Seminars in Arthritis and Rheumatism 66 (2024) 152420

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

OMERACT validation of a deep learning algorithm for automated absolute quantification of knee joint effusion versus manual semi-quantitative assessment

Banafshe Felfeliyan [a,*], Stephanie Wichuk [a], Abhilash R. Hareendranathan [a], Robert G. Lambert [a,b], Walter P. Maksymowych [c,d], Jacob Jaremko [a,b]

[a] Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, Canada
[b] Medical Imaging Consultants, Edmonton, Alberta, Canada
[c] Department of Medicine, University of Alberta, Edmonton, Canada
[d] CARE Arthritis, Edmonton, Alberta, Canada

ARTICLE INFO

Keywords:
Osteoarthritis
Effusion-synovitis
MRI
OMERACT
Deep learning

ABSTRACT

Objective: To begin evaluating deep learning (DL)-automated quantification of knee joint effusion-synovitis via the OMERACT filter.
Methods: A DL algorithm previously trained on Osteoarthritis Initiative (OAI) knee MRI automatically quantified effusion volume in MRI of 53 OAI subjects, which were also scored semi-quantitatively via KIMRISS and MOAKS by 2–6 readers.
Results: DL-measured knee effusion correlated significantly with experts' assessments (Kendall's tau 0.34–0.43)
Conclusion: The close correlation of automated DL knee joint effusion quantification to KIMRISS manual semi-quantitative scoring demonstrated its criterion validity. Further assessments of discrimination and truth vs. clinical outcomes are still needed to fully satisfy OMERACT filter requirements.

Introduction

Knee effusion-synovitis (E-S) is an important pathophysiological manifestation of arthritis. E-S visible on non-contrast fluid-sensitive MRI sequences represents an attractive target for therapeutic interventions in arthritis, as it is associated with stiffness, pain, and disease progression [1,2]. In OMERACT terms, assessment of E-S may be a useful component for evaluating the inflammation and disease activity in arthritis [3]. This facilitates validating imaging biomarkers against patient-reported outcomes, disease monitoring, treatment evaluation, and clinical research enhancing the arthritis management. The OMERACT Filter 2.1 evaluates imaging instruments as to truth, feasibility and discrimination [3]. Currently, E-S is assessed holistically by a radiologist, or by semi-quantitative tools such as MRI OA Knee Score (MOAKS) [4]; these approaches are susceptible to high inter-user variability, and because MOAKS effusion scores are ordinal from 0 to 3, discrimination is limited to substantial changes. A more granular semi-quantitative scoring system like the Knee Inflammation MRI Scoring System (KIMRISS) is attractive as it provides a broader scoring range [5]. KIMRISS showed less variability and improved discrimination in OA vs. MOAKS [5]. However, KIMRISS manual measurement is user-dependent and time-consuming in large datasets [5], limiting feasibility. Volumetric quantitative measurement (VQM), which determines effusion volume directly by counting MRI voxels, has high face validity but is tedious when performed manually. At OMERACT 2021 it was shown that artificial intelligence (AI), particularly a deep learning (DL) algorithm, could automate VQM with high correlation to human expert measurements at the hip [6].

Most existing DL techniques rely on extensive training on large datasets, meticulously annotated with ``ground-truth'' labels. For effusion detection, accurate labeling is a costly process limiting feasibility, since it requires expertise: joint fluid pools in complex articular recesses and volume-averaging artifacts are common.

At OMERACT 2023, within the OMERACT MRI in Arthritis Working Group we began applying the OMERACT filter [3] to the automated DL-based joint effusion VQM (DL-ES-VQM) at the knee.

The aim is to evaluate if DL-ES-VQM meets OMERACT standards for feasibility and convergent validity.

[*] Corresponding author.
E-mail address: banfel@ualberta.ca (B. Felfeliyan).

https://doi.org/10.1016/j.semarthrit.2024.152420

Available online 17 February 2024

Materials and methods

We designed our evaluation of DL-ES-VQM to conform to the OMERACT Filter 2.1 Instrument Selection Algorithm (OFISA) [7]. OFISA includes the conventional components found in the original OMERACT Filter [3], namely Truth, Feasibility, and Discrimination [7]. The target domain of E-S is well-established for MRI-based instruments in OA [8,9] and addressed through construct/criterion validity testing using the KIMRISS score [5]. Specifically, automated DL effusion measurement ought to be feasible with small labeled training data sets, closely approximate imaging truth vs. human expert effusion measurement [10], demonstrate high discrimination, and show evidence it reflects the target domain via correlation to clinical measures of life impact [8]. This exercise mainly serves to address feasibility and preliminary criterion validity.

Materials

We employed MOAKS [4] and KIMRISS [5] semi-quantitative scoring systems to validate the results obtained from the DL approach. MOAKS evaluates effusion inflammatory phenotypes using effusion-synovitis (E-S) and Hoffa's Synovitis (H-S) scores, derived respectively from axial and sagittal views, with a score range of 0–3 based on size. KIMRISS, an OMERACT-validated system, scores E-S and H-S in more granular fashion, assessing presence and size per MRI slice on a 0–4 scale, for a range of 0–100 for 25 slices.

The Data for DL training and assessment were retrieved from the publicly available Osteoarthritis Initiative (OAI) dataset (http://www.oai.ucsf.edu/), including baseline and 1-year Sagittal IW TSE (intermediate weighted turbo spin-echo) MRI of 53 randomly selected subjects with baseline MOAKS inflammation score >1. Scores was evaluated at baseline and one-year follow-up using KIMRISS (by 2–6 readers) and MOAKS (by 1–2 readers). Readers were blinded to chronology of scans. Our study utilized summations of MOAKS inflammatory phenotypes' paired scores (E-S + H-S scores) for statistical analysis. For subjects with multiple scores available for the same MRI, the mean value was used for analysis.

Deep learning tool

We deployed a DL approach called Self Supervised-MRCNN (SS-MRCNN) developed for DL-ES-VQM [11]. The SS-MRCNN was developed to address the challenge of training with very limited labeled data, using a unique two-phase training approach built upon Mask R-CNN [12] DL architecture: self-supervised learning (SSL), and fine-tuning.

The SSL is a branch of DL, that aims to learn data representation from the data itself by learning to perform some auxiliary (pre-text) tasks related to the downstream task (E-S segmentation). During the SSL pre-training phase, we exposed the model to unlabeled knee MRI scans distorted in arbitrary area. The network aims to learn valuable visual features by comparing the distorted images with the original ones and learning to recognize and correcting these distortions within the images. The model learns to ``see through'' the distortions to understand the scans data distribution. For the fine-tuning phase, we trained the network using selected slices from 23 scans from OAI data (distinct from the 53 scans in the test set), focusing on effusion presence labeled by expert radiologists.

Statistics

Criterion validity of DL-ES-VQM was assessed through a comparison of baseline and 1-year change (=Δ) in volume with MOAKS E-S and KIMRISS E-S using Kendall's tau correlations [13] to account for non-normality in population values and handle the presence of outliers in the data [14]. Differences in DL-ES-VQM between Kelgren–Lawrence (KL) grades were evaluated via Kruskal–Walli's test (Table 1).

Table 1
Characteristics of cases whose scans were evaluated in this exercise.

Variable		Baseline					1 year follow-up
Age		62.0 (8.5)					
Males no. (%)		18 (34.0 %)					
Scores:KL grade no. (%)	***Scores***	***0***	***1***	***2***	***3***	***4***	
	no.	6	15	21	5	6	–
	%	11.3 %	28.3 %	39.6 %	9.4 %	11.3 %	
		Mean (SD) [median (IQR)]					**Mean (SD) [median (IQR)]**
KIMRISS	Effusion	36.2 (14.9) [34.8 (25.4–48.6)]					35.7 (14.6) [34.4 (24.2–46.8)]
MOAKS	ES	1.5 (0.8) [1.0 (1.0–2.0)]					1.6 (0.9) [2.0 (1.0–2.0)]
	HS	1.1 (1.0) [0.7 (1.0–2.0)]					1.1 (0.7) [1.0 (1.0–2.0)]

Results

The DL-ES-VQM calculation took 35 s/scan. The qualitative outcomes were visually congruent with human segmentation (Fig. 1).

Effusion volumes were relatively high in our data set, and consistent between methods (Table 2). Median (IQR) volume by DL-ES-VQM at baseline was 14.8 (9.3–26.0) ml, with a median (IQR) 1-year Δ of 0.51 (−5.6–6.1) ml. DL-ES-VQM was highest in cases with KL = 2 [median (IQR) 20.4 (9.5–26.8)] and lowest in cases with KL = 0 or 4 [median (IQR) 12.4 (9.4–14.8) and 9.9 (6.0–18.9), for grades 0 and 4, respectively]. Differences in DL-ES-VQM between KL grades did not reach significance in this dataset ($p = 0.638$).

There was significant moderate positive correlation between DL-ES-VQM and MOAKS E-S baseline and change scores [Kendall's tau (95 % CI) 0.35 (0.17–0.52); $p = 0.003$ and 0.24 (0.03–0.5), $p = 0.010$ for baseline and change, respectively]. DL-ES-VQM was also strongly correlated with KIMRISS baseline and change scores [Kendall's tau 0.58 (0.43–0.66), $p < 0.0001$ and 0.61 (0.40–0.75), $p < 0.0001$ for baseline and change, respectively]. A comparison between distributions of the DL-ES-VQM, the KIMRISS effusion score, and the corresponding MOAKS score presented in Fig. 2 also is another indication of the strong correlation between KIMRISS and DL predicted values.

Discussion

We found that automatic knee joint effusion measurement by DL (DL-ES-VQM) was feasible by our self-supervised tool using only minimal labeled training data (selected slices from only 23 knees). These DL-ES-VQM measurements correlated closely to manual semi-quantitative effusion scoring (MOAKS, KIMRISS), and showed the expected significant association between higher radiographic KL grades of OA (3–4) and larger effusions. This supports feasibility and criterion validity of the proposed fully automated method for quantifying knee effusion, two important parts of the OMERACT filter.

Our DL approach uses an MRI sequence routinely acquired in clinical knee MRI (sagittal TSE). Training AI models requires a GPU-enabled computer, however once network is trained a typical desktop computer can generate results rapidly in less than one second per slice. New AI ecosystems support model deployment across various systems. With commonly available resources, joint fluid volume could potentially become routine in arthritis MRI reporting.

Our DL method yields a quantitative value (in mL) for effusion volume, offering more direct effusion assessment than semi-quantitative scoring (MOAKS, KIMRISS), potentially enhancing discrimination between groups as per the OMERACT filter requirements; however,

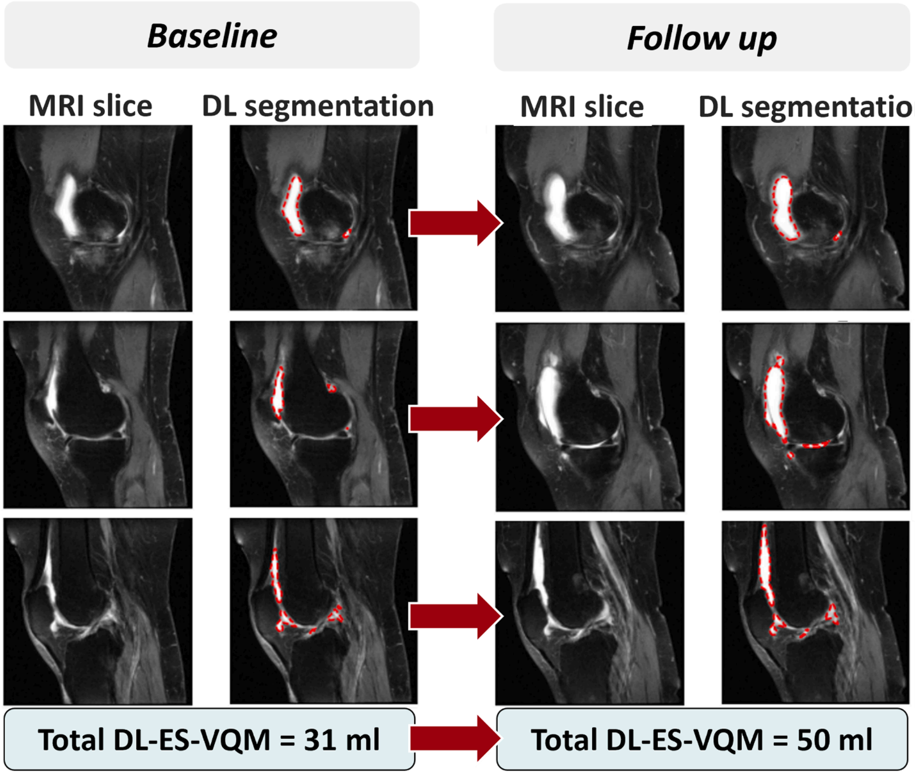

Fig. 1. Deep learning effusion segmentation results for a subject at baseline and follow-up (follow-up scan registered to baseline to illustrate detail of DL-ES-VQM's precision).

Table 2

Statistical Analysis results. (a) DL-ES-VQM values in different KL scores. (b) Kendall's Tau correlation analysis between different effusion assessment methods—DL-ES-VQM, MOAKS ES, and KIMRISS—at both baseline and change within the one-year follow-up.

(a)

Time-point	Variable	All cases *[median (IQR)]*	Kellgren–Lawrence grade at baseline *[median (IQR)]*			
			0 ($n = 6$)	1 ($n = 15$)	2 ($n = 21$)	≥3 ($n = 11$)
Baseline	DL-ES-VQM	14.8 (9.3–26.0)	12.4 (9.4–14.8)	14.7 (9.9–28.5)	20.4 (9.5–26.8)	12.1 (6.4–18.7)
1y Change	Δ DL-ES-VQM	0.5 (−5.6 to 6.1)	4.0 (1.3–8.6)	−3.3 (−15.4–4.5)	1.6 (−4.2–5.6)	0.2 (−10.1–7.4)

(b)

Time-point	Variable1	Variable2	Kendall's tau (95 % CI)	*p*-value	*Correlation status
Baseline	DL-ES-VQM	MOAKS ES	0.35 (0.17–0.52)	0.0003	Significant moderate positive
1y Change	ΔDL-ES-VQM	ΔMOAKS ES	0.24 (0.02–0.45)	*0.010*	Significant moderate positive
Baseline	DL-ES-VQM	KIMRISS	0.58 (0.43–0.66)	<0.0001	Significant moderate positive
1y Change	ΔDL-ES-VQM	ΔKIMRISS	0.61 (0.40–0.75)	<0.0001	Significant strong positive
Baseline	KIMRISS	MOAKS	0.35 (0.16–0.51)	0.0002	Significant moderate positive
1y Change	Δ KIMRISS	ΔMOAKS	0.17 (−0.03–0.37)	0.075	Not significant

uncertainty remains regarding measurement errors margin in this study [15]. Furthermore, sensitivity to change and clinical relevance were not readily assessed in the available observational OAI data set because of the lack of disease-modifying interventions. Hence assessment of discrimination requires additional study in clinical-trial data sets.

The impact of knee arthritis involves various factors, with knee effusion volume being just one aspect. When utilizing tools like multivariate regression, the diverse scoring ranges of DL-ES-VQM, KIMRISS, and MOAKS limit direct reliability comparisons based on ICC measurements, especially for criterion validity within the truth. Additional statistical and clinical evaluation, including comparison with manual segmentation, are required in larger datasets to assess the DL-ES-VQM criterion validity fully and to evaluate the technique's reliability across different MRI sequences. Our study demonstrates alignment with OMERACT filter components, including Feasibility, Criterion Validity, Construct Validity, and Sensitivity to Change, limitations hinder determining satisfaction of the truth component to its full extent.

Conclusion

The automated knee joint effusion measurement (DL-ES-VQM) was highly feasible with a self-supervised AI network and showed high criterion validity vs. established joint effusion scoring instruments in OA. Ultimately, such a tool could efficiently measure effusions in the millions of knee MRI performed annually worldwide, with potential to enhance both clinical care and clinical trials in arthritis.

Future work to complete the evaluation of this tool via the OMERACT filter will include additional assessment of truth in terms of life impact (clinical outcomes) and of discrimination (clinical-trial data).

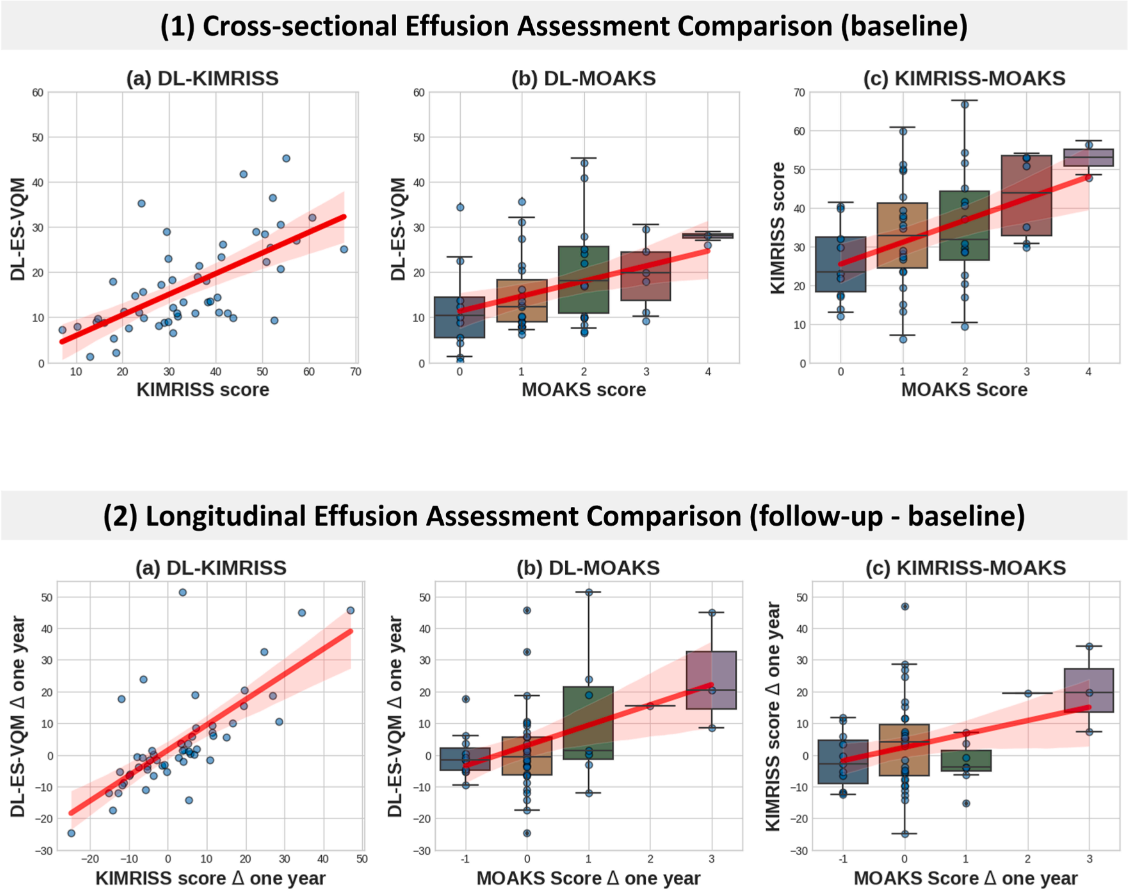

Fig. 2. (1) Correlation plots and boxplots between the DL-ES-VQM prediction and scoring tools, (1.a) DL-ES-VQM vs. KIMRISS effusion score, (1.b) DL-ES-VQM vs. MOAKS effusion + Hoffa-synovitis score, box plots show DL effusion prediction volume distribution across MOAKS score, and (1.c) KIMRISS effusion score vs. MOAKS effusion + Hoffa-synovitis score. (2) Correlation plots and boxplots between the one-year change DL-ES-VQM prediction and scoring tools (2.a) DL-ES-VQM vs. KIMRISS effusion score, (2.b) DL-ES-VQM vs. MOAKS effusion + Hoffa-synovitis score, and (2.c) KIMRISS effusion score vs. MOAKS effusion + Hoffa-synovitis score.

CRediT authorship contribution statement

Banafshe Felfeliyan: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Stephanie Wichuk:** Data curation, Formal analysis, Methodology, Validation. **Abhilash R. Hareendranathan:** Conceptualization, Writing – review & editing, Funding acquisition. **Robert G. Lambert:** Writing – review & editing, Validation. **Walter P. Maksymowych:** Writing – review & editing, Validation. **Jacob Jaremko:** Supervision, Project administration, Funding acquisition, Conceptualization, Writing – review & editing, Investigation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jacob Jaremko reports financial support was provided by Medical Imaging Consultants (MIC), Edmonton, Canada. Banafshe Felfeliyan reports on a relationship with Alberta Innovates that includes funding grants. Walter P. Maksymowych is Chief Medical Officer CARE Arthritis Limited. Robert G. Lambart has received consulting fees from CARE Arthritis, Image Analysis Group and Calyx. We thank the members of the OMERACT MRI in Arthritis Working Group for their participation and support in this project.

Acknowledgment

BF is supported by an Alberta Innovates Postdoctoral Fellowship for Data-Enabled Innovation. JJ is supported by a Canada CIFAR AI Chair. Academic time for JJ, RL and VQL is made available by Medical Imaging Consultants (MIC), Edmonton, Canada. Part of this project is funded by the Alberta Machine Intelligence Institute. We thank Dr. Janet Ronsky and the University of Calgary for providing the resources to develop the DL algorithm. We thank the members of the OMERACT MRI in Arthritis Working Group for their participation and support in this project. We also thank the participants of the Special Interest Group (SIG) for their insights and feedback.

References

[1] Roemer FW, et al. Structural phenotypes of knee osteoarthritis: potential clinical and research relevance. Skeletal Radiol 2022:1–10. https://doi.org/10.1007/s00256-022-04191-6.

[2] Van Spil WE, Kubassova O, Boesen M, Bay-Jensen AC, Mobasheri A. Osteoarthritis phenotypes and novel therapeutic targets. Biochem Pharmacol 2019;165:41–8. https://doi.org/10.1016/j.bcp.2019.02.037. Elsevier.

[3] Boers M, et al. OMERACT filter 2.1: elaboration of the conceptual framework for outcome measurement in health intervention studies. J Rheumatol 2019;46(8): 1021–7. https://doi.org/10.3899/JRHEUM.181096.

[4] Guermazi A, Roemer FW, Haugen IK, Crema MD, Hayashi D. MRI-based semiquantitative scoring of joint pathology in osteoarthritis. Nat Rev Rheumatol 2013;9(4):236–51. https://doi.org/10.1038/nrrheum.2012.223.

[5] Jaremko JL, et al. Validation of a knowledge transfer tool for the knee inflammation MRI scoring system for bone marrow lesions according to the OMERACT filter: data from the osteoarthritis initiative. J Rheumatol 2017;44(11): 1718–22. https://doi.org/10.3899/jrheum.161102.

[6] Jaremko JL, et al. Volumetric quantitative measurement of hip effusions by manual versus automated artificial intelligence techniques: an OMERACT preliminary validation study. Semin Arthritis Rheum 2021. https://doi.org/10.1016/j.semarthrit.2021.03.009.

[7] D'Agostino MA, et al. Improving domain definition and outcome instrument selection: lessons learned for OMERACT from imaging. Semin Arthritis Rheum 2021;51(5):1125–33. https://doi.org/10.1016/j.semarthrit.2021.08.004.

[8] Hunter DJ, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). Osteoarthr Cartil 2011;19(8):990–1002. https://doi.org/10.1016/j.joca.2011.05.004.

[9] Peterfy CG, et al. Whole-organ magnetic resonance imaging score (WORMS) of the knee in osteoarthritis. Osteoarthr Cartil 2004;12(3):177–90. https://doi.org/10.1016/j.joca.2003.11.003.

[10] Antoniou T, Mamdani M. Evaluation of machine learning solutions in medicine. CMAJ 2021;193(36):E1425–9. https://doi.org/10.1503/CMAJ.210036/TAB-RELATED-CONTENT.

[11] Felfeliyan B, et al. Self-supervised-RCNN for medical image segmentation with limited data annotation. Comput Med Imaging Graph 2023;109. https://doi.org/10.1016/j.compmedimag.2023.102297.

[12] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 2020;42(2):386–97. https://doi.org/10.1109/TPAMI.2018.2844175.

[13] Kendall MG. Rank correlation methods. 4th ed. Griffin; 1976.

[14] Bland M. An introduction to medical statistics. Oxford University Press, Incorporated; 2015.

[15] McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA 2014;312(13):1342–3. https://doi.org/10.1001/JAMA.2014.13128.