Osteoarthritis and Cartilage Open 5 (2023) 100378

Contents lists available at ScienceDirect

Osteoarthritis and Cartilage Open

journal homepage: www.elsevier.com/journals/osteoarthritis-and-cartilage-open/2665-9131

Brief Report

Automated segmentation and prediction of intervertebral disc morphology and uniaxial deformations from MRI

James A. Coppock [a,b], Nicole E. Zimmer [a,b], Charles E. Spritzer [c], Adam P. Goode [a,d,e], Louis E. DeFrate [a,b,f,*]

[a] Department of Orthopedic Surgery, Duke University School of Medicine, United States
[b] Department of Biomedical Engineering, Duke University, United States
[c] Department of Radiology, Duke University School of Medicine, United States
[d] Duke Clinical Research Institute, Duke University School of Medicine, United States
[e] Department of Population Health Sciences, Duke University, United States
[f] Department of Mechanical Engineering and Materials Science, Duke University, United States

ARTICLE INFO

Keywords:
Medical image segmentation
Machine learning
Computer aided diagnosis
Low back pain
Intervertebral disc degeneration
Intervertebral disc mechanics

ABSTRACT

Objective: The measurement of *in vivo* intervertebral disc (IVD) mechanics may be used to understand the etiology of IVD degeneration and low back pain (LBP). To this end, our lab has developed methods to measure IVD morphology and uniaxial compressive deformation (% change in IVD height) resulting from dynamic activity, *in vivo*, using magnetic resonance images (MRI). However, due to the time-intensive nature of manual image segmentation, we sought to validate an image segmentation algorithm that could accurately and reliably reproduce models of *in vivo* tissue mechanics.

Design: Therefore, we developed and evaluated two commonly employed deep learning architectures (2D and 3D U-Net) for the segmentation of IVDs from MRI. The performance of these models was evaluated for morphological accuracy by comparing predicted IVD segmentations (Dice similarity coefficient, mDSC; average surface distance, ASD) to manual (ground truth) measures. Likewise, functional reliability and precision were assessed by evaluating the intraclass correlation coefficient (ICC) and standard error of measurement (SE_m) of predicted and manually derived deformation measures.

Results: Peak model performance was obtained using the 3D U-net architecture, yielding a maximum mDSC = 0.9824 and component-wise ASD_x = 0.0683 mm; ASD_y = 0.0335 mm; ASD_z = 0.0329 mm. Functional model performance demonstrated excellent reliability ICC = 0.926 and precision SE_m = 0.42%.

Conclusions: This study demonstrated that a deep learning framework can precisely and reliably automate measures of IVD function, drastically improving the throughput of these time-intensive methods.

1. Introduction

Changes in intervertebral disc (IVD) mechanics may be related to the future development of discogenic low back pain (LBP) [1–3]. To this point, recent studies examining IVD mechanics *in vivo* have demonstrated that factors associated with the development of IVD degeneration and LBP (e.g., IVD composition, BMI) are predictive of IVD function in response to dynamic activity. As such, it is believed that studying relationships between IVD mechanics and composition *in vivo* may be useful for evaluating the etiology of LBP as it pertains to IVD function [4,5]. However, because recent work examining *in vivo* IVD function often involves time-intensive manual image segmentations, the translational viability of these methods remains limited.

Deep learning algorithms have the potential to ameliorate this bottleneck by automating the segmentation process. In particular, u-net-based algorithms have proven to be powerful tools for efficiently segmenting and classifying anatomy, including in the spine [6–8]. However, limited data exists examining the utility of such models as a means for deriving measures of IVD tissue mechanics directly from MR images, without the need for manual intervention.

Thus, the objective of this study was to develop a deep learning model to accurately segment IVD morphology. Subsequently, we sought to

* Corresponding author. Duke University, Box 3093, Durham, NC, 27708, United States. Tel.: +(919) 681 9959; fax: +(919) 681 8490.
E-mail address: Lou.DeFrate@duke.edu (L.E. DeFrate).

https://doi.org/10.1016/j.ocarto.2023.100378
Received 9 January 2023; Accepted 26 May 2023

evaluate the reliability and precision of this method as a means for deriving measures of uniaxial compressive deformations (% change in IVD height) resulting from dynamic activity. To do so, we utilized two commonly employed deep learning architectures (2D and 3D U-net) to segment IVD tissues from MR images [6]. The performance of these models was evaluated by comparing predicted IVD morphology (dice overlap, surface distance) and functional (deformation) metrics to manual (ground truth) measures. We hypothesized that automated segmentations would provide sufficient morphological accuracy to reliably and precisely measure uniaxial compressive deformations from MR images taken before and after a dynamic walking activity.

2. Methods

2.1. Study design

The present study utilizes pre -and post-exercise imaging data from 25 asymptomatic subjects (i.e., with no history of back pain, injury, or surgery) who participated in an IRB approved study (Age (y): 19–63; Height (m): 1.57–1.91; Mass (kg): 54.3–121.6; BMI (kg/m^2): 19.3–37.1). Imaging data from 18 of the 25 subjects in this study have been used to evaluate IVD function in two previously published works [4,5]. Briefly, to minimize the effects of diurnal changes in IVD height, subjects were instructed to lay supine for 45 min prior to the baseline (pre-exercise) MR imaging scan [1]. Following the baseline scan, subjects walked on a treadmill for 30 min at a constant speed after which they immediately returned to the scanner to complete post-exercise imaging.

2.2. Image processing

MR images were collected on a single 3.0-T (Tim Trio, Siemens Medical Solutions) and 12-channel spine matrix coil. MR images used for this analysis were acquired using a Sagittal 3D T2-weighted SPACE (TE/TR:223/2500 ms; resolution, 0.875 × 0.875 × 0.875 mm; matrix, 320 × 320 × 80 pixels [3]) sequence. Ground truth (manual) binary segmentations were performed by two investigators (with 2 and 4 years of experience) and independently reviewed by a musculoskeletal radiologist for accuracy (>30 years of experience). Segmentations [4,5] of the L1-L2–L5-S1 IVDs from SPACE MR volumes were performed for both the pre- and post-exercise scan as previously described [4]. IVD images (2D: 64x64) and volumes (3D: 64x64x80) were manually localized (such that only one IVD is present in each image) and stored in separate files for subsequent model development. Thus, because images were cropped to contain only a single IVD, each subject effectively contributed 10 IVDs (i.e., L1-L2 – L5-S1; pre- and post-exercise) to the dataset.

2.3. Model development and training

Two basic U-net architectures (2D/3D) [6] were developed to evaluate this problem. All model architectures were constructed with five encoder-decoder steps using 16, 32, 64, 128 and 256 filters at each level, respectively. 2D and 3D model architectures were trained and evaluated using deidentified bitmap (2D), and NIfTI-1 images (3D). Within each model architecture group, a grid search optimization was performed to examine the effect of hyperparameters: learning rate (Adam optimizer: 1e-1,1e-2,1e-3,1e-4), kernel size (3,5,7,9,11,13) and training batch size (2D: 4,8,16; 3D: 1,2,4) on model performance. Validation and test batch sizes were held constant at n = 2 (20 IVDs) and n = 1 subjects (10 IVDs), respectively, for 2D and 3D model architectures. Model predictions were converted into predicted class probabilities and then binarized using the argmax transformation.

Model segmentation performance was assessed during validation and testing steps by measuring the mean Dice similarity coefficient (mDSC) (equation (1)), where true positive (TP), true negative (TN), false positive (FP) and false negative (FN) pixels denote the agreement between ground truth and predicted labels. Dice coefficient calculations are inclusive of the background class. Similarly, a Dice coefficient loss function (equation (2)) was utilized during model training. For the loss function, background class predictions were included in calculation of Dice loss.

Subject assignment to training, testing and validation groups was done randomly; however, training, validation and testing data sets were held constant across all model permutations. Three subjects' data were withheld from training (n = 22 subjects) and divided into validation (n = 2 subjects) and testing (n = 1 subject) datasets. All models were developed, trained, tested and analyzed using python (3.8) packages pytorch (1.10.2), monai (0.9.dev2221), nibabel (3.2.2) and pydicom (2.3.0). Training was performed for n = 2000 epochs with early stopping initiated after 200 epochs without improvement (minimum mDSC delta = 1e-4). Training and grid search optimization was conducted using a high-performance computing cluster.

$$mDSC = \frac{1}{n}\sum_{i=1}^{n}\frac{2TP}{(2TP + FP + FN)} \tag{1}$$

$$DSC_{loss} = 1 - mDSC \tag{2}$$

$$ASD_i = \frac{1}{n}\sum_{b\in Pred}\min_{b\in Pred}\|a_i - b_i\|_1 \tag{3}$$

$$Deformation = \frac{H_{post} - H_{pre}}{H_{pre}} \times 100 \tag{4}$$

$$SE_m = \sqrt{MSE} \tag{5}$$

2.4. IVD morphology analysis

IVD morphological performance was assessed using mDSC and the component-wise (i.e., x, y, z) average surface distance (ASD; equation (3)). The component-wise ASD was defined as the average minimum L_1 distance (mm) between the nearest neighboring point on the boundary of the ground truth (a_i) and predicted (b_i) segmentations. The x, y and z ASD components were defined along the anterior-posterior, superior-inferior and medial-lateral, respectively.

2.5. IVD deformation analysis

Predicted segmentations from the top performing models during testing were then used to evaluate uniaxial IVD deformation. IVD deformation was defined as the percent change in IVD height from the pre- (H_{pre}) to the post-exercise (H_{post}) MR scan (equation (4)). To do so, surface contours were extracted from the binary segmentations, yielding point clouds of the IVD. Surface models were then constructed using a modified Poisson reconstruction algorithm as implemented in open3D (v0.15.1). IVD height was then analyzed by measuring the average distance between the superior and inferior surfaces of each IVD using a custom algorithm which has been previously described and validated [4]. Reliability and precision of automated deformation measurements (L1-L2 – L5-S1; n = 5 IVD pairs) were compared to manually deformations derived from manual (ground truth) image segmentation volumes using an intraclass correlation coefficient (ICC; ICC(2,k)) and standard error of measurement (SE_m (RMSE); equation (5)), respectively. SE_m calculations were derived from the mean-squared error (*MSE*) term of a repeated measures ANOVA to determine differences between automated and ground truth IVD [9].

2.6. Throughput analysis

Each of the five IVD pairs during testing was segmented, processed, and stored 100 times to evaluate the mean time required to segment and surface reconstruct the surface of an IVD pair. Individual segmentations

and deformation analyses were performed in series on CPU (Intel® Xeon® W-2295 CPU; 128 GB RAM).

3. Results

Visual comparisons between ground truth and predicted segmentations can be seen in Fig. 1. Surface reconstructions of the manual and automatically derived segmentations are depicted in Peak model performance was obtained using the 3D U-net architecture kernel size of 3, learning rate of 1e-3 and training batch size of 1 yielding a maximum mDSC of 0.9824 and minimum ASD_i of 0.033–0.068 mm (Table 1). In comparison, the 2D model architecture achieved peak test performance (mDSC = 0.959, ASD_i = 0.054–0.089 mm) using a kernel size of 9, learning rate of 1e-2 and training batch size of 4. Deformation analyses from five IVD pairs (n = 1 subject, L1-L2 – L5-S1) yielded an ICC(2, k) of 0.926 and corresponding SE_m of 0.42% for predictions stemming from the optimized 3D U-Net architecture. Processing time for the segmentation and surface reconstruction of a pre-post-exercise IVD pair took, on average, 4.2 s per pair.

4. Discussion

The present study utilized deep learning to automate the segmentation of IVDs from MR images. We quantified deformation reliability (ICC) and precision (SE_m) in addition to measures of IVD morphology (mDSC, ASD). Using the 3D model, it was observed that mean surface deviations between ground truth and predicted segmentations were minimal, particularly along the superior-inferior boundary of the IVD. This result was reflected in the deformation analysis, which found automated measures to be comparably reliable and precise when compared to ground truth measures. Importantly, this study demonstrated that a relatively simplistic deep learning framework can reliably automate measures of IVD function, improving the throughput of these time-intensive methods.

Morphological accuracy was quantified using two methods, mDSC and the component-wise ASD. Overall, machine learned predictions demonstrated excellent morphological similarity to ground truth, manually derived, segmentations. Encouragingly, both 2D and 3D model

Table 1
IVD morphological accuracy, reliability, and precision comparison by model architecture.

Model	Morphology				Deformation	
	mDSC	ASD_x (mm)	ASD_y (mm)	ASD_z (mm)	ICC(2,k)	SE_m (%)
2D U-Net	*0.959*	*0.054*	*0.089*	–	*0.647*	*0.820*
3D U-Net	***0.982***	***0.068***	***0.034***	***0.033***	***0.926***	***0.420***

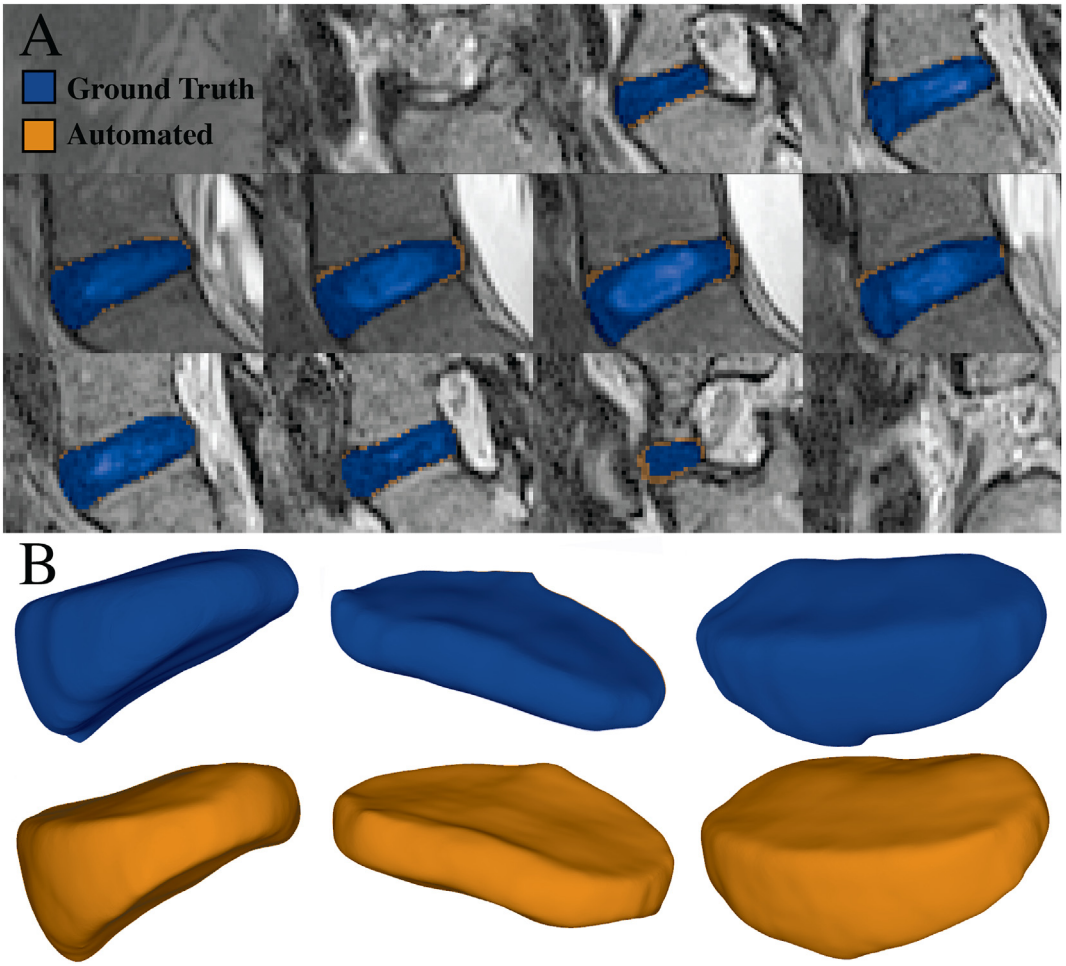

Fig. 1. (A) Segmentation and reconstruction comparison between ground truth (manual segmentation) and predicted (machine learned) IVD masks from the best performing 3D U-Net. Each panel depicts a different slice of the predicted (green) and ground truth (blue) segmentation mask from a single IVD volume. Test performance (mDSC = 0.9824) indicates that 98.24% of voxels in the ground truth are contained in the predicted segmentation. (B) 3D Surface reconstruction comparison. Ground truth (top row) and predicted (bottom row) segmentations from the 3D U-Net are depicted above. Surface reconstructions were performed using a Poisson reconstruction algorithm.

architectures achieved morphological accuracy (Table 1) comparable to those of previously reported autosegmentation-derived models [7,8,10]. Namely, prior works have reported mDSCs of 0.89–0.94[7, 8, 10], whereas the two models developed here achieved mDSCs on the order of 0.96–0.98.

Additionally, we expressed ASD in a component-wise manner using the L_1-norm (as opposed to the more common L_2-norm) to enable directional characterization of the errors in the predicted masks. Overall, while errors along the superior-inferior ($ASD_y = 0.034$ mm) and medial-lateral directions ($ASD_z = 0.033$ mm) were smaller than those in the anterior-posterior direction ($ASD_x = 0.068$ mm), all component-wise ASDs are within 1% of total IVD length along their respective directions. Variations in ASD may be related to differences in contrast between tissues at the IVD boundary which render segmentation more challenging. However, because our measures of IVD function (e.g., uniaxial deformations) rely on precise quantification of IVD height, as opposed to width, minimizing segmentation errors in the superior-inferior direction is more crucial to the immediate success of this automation process [4,5].

The results of the automated deformation analysis supported this finding, whereby the 3D model demonstrated excellent reliability (ICC >0.9) [11] with a corresponding $SE_m = 0.42\%$, which was roughly half that of the 2D model's $SE_m = 0.820\%$. Repeatability of IVD height measurements have previously been estimated to be within 1% of measured IVD height both between- (inter-rater) and within-raters (day-to-day) [4]. Thus, provided that automated segmentation algorithms are deterministic when deployed, the results of the present study suggest that automated evaluation of IVD uniaxial deformations made using this new technique may be comparably precise and potentially more reliable than those made using manual methodologies.

In addition to providing both precise and reliable IVD segmentations, we estimate that this method will greatly improve throughput compared to manual techniques. Currently, careful analysis of a pre- and post-exercise SPACE image set takes, on average, 10 h for manual segmentations alone. In contrast, the present methods took an average of only 21 s (total) to segment, reconstruct and store 10 IVDs, representing an increase in throughput of over 2 orders of magnitude.

To derive these functional models, we exclusively utilize anatomic T2-weighted SPACE images. This sequence provides excellent contrast for identifying IVD morphology and has been utilized to classify the extent of IVD degeneration via Pfirrmann grading [12]. Traditionally, the use of limited scanner and sequence modalities during model development has been seen as a limitation to the external generalizability of a deep learning model. However, because this model is intended to accomplish a specific purpose, the need for this model to be robust to variability imposed by imaging factors (e.g., scanner hardware, sequence modality) outside the scope of its intended use largely mitigates this limitation at present. Nevertheless, future studies building upon this work will aim to improve the generalizability of the model presented here. Specifically, the incorporation of symptomatic subjects, and more extensive model validation (e.g., k-fold cross-validation, model architecture) procedures may help improve the external validity of the current technique. Moreover, because IVD composition is believed to play an integral role in regulating IVD function [13–15] and development of LBP [1–3,14,16], extending the capabilities of this model to process and analyze quantitative MR imaging data (e.g., T1rho and T2map relaxation imaging) is of great interest.

In conclusion, the present study represents a first step towards the development of segmentation models for the evaluation of *in vivo* IVD function. Encouragingly, we demonstrated that IVD function can be precisely and reliably assessed using a relatively simple deep learning framework. Moreover, this fully automated model substantially decreases the time required to analyze *in vivo* IVD deformations in response to exercise, without sacrificing measurement reliability. Thus, this machine learning algorithm has the potential to greatly increase research throughput for investigating IVD function.

Author contributions

JAC: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – Original draft preparation, Writing – Reviewing and Editing; **NEZ:** Validation, Data curation, Writing – Reviewing and Editing; **CES:** Methodology, Investigation, Writing - Review & Editing; **APG:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Methodology, Investigation, Writing - Review & Editing; **LED:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Investigation, Writing - Review & Editing.

Role of the funding source

This work was supported by NIH grants R01AR074800, R01AR065527, R01AR075399, R01AR071440.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank our clinical research coordinator Stephanie Danyluk as well as Jean Shaffer and Raven Boykin at the Duke Center for Advanced Magnetic Resonance Development for their continued help and support.

References

[1] A. Borthakur, P.M. Maurer, M. Fenty, C. Wang, R. Berger, J. Yoder, et al., T1ρ MRI and discography pressure as novel biomarkers for disc degeneration and low back pain, Spine 36 (2011) 2190–2196.
[2] K. Fujii, M. Yamazaki, J.D. Kang, M.V. Risbud, S.K. Cho, S.A. Qureshi, et al., Discogenic back pain: literature Review of definition, diagnosis, and treatment, JBMR Plus 3 (2019), e10180.
[3] A.M.R. Groh, D.E. Fournier, M.C. Battié, C.A. Séguin, Innervation of the human intervertebral disc: a scoping Review, Pain Med. 22 (2021) 1281–1304.
[4] J.A. Coppock, S.T. Danyluk, Z.A. Englander, C.E. Spritzer, A.P. Goode, L.E. DeFrate, Increasing BMI increases lumbar intervertebral disc deformation following a treadmill walking stress test, J. Biomech. 121 (2021), 110392.
[5] J.A. Coppock, N.E. Zimmer, Z.A. Englander, S.T. Danyluk, A.S. Kosinski, C.E. Spritzer, et al., In vivo intervertebral disc mechanical deformation following a treadmill walking "stress test" is inversely related to T1rho relaxation time, Osteoarthritis Cartilage 31 (2022) 126–133.
[6] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention, Pt Iii 9351 (2015) 234–241.
[7] J. Dolz, C. Desrosiers, I.B. Ayed, IVD-net: Intervertebral Disc Localization and Segmentation in MRI with a Multi-Modal UNet, 2018.
[8] C. Wang, Y. Guo, W. Chen, Z. Yu, Fully Automatic Intervertebral Disc Segmentation Using Multimodal 3D U-Net. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), IEEE, 2019.
[9] J.P. Weir, Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM, J Strength Cond Res 19 (2005) 231–240.
[10] G. Zeng, D. Belavy, S. Li, G. Zheng, Evaluation and comparison of automatic intervertebral disc localization and segmentation methods with 3D multi-modality MR images: a grand challenge, in: Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 163–171.
[11] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, Journal of Chiropractic Medicine 15 (2016) 155–163.
[12] C.W. Pfirrmann, A. Metzdorf, M. Zanetti, J. Hodler, N. Boos, Magnetic resonance classification of lumbar intervertebral disc degeneration, Spine 26 (2001) 1873–1878.

[13] G.D. O'Connell, N.T. Jacobs, S. Sen, E.J. Vresilovic, D.M. Elliott, Axial creep loading and unloaded recovery of the human intervertebral disc and the effect of degeneration, J. Mech. Behav. Biomed. Mater. 4 (2011) 933–942.
[14] P.-P.A. Vergroesen, I. Kingma, K.S. Emanuel, R.J.W. Hoogendoorn, T.J. Welting, B.J. Van Royen, et al., Mechanics and biology in intervertebral disc degeneration: a vicious circle, Osteoarthritis Cartilage 23 (2015) 1057–1070.
[15] S. Tavana, S.D. Masouros, N. Baxan, B.A. Freedman, U.N. Hansen, N. Newell, The effect of degeneration on internal strains and the mechanism of failure in human intervertebral discs analyzed using digital volume correlation (DVC) and ultra-high field MRI, Front. Bioeng. Biotechnol. 8 (2021), 610907.
[16] F.-J. Lyu, H. Cui, H. Pan, K.M. Cheung, X. Cao, J.C. Iatridis, et al., Painful intervertebral disc degeneration and inflammation: from laboratory evidence to clinical interventions, Bone Research 9 (2021) 7.