BJR https://doi.org/10.1259/bjr.20210141

Received: 26 January 2021 | Revised: 20 June 2021 | Accepted: 14 July 2021

Cite this article as:
Schomöller A, Risch L, Kaplick H, Wochatz M, Engel T, Schraplau A, et al. Inter-rater and inter-session reliability of lumbar paraspinal muscle composition in a mobile MRI device. *Br J Radiol* 2021; **94**: 20210141.

FULL PAPER

Inter-rater and inter-session reliability of lumbar paraspinal muscle composition in a mobile MRI device

ANNE SCHOMÖLLER, LUCIE RISCH, HANNES KAPLICK, MONIQUE WOCHATZ, TILMAN ENGEL, ANNE SCHRAPLAU, DOMINIK SONNENBURG, ALEXANDER HUPPERTZ and FRANK MAYER

University of Potsdam, University Outpatient Clinic, Sports Medicine and Sports Orthopaedics, Potsdam, Germany

Address correspondence to: Mrs Anne Schomöller
E-mail: *anne-schomoeller@gmx.de*

Objective: To assess the reliability of measurements of paraspinal muscle transverse relaxation times (T2 times) between two observers and within one observer on different time points.
Methods: 14 participants (9f/5m, 33 ± 5 years, 176 ± 10 cm, 73 ± 12 kg) underwent 2 consecutive MRI scans (M1,M2) on the same day, followed by 1 MRI scan 13–14 days later (M3) in a mobile 1.5 Tesla MRI. T2 times were calculated in T_2 weighted turbo spin-echo-sequences at the spinal level of the third lumbar vertebrae (11 slices, 2 mm slice thickness, 1 mm interslice gap, echo times: 20, 40, 60, 80, 100 ms) for M. erector spinae (ES) and M. multifidius (MF). The following reliability parameter were calculated for the agreement of T2 times between two different investigators (OBS1 & OBS2) on the same MRI (inter-rater reliability, IR) and by one investigator between different MRI of the same participant (intersession variability, IS): Test–Retest Variability (TRV, Differences/Mean*100); Coefficient of Variation (CV, Standard deviation/Mean*100); Bland–Altman Analysis (systematic bias = Mean of the Differences; Upper/Lower Limits of Agreement = Bias+/−1.96*SD); Intraclass Correlation Coefficient 3.1 (ICC) with absolute agreement, as well as its 95% confidence interval.
Results: Mean TRV for IR was 2.6% for ES and 4.2% for MF. Mean TRV for IS was 3.5% (ES) and 5.1% (MF). Mean CV for IR was 1.9 (ES) and 3.0 (MF). Mean CV for IS was 2.5% (ES) and 3.6% (MF). A systematic bias of 1.3 ms (ES) and 2.1 ms (MF) were detected for IR and a systematic bias of 0.4 ms (ES) and 0.07 ms (MF) for IS. ICC for IR was 0.94 (ES) and 0.87 (MF). ICC for IS was 0.88 (ES) and 0.82 (MF).
Conclusion: Reliable assessment of paraspinal muscle T2 time justifies its use for scientific purposes. The applied technique could be recommended to use for future studies that aim to assess changes of T2 times, *e.g.* after an intense bout of eccentric exercises.

INTRODUCTION

Transverse relaxation time (T2 time) is defined by the duration that it takes the magnetic resonance signal in the transverse plane to reach 37% of the original signal.[1] Signal intensities and thereby calculated T2 times in MRI deliver information about morphology and intramuscular tissue constitution[2] and are sensitive to changes of fluid concentration[1] or fatty infiltration of a muscle.[2] In a group of patients with unilateral acute or chronic low back pain (LBP), higher signal intensities were found in the painful side for M.erector spinae (ES) and M.multifidius (MF) in chronic LBP patients.[3] On the other hand, T2 time also increases after intense bouts of eccentric ecercises due to local inflammatory muscle edema.[4,5] Increases of 187% in T2 times were detected 24 h and 7 days post-eccentric exercise in M. semitendinosus[5] as well as 24 h, 48 h and 72 h after eccentric quadriceps exercises (7% increase).[4]

Methodological differences, such as inclusion or exclusion of fatty infiltration within a muscle, the use of different software tools and analysis methods and doubtful reliability in clinical imaging analysis between different observers exist in MRI analysis. This makes the comparison between studies difficult. However, reliability and standardisation of MRI recording, positioning of the participant and scan analysis is crucial and should be discussed with respect to the sequences used, as well as to the location adressed. The agreement within one and between two observers was assessed in previous studies.[2,6,7] A study including fatty infiltrated cells in the analysis found high intrarater reliability of T2 times for MF (ICC:0.93, SEM:5.01) and ES (ICC:0.96, SEM:4.73) and a marginally weaker interrater reliability (MF:ICC:0.89, SEM:6.56; ES: ICC:0.92, SEM: 5.96) in chronic LBP patients.[2] Intra- and interrater reliability was tested for signal intensities in MF in

non-clinical populations and delivered satisfying results as well (ICC:0.89–0.96).[6,7]

Additionally to observer-dependency, day-to-day variability of the MRI scanner or daily physiological variations within the participant might contribute to differing intersession and interday results, which is specifically detrimental in longitudinal study designs or intersubject comparisons. To the best of our knowledge, no study evaluated the influence of daily measurement variations of mobile MRI devices so far, although mobile measurement tools have already been used for scientific purposes.[8–11] A mobile MRI truck was used in an observational cohort study that tracked adaptive responses in tissues (*e.g.* muscles, tendons), organs (*e.g.* heart) and of body composition during the course of an ultra-long-distance running race[8] and in studies dealing with different clinical populations (Multiple Sclerosis patients,[9,10] patients with spinal stenosis[11]). As location-independence and the opportunity to reach a broader patient/participant clientele with mobile measurement tools are advantageous, specifically in rural areas, more focus should be put on reliability testing of mobile research devices.

Consequently, the aim of this study was to analyse the reliability of paraspinal muscle T2 times between two observers and within one observer on different measurement occasions in a mobile measurement MRI device.

METHODS

Participants

14 participants were included in this study. Anthropometric data are displayed in Table 1. Participants were included when matching the criteria of >18 years of age, absence of structural spinal pathology and no infection during the week prior to both measurement days.

Participants were advised to maintain their usual physical activity habits 4 days prior to MRI measurement days to avoid activity-dependent changes in muscle constitution. The participants were informed about all procedures in the MRI truck and provided informed consent on their first visit to the Outpatient Clinic. All procedures were conducted in line with ethical standards. A physician excluded participants with contra indications for the MRI scan prior to each MRI-measurement day.

STUDY DESIGN

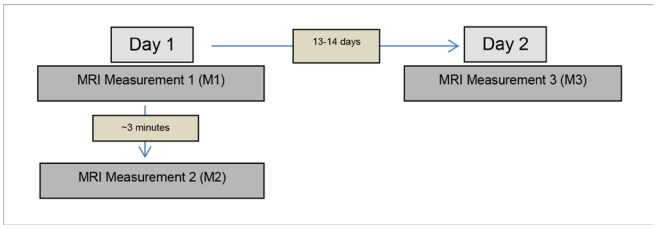
On Day 1, the participants were placed in the MRI in supine position twice (for measurements M1 and M2) with an interval of ~5 min between subsequent measurements. The participants stood up between both session and were repositioned by the same examiner for the second scan.

Table 1. Sample characteristics.

Participants	Age (y)	Body height (cm)	Body weight (kg)
14 (9f,5m)	33.4 ± 4.9	175.8 ± 10.0	73.3 ± 11.9

cm, centimeters; f, female; kg, kilograms; m, male; y, years of age. Anthropometrics are displayed with means ± standard deviations.

Figure 1. Flow chart of the study design.



On Day 2, the participants were measured once (measurement M3) with the identical settings as on Day 1. Time frames between Day 1 and Day 2 were 13–14 days (compare Figure 1).

MRI measurements

MRI was performed in a mobile 1.5 Tesla large bore (70 cm diameter) system (Ingenia, Philips Healthcare, Best, The Netherlands), equipped with a high-performance gradients system (maximal slew rate 120 T/m/s, maximal amplitude 33 mT/m). The system was installed in a dedicated trailer (Lamboo Medical, Zoetermeer, The Netherlands) and was moved on a daily base within the federal state (Figure 2). Between the repetitive measurements M1/M2 and M3, the trailer was moved eight times and travelled approximately 800 km between 5 different hospitals.

MRI was acquired on the spinal level of the third lumbar vertebrae using a surface coil (ds flex coverage posterior coil, Philips). T_2 weighted turbospin echo (TSE) sequences (repetition time: 2500 msec, Flip angle: 90°, field of view: 155 × 198 mm, aquisition matrix: 156 × 146, slices: 11, slice thickness: 2 mm, intersection gap: 1 mm, slice package: 32 mm, echo times: 20, 40, 60, 80, 100 ms) were used. Volume shim was activated before each MRI measurement and a saturation pulse was locally placed on ventral organs to exclude interfering signals. For standardisation, scans always started with the first slice covering the inferior vertebral endplate of the third lumbar vertebrae.

MRI scan analysis

The software ImageJ was used for analysis. T2 times were analysed in slices 1, 3, 5, 8 and 11. As preliminary data of three participants revealed no difference of T2 times in adjacent slices, just 5 out of 11 slices were analysed. The respective slices were chosen because they altogether cover the whole height of the third lumbar vertebrae. T2 times were assessed in the right body side. T2 picture generation and region of interest (ROI) definition were conducted by two different observers (OBS1, OBS2) in the five slices. Both raters were blinded to participants and to each other's results. Analysis was applied by both observers to all scans. Figure 3a displays an exemplary MRI scan image of slice one at echo time 20 ms. The threshold between muscle tissue and other tissue was set manually in all five echo time scans by both raters in order to just include muscle tissue in the T2 time calculation (3b). Figure 3b,c shows the differentiation between muscle tissue (included in T2 time calculation) and other tissue. After setting the threshold as exemplary described in all five echo time scans (20, 40, 60, 80, 100 ms) in the respective slice, a T2 picture of the MRI scan was generated (3e) with use of a calculation plug in for MRI analysis. Polygonal shaped ROIs were drawn in the

Figure 2. Truck trailer with mobile 1.5 T MRI system.

raw MRI scan (3d) in ES (including M. iliocostalis lumborum and M. longissimus lumborum) and MF and copied into the plug-in generated T2-time picture (3e).

STATISTICS

Descriptive and inferential analysis of reliability parameters were calculated in Microsoft Excel and SPSS [Test–Retest Variability (TRV) = Differences/Mean*100; Coefficient of Variation (CV) = Standard deviation/Mean*100; Bland–Altman Analysis: systematic bias = Mean of the Differences; Upper/Lower Limits of Agreement = Bias+/−1.96*SD]. Intraclass Correlation Coefficient 3.1 (ICC) with absolute agreement was calculated in SPSS, as well as the ICC's 95% confidence interval. Means and standard deviations of the reliability parameters result from the mean of the parameters of the five slices (slices 1, 3, 5, 8, 11).

Table 2. T2 times of ES and MF on measurement occasion M1-M3.

Measurement (rater)	T2 time (in ms) M. Erector spinae (ES)	T2 time (in ms) M. Multifidius (MF)
M1 (OBS2)	53.5 ± 3.9	58.8 ± 4.4
M1 (OBS1)	54.6 ± 4.2	60.8 ± 5.0
M2 (OBS1)	53.7 ± 3.2	60.4 ± 5.0
M3 (OBS1)	54.6 ± 3.7	60.8 ± 5.2

ES, M.erector spinae; MF, M.multifidius; OBS, observer.
T2 times are displayed with means ± standard deviations in milliseconds (ms).

RESULTS

The means and standard deviations of the assessed T2 times for ES and MF are displayed in Table 2. Higher T2 times were found in MF compared to ES at all measurement times (M1–M3) and by both investigators (OBS1 and OBS2). For ES, T2 time was lowest at M1 (assessed by OBS2) with 53.5 ± 3.9 ms, followed by M2 (OBS1) with 53.7 ± 3.2 ms, M3 (OBS1) with 54.6 ± 3.7 ms and M1 (OBS1) with 54.6 ± 4.2 ms. For MF, T2 time was also lowest at M1 (OBS2) with 58.8 ± 4.4 ms, followed by M2 (OBS1) with 60.4 ± 5.0 ms and M1 (OBS1) with 60.8 ± 5.0 ms and M3 (OBS1) with 60.8 ± 5.2 ms.

Table 3 displays the reliability parameters (for IR and IS) TRV, CV, and ICC with its confidence interval. For IR and for IS, the reliability for the assessments of T2 time was better for ES

Figure 3. MRI scan processing. a: Raw axial MRI scan at slice one at echo time 20 ms, b: threshold adaptation with use of pixel distribution map, differentiating between muscle tissue (included in T2 time calculation) and other tissue, c: graphic display of pixel distribution in b; pixel in square are included in T2 time calculation, d: exemplary ROIs for MF and ES, drawn in Figure 3a and copied into figure 3e, e: T2 time mapped picture; bright: included tissue for T2 time analysis of ES and MF; dark: excluded tissue. ROI, region of interest.

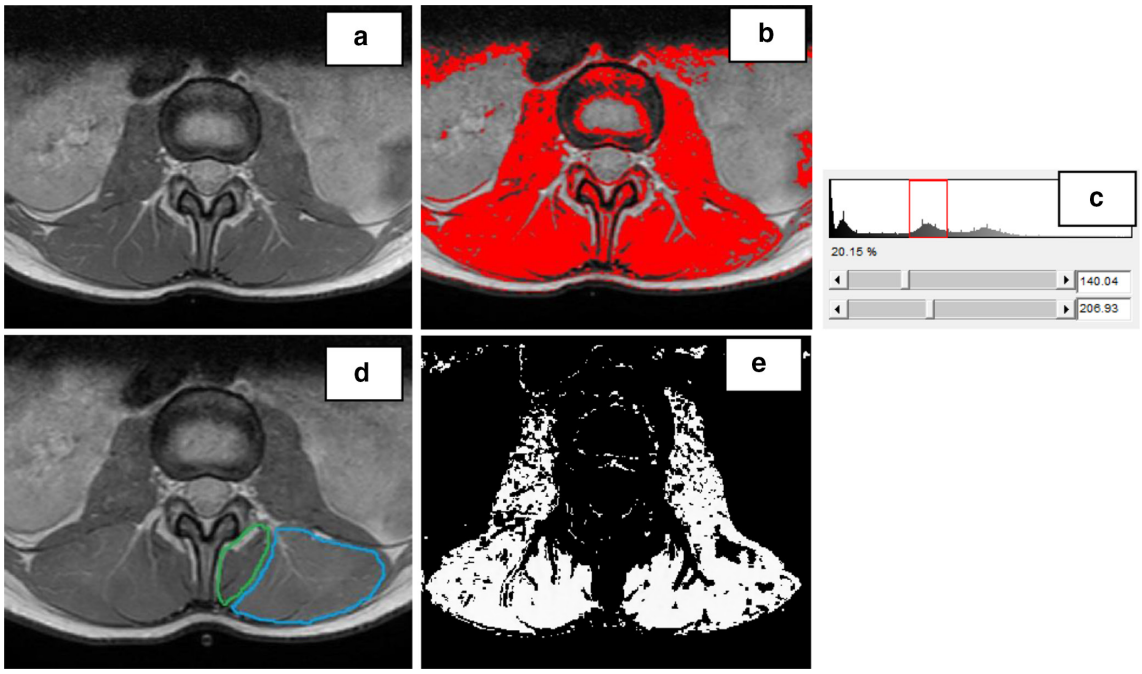


Table 3. Reliability parameters.

Reliability	Measurements (rater)	Muscle	Test–Retest Variability (TRV in %)	Coefficient of Variation (CV in %)	Intraclass Correlation Coefficient (ICC)	Confidence Interval (CI) of ICC
Inter-rater reliability (**IR**)	M1(OBS1)-M1(OBS2)	ES	2.6 ± 1.3	1.9 ± 0.9	0.94	0.70–0.98
		MF	4.2 ± 1.6	3.0 ± 1.1	0.87	0.44–0.95
Intersession reliability 1 (**IS1**)	M1(OBS1)-M2(OBS1)	ES	3.6 ± 1.8	2.5 ± 1.3	0.87	0.78–0.92
		MF	5.3 ± 2.0	3.7 ± 1.4	0.82	0.69–0.89
Intersession reliability 2 (**IS2**)	M1(OBS1)-M3(OBS1)	ES	3.5 ± 1.5	2.5 ± 1.1	0.88	0.81–0.92
		MF	4.9 ± 2.2	3.5 ± 1.6	0.81	0.69–0.88

CI, Confidence Interval; CV, Coefficient of Variation; ICC, Intraclass Correlation Coefficient; OBS1, observer 1; OBS2, observer 2; TRV, Test-Retest Variability.
Inter rater reliability (IR) and session-to-session variability (IS) of T2 times for M. erector spinae (ES) and M. multifidius (MF).

compared to MF, shown by all reliability parameter. IR delivered more reliable results (TRV: 2.6 ± 1.3 (ES); 4.2 ± 1.6 (MF), CV: 1.9 ± 0.9 (ES); 0.870 (MF), ICC: 0.944 (ES); 0.870 (MF)) compared to IS1 (TRV: 3.6 ± 1.8 (ES); 5.3 ± 2.0 (MF), CV: 2.5 ± 1.3 (ES); 3.7 ± 1.4 (MF), ICC: 0.869 (ES); 0.869 (MF)) and to IS2 (TRV: 3.5 ± 1.5 (ES); 4.9 ± 2.2 (MF), CV: 2.5 ± 1.1 (ES); 3.5 ± 1.6 (MF), ICC: 0.884 (ES); 0.813 (MF)). For IS, the reliability between two scans on the same day (IS1) was similar to the reliability of two scans with a time frame of two weeks between the scans (IS2).

For IR, Blant–Altman Analysis revealed a systematic bias of 1.3 ms (ES) and 2.1 ms (MF). For IS1, a systematic bias of 0.5 ms (ES) and 0.03 ms (MF) was detected. For IS2, a systematic bias of 0.2 ms (ES) and 0.1 ms (MF) was found. The random error of the assessments of IR and IS and for both muscles is shown by the lower and upper limits of agreement (LoA) in Figure 4a–f.

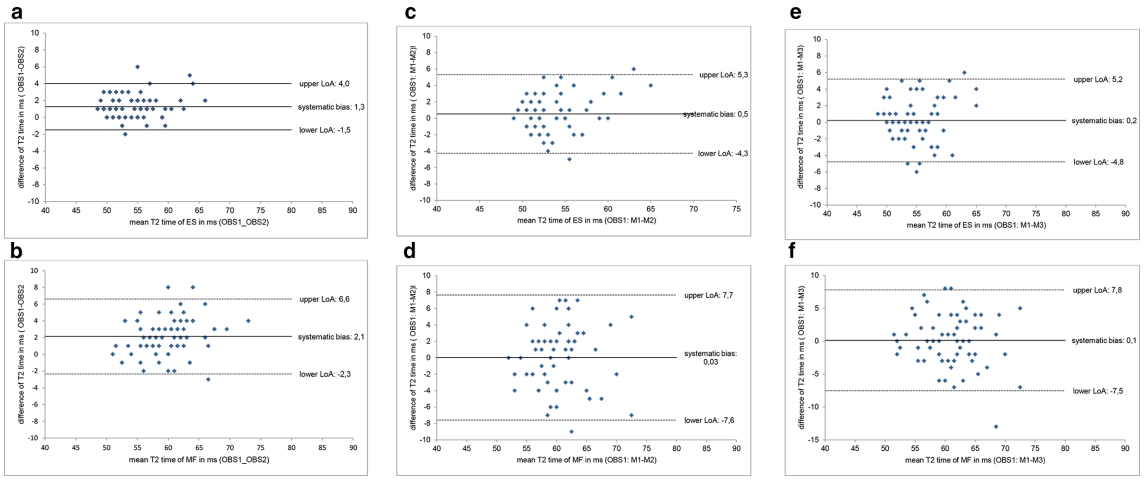
DISCUSSION

This study yielded at determining the inter-rater reliability and intersession reliability of quantitative paraspinal muscle constitution. Highly reliable results were found for IR, indicated by a low TRV (ES/MF: 2.6 ± 1.3/4.2 ± 1.6), a low CV (ES/MF: 1.9 ± 0.9/3.0 ± 1.1) and a high ICC (ES/MF: 0.94/0.87). Slightly weaker, but still reliable results were found for IS1 (TRV: ES/MF: 3.6 ± 1.8/5.3 ± 2.0; CV: 2.5 ± 1.3/3.7 ± 1.4; ICC: 0.87/0.82) and for IS2 (TRV: ES/MF: 3.5 ± 1.5/4.9 ± 2.2; CV: 2.5 ± 1.1/3.5 ± 1.6; ICC: 0.88/0.81).

For IR, a systematic bias of 1.3 ms (ES) and 2.1 ms (MF) was found, whereas there was no relevant systematic bias for IS1 and IS2.

Homogenous T2 times were found within the sample for ES and MF with low variability. One explanation for this homogeneity

Figure 4. a. Bland–Altman Plot for IR of ES at measurement 1. b. Bland–Altman Plot for IR of MF at M1. c. Bland–Altman Plot for intersession variability of the same test day (IS1) of ES. d. Blant–Altman Plot for intersession variability of the same test day (IS1) of MF. e. Bland–Altman Plot for intersession variability of different test days (IS2) of ES. f. Bland–Altman Plot for intersession variability of different test days (IS2) of MF. ES, M. erector spinae; IR, inter-rater reliability; MF, M. multifidius,



could be the methodological approach of solely including muscle cells in T2 time calculations. Still, the pixel distribution-based differentiation between muscle and fat cells are for some pixels subject of interpretation as borders from muscle to fat-infiltrated cells are not clearly defined. This lack of definition might explain the systematic error of 1.3 ms (ES) and 2.1 ms (MF) between the two raters. Also, the slightly better reliability for T2 time in ES compared to MF possibly mirrors the difficulty of differentiating between muscle and fat cells, as the majority of participants had visibly more fatty infiltrated vessels in MF than in ES. In line with this, one previous study[2] also found slightly better reliability results (higher ICC, lower SEM) for ES compared to MF.

Higher muscle T2 times were assessed for MF compared to ES by both raters at all measurement time points. This result underlines that absolute T2 time comparisons between different muscles could not be recommended when using the applied technique. Previous studies that assessed reliability in T2 measurements[2,6] did not report their absolute T2 times so that no comparison could be made.

The inter-rater comparison delivered more reliable results than the reliability testing between different MRI sessions analysed by the same investigator. Also, there was no difference in the reliability between measurements performed on the same test day twice compared to measurements on different test days (compare IS1 and IS2 in Table 3). Hence, one might speculate that positioning of the participant in the MRI and/or software-based local determination of the first slice has a greater influence on the outcome than observer-dependency. This result underlines the importance of high standardisation in MRI measurements to make repetition of measurements reliable in inter subject comparisons or in longitudinal study designs. Still, intersession differences within one participant are rather small, satisfying reliability outcomes for IS are underlining this finding. Besides, the low systematic bias within one rater between M1 and M3 hints at the absence of meaningful daily physiological variations within one participant.

In comparison to previous research,[2,6,7] the ICC for ES and MF was slighlty weaker in our study (ICC range: 0.813–0.944), but still represents acceptable variability in outcome measures.

One disadvantage of the methodological procedure of this study is the time-consuming threshold adaptation in the original MRI scans and the calculation of T2 times with the calculation plug in. As this study proves reliable results for the applied technique, it is recommended using exactly this technique for further studies. But, as there was no clear difference in T2 times between different slices, it could be considered to downgrade the number of analysed slices to 2–3 instead of 5 slices for time-saving reasons in future studies. One further limitation is that the results of this study refer to a sample of normal participants only. Measurements in a sample of participants with diseased muscles will possibly result in a lower reliability and methodological difficilties when using the applied technique, which has to be considered in clinical studies.

This study proves that T2 time MRI measurements are a reliable tool to assess quantitative muscle structures with just small inter-rater differences and a low intersession variability in a mobile measurement setting. Hence, the applied technique to assess T2 times is recommended to use for future studies that aim to assess changes of T2 times, *e.g.* after an intense bout of eccentric exercises.

The use of mobile diagnostic devices, such as a mobile MRI, should be enhanced in future studies and clinical practice as their advantage of location-independency could lead to new pathways for sharing research with broader communities. However, it is not clear if and how the mobile unit measurements differ from standard MRI measurements and thus if the reliability of measures in a mobile device are as good as in a standard device. The movement of the MRI trailer could have had a negative impact on the measurements.

CONCLUSION

Reliable assessment of paraspinal muscle T2 times makes this outcome parameter valid for application in scientific research. The use of the mobile MRI device proves great reliability and thus supports the integration of mobile diagnostic devices in diagnostics and research.

ACKNOWLEDGEMENTS

We thank Bianca Lienerth and Carina Schücke for their technical assistance and support in conducting the measurements.

FUNDING

The project has been funded by the Federal State of Brandenburg, Germany. Open Access funding enabled and organized by Projekt DEAL.

REFERENCES

1. Cagnie B, Elliott JM, O'Leary S, D'hooge R, Dickx N, Danneels LA. Muscle functional MRI as an imaging tool to evaluate muscle activity. *J Orthop Sports Phys Ther* 2011; **41**: 896–903. doi: https://doi.org/10.2519/jospt.2011.3586
2. Hu Z-J, He J, Zhao F-D, Fang X-Q, Zhou L-N, Fan S-W. An assessment of the intra- and inter-reliability of the lumbar paraspinal muscle parameters using CT scan and magnetic resonance imaging. *Spine* 2011; **36**: 868–74. doi: https://doi.org/10.1097/BRS.0b013e3181ef6b51
3. Wan Q, Lin C, Li X, Zeng W, Ma C. Mri assessment of paraspinal muscles in patients with acute and chronic unilateral low back pain. *Br J Radiol* 2015; **88**: 20140546. doi: https://doi.org/10.1259/bjr.20140546
4. Maeo S, Saito A, Otsuka S, Shan X, Kanehisa H, Kawakami Y. Localization of muscle damage within the quadriceps femoris induced by different types of eccentric exercises. *Scand J Med Sci Sports* 2018; **28**:

95–106. doi: https://doi.org/10.1111/sms.12880
5. Carmona G, Mendiguchía J, Alomar X, Padullés JM, Serrano D, Nescolarde L, et al. Time course and association of functional and biochemical markers in severe semitendinosus damage following intensive eccentric leg curls: differences between and within subjects. *Front Physiol* 2018; **9**: 1–16. doi: https://doi.org/10.3389/fphys.2018.00054
6. Fan S, Hu Z, Zhao F, Zhao X, Huang Y, Fang X. Multifidus muscle changes and clinical effects of one-level posterior lumbar interbody fusion: minimally invasive procedure versus conventional open approach. *Eur Spine J* 2010; **19**: 316–24. doi: https://doi.org/10.1007/s00586-009-1191-6
7. Kim D-Y, Lee S-H, Chung SK, Lee H-Y, Sang KC. Comparison of multifidus muscle atrophy and trunk extension muscle strength: percutaneous versus open pedicle screw fixation. *Spine* 2005; **30**: 123–9.
8. Schütz UHW, Schmidt-Trucksäss A, Knechtle B, Machann J, Wiedelbach H, Ehrhardt M, et al. The TransEurope FootRace project: longitudinal data acquisition in a cluster randomized mobile MRI observational cohort study on 44 endurance runners at a 64-stage 4,486 Km transcontinental ultramarathon. *BMC Med* 2012; **10**: 78. doi: https://doi.org/10.1186/1741-7015-10-78
9. Bonavita S, Dinacci D, Lavorgna L, Savettieri G, Quattrone A, Livrea P, et al. Treatment of multiple sclerosis with interferon beta in clinical practice: 2-year follow-up data from the South Italy mobile MRI project. *Neurol Sci* 2006; **27**(S5): s365–8. doi: https://doi.org/10.1007/s10072-006-0696-6
10. De Stefano N, Cocco E, Lai M, Battaglini M, Spissu A, Marchi P, et al. Imaging brain damage in first-degree relatives of sporadic and familial multiple sclerosis. *Ann Neurol* 2006; **59**: 634–9. doi: https://doi.org/10.1002/ana.20767
11. Ishimoto Y, Yoshimura N, Muraki S, Yamada H, Nagata K, Hashizume H, et al. Associations between radiographic lumbar spinal stenosis and clinical symptoms in the general population: the Wakayama spine study. *Osteoarthritis Cartilage* 2013; **21**: 783–8. doi: https://doi.org/10.1016/j.joca.2013.02.656