power for genomic and epidemiological analysis. Although not based on the imaging data, two

373 novel phenotyping methods leveraging deep learning to impute missing data in the UKB and to
374 generate disease liability scores from binary case-control data in the EHR have shown significant
375 boosts in statistical power for genomic studies^{31,32}. Broadly, these and other approaches suggest
376 that analysis of biobank data could benefit from quantitative refinement of disease phenotypes
377 using alternative approaches.

378
379 One potential limitation of our study is that knee joint space narrowing is both causal and
380 symptomatic in knee OA progression. As arthritis progresses, the joint space narrows due to the
381 breakdown of cartilage, causing a resulting increase in pain and difficulty with movement. This
382 narrowing of the joint space can also cause further damage, due to increased contact pressure at
383 the affected joint. This makes it difficult to understand the root cause of knee OA with respect to
384 the mJSW endophenotype, because joint space narrowing can both be a result of OA and a
385 contributing cause to the progression of the condition. While the DL-binary method discovered
386 two-fold more cases than what is annotated in the ICD-10 record, it is still likely to be an
387 undercount due to our choice to use a particular instantiation of the model to limit the false
388 positive rate as much as possible (**Fig. 1e,f**). Thus, despite improving the case-control ratio in the
389 dataset, there may still be additional cases undetected by either method which could further
390 improve statistical power in GWAS. Third, all GWAS in this work were restricted to individuals
391 with European ancestry. Thus, the transferability of the specific findings in this genetic analysis
392 (i.e. loci discovered from mJSW GWAS, trait heritability, and genetic correlation) across
393 ancestries is not warranted without follow-up analyses.

394
395 Taken together, our study provides a proof-of-concept for the utility of quantitative
396 phenotyping in biobank scale settings where a direct measurement of disease severity for a
397 complex disease phenotype is possible. The results of this work suggest that this concept extends
398 not only to other musculoskeletal diseases in which radiography is one of the primary methods
399 for diagnosis (for example, directly measuring spinal curvature as opposed to scoliosis
400 diagnosis), but to other analyses in which one can derive a quantitative alternative to case-control
401 disease phenotyping.

402 **Methods**

403 **UKB participants and dataset**

404 All analyses were conducted with data from the UKB unless otherwise stated. The UKB
405 is a richly phenotyped, prospective, population-based cohort that recruited 500,000 individuals
406 aged 40–69 (mean 58) in the UK via mailers from 2006 to 2010¹
407 (<https://www.nature.com/articles/s41586-018-0579-z>). In total, we analyzed data from 402,000
408 participants with genetic data of self-identified white British ancestry who had not withdrawn
409 consent as of February 22, 2022. Of this genotyped cohort, 42,284 had available DXA imaging
410 data. Access was provided under application number 65439.

411 **Dual-energy X-ray Absorptiometry (DXA) Imaging**

412 The UKB has released DXA imaging data for a total of 50,000 participants as part of a
413 bulk data field ID. The DXA images were collected using a Lunar iDXA instrument¹ (GE
414 healthcare) in DICOM format. A series of 8 images were taken for each patient: two whole body
415 images - one of the skeleton and one of the adipose tissue, the lumbar spine, the lateral spine
416 from L4 to T4, each knee, and each hip. Dual-energy X-ray absorptiometry (DXA) images were
417 downloaded from the UKB bulk data. The bulk download resulted in 42,284 zip files, each
418 corresponding to a specific identifier otherwise known as each subject's EID. The uncompressed
419 directories corresponding to each imaged subject contained several DXA images of the
420 individual as described above. For this analysis, only images of the right and left knees from the
421 AP view were used. It is important to note that all subjects in this analysis were instructed to lay
422 flat on the DXA scanner machine during imaging, so that all resulting images are non-weight
423 bearing.

424 **Phenotype and clinical data acquisition**

425 The binary classification of patient disease phenotypes was obtained from a combination
426 of primary and secondary ICD-10 codes. ICD-10 codes were truncated to only be the initial three
427 characters. Patients received a "one" if a disease code appeared in their hospital records, and a
428 "zero" otherwise. Reports of a fracture within the last 5 years of any visit (instance 0 to 3) was
429 considered a case. Our classification of fractures increases case counts while excluding any
430 childhood incidence.

431 **Computing infrastructure**

432 We carried out all training using the Python programming language (www.python.org,
433 version 3.7.7) with the PyTorch³³ and Fastai version 1³⁴ (<https://github.com/fastai/fastai1>)
434 libraries on NVIDIA 1080-TI GPUs on the Maverick2 system and NVIDIA Quadro RTX 5000
435 GPUs on the Frontera system of the Texas Advanced Computing Cluster using the CUDA 11.1
436 toolkit.

439 DXA scan image quality control and standardization

440 DXA images in DICOM format were first organized by anatomy following the manifest
441 files located in each directory output by the imaging machine. DXA scans were subject to further
442 quality control following the methods described in Kun et al., 2022¹⁵. Following initial data
443 cleaning, AP view knee DXA scans were converted from DICOM to JPG format using the
444 pydicom library³⁵. To prepare a uniform set of images for segmentation, the numpy³⁶ and
445 opencv-python³⁷ libraries was used to pad images to a standard width and height (800 × 1000
446 pixels), and outlier images that had resolutions outside of this standardized range were removed
447 from all downstream analyses. Padded images were subject to further image resizing during
448 training of the U-net architecture²³ for segmentation (using a progressive resizing technique), but
449 not during training of the classification model.

450 Binary classification: DXA scan annotation procedure

451 546 images were sampled from the UKB for orthopedic surgeons to annotate. Images
452 were sampled with reference to the ICD-10 code M17 to create a balanced dataset for training
453 and validation. The KL grade²¹ based phenotype (DL-binary) was defined taking the following
454 observations as input: presence or absence of osteophytosis, visible sclerosis of bone, and
455 narrowing of the inter-bone joint space between the femur and tibia). Participating surgeons were
456 instructed to annotate images as 0 or 1 based on whether or not each image qualified as KL grade
457 3 or greater, meaning that based on the radiographic evidence of knee OA the individual would
458 be a candidate for joint replacement surgery. We considered 0 to be a control (but not necessarily
459 devoid of any radiographic OA symptoms) and 1 to be a case (KL grade 3 or 4) warranting joint
460 replacement. Surgeons went through a series of two rounds of independent grading on two
461 datasets. Following a review of inter-rater reliability, the three annotating physicians went
462 through a series of consensus grading, resulting in the final DL-binary dataset used for training
463 and validation. For the DL-binary classification model, 80% of the data was reserved for training
464 and 20% was reserved for validation.

465 Binary classification: Normalization and data augmentation

466 Prior to performing binary classification, images were scaled to 224 × 224 pixels and
467 normalized using ImageNet statistics. The ResNet-101 convolutional neural network (CNN)
468 weights were initialized using the Kaiming normal method²². While training, multiple
469 transformations were applied to the input images to regularize the model. These included a
470 padding process as described above, as well as other transformations such as vertical flipping of
471 the image, random rotation, zooming, warping, light and contrast change. This data
472 augmentation was performed to improve the model's ability to generalize in its predictions
473 relative to variation in contrast and other image artifacts common to DXA scanning³⁸.

474 Binary classification: Network architecture and model training

475 We constructed a ResNet-101 CNN²² for our binary DXA image classifier, implementing
476 transfer learning to reduce the amount of training time and resources for our classification task,
477 using a pre-trained model obtained from training on the ImageNet³⁹ (image-net.org) dataset and

478 transferring the weights from this model to earlier layers of the network. We applied batch
479 normalization and ReLU after each layer of the CNN to reduce overfitting and provide additional
480 regularization using the Fastai version 1³⁴ and PyTorch³³ default parameters, and dropout was
481 applied to the fully connected portion of the network. The output of the model is a binary
482 classification for each DXA scan derived image passed in, a one-dimensional tensor containing
483 values of 0 or 1 (control and case status), produced from passing the final layer of the network
484 (the classification head) through the sigmoid and argmax activation functions. The batch size for
485 all models was 64. We first plotted cross entropy loss as a function of learning rate in order to
486 select the optimal hyperparameters. We trained the model for 42 epochs with discriminative
487 learning rates ranging from 1×10^{-3} to 1×10^{-6} .

488 Image segmentation: DXA scan annotation procedure

489 We collected human generated annotations of each anatomical structure present in 63
490 DXA scans of the knee (40 training, 23 validation). Annotations were produced at the pixel level
491 for each of the following segments of an AP knee DXA scan the: (1) femur, (2) tibia and (3)
492 fibula. All annotations were reviewed by an orthopedic surgeon prior to training.

493 Image segmentation: Network architecture and model training

494 We trained a U-net architecture²³ with a 34-layer ResNet encoder²² to perform semantic
495 segmentation of the knee joint, annotating the femur, tibia, and fibula coded as 1, 2 and 3
496 respectively at pixel-level resolution. We used a batch size of 4 for the segmentation model. We
497 used the same transfer learning approach with the ImageNet dataset as described for the binary
498 classifier, as well as a progressive upsampling strategy during training. First we down sampled
499 masks to half their size, trained for 28 epochs, saved the model, then restarted the kernel and
500 trained the saved model on regular now upsampled mask. This training procedure was used to
501 efficiently utilize memory and reduce the model's time to convergence. As described previously,
502 we plotted cross entropy loss as a function of learning rate in order to select the optimal
503 hyperparameters.

504 Image segmentation: Measurement and quality control

505 After performing segmentation, we computed the minimum inter-bone knee joint space
506 distance in pixels (of either leg), abbreviated as mJSW. Segmentation masks were processed
507 using software developed for this analysis, written in python using the numpy³⁶ and opencv-
508 python³⁷ libraries (https://github.com/briannaflynn/UKB_knee_segmentation). Labeled polygons
509 within each segmentation mask were processed independently, converted to an identity matrix of
510 ones and zeros (ones being the polygon processed, for example the femur, tibia or fibula). From
511 this identity matrix, two matrices were produced from indexes produced and along the x and y
512 axes. These indexes were used in the computation of basic features of the polygon such as
513 maximum width, and maximum height. Indices were saved from this process and were later used
514 to compute measurements of joint space width between the femur and tibia.

515
516 A major issue in combining our analysis across input pixel ratios was that these pixel
517 ratios represented different resolution scalings due to variable distance between the scanner and

518 the patient as a function of DXA scanner type and the size of the patient. To control for this
519 scaling issue and to standardize the images, we chose to regress our mJSW measurements across
520 all image resolutions with height obtained from the UKB. The estimates obtained from this
521 regression were used to obtain a scaling factor for each image resolution that were then used for
522 measurement normalization. We validated this regression and normalization procedure by
523 comparing measurements taken on individuals who had DXA scans taken at two imaging
524 assessments at different resolutions.

525 Genetic QC

526 For all genome-wide association analyses, we filtered the participants to Caucasian
527 individuals (FID 22006) from the white, British population as determined by genetic PCA (FID
528 21000) and participant surveys. We removed individuals whose reported sex (FID 31) did not
529 match genetic sex (FID 22001), had evidence of aneuploidy on the sex chromosomes (FID
530 222019), were outliers of heterozygosity or genotype missingness rates as determined by UKB
531 quality control of sample processing and preparation of DNA for genotyping (FID 22027), or
532 had more than nine third-degree relatives or any of unknown kinship (FID 220021). In total
533 402,233 individuals remained. We further filtered to imaged participants (FID 20158) with
534 complete DXA measurements (FID 12254); 33,475 remained.

536 Imputed genetic data for 487,253 individuals was downloaded from UKB for
537 chromosomes 1 through 22 (FID 22828) then filtered to the quality-controlled subset using
538 PLINK2⁴⁰. All duplicate single nucleotide polymorphisms (SNPs) were excluded (--rm-dup
539 'exclude-all') and restricted to only biallelic sites (--snps-only 'just-acgt') with a maximum of 2
540 alleles (--max-alleles 2), a minor allele frequency of 0.1% (--maf 0.001), an individual
541 missingness rates no more than 2.5% (FID 22005), and genotype missingness of no more than
542 5% (--maxMissingPerSnp 0.05). In total 14,846,570 SNPs remained in the imputed dataset. Non-
543 imputed genetic data did not contain duplicate or multiallelic SNPs but were filtered to the
544 quality-controlled subset; 703,993 SNPs remained.

545 GWAS

546 GWAS was carried out using PLINK2, with a minor allele frequency of 0.001, a
547 missingness per SNP of 5%, and a missingness per individual of 2.5%. Covariates were the first
548 20 genetic principal components provided by UKB (FID 22009), sex (FID 31), age (FID 21022),
549 BMI (FID 21001) and standing height taken at the imaging assessment, instance 2 (FID 50). The
550 final population size for all GWAS after both genetic and imaging QC was 29,257, and all
551 GWASs had the same number of SNPs: 12,129,706. SNPs in each resulting GWAS were
552 clumped using --clump with a significance threshold of 5.0×10^{-8} , a secondary significance
553 threshold of 1.0×10^{-4} for clumped SNPs, an r^2 threshold of 0.1, and a 250 kb threshold of
554 physical distance. SNPs were assigned to genes with --clump-verbose --clump-range glist-hg19.

555 Heritability and genetic correlation

556 LD Score v1.0.1 was used to compute linkage disequilibrium regression scores per
557 chromosome with a window size of 1 cM²⁴ with the non-imputed genetic data. The heritability of

558 each phenotype was then assessed using LD score regression²⁴ with the same covariates as the
559 GWAS. We examined the pairwise genetic correlation of the DL-binary and mJSW using GCTA
560 version 1.93.2 beta for Linux⁴¹. We created the genetic relationship matrix for our quality-
561 controlled subset with a minor allele frequency of 0.001, and then ran GCTA, using the first 20
562 genetic principal components provided by UKB (FID 22009), sex, age, BMI and standing height
563 as covariates.

564 Polygenic Risk Scoring

565 PRSs were computed with the IDP GWAS summary statistics in PLINK (v1.9) using the
566 clumping and thresholding method. GWAS were clumped using an r^2 threshold of 0.1 and a 250
567 kb threshold of physical distance for clumping. Significance thresholds of 1, 0.1, 1×10^{-2} , 1×10^{-3} ,
568 1×10^{-4} , 1×10^{-5} , and 1×10^{-6} were used to compute PRSs for all three phenotypes run in
569 GWAS. We then regressed ICD-10 code diagnosis of knee OA on the z-scores generated from
570 each PRS obtained for each phenotype in all genotyped non-imaged individuals of white British
571 ancestry (who had also undergone genetic QC), $n = 371,723$. In our logistic regressions we also
572 controlled for age, sex, height, BMI and the first 20 principal components as covariates for all
573 phenotypes.

574 Transcriptome Analysis

575 To connect the genetics of joint space with the biology underlying synovial membrane
576 differences in individuals with knee OA, as synovial fluid is functionally important in OA
577 progression and inflammation^{42,43}. We looked for enrichment of genes associated with our
578 mJSW GWAS in gene expression data obtained from synovial tissues in 12 knee OA patients⁴⁴.
579 A set of inflamed as well normal synovial tissue was obtained from each patient for a set of 24
580 data points. This microarray gene expression data was obtained from the GEO data repository
581 GSE46750, and the data were quantile normalized and log-transformed. Gene level p-values for
582 our skeletal phenotype GWAS were first calculated using the positional mapping tool with
583 default settings in SNP2GENE (version 1.3.7)⁴⁵. We then performed gene property analysis in
584 MAGMA (version 1.10)⁴⁶ to determine associations between genes implicated in our mJSW
585 GWAS and genes expressed in normal as well as inflamed synovial tissue. We found enrichment
586 in our mJSW genes for gene expression from both normal and synovial tissue obtained from
587 knee OA patients but no difference in enrichment from differential expression of normal and
588 inflamed tissue **Table 1**.

589

| VARIABLE | ESTIMATE | ESTIMATE_STD | SE | P-VALUE |
|------------------|----------|--------------|-------|----------|
| normal_avg | 0.017 | 0.029 | 0.004 | 2.84E-05 |
| inflammatory_avg | 0.017 | 0.029 | 0.004 | 3.05E-05 |

590

591 **Table 1.** Results of enrichment analysis for genes significantly associated with the mJSW
592 phenotype from gene expression data obtained from normal (normal_avg) and inflamed
593 (inflammatory_avg) synovial tissue. The statistics are produced from a one-sided association test
594 between the phenotype and the 12 normal and 12 inflammatory data points.

595

596 Data availability

597
598 Deep learning and image processing tools can be found at tools can be found at
599 https://github.com/briannaflynn/UKB_knee_segmentation and
600 <https://github.com/briannaflynn/dxaconv/>. GWAS summary statistics are available at
601 <https://utexas.box.com/s/8stbz74t9hrx7fdbgl0gcqi66miodl92>. Individual level information of
602 image derived phenotypes has been reported back to the UKB and will be available upon
603 publication.
604

605 References

- 606
607 1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
608 *Nature* **562**, 203–209 (2018).
609 2. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J.*
610 *Epidemiol.* **27**, S2–S8 (2017).
611 3. Kurki, M. I. *et al.* FinnGen: Unique genetic insights from combining isolated population
612 and national health register data. *bioRxiv* (2022) doi:10.1101/2022.03.03.22271360.
613 4. Bernabeu, E. *et al.* Reply to: Genotype by sex interactions in ankylosing spondylitis. *Nat.*
614 *Genet.* **55**, 17–18 (2023).
615 5. Videm, V., Thomas, R., Brown, M. A. & Hoff, M. Self-reported Diagnosis of Rheumatoid
616 Arthritis or Ankylosing Spondylitis Has Low Accuracy: Data from the Nord-Trøndelag
617 Health Study. *J. Rheumatol.* **44**, 1134–1141 (2017).
618 6. Birmpili, P. *et al.* Evaluation of the ICD-10 system in coding revascularisation procedures
619 in patients with peripheral arterial disease in England: A retrospective cohort study using
620 national administrative and clinical databases. *EClinicalMedicine* **55**, 101738 (2023).
621 7. Lanyon, P., O'Reilly, S., Jones, A. & Doherty, M. Radiographic assessment of
622 symptomatic knee osteoarthritis in the community: definitions and normal joint space.
623 *Ann. Rheum. Dis.* **57**, 595–601 (1998).
624 8. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays
625 with Deep Learning. *arXiv [cs.CV]* (2017).
626 9. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S. & Jamalipour Soufi, G. Deep-COVID:
627 Predicting COVID-19 from chest X-ray images using deep transfer learning. *Med. Image*
628 *Anal.* **65**, 101794 (2020).
629 10. Currant, H. *et al.* Genetic variation affects morphological retinal phenotypes extracted
630 from UK Biobank optical coherence tomography images. *PLoS Genet.* **17**, e1009497
631 (2021).
632 11. Agrawal, S. *et al.* Association of machine learning-derived measures of body fat
633 distribution with cardiometabolic diseases in >40,000 individuals. *bioRxiv* (2021)
634 doi:10.1101/2021.05.07.21256854.
635 12. Bai, W. *et al.* A population-based phenome-wide association study of cardiac and aortic
636 structure and function. *Nat. Med.* **26**, 1654–1662 (2020).
637 13. Pirruccello, J. P. *et al.* Deep learning enables genetic analysis of the human thoracic aorta.
638 *bioRxiv* (2020) doi:10.1101/2020.05.12.091934.
639 14. Grasby, K. L. *et al.* The genetic architecture of the human cerebral cortex. *Science (80-.)*.
640 **367**, (2020).

- 641 15. Kun, E. *et al.* The genetic architecture of the human skeletal form. *bioRxiv* (2023)
642 doi:10.1101/2023.01.03.521284.
- 643 16. Faber, B. G. *et al.* A novel semi-automated classifier of hip osteoarthritis on DXA images
644 shows expected relationships with clinical outcomes in UK Biobank. *Rheumatol.* **61**,
645 3586–3595 (2022).
- 646 17. Frysz, M. *et al.* Machine Learning-Derived Acetabular Dysplasia and Cam Morphology
647 Are Features of Severe Hip Osteoarthritis: Findings From UK Biobank. *J. Bone Miner.*
648 *Res.* **37**, 1720–1732 (2022).
- 649 18. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics—2019 Update: A Report From
650 the American Heart Association. *Circulation* **139**, e56–e528 (2019).
- 651 19. Domanski, M., Lloyd-Jones, D., Fuster, V. & Grundy, S. Can we dramatically reduce the
652 incidence of coronary heart disease? *Nat. Rev. Cardiol.* **8**, 721–725 (2011).
- 653 20. Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: equating the
654 power of case-control and quantitative trait association studies. *Genet. Epidemiol.* **34**,
655 254–257 (2010).
- 656 21. Kohn, M. D., Sassoon, A. A. & Fernando, N. D. Classifications in Brief: Kellgren-
657 Lawrence Classification of Osteoarthritis. *Clin. Orthop. Relat. Res.* **474**, (2016).
- 658 22. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition.
659 *CoRR abs/1512.0*, (2015).
- 660 23. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical
661 Image Segmentation BT - Medical Image Computing and Computer-Assisted
662 Intervention – MICCAI 2015. in (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi,
663 A. F.) 234–241 (Springer International Publishing, 2015).
- 664 24. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
665 polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- 666 25. Boer, C. G. *et al.* Deciphering osteoarthritis genetics across 826,690 individuals from 9
667 populations. *Cell* **184**, 4784-4818.e17 (2021).
- 668 26. Doré, A. L. *et al.* Lower-extremity osteoarthritis and the risk of falls in a community-
669 based longitudinal study of adults with and without osteoarthritis. *Arthritis Care Res.* **67**,
670 633–639 (2015).
- 671 27. Burns, E. & Kakara, R. Deaths from Falls Among Persons Aged ≥ 65 Years - United
672 States, 2007-2016. *MMWR Morb. Mortal. Wkly. Rep.* **67**, 509–514 (2018).
- 673 28. Takuwa, H., Uchio, Y. & Ikegawa, S. Genome-wide association study of knee
674 osteoarthritis: present and future. *Ann. Jt.* **3**, 64 (2018).
- 675 29. Al-Absi, H. R. H., Islam, M. T., Refaee, M. A., Chowdhury, M. E. H. & Alam, T.
676 Cardiovascular Disease Diagnosis from DXA Scan and Retinal Images Using Deep
677 Learning. *Sensors* **22**, (2022).
- 678 30. Sethi, A. *et al.* Calcification of the abdominal aorta is an under-appreciated cardiovascular
679 disease risk factor in the general population. *Front Cardiovasc Med* **9**, 1003246 (2022).
- 680 31. An, U. *et al.* Deep Learning-based Phenotype Imputation on Population-scale Biobank
681 Data Increases Genetic Discoveries. *bioRxiv* (2022) doi:10.1101/2022.08.15.503991.
- 682 32. Yang, L., Sadler, M. C. & Altman, R. B. Genetic association studies using disease
683 liabilities from deep neural networks. *medRxiv* (2023) doi:10.1101/2023.01.18.23284383.
- 684 33. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning
685 Library. *arXiv [cs.LG]* (2019).
- 686 34. Howard, J. & Gugger, S. fastai: A Layered API for Deep Learning. *arXiv [cs.LG]* (2020).

- 687 35. Mason, D. *et al.* pydicom/pydicom: pydicom 2.3.0. (2022) doi:10.5281/zenodo.6394735.
688 36. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
689 37. Bradski, G. The OpenCV Library. *Dr. Dobb's J. Softw. Tools Prof. Program*.
690 38. Martineau, P., Bazarjani, S. & Zuckier, L. S. Artifacts and Incidental Findings
691 Encountered on Dual-Energy X-Ray Absorptiometry: Atlas and Analysis. *Semin. Nucl.*
692 *Med.* **45**, 458–469 (2015).
693 39. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE*
694 *Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
695 doi:10.1109/CVPR.2009.5206848.
696 40. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
697 datasets. *Gigascience* **4**, 7 (2015).
698 41. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide
699 complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
700 42. Benito, M. J., Veale, D. J., FitzGerald, O., van den Berg, W. B. & Bresnihan, B. Synovial
701 tissue inflammation in early and late osteoarthritis. *Ann. Rheum. Dis.* **64**, 1263–1267
702 (2005).
703 43. Sellam, J. & Berenbaum, F. The role of synovitis in pathophysiology and clinical
704 symptoms of osteoarthritis. *Nat. Rev. Rheumatol.* **6**, 625–635 (2010).
705 44. Lambert, C. *et al.* Gene expression pattern of cells from inflamed and normal areas of
706 osteoarthritis synovial membrane. *Arthritis Rheumatol* **66**, 960–968 (2014).
707 45. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
708 annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
709 46. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-
710 Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
711

712 **Acknowledgements**

713
714 This research has been conducted using the UKB Resource under Application Number 65439.
715 We thank Olivia Smith for insightful discussions and comments regarding bioinformatic
716 analysis.
717

718 V.M.N was supported on a grant from the Allen Discovery Center program, a Paul G. Allen
719 Frontiers Group advised program of the Paul G. Allen Family Foundation and a Good Systems
720 for Ethical AI grant from the University of Texas at Austin. B.F. was supported on an NSF
721 Graduate Research Fellowship DGE 2137420. E.M.J. and B.F were supported by an NIH T32
722 grant 5T32LMO012414. B.F. was also supported on a UT Austin Provost's Graduate Excellence
723 Fellowship.

724 **Author contributions**

725 B.F., E.J. and V. M. N. wrote the paper with input from all co-authors. B.F., E.J., E.K., A.G,
726 K.K., K.A. and E.L. performed analysis. T.J., Z.T., P.J. and V.M.N. supervised the analysis.
727

728

729

730 **Ethics declarations**

731 Competing interests

732 The authors declare no competing interests.