

SUPPLEMENTARY TABLE 1. Data extraction table

Keller et al. (2004)																									
Study design	Randomized controlled trial																								
Patient characteristics	N = 61* Clinical status: CLBP and post laminectomy syndrome. Age: 42 years (SD unknown) Sex (M/F): 29/32																								
Sample size	Exercise group: N = 52 Lumbar fusion group: N = 60 Dropouts: N = 12																								
Interventions	Exercise group: Supervised exercise program, consisting cognitive therapy and exercises for muscle strength, endurance and coordination lasting for 8 weeks. Lumbar fusion group: Lumbar fusion, exercises were not recommended for the first 3 months after surgery.																								
Evaluation technique	Image acquisition: Computed Tomography Imaging Approach: Axial unislice vertebral level T12-L1 and L3-L4. Imaging was performed at baseline and follow-up (1 year)																								
Outcome measure	Muscle density quantified in Hounsfield Units within a homogenous part of the centre of the erector spinae. Muscle(s) of interest: Erector Spinae																								
Results	<table border="0"> <tr> <td>Exercise group L3-L4</td> <td>Lumbar fusion group: L3-L4</td> </tr> <tr> <td>Baseline 53.9 (9.6)</td> <td>Baseline 55.1 (13.6)</td> </tr> <tr> <td>Follow up 53.3 (9.1)</td> <td>Follow up 49.4 (15.4)</td> </tr> <tr> <td>Mean difference: -0.6 (p > 0.05)</td> <td>Mean difference: -5.9 (p < 0.01)</td> </tr> <tr> <td colspan="2" style="text-align: center;">Mean difference between groups</td> </tr> <tr> <td colspan="2" style="text-align: center;">5.3 (P<0.05)</td> </tr> <tr> <td>Exercise group T12-L1</td> <td>Lumbar fusion group: T12-L1</td> </tr> <tr> <td>Baseline 55.6 (8.5)</td> <td>Baseline 63.8 (18.3)</td> </tr> <tr> <td>Follow up 63.1 (12.8)</td> <td>Follow up 68.8 (17.6)</td> </tr> <tr> <td>Mean difference: 7.4 (p < 0.01)</td> <td>Mean difference: 5.0 (p > 0.05)</td> </tr> <tr> <td colspan="2" style="text-align: center;">Mean difference between groups</td> </tr> <tr> <td colspan="2" style="text-align: center;">2.4 (P>0.05)</td> </tr> </table> <p>Main conclusion: Statistical significant difference in fatty infiltration at T12-L1, not L3-L4 by means of exercise therapy.</p>	Exercise group L3-L4	Lumbar fusion group: L3-L4	Baseline 53.9 (9.6)	Baseline 55.1 (13.6)	Follow up 53.3 (9.1)	Follow up 49.4 (15.4)	Mean difference: -0.6 (p > 0.05)	Mean difference: -5.9 (p < 0.01)	Mean difference between groups		5.3 (P<0.05)		Exercise group T12-L1	Lumbar fusion group: T12-L1	Baseline 55.6 (8.5)	Baseline 63.8 (18.3)	Follow up 63.1 (12.8)	Follow up 68.8 (17.6)	Mean difference: 7.4 (p < 0.01)	Mean difference: 5.0 (p > 0.05)	Mean difference between groups		2.4 (P>0.05)	
Exercise group L3-L4	Lumbar fusion group: L3-L4																								
Baseline 53.9 (9.6)	Baseline 55.1 (13.6)																								
Follow up 53.3 (9.1)	Follow up 49.4 (15.4)																								
Mean difference: -0.6 (p > 0.05)	Mean difference: -5.9 (p < 0.01)																								
Mean difference between groups																									
5.3 (P<0.05)																									
Exercise group T12-L1	Lumbar fusion group: T12-L1																								
Baseline 55.6 (8.5)	Baseline 63.8 (18.3)																								
Follow up 63.1 (12.8)	Follow up 68.8 (17.6)																								
Mean difference: 7.4 (p < 0.01)	Mean difference: 5.0 (p > 0.05)																								
Mean difference between groups																									
2.4 (P>0.05)																									
Risk of Bias	High																								
Additional comments	*124 patients were included in total; 61 out of 124 people underwent pre-post imaging. The vertebral level T12-L1 was only captured in 41 out of 61 people.																								

Storheim et al. (2003)

Study design	Randomized controlled trial																				
Patient characteristics	N = 24 Clinical status: Subacute LBP, symptoms between 8-12 weeks. Age: 44.9 (10.3) Sex (M/F): 12/12																				
Sample size	Exercise group: N = 11 Control group: N = 13																				
Interventions	Exercise group: Treatment program lasting 15 weeks, with preferably three sessions per week lasting for 1 hour. The training program consisted exercises based on the Norwegian Aerobic Fitness model. Control group: Treated by their general practitioner with no treatment or referral restrictions.																				
Evaluation technique	Image acquisition: Computed Tomography Approach: Axial unislice vertebral level L4-L5 and L3-L4. CT-imaging was performed at baseline and 16 weeks.																				
Outcome measure	Muscle density quantified in Hounsfield Units within a homogenous part of the centre of the erector spinae. Muscle(s) of interest: Erector Spinae																				
Results	<table><tr><td>Exercise group L4-L5</td><td>Control group L4-L5</td></tr><tr><td>Baseline 57.1 (13.3)</td><td>Baseline 66.2 (38.7)</td></tr><tr><td>Follow-up 61.2 (15.4)</td><td>Follow-up 62.4 (30.7)</td></tr><tr><td colspan="2">Mean difference in change (95%CI)</td></tr><tr><td colspan="2">8.0 (-2.1 to 18.1) (p = NS)</td></tr><tr><td>Exercise Group L3-L4</td><td>Control Group L3-L4</td></tr><tr><td>Baseline 60.4 (11.1)</td><td>Baseline 70.7 (29.3)</td></tr><tr><td>Follow up 66.7 (19.7)</td><td>Follow up 67.9 (25.5)</td></tr><tr><td colspan="2">Mean difference in change (95%CI)</td></tr><tr><td colspan="2">9.2 (-0.6 to 19.0) (p = NS)</td></tr></table> <p>Main conclusion: No statistically significant differences in fatty infiltration by means of exercise.</p>	Exercise group L4-L5	Control group L4-L5	Baseline 57.1 (13.3)	Baseline 66.2 (38.7)	Follow-up 61.2 (15.4)	Follow-up 62.4 (30.7)	Mean difference in change (95%CI)		8.0 (-2.1 to 18.1) (p = NS)		Exercise Group L3-L4	Control Group L3-L4	Baseline 60.4 (11.1)	Baseline 70.7 (29.3)	Follow up 66.7 (19.7)	Follow up 67.9 (25.5)	Mean difference in change (95%CI)		9.2 (-0.6 to 19.0) (p = NS)	
Exercise group L4-L5	Control group L4-L5																				
Baseline 57.1 (13.3)	Baseline 66.2 (38.7)																				
Follow-up 61.2 (15.4)	Follow-up 62.4 (30.7)																				
Mean difference in change (95%CI)																					
8.0 (-2.1 to 18.1) (p = NS)																					
Exercise Group L3-L4	Control Group L3-L4																				
Baseline 60.4 (11.1)	Baseline 70.7 (29.3)																				
Follow up 66.7 (19.7)	Follow up 67.9 (25.5)																				
Mean difference in change (95%CI)																					
9.2 (-0.6 to 19.0) (p = NS)																					
Risk of Bias	Low																				
Additional comments	-																				

Berry et al. (2019)

Study design	Single-arm Cohort Study
Patient characteristics	N = 14 Clinical status: CLBP (including stenosis and spondylolysis) Age: 52.8 (14.8) years Sex (M/F): 7/7
Sample size	Exercise group: N = 14
Interventions	Exercise group: Machine based High-intensity training for 10 weeks (20 sessions).
Evaluation technique	Image acquisition: Magnetic Resonance Imaging Sequence: T1-weighted, Approach: Axial unislice disc level L4. Threshold method: Gaussian mixture model. MRI was performed at baseline and 10 weeks.
Outcome measure	Fat fraction (muscle-to-fat index) Muscle(s) of interest: Lumbar Multifidus and Erector Spinae
Results	Fat fraction multifidus (-0.007; $F(1,23) = 0.331$; $p=0.570$) Fat fraction erector spinae (-0.013; $F(1,23) = 1.079$; $p=0.310$) Main conclusion: No statistically significant differences in fatty infiltration by means of exercise.
Risk of Bias	Moderate
Additional comments	-

Mooney et al. (1997)

Study design	Non-randomized controlled trial
Patient characteristics	N = 16 Clinical status: LBP Age LBP group: 45-64 years Age Control Group: 35 years Sex LBP group (M/F): 4/4 Sex Control group (M/F): 8/0
Sample size	Exercise group: N = 8 Control group: N = 8
Interventions	Exercise group: Treatment program lasting 8 weeks, with two exercise sessions each week. The training included both concentric and eccentric isolated lumbar extension isotonic exercise. Control group: Eight age-related healthy male subjects were similarly assigned to the exercise program.
Evaluation technique	Image acquisition: Magnetic Resonance Imaging Sequence: T1-weighted, Approach: Axial multi-slice between L3 endplate and lower endplate L5. Magnetic Resonance Imaging was performed at baseline and 8 weeks.
Outcome measure	Qualitative grading scale (normal, mild, moderate, severe). Muscle(s) of interest: Lumbar Multifidus and Erector Spinae
Results	Main conclusion: 4/5 patients with severe fatty infiltration in the lumbar extensor muscles had a decrease in the degree of infiltration but no change in lean muscle mass. There were no changes in fat infiltration or muscle mass among the other patients with moderate fatty infiltration levels at onset.
Risk of Bias	Critical
Additional comments	No statistical evaluation was performed due to a small sample size (N=8 exercise, N=8 control).

Welch et al. (2015)

Study design	Single-arm Cohort Study
Patient characteristics	N = 30 Clinical status: CLBP (symptoms longer than 3 months) Age females: 39.6 (12.4) years Age males: 39.7 (9.7) years Sex M/F: 19/11
Sample size	Exercise group: N = 30
Interventions	Exercise group: Free-weight based resistance training for 16 weeks (48 sessions)
Evaluation technique	Image acquisition: Magnetic Resonance Imaging Sequence: T2-weighted. Approach: Multi-slice between L3-L4, L4-L5 and L5-S1. Threshold method: Interactive segmentation tool using Otsu's method. Magnetic Resonance Imaging was performed at baseline and 16 weeks.
Outcome measure	Percentage fat infiltration calculated by a standalone graphical user interface. Muscle(s) of interest: Erector Spinae and Lumbar Multifidus
Results	L3-L4 T0 left (% FI): 13.0 (8.2) L3-L4 T0 right (% FI): 12.1 (6.1) L3-L4 T16 left (% FI): 10.0 (6.3) L3-L4 T16 right (% FI): 9.4 (5.3) Mean difference: -23% (p<0.05) Mean difference: -22% (p<0.05) L4-L5 T0 left (% FI): 14.3 (7.0) L4-L5 T0 right (% FI): 13.6 (5.6) L4-L5 T16 left (% FI): 11.8 (6.0) L4-L5 T16 right (% FI): 11.7 (5.6) Mean difference: -18% (p<0.05) Mean difference: -14% (p<0.05) L5-S1 T0 left (% FI): 18.0 (5.9) L5-S1 T0 right (% FI): 17.8 (6.2) L5-S1 T16 left (% FI): 17.3 (7.0) L5-S1 T16 right (% FI): 16.3 (7.2) Mean difference: - 3% (p>0.05) Mean difference: -8% (p>0.05) Main conclusion: statistically significant reduction L3-L4, L4-L5, not L5-S1.
Risk of Bias	Critical
Additional comments	

Willemink et al. (2012)

Study design	Single-arm Cohort Study
Patient characteristics	N = 16 Clinical status: CLBP (symptoms longer than 12 weeks) Age: 46.2 (9.7) years Gender M/F: 16/0
Sample size	Exercise group: N = 16 Drop-outs: -
Interventions	Exercise group: Machine based isolated lumbar extensor training for 12 weeks (10 sessions). Afterwards until 24 weeks the training continued at a frequency that was tailored to the patient's convenience.
Evaluation technique	Image acquisition: Magnetic Resonance Imaging Sequence: Balanced fast-field echo Approach: Multi-slice between L3-L4, L4-L5 and L5-S1. Threshold method: Custom written Matlab script. MRI was performed at baseline, 12 and 24 weeks.
Outcome measure	Area of Fatty Infiltration (AFI) Muscle(s) of interest: Lumbar Multifidus
Results	L3-L4 AFI (cm ²) T0: 1.08 (0.55) L3-L4 AFI (cm ²) T12: 1.02 (0.52) L3-L4 AFI (cm ²) T24: 1.01 (0.61) Mean difference: L3-L4 = -6.5%(p=0.475) L4-L5 AFI (cm ²) T0: 1.80 (0.81) L4-L5 AFI (cm ²) T12: 1.79 (1.08) L4-L5 AFI (cm ²) T24: 1.76 (1.01) Mean difference L4-L5 = - 2.2%(p=0.820) L5-S1 AFI (cm ²) T0: 2.27 (0.99) L5-S1 AFI (cm ²) T12: 2.14 (0.98) L5-S1 AFI (cm ²) T24: 2.17 (1.00) Mean difference L5-S1 = -4.4% (p=0.155) Main conclusion: No statistically significant reductions in fatty infiltration by means of exercise.
Risk of Bias	Moderate
Additional comments	-

CLBP: Chronic Low Back pain; F: Female; M: male; N: number of participants; M: Male.

Supplementary TABLE 2A: Cochrane Risk of Bias Tool for Randomized Clinical Trials

	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Blinding of participants and personnel (performance bias)	Blinding of outcome assessment (detection bias)	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)	Other bias
Keller et al. (2004)	L	L	H	?	H	H	L
Storheim et al. (2003)	L	L	H	L	L	L	L

H = high risk of Bias, L = low risk of Bias, ? = unknown

Supplementary TABLE 2B: ROBINS-I Tool for Non-Randomized Clinical Trials

	Bias due to confounding	Bias in selection of participants into the study	Bias in classification of interventions	Bias due to deviations from intended interventions	Bias due to missing data	Bias in measurement of outcomes	Bias in selection of the reported result
Berry et al. (2019)	M	L	L	L	L	L	M
Mooney et al. (1999)	M	C	L	L	M	C	C
Welch et al. (2015)	M	L	L	L	C	L	C
Willemink et al. (2012)	M	M	L	L	L	L	L

H = high risk of Bias, M = moderate risk of bias L = low risk of Bias, ? = unknown

SUPPLEMENTARY TABLE 3. Tidier checklist

	1. Intervention	2. Rationale	3. Materials	4. Procedure	5. Persons	6. Modes	7. Location	8A. Sessions (N)	8B. Frequency N/Week	8C. Training (Weeks)	8D. Intensity	9. Tailoring	10. Modification	11. Adherence	12. Delivered
Berry et al. 2019	Y	Y	Y	Y	PT	Y	Y	20	2	10	Y	Y	N	N	N
Keller et al. 2004	Y	N	N	N	O	Y	Y	?	?	5	N	N	N	N	N
Mooney et al. 1997	Y	Y	Y	Y	?	Y	Y	16	2	8	Y	Y	N	N	N
Storheim et al. 2003	Y	N	N	N	PT	Y	Y	27	2-3	15	Y	Y	Y	Y	Y
Welch et al. 2015	Y	Y	Y	Y	?	Y	Y	48	3	16	Y	Y	Y	Y	N
Willemink et al. 2012	Y	Y	Y	Y	PT	Y	Y	10	1	12	Y	Y	N	N	N

PT: Physical Therapist; O: Other; ?: Unknown; Y: Yes; N: No.