Arthroplasty Today 29 (2024) 101503

Contents lists available at ScienceDirect

Arthroplasty Today

journal homepage: http://www.arthroplastytoday.org/

Original Research

Analyzing Racial Differences in Imaging Joint Replacement Registries Using Generative Artificial Intelligence: Advancing Orthopaedic Data Equity

Bardia Khosravi, MD, MPH, MHPE[a, b, 1], Pouria Rouzrokh, MD, MPH, MHPE[a, b, 1], Bradley J. Erickson, MD, PhD[b], Hillary W. Garner, MD[c], Doris E. Wenger, MD[b], Michael J. Taunton, MD[a], Cody C. Wyles, MD[a, d, *]

[a] Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN, USA
[b] Department of Radiology, Mayo Clinic, Rochester, MN, USA
[c] Department of Radiology, Mayo Clinic, Jacksonville, FL, USA
[d] Department of Clinical Anatomy, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 12 May 2024
Received in revised form
17 June 2024
Accepted 9 August 2024
Available online xxx

Keywords:
Generative AI
Explainability
Dataset curation
Equity
Bias

ABSTRACT

Background: Discrepancies in medical data sets can perpetuate bias, especially when training deep learning models, potentially leading to biased outcomes in clinical applications. Understanding these biases is crucial for the development of equitable healthcare technologies. This study employs generative deep learning technology to explore and understand radiographic differences based on race among patients undergoing total hip arthroplasty.

Methods: Utilizing a large institutional registry, we retrospectively analyzed pelvic radiographs from total hip arthroplasty patients, characterized by demographics and image features. Denoising diffusion probabilistic models generated radiographs conditioned on demographic and imaging characteristics. Fréchet Inception Distance assessed the generated image quality, showing the diversity and realism of the generated images. Sixty transition videos were generated that showed transforming White pelvises to their closest African American counterparts and vice versa while controlling for patients' sex, age, and body mass index. Two expert surgeons and 2 radiologists carefully studied these videos to understand the systematic differences that are present in the 2 races' radiographs.

Results: Our data set included 480,407 pelvic radiographs, with a predominance of White patients over African Americans. The generative denoising diffusion probabilistic model created high-quality images and reached an Fréchet Inception Distance of 6.8. Experts identified 6 characteristics differentiating races, including interacetabular distance, osteoarthritis degree, obturator foramina shape, femoral neck-shaft angle, pelvic ring shape, and femoral cortical thickness.

Conclusions: This study demonstrates the potential of generative models for understanding disparities in medical imaging data sets. By visualizing race-based differences, this method aids in identifying bias in downstream tasks, fostering the development of fairer healthcare practices.

Introduction

Deep learning (DL) is a field of artificial intelligence (AI) that utilizes neural networks to learn patterns from the training data. As this is an automated process, the developers have little to no

* Corresponding author. Department of Orthopedic Surgery, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. Tel.: +1 507 284 2884.
E-mail address: Wyles.Cody@mayo.edu
[1] Equal contribution.

https://doi.org/10.1016/j.artd.2024.101503

control on the specific features picked up by the model during training. For example, studies have shown that models can sometimes latch onto unexpected or unintended features as shortcuts for making predictions. In the context of medical imaging, a model trained to detect pneumothorax might learn to associate the presence of a chest tube with the condition, rather than focusing on the actual radiographic signs of pneumothorax [1].

In orthopaedic surgery, DL models have been used for classification, segmentation, and object detection tasks, such as image anonymization, segmenting anatomical structures, and detecting implants on radiographs [2-4]. However, disparities in population and medical databases can introduce bias, particularly when training DL models, as these biases may be propagated without detection. For example, it has been shown that models trained on a publicly available chest radiograph data set have a higher tendency to underdiagnose African American patients with pneumonia and other pulmonary conditions [5]. More recently, it has been shown that DL models can with very high accuracy pick up patients' race from different medical imaging modalities, including hand radiographs, mammograms, and chest radiographs [6]. Identifying patient race from imaging can help uncover potential disparities in data sets and models, which may influence the accuracy and fairness of AI-assisted diagnostic and treatment planning tools in orthopaedic surgery. Understanding and characterizing these disparities and differences is crucial for developing more equitable AI models and improving patient care.

Imaging registries play a pivotal role in orthopaedic surgery, offering invaluable insights into patient outcomes, procedural efficacy, and long-term trends in surgical care by providing large-scale, standardized data on patient demographics, diagnoses, treatments, and follow-up [7,8]. Detecting racial differences within these registries is crucial, as it allows for the identification of disparities in disease prevalence, treatment outcomes, and access to care. For example, identifying a higher prevalence of advanced osteoarthritis among certain racial groups in an imaging registry may suggest disparities in access to timely diagnosis and treatment. While the exact causes of these disparities cannot be determined from the imaging data alone, our work provides a foundation for further investigation into the factors contributing to these differences and the development of targeted interventions to address them. Understanding these variations is also essential for developing targeted interventions aimed at improving healthcare equity and ensuring all patients receive optimal treatment regardless of their racial background.

Generative AI models, which create new synthetic data based on patterns learned from real data sets, offer a novel approach to understanding disparities in medical imaging data sets [9]. It is of utmost importance to understand systematic differences in large imaging data sets that might lead to bias in downstream tasks, such as automated diagnosis, treatment planning, or outcome prediction. For instance, if a data set contains a disproportionate number of images from a specific demographic group with a higher prevalence of advanced osteoarthritis, a model trained on these data may learn to associate certain anatomical features or imaging characteristics related to advanced osteoarthritis with that particular demographic group. Consequently, when the model is applied to a more diverse patient population, it may make biased predictions or recommendations based on these learned associations.

Several medical data set exploration techniques have been introduced to address this issue, but they often fall short when working with imaging data. Generative AI refers to the use of DL algorithms to generate new data that is similar to a training set and can be used for data exploration to understand hidden biases [10,11]. This study used generative DL technology to understand radiographic differences based on race among patients undergoing total hip arthroplasty in a tertiary referral center in the United States. By identifying and characterizing these differences, we hope to lay the groundwork for developing fairer AI models and improving patient care in orthopaedic surgery [12].

Material and methods

A large institutional registry of total hip arthroplasty patients was used to retrospectively assess pelvic radiographs from January 1, 1997, to October 1, 2021. Patients were characterized by age, sex, self-reported race, and body mass index, based on our clinical registry. All radiographs pertaining to the patients in the registry were automatically characterized using an already validated DL tool. This tool enabled us to extract image laterality, projection and presence or absence of implants [13].

We used denoising diffusion probabilistic models (DDPMs) to generate radiographs conditioned on demographic and image characteristics, including laterality and projection plane [14]. The DDPM was trained through a dual-phase process: an initial forward diffusion phase and a subsequent reverse diffusion phase, Figure 1 [15]. During the forward diffusion, incremental amounts of Gaussian noise are applied to an image in a sequential manner. The reverse diffusion process is more complex and involves estimating the noise addition between successive steps. The reverse diffusion process in our DL model employs an architecture akin to U-Net. This architecture accepts a noisy image version as its input and outputs a corresponding image with reduced noise levels.

In the context of image processing, *noise* refers to random variations or distortions that can appear in an image. These variations can be thought of as tiny specks or dots that are scattered throughout the image, making it appear grainy or less clear. Gaussian noise is a specific type of noise that follows a particular mathematical pattern called a Gaussian distribution. This distribution is often referred to as a "bell curve" due to its shape Gaussian noise adds random values to each pixel in an image. In the diffusion process, Gaussian noise is gradually added to an image over a series of steps. At each step, the noise makes the image progressively more distorted and less recognizable. The goal of the diffusion model is to learn how to reverse this process and remove the noise, gradually transforming a heavily distorted image back into a clear, recognizable one.

For conditioning the model on demographic and imaging factors, we employed the classifier-free guidance technique [16]. To produce diagnostic-quality images, we scaled up the size of the generated images, which were 256 pixels in each dimension. Consequently, we trained a diffusion-based super-resolution model. The U-Net model's input was conditioned on a low-resolution version of real images. This network was trained to generate sharp, high-resolution 1024 $\times$ 1024-pixel images from these low-resolution inputs. To ensure the reproducibility of our work, all training scripts are available at https://github.com/BardiaKh/Mediffusion. All model hyperparameters are summarized in Table 1. All models were trained on 4 NVIDIA (Santa Clara, CA, USA) A100 graphical processing units. Additional technical details can be found in the supplementary materials.

For inference, we used implicit sampling with 200 sampling steps for faster sampling during inference, which has been shown to preserve image quality [17]. Following model training, 60 high-resolution videos were generated of an anteroposterior pelvis radiograph from a White patient being transformed into its closest African American counterpart and vice versa, while controlling for age, sex, body mass index, and imaging characteristics Figure 2.

It has been previously studied that by using implicit sampling and starting from the same initial noise, the generated images from different classes would be very close to each other with minor *style*

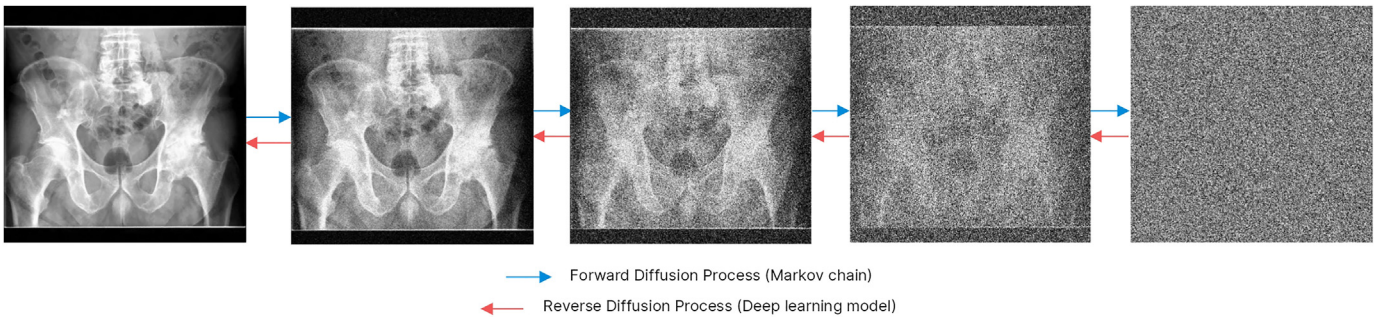

Figure 1. Demonstration of forward and reverse diffusion processes.

changes [18]. In the context of our study, *style* refers to the subtle, characteristic visual features that distinguish the generated radiographs of each demographic group while preserving the overall anatomical structures. To create counterfactuals, we used the same initiating noise and interpolated the class (African American vs White) embedding space in 180 frames. This ensured that the created counterfactual images are closest to each other, enabling studying differences that are related to the population rather than individual differences. Finally, all images were passed through the super-resolution model to create 1024 × 1024 images to make them ready for expert inspection.

The quality of the generated images was evaluated using Fréchet inception distance, which is a metric quantifying the diversity and realism of the generated images [19]. The lower this metric the better the generated images. To evaluate race-specific differences, 2 musculoskeletal radiologists with fellowship training in musculoskeletal imaging and 2 orthopaedic surgeons with fellowship training in joint reconstruction examined these interpolation videos and were asked to characterize systematic differences between images. They first inspected 30 videos, then convened to decide on the evaluation criteria. The evaluation criteria were then converted into a 5-point Likert scale and used for independent examination on 30 previously unseen videos, after a washout period of at least 28 days. To evaluate the ordinal rankings, Gwet's AC1 (GAC) was used to measure agreement among readers, with values > 0.60 designating substantial agreement and significance was set at $P < .05$.

Results

The training set consisted of 480,407 pelvic radiographs from 15,127 unique patients (52% female and 2% African American, 97% White, 1% other). Training the base generative model spanned 103 hours, while the super-resolution model took 47 hours to be fully trained. The generative models reached an Fréchet inception distance of 6.8, representing excellent image quality and diversity. Creating each 180-frame high-resolution (1024 × 1024) video took 101 minutes. Examples of the generated videos can be found at https://bit.ly/OSAIL-GenAI-Race.

Expert evaluators identified 6 characteristics that were systematically and consistently different between the 2 races, Table 2. The group found that African American patients, when compared to White patients, demonstrated (1) decreased interacetabular distance (GAC: 0.83; *P* value <.001), (2) higher degree of osteoarthritis (GAC: 0.82; *P* value <.001), (3) more elliptical obturator foramen (GAC: 0.80; *P* value <.001), (4) a decreased femoral neck-shaft angle (GAC: 0.76; *P* value <.001), (5) elongated pelvis ring (GAC: 0.61; *P* value <.001), and (6) increased femoral metaphyseal cortical thickness (GAC: 0.60; *P* value <.001). Table 3 summarizes these findings.

Table 1
Model hyperparameters.

Hyperparameter	Image generation model	Super-resolution model
Input shape	1 × 256 × 256	2 × 1024 × 1024[a]
Noise schedule (β)	Cosine	Cosine
Total steps (T)	1000	1000
Attention resolution	32, 16, 8	32
Embedding channels	512	512
U-net channels	128	64
U-net channel multiplier	1, 1, 2, 2, 4, 4	1, 1, 2, 2, 4, 4
Number of residual blocks	2	2
Number of attention heads	4	4

[a] The low-resolution image is added as a separate condition channel, making the input have two channels.

Discussion

Understanding disparities in large medical imaging data sets is crucial, as these can lead to biased downstream models. Generative models can be used to explain complex relationships in the underlying data, by creating counterfactual images [20]. Compared to generative counterfactuals, conventional methods for inspecting individual radiographs require a substantially higher volume and workload, due to the necessity of distinguishing individual variations from broader population differences. In this study, we introduce a novel approach for data set explainability, utilizing generative models to analyze racial differences and disparities within a large imaging registry. This method is advantageous because these models provide a comprehensive overview of the population distribution.

In line with previous findings, we found that African Americans have a shorter interacetabular distance [21]. The more elongated pelvis and a more elliptical obturator foramen that are caused by a change in the x-ray projection plane can be indicative of increased lumbar lordosis, which is previously described [22]. However, this projection change can also be due to a reported increased pelvic incident or other factors that cannot be assessed in anterior posterior radiographs alone [23,24]. Additionally, we observed a higher cortical thickness in the femoral neck region of the African American population of our study. This phenomenon has been previously reported in Study of Osteoporotic Fractures and the Baltimore Men's Osteoporosis Study in which they report a higher adjusted bone mineral density in African Americans compared to White patients [25-27]. However, compared to other findings, femoral cortical thickness had the lowest agreement among the readers. By doing a subanalysis, we see a much higher agreement between the 2 orthopaedic surgeons (GAC: 0.72). We believe this is because the surgeons were primed to look for this finding based on their

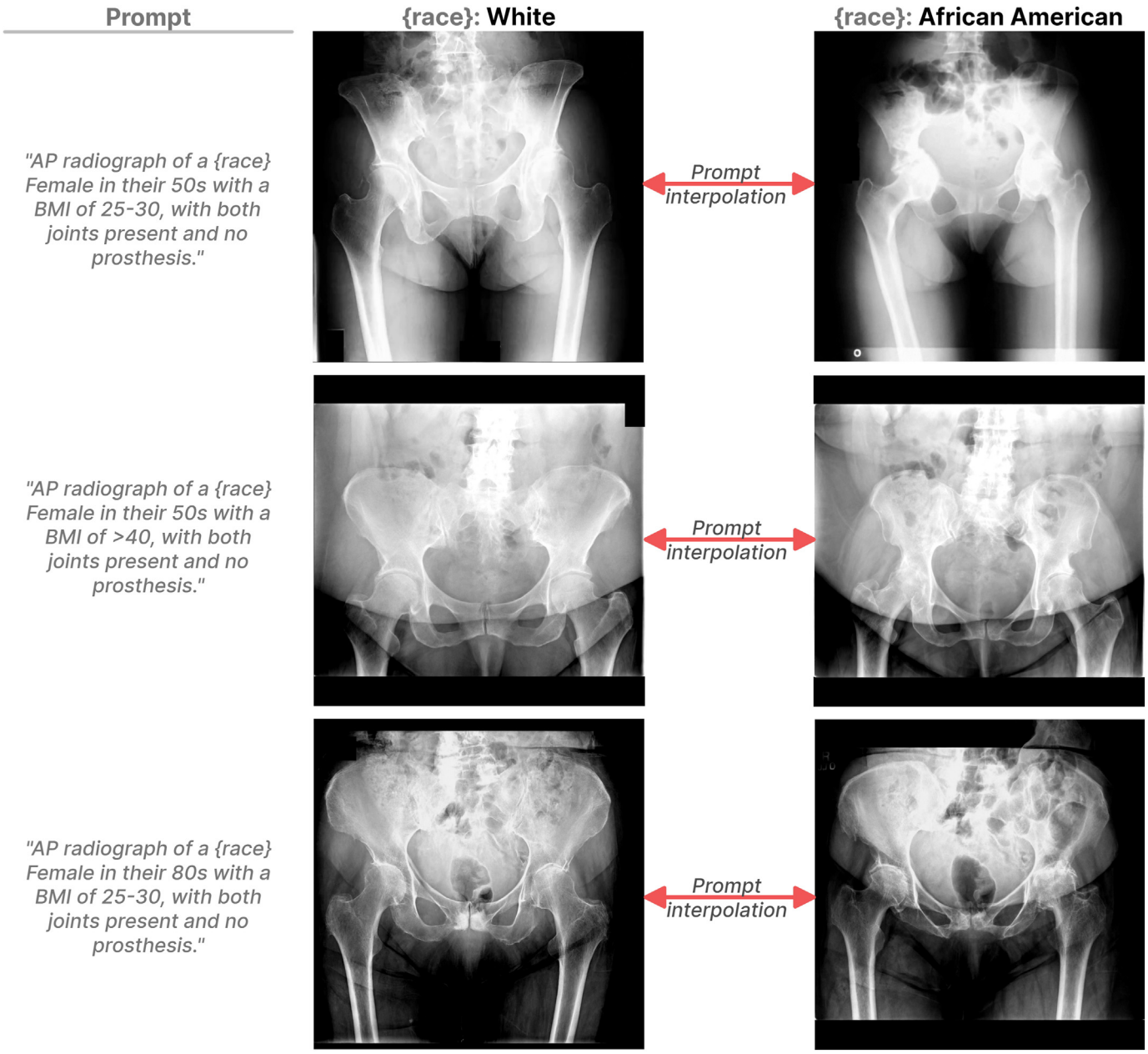

Figure 2. Examples of the generated images. We used embedding interpolation with the same initial noise, to have the closest pelvis radiograph to a given image but from a different race.

experience with higher bone mineral density in African American patients. This consistency with the previous results is indeed reassuring and strengthens the credibility of our results.

Importantly, we found that African Americans presenting for THA have a significantly higher degree of osteoarthritis, which is not an actual anthropomorphic difference, but reflects a more troubling social medicine phenomenon of reduced access and delays in care [21,28]. It has been previously shown in the Johnston County Osteoarthritis Project that African American patients have a more severe osteoarthritis in hip and knee joint radiographs [29,30].

This finding aligns with broader research indicating that these disparities are not merely clinical but can be rooted in social determinants of health [31]. African Americans' higher prevalence of severe osteoarthritis can be attributed to a confluence of factors including socioeconomic disparities, barriers to accessing health care, and delays in receiving timely and effective treatment [32]. This multifaceted issue underscores the critical need for comprehensive strategies that address the underlying social and economic barriers to care, aiming to mitigate the impact of these disparities on the health outcomes of African Americans.

Our work demonstrates the potential of using generative models to identify and characterize racial differences in imaging datasets, which can inform the development of more equitable AI models. For instance, by understanding the higher prevalence of severe osteoarthritis among African American patients in our data set, we can take proactive steps to rebalance the data used for training AI algorithms. This might involve ensuring that the training data includes a balanced representation of mild, moderate, and severe osteoarthritis cases from both African American and

Table 2
Frequency of selected categories for each criteria.

Criteria/Category	Median (IQR)/Frequency
Pelvis ring	2 (1)
1. Becomes much more elongated	44%
2. Becomes slightly elongated	33%
3. Does not change	11%
4. Becomes slightly more round	12%
5. Becomes much more round	0%
Obturator Foramen	1 (1)
1. Becomes much more elliptical	57%
2. Becomes slightly elliptical	28%
3. Does not change	15%
4. Becomes slightly more round	0%
5. Becomes much more round	0%
Interacetabular distance	1 (1)
1. Decreases significantly	66%
2. Decreases slightly	28%
3. Does not change	5%
4. Increases slightly	1%
5. Increases significantly	0%
Femoral neck-shaft angle	2 (1)
1. Decreases significantly	43%
2. Decreases slightly	43%
3. Does not change	14%
4. Increases slightly	0%
5. Increases significantly	0%
Medial femoral metaphyseal cortical thickness	2 (1)
1. Decreases significantly	1%
2. Decreases slightly	6%
3. Does not change	26%
4. Increases slightly	27%
5. Increases significantly	40%
Osteoarthritis severity	5 (1)
1. Decreases significantly	2%
2. Decreases slightly	6%
3. Does not change	13%
4. Increases slightly	18%
5. Increases significantly	61%

The question was formulated as, "With transitioning from White to African American, how does the {criteria} change?" For each category, we report their aggregated reported frequency, and for each criterion, we show the median (interquartile range; IQR). The reported median is the code corresponding to the category.

White populations. By carefully curating the training data to account for these differences, we can help the model learn to perform consistently across different demographic groups, mitigating the risk of biased predictions. This example further illustrates the practical utility of our approach in identifying disparities and informing strategies to address them in the development of AI tools for orthopaedic applications Another implication of our work is for synthetic image supplementation, where generated images, usually by a DL model, are added to real images to improve the model's performance [33-35]. We should be cognizant of the underlying data disparities and control for them in the generation process, for instance, by decoupling race and osteoarthritis severity. We believe that understanding these differences should inform the design of AI algorithms, in order to have fairer models.

Table 3
Identified racial differences in pelvic radiographs. *P* value in all instances were <0.001.

Characteristic	African American	White	Agreement (Gwet's AC1)
Interacetabular distance	Decreased	Increased	0.83
Osteoarthritis	Higher grade	Lower grade	0.82
Obturator foramen shape[a]	More elliptical	More circular	0.80
Femoral neck-shaft angle	Decreased	Increased	0.76
Pelvic ring shape[a]	More elongated	More circular	0.61
Femoral metaphyseal cortical thickness	Increased	Decreased	0.60

[a] Indicative of change in projection plain.

Our study has several limitations that should be considered when interpreting the results. First, our data set represents the demographics of a tertiary referral center in the United States, which may not be representative of the general population or other healthcare settings. The low representation of African American patients in our data set reflects the demographic composition of the patient population at our institution, which may limit the generalizability of our findings to other settings with different patient distributions. Second, we relied on self-reported race, which may not capture the full spectrum of patient diversity. Future studies should consider using more granular and objective measures of race and ethnicity to better characterize population differences. Third, while our generative models produced high-quality images, there is still room for improvement to create a more diagnostic-grade image, facilitating the evaluation of more granular differences. Finally, validation of these findings in independent datasets with other projection planes and other imaging modalities is necessary to confirm their robustness.

Conclusions

The study presents a promising approach for enhancing our understanding of underlying differences in medical imaging data sets by employing generative DL models. This method, through the creation of counterfactual images, offers a novel way to visualize and analyze race-based disparities within large imaging registries. By enabling a more nuanced exploration of demographic and image characteristics, this approach can serve as a valuable tool for researchers and clinicians aiming to identify and mitigate disparities in medical data sets. While these findings suggest potential anatomical differences between African American and White patients in our data set, it is important to note that they may not be representative of the entire population and should be interpreted with caution. Further research is needed to validate these results in larger, more diverse data sets and to investigate their potential impact on clinical outcomes and treatment decision-making. The proposed approach is much less resource-intensive than running a prospective cohort study to understand these disparate differences. It lays the groundwork for further studies to explore and understand these disparities, contributing to the development of more equitable healthcare practices.

Funding

This work was supported by the Mayo Foundation Presidential Fund.

Conflicts of interest

The authors declare there are no conflicts of interest.

For full disclosure statements refer to https://doi.org/10.1016/j.artd.2024.101503.

CRediT authorship contribution statement

Bardia Khosravi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pouria Rouzrokh:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bradley J. Erickson:** Writing – review & editing, Validation, Resources,

Methodology, Investigation, Conceptualization. **Hillary W. Garner:** Writing – review & editing, Validation, Methodology, Investigation. **Doris E. Wenger:** Writing – review & editing, Validation, Methodology, Investigation. **Michael J. Taunton:** Writing – review & editing, Validation, Methodology, Investigation. **Cody C. Wyles:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

References

[1] Rueckel J, Trappmann L, Schachtner B, Wesp P, Hoppe BF, Fink N, et al. Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs. Invest Radiol 2020;55:792–8.

[2] Rouzrokh P, Mickley JP, Khosravi B, Faghani S, Moassefi M, Schulz WR, et al. THA-AID: deep learning tool for total hip arthroplasty automatic implant detection with uncertainty and outlier quantification. J Arthroplasty 2024;39:966–973.e17.

[3] Rouzrokh P, Wyles CC, Philbrick KA, Ramazanian T, Weston AD, Cai JC, et al. A deep learning tool for automated radiographic measurement of acetabular component inclination and version after total hip arthroplasty. J Arthroplasty 2021;36:2510–2517.e6.

[4] Khosravi B, Mickley JP, Rouzrokh P, Taunton MJ, Larson AN, Erickson BJ, et al. Anonymizing radiographs using an object detection deep learning algorithm. Radiol Artif Intell 2023;5:e230085.

[5] Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27:2176–82.

[6] Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4:e406–14.

[7] Khosravi B, Rouzrokh P, Maradit Kremers H, Larson DR, Johnson QJ, Faghani S, et al. Patient-specific hip arthroplasty dislocation risk calculator: an explainable multimodal machine learning-based approach. Radiol Artif Intell 2022;4:e220067.

[8] Wyles CC, Maradit-Kremers H, Fruth KM, Larson DR, Khosravi B, Rouzrokh P, et al. Frank stinchfield award: creation of a patient-specific total hip arthroplasty periprosthetic fracture risk calculator. J Arthroplasty 2023;38:S2–10.

[9] Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating bias in radiology machine learning: 1. Data handling. Radiol Artif Intell 2022;4:e210290.

[10] Luccioni AS, Akiki C, Mitchell M, Jernite Y. Stable bias: analyzing societal representations in diffusion models [Internet]. arXiv [cs.CY]. http://arxiv.org/abs/2303.11408; 2023. [Accessed 16 June 2024].

[11] Khosravi B, Rouzrokh P, Mickley JP, Faghani S, Larson AN, Garner HW, et al. Creating high fidelity synthetic pelvis radiographs using generative adversarial networks: unlocking the potential of deep learning models without patient privacy concerns. J Arthroplasty 2023;38:2037–20343.e1.

[12] Baumgartner R, Arora P, Bath C, Burljaev D, Ciereszko K, Custers B, et al. Fair and equitable AI in biomedical research and healthcare: social science perspectives. Artif Intell Med 2023;144:102658.

[13] Rouzrokh P, Khosravi B, Johnson QJ, Faghani S, Vera Garcia DV, Erickson BJ, et al. Applying deep learning to establish a total hip arthroplasty radiography registry: a stepwise approach. J Bone Joint Surg Am 2022;104:1649–58.

[14] Khosravi B, Rouzrokh P, Mickley JP, Faghani S, Mulford K, Yang L, et al. Few-shot biomedical image segmentation using diffusion models: beyond image generation. Comput Methods Programs Biomed 2023;242:107832.

[15] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [internet]. arXiv [cs.LG]. http://arxiv.org/abs/2006.11239; 2020. [Accessed 16 June 2024].

[16] Ho J, Salimans T. Classifier-free diffusion guidance [internet]. arXiv [cs.LG]. http://arxiv.org/abs/2207.12598; 2022. [Accessed 16 June 2024].

[17] Song J, Meng C, Ermon S. Denoising diffusion implicit models [internet]. arXiv [cs.LG]. http://arxiv.org/abs/2010.02502; 2020. [Accessed 16 June 2024].

[18] Wu CH, De la Torre F. Unifying diffusion models' latent space, with applications to CycleDiffusion and guidance [Internet]. arXiv [cs.CV]. http://arxiv.org/abs/2210.05559; 2022. [Accessed 16 June 2024].

[19] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium [Internet]. arXiv [cs.LG]. http://arxiv.org/abs/1706.08500; 2017. [Accessed 16 June 2024].

[20] Cohen JP, Brooks R, En S, Zucker E, Pareek A, Lungren MP, et al. Gifsplanation via Latent Shift: a simple autoencoder approach to counterfactual generation for chest X-rays [Internet]. arXiv [cs.CV]. http://arxiv.org/abs/2102.09475; 2021. [Accessed 16 June 2024].

[21] Edwards K, Leyland KM, Sanchez-Santos MT, Arden CP, Spector TD, Nelson AE, et al. Differences between race and sex in measures of hip morphology: a population-based comparative study. Osteoarthritis Cartilage 2020;28:189–200.

[22] Hanson P, Magnusson SP, Simonsen EB. Differences in sacral angulation and lumbosacral curvature in black and white young men and women. Acta Anat 1998;162:226–31.

[23] Arima H, Dimar JR 2nd, Glassman SD, Yamato Y, Matsuyama Y, Mac-Thiong JM, et al. Differences in lumbar and pelvic parameters among African American, Caucasian and Asian populations. Eur Spine J 2018;27:2990–8.

[24] Merrill RK, Kim JS, Leven DM, Kim JH, Meaike JJ, Bronheim RS, et al. Differences in fundamental sagittal pelvic parameters based on age, sex, and race. Clin Spine Surg 2018;31:E109–14.

[25] Hochberg MC. Racial differences in bone strength. Trans Am Clin Climatol Assoc 2007;118:305–15.

[26] Black DM, Cummings SR, Genant HK, Nevitt MC, Palermo L, Browner W. Axial and appendicular bone density predict fractures in older women. J Bone Miner Res 1992;7:633–8.

[27] Tracy JK, Meyer WA, Flores RH, Wilson PD, Hochberg MC. Racial differences in rate of decline in bone mass in older men: the Baltimore men's osteoporosis study. J Bone Miner Res 2005;20:1228–34.

[28] Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021;27:136–40.

[29] Nelson AE, Braga L, Renner JB, Atashili J, Woodard J, Hochberg MC, et al. Characterization of individual radiographic features of hip osteoarthritis in African American and White women and men: the Johnston County Osteoarthritis Project. Arthritis Care Res 2010;62:190–7.

[30] Jordan JM. An ongoing assessment of osteoarthritis in african Americans and caucasians in North Carolina: the Johnston county osteoarthritis Project. Trans Am Clin Climatol Assoc 2015;126:77–86.

[31] Callahan LF, Cleveland RJ, Allen KD, Golightly Y. Racial/ethnic, socioeconomic, and geographic disparities in the epidemiology of knee and hip osteoarthritis. Rheum Dis Clin North Am 2021;47:1–20.

[32] Moss AS, Murphy LB, Helmick CG, Schwartz TA, Barbour KE, Renner JB, et al. Annual incidence rates of hip symptoms and three hip OA outcomes from a U.S. population-based cohort study: the Johnston County Osteoarthritis Project. Osteoarthritis Cartilage 2016;24:1518–27.

[33] Khosravi B, Li F, Dapamede T, Rouzrokh P, Gamble CU, Trivedi HM, et al. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. EBioMedicine 2024;104:105174.

[34] Rouzrokh P, Khosravi B, Faghani S, Mulford KL, Taunton MJ, Erickson BJ, et al. RadRotator: 3D rotation of radiographs with diffusion models [internet]. arXiv [eess.IV]. http://arxiv.org/abs/2404.13000; 2024. [Accessed 16 June 2024].

[35] Rouzrokh P, Khosravi B, Mickley JP, Erickson BJ, Taunton MJ, Wyles CC. THA-net: a deep learning solution for next-generation templating and patient-specific surgical execution. J Arthroplasty 2023;39:727–733.e4. https://doi.org/10.1016/j.arth.2023.08.063.

Supplemental Material

Training details

The denoising diffusion probabilistic models (DDPM) are trained through a dual-phase process: an initial forward diffusion phase and a subsequent reverse diffusion phase (1). During the forward diffusion, incremental amounts of Gaussian noise are applied to an image in a sequential manner. Over a specified number of steps (total timesteps, T), the original image progressively transforms into isotropic Gaussian noise. This transformation adheres to a Markov process, guided by a predetermined noise schedule. For instance, to reach a noise level at timestep 100 ($t = 100$), the model sequentially progresses through the first 99 timesteps. Notably, due to the Gaussian nature of the noise, it's possible to calculate the appearance of an image at any given timestep directly, bypassing the need for sequential noise addition. This is achieved using a specific formula with the initial input (x_0), allowing the calculation of x0's noisier version at any chosen timestep using the following formula:

$$x_t = \sqrt{\overline{\alpha_t}}\, x_0 + \sqrt{1 - \overline{\alpha_t}}\ \epsilon$$

where:

$$\alpha_t = 1 - \beta_t \ ; \ \overline{\alpha_t} = \prod_{s=0}^{t} \alpha_s$$

The noise schedule (β) is defined prior to training, which then sets the values for αt and $\overline{\alpha_t}$ (cumulative product from α_0 to α_t for each timestep t). At each training stage, noise ϵ (sourced from a Gaussian distribution) is randomly selected for each image in the batch. This noise is then integrated with the formula mentioned earlier to generate x_t. The reverse diffusion process, in contrast, is more complex and involves estimating the noise addition (ϵ) between successive steps. This phase requires the training of a sophisticated deep learning model, commonly known as a diffusion model. The primary training objective is to diminish the mean squared error loss between the model-predicted noise and the actual noise (ϵ), originally computed during the forward diffusion.

The aim during training is to minimize the mean squared error loss between the noise predicted by the diffusion model and the original noise, computed in advance using the forward diffusion process. In the inference stage, the process starts with random noise, which is then processed by the DL model T times to execute a reverse diffusion process. In each iteration, the model gradually eliminates a portion of the noise, eventually yielding a clearer image similar to those in the training data set. In our research, we used $T = 1000$ and a cosine noise schedule for defining forward diffusion parameters.

The reverse diffusion process in our deep learning model employs an architecture akin to U-Net. This architecture accepts a noisy image version as its input and outputs a corresponding image with reduced noise levels. Our DDPM offers several enhancements over the conventional U-Net models typically used in biomedical image segmentation (2). Key improvements include the integration of multiple residual blocks within each layer and the incorporation of self-attention modules, which significantly enhance image quality. Additionally, the model is designed to include the timestep number and a conditioning vector within its structure. The conditioning vector carries class-specific information.

Due to the model's awareness of image attributes through the conditioning vector, it is capable of generating images with predetermined characteristics. This is achieved by supplying the model with desired class information during the inference stage. Such a design enables more precise and tailored image generation, particularly beneficial in scenarios where specific image traits are required.

For conditioning the model on demographic and imaging factors, we employed the classifier-free guidance technique. This technique incorporates a learned embedding for null classes and randomly selects a subset of images during training to substitute their class embedding with this null embedding. We implemented a dropout rate of 10% and a guidance score of 4.0 for all inference instances. The model underwent training for over 275,000 iterations with a batch size of 64 on four A100 graphical processing units, generating 256 $\times$ 256-pixel images. This training spanned 103 hours.

To produce diagnostic-quality images, we scaled up the size of the generated images. Consequently, we trained a diffusion-based super-resolution model. The U-Net model's input was conditioned on a low-resolution version of authentic images. This network was trained to generate sharp, high-resolution 1024 $\times$ 1024-pixel images from these low-resolution inputs. The training of this model encompassed 100,000 steps, utilizing a batch size of 5 across four A100 graphical processing units, completed over a 47-hour training period.

To streamline the training process, we used an open-source package called *Mediffusion*, developed by the authors of this present study, to streamline the model development and reproducibility of our work. The training and inference codes can be found at https://github.com/BardiaKh/Mediffusion.

References

[1] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models [Internet]. arXiv [cs.LG]. http://arxiv.org/abs/2006.11239; 2020.

[2] Khosravi B, Li F, Dapamede T, Rouzrokh P, Gamble CU, Trivedi HM, et al. Synthetically Enhanced: Unveiling Synthetic Data's Potential in Medical Imaging Research [Internet]. arXiv [cs.CV]. http://arxiv.org/abs/2311.09402; 2023.