Doi: 10.51224/SRXIV.308 | SportRχiv Preprint version 1

Effect size thresholds to interpret comparisons with exercise interventions for tendinopathy: A systematic review with meta-analysis.

Corresponding author:
Paul A Swinton, Robert Gordon University, School of Health Sciences, Aberdeen, UK.
Email: p.swinton@rgu.ac.uk

Co-authors:

Joanna SC Shim, Robert Gordon University, School of Health Sciences, Aberdeen, UK.

Anastasia V Pavlova, Robert Gordon University, School of Health Sciences, Aberdeen, UK.

Colin MacLean, Robert Gordon University, Library Services, Aberdeen, UK

David Brandie, Sportscotland Institute of Sport, Stirling, UK

Victoria Tzortziou Brown, Wolfson Institute of Population Health, Queen Mary University of London, London, UK

Dylan Morrissey, Barts and The London School of Medicine and Dentistry Blizard Institute, London, UK

Lyndsay Alexander, School of Health Sciences, Robert Gordon University, Aberdeen, UK

Kay Cooper, School of Health Sciences, Robert Gordon University, Aberdeen, UK

Twitter Handles:
@PaulSwinton9; @DrDylanM; @VictoriaTzB; @lynzalexander; @AHPkaycooper

Abstract

Introduction: The purpose of this systematic review with meta-analysis was to develop tendinopathy-specific effect size thresholds to aid interpretation of comparative effectiveness of exercise therapies.

Methods: A comprehensive systematic literature search for studies comparing exercise, non-active controls, and non-exercise interventions for tendinopathy management was conducted. Trials with participants diagnosed with rotator cuff, lateral elbow, gluteal, patellar or Achilles tendinopathy of any severity or duration were included. Standardised mean difference comparative effect sizes were combined with Bayesian hierarchical models. Non symmetric (exercise vs. non-active), and symmetric (exercise vs. exercise, and exercise vs non-exercise) distributions centred on zero were constructed with small, medium, and large thresholds calculated to capture 25, 50, and 75% of the comparative effect-size distributions, respectively. Analyses were combined across all tendinopathy locations and separated according to outcome domains including disability, pain, physical function capacity and range of motion where sufficient data were available.

Results: Data were extracted from 96 studies and yielded 130 pairwise comparisons. When pooling data across all outcomes, 98 comparative effect sizes showed small: 0.11 [95%CrI:0.09-0.13], medium: 0.25 [95%CrI:0.23-0.27], and large: 0.46 [95%CrI:0.44-0.49] effect sizes for exercise therapy compared with non-active control. Symmetric distributions calculated from 636 effect sizes showed lower magnitude thresholds when comparing exercise therapies (small: 0.11 [95%CrI:0.09-0.13], medium: 0.25 [95%CrI:0.23-0.27], and large: 0.46 [95%CrI:0.44-0.49]) than comparing exercise therapies with non-exercise interventions (254 effect sizes; small: 0.17 [95%CrI:0.13-0.21], medium: 0.37 [95%CrI:0.33-0.41], and large: 0.70 [95%CrI:0.64-0.75]). Analyses showed greater magnitude effect sizes for pain and disability outcomes than physical function capacity and range of motion.

Conclusion: Exercise therapy generally results in improvements beyond natural healing processes and expectancy effects present in non-active controls. Comparative effect sizes between exercise therapies are, however, generally low in magnitude indicating therapy variations yield relatively minor incremental effects. These findings should inform the clinical interpretation of existing and future trials, and guide study design of future research – such as the necessity for larger samples than have been previously used –. It is recommended that small, medium, and large threshold values presented in this review be used instead of Cohen’s general values.

Key words: Sample size, Power, Exercise therapy, Applied statistics, Bayesian

1.0 Introduction

Tendinopathy is a musculoskeletal condition frequently experienced by athletic, active, and sedentary populations that is characterised by pain, reduced function, and disability [1-4]. Management of tendinopathy includes a range of interventions, with exercise therapy comprised predominantly of resistance exercise being the most common and viewed as the mainstay of conservative management [5,6]. In a recent scoping review mapping research investigating exercise therapies and tendinopathy, it was identified that an extensive research base has been developed with randomised controlled trials (RCTs) the most common design [6]. The scoping review identified that over 90% of the research base investigated either rotator cuff related shoulder pain (RCRSP), lateral elbow tendinopathy, patellar tendinopathy, or Achilles tendinopathy [6]. The most common outcome domains investigated included pain, disability, self-reported function, physical function capacity, quality of life and range of motion (ROM) at the shoulder [6]. The scoping review also identified that exercise therapies were frequently compared against each other or were used as a comparator to other therapies including injection, laser, extracorporeal shockwave, manual therapy, and splinting/taping [6]. Less frequently, experimental studies included non-active controls including natural history groups (such as wait-and-see), placebo, and sham treatments [6]. These differing comparisons provide distinct insights for both clinicians and researchers.

Drawing robust conclusions from interventional research is challenging due to multiple reasons including the lack of condition-specific effect size interpretation guidance to refer to. To better interpret the findings from interventional research there are several key concepts (Table 1). The first is to identify the difference between 'statistical significance' and 'practical significance'. Musculoskeletal research has typically evaluated findings according to statistical significance, but this is slowly shifting towards practical significance and a focus on presentation of the magnitude of effects [7]. This is typically achieved with effect sizes that quantify some feature of the mean change within groups or the mean difference between groups (Table 1).

A second important conceptual point is to differentiate between non-comparative and comparative effectiveness. Attempts were made in a recent meta-analysis of tendinopathy and exercise therapy to quantify non-comparative effectiveness (Table 1) across a range of tendinopathies and outcome domains [8]. The meta-analysis used non-comparative standardized mean difference effect sizes extracted from 114 studies comprising 4104 participants. Effect sizes were calculated from the exercise therapy group alone, and described the change from a given assessment point (e.g. mid-intervention, post-intervention or follow-up) relative to baseline. The results showed that effect sizes were similar across tendinopathies, but substantial differences were identified across outcome domains [8]. Rather than focus solely on mean values, Swinton et al [8] modelled the majority of the effect size distributions, estimating the 1st, 2nd and 3rd quartiles which they qualitatively labelled as "small", "medium", and "large". Whilst these estimates provide clinicians and researchers with benchmarks to assess the effectiveness of exercise therapies and highlight the importance of outcome domain in the magnitude of change that can be expected, the non-comparative nature of the effect sizes are limited, and it is not clear for example how much of any effect is due to natural recovery processes independent of the therapy.

A third important conceptual point is to identify the different interpretations that can be made based on the comparative interventions studied (Table 1). Comparisons of exercise therapies with non-active controls provide a measure of effectiveness whilst controlling for improvements related to natural healing processes and expectancy effects that may be common in subjective measurements, including self-rated assessments of function and pain [9,10]. In contrast, comparisons of different exercise therapies frequently include relatively minor adjustments to the intervention, including contraction modes [11-13], dosage [14-16], or the influence of setting [17]. These types of comparisons are reflective of clinicians' and researchers' attempts to optimise exercise therapies and systematically establish which factors can be manipulated to produce the greatest improvements. Finally, comparisons of exercise therapies and non-exercise interventions may reflect recommendations to include usual care as a control group in clinical trials to determine whether standard practice should be updated [18]. The purpose of this systematic review with meta-analysis was to model comparative effect size distributions obtained: 1) between exercise therapies; 2) between exercise therapies and non-active controls (e.g. wait-and-see, placebo, and sham); and 3) between exercise therapies and non-exercise interventions. In addition to providing clinicians and researchers with important information regarding expected differences, comparative effect sizes are integral in determining sample sizes for future experimental trials.

Table 1: Definitions of key terms required to interpret interventional research in tendinopathy.

Term	Definition
Statistical significance	Process used to claim whether a non-zero change in outcome, or non-zero difference between interventions would occur on average in the population studied.
Practical significance	Process used to estimate and interpret the magnitude of change in outcome from a single intervention, or magnitude of difference between interventions that should be expected on average in the population studied.
Comparative effect size	Statistic used to quantify the magnitude of difference between interventions that should be expected on average in the population studied.
Non-comparative effect size	Statistic used to quantify the magnitude of change in outcome from a single intervention that should be expected on average in the population studied.
Non-active controls	Participants allocated to either: 1) wait-and-see (generally single visit to professional where practical solutions including ergonomics advice and pain relief are discussed with individuals encouraged to await further spontaneous improvement); or 2) placebo/sham (participants allocated to medicine [placebo] or procedure [sham], not expected to have a therapeutic effect, but may provide psychological benefits).
Exercise therapies	Use of a range of exercise types including resistance, flexibility, proprioception, plyometric and vibration to create a therapeutic effect improving outcomes across a range of health domains affected by tendinopathy [19].
Non-exercise interventions	Interventions frequently used to improve improving outcomes across a range of health domains affected by tendinopathy [19] that do not include any exercise components. Interventions included electrotherapy, biomechanical alterations, manual therapy, injection therapy, and surgery.

2.0 Methods

This review was part of a project funded by the National Institute for Health Research (NIHR) (Health Technology Assessment (HTA) 129388 Exercise therapy for the treatment of tendinopathies) and follows on from a systematic review with meta-analysis quantifying non-comparative effect size distributions [8]. The review is reported according to the PRISMA 2020 statement with checklist provided in online supplementary file 1.

2.1 Protocol deviations

Multiple protocol deviations occurred due to pragmatic considerations and reflecting on processes from previous work packages in the larger project. Extraction of data was performed in duplicate rather than performed individually with an assessment of reliability as originally stated. Originally, it was intended to conduct risk of bias using the ROBINS-I tool for quasi-experimental studies [20]. As outlined in the following sections, due to pragmatic considerations Cochrane's Risk of Bias tool was used for both randomised and non-randomised studies [21].

2.2 Inclusion criteria

2.2.1 Participants

This review included people of any age or gender with a diagnosis of RCRSP, lateral elbow, patellar, Achilles or gluteal tendinopathy of any severity or duration. We accepted trial authors' diagnoses of tendinopathy in the absence of full thickness or large tears.

2.2.2 Intervention

The primary intervention assessed was exercise therapy which was comprised of five different therapy classes (resistance, plyometric, vibration, flexibility, and proprioception). Definitions for each therapy class are presented in online supplementary file 2. Many of the exercise therapies included multiple classes, and so for each therapy the dominant class was identified based on exercise volume or the primary goal (e.g. strengthening or mobility) stated by the authors. We included exercise therapies delivered in a range of settings and by a range of health, exercise professionals or support workers. We also included both supervised and unsupervised exercise therapies.

2.2.3 Comparator

Comparisons were conducted: 1) between exercise-only trial arms; 2) between exercise and non-active trial arms; and 3) between exercise and non-exercise trial arms. Non-active controls included wait-and-see, placebo, and sham. Non-exercise interventions included injection, electrotherapy, biomechanical modifications, manual-therapy or surgery. Definitions of each class are presented in supplementary file 2.

2.2.4 Outcomes

Based on the results of our initial scoping review [6] and subsequent stakeholder workshops, we extracted data from outcomes that assessed six domains: 1) disability; 2) physical function capacity; 3) function; 4) pain (on loading/activity, over a specified time, or without further specification); 5) quality of life; and 6) ROM (shoulder joint only). Definitions of each domain and example tools used to measure the outcomes are presented in online supplementary file 3.

2.2.5 Types of studies

We included randomised controlled trials (RCTs) and non-RCTs that included treatment arms that matched our comparator and other inclusion criteria.

2.2.6 Context

The context included primary care, secondary care or community locations in nations defined as very high or high on the Human Development Index (top 62 countries at the time of protocol development) for the findings to be relevant to the UK context [22].

2.2.7 Exclusion criteria

We excluded studies where sufficient information was not available to code treatment arm interventions according to exercise, non-active and non-exercise criteria. We also excluded trial arms that included any combination of exercise, non-active or non-exercise interventions.

2.3 Search strategy

The search strategy used for this study was part of a larger search conducted to scope the entire exercise therapy for tendinopathy management research base and comprised three steps. Firstly, a limited search of MEDLINE and CINAHL using initial keywords (MH tendinopathy OR TX tendin* OR TX tendon*) AND (MH exercise OR TX exercis*) was conducted with analysis of the text words in the titles/abstracts and those used to describe articles to develop a full search strategy. Secondly, the full search strategy was adapted to each database and applied systematically to: MEDLINE, CINAHL, AMED, EMBase, SPORTDiscus, Cochrane library (Controlled trials, Systematic reviews), JBI Evidence Synthesis, PEDRo, and Epistemonikos. The following trial registries were also searched: ClinicalTrials.gov, ISRCTN Registry, The Research Registry, EU-CTR (European Union Clinical trials Registry), ANZCTR (Australia and New

Zealand Clinical trials Registry). Finally, the third step involved conducting a search of cited and citing articles using Scopus and hand-searching a total of 130 systematic reviews that were identified to include information relevant to exercise therapy and tendinopathy. No limit was placed on language, with research studies published in languages other than English translated via Google Translate or via international collaborations of the review team members. Searches were initiated from 1998 as (i) the heavy load eccentric calf-training protocol for Achilles tendinosis by Alfredson et al [23] was published in 1998 and may be considered seminal work in the field of tendinopathy, and (ii) there has been a proliferation of research on exercise interventions for tendinopathies post 1998. Search terms are presented in Supplementary file 4 according to the last date of the search which was conducted on 25/03/2022.

2.4 Study selection and Data extraction

Two independent reviewers screened titles and abstracts followed by full-text copies. Conflicts were resolved by a third reviewer with all screening conducted within the Covidence (Melbourne, Australia) platform. Data were extracted independently by eight members of the review team into pre-piloted excel sheets and coded as described in the codebook presented in the Supplementary file 5. Each entry was then independently checked.

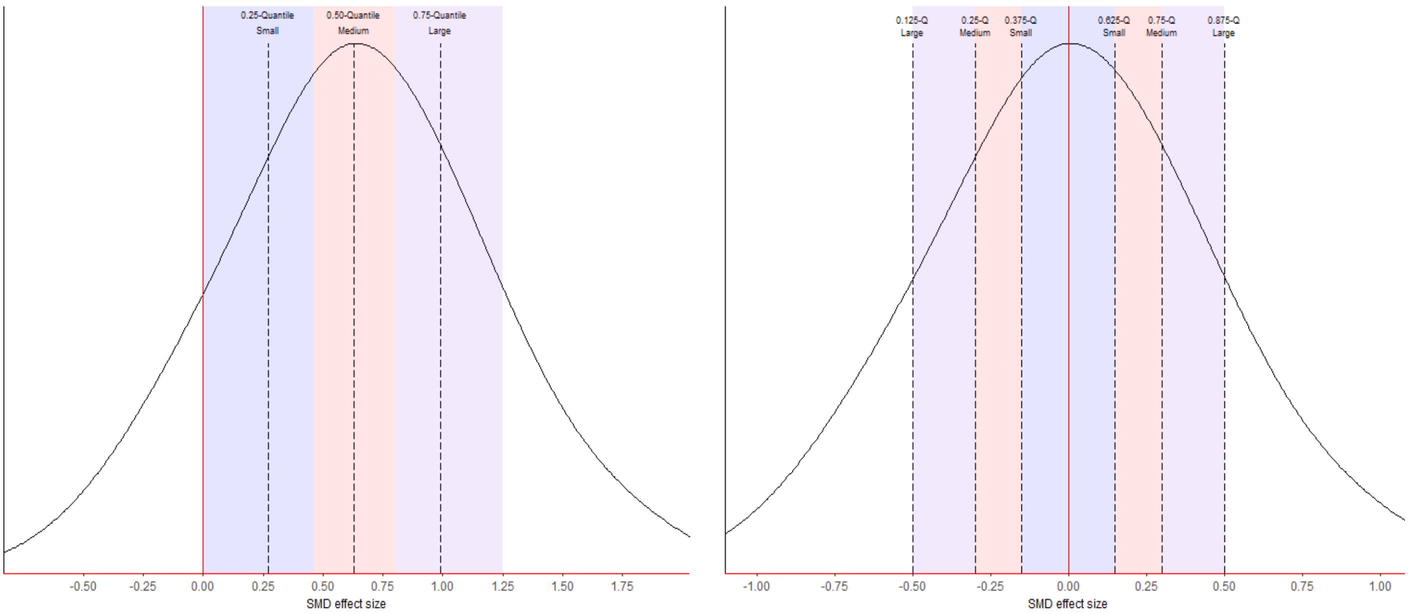
2.5 Risk of bias assessment

We used Cochrane's Risk of Bias (RoB) tool [21] to assess six domains: 1) selection bias (random sequence generation & allocation concealment); 2) performance bias (blinding of participants); 3) detection bias (blinding of outcome assessors); 4) attrition bias (incomplete outcome data); 5) reporting bias (selective reporting); and 6) other biases. RoB was recorded for each outcome and time point within each study. The Cochrane's RoB tool [21] was selected as a recent review of popular tools in tendinopathy management highlighted none were superior [24] and Cochrane's RoB tool [21] could be semi-automated with RobotReviewer [25], a machine learning system software. RobotReviewer was used to make initial assessments on selection bias and performance bias domains, with manual validation made on the relevant free texts extracted to support the final selection of low, high, or unclear RoB. This semi-automated process was more efficient and provided an additional element of consistency in the review process.

2.6 Meta-analysis

Comparative effect sizes were calculated for studies comparing at least two trial arms that matched the inclusion criteria identified. For exercise vs. non-active controls, comparative effect sizes were calculated to assess a hierarchy such that positive values favoured exercise therapy. Given previous research has established substantive non-comparative effect sizes for exercise therapy [8], comparison with non-active controls was included to establish the contribution of natural healing processes and expectancy effects. In contrast, no attempt was made to rank or create hierarchies when comparing across exercise therapies, or between exercise therapies and non-exercise interventions. The "direction" of comparative effect sizes was considered random such that across the large database and with additional bootstrapping procedures the effect size distribution would be centred on zero reflecting two-tailed hypothesis testing from a frequentist perspective and "sceptical" priors from a Bayesian perspective [26]. That is, when comparing two interventions both believed to be effective, a standard perspective is to hold that the difference is zero unless evidence to the contrary emerges. Schematics illustrating the differences between comparative effect size distributions where the mean is centred on zero, and where most exercise-therapies are expected to generate greater improvements than not-active controls is presented in Figure 1. Previously, 0.25-, 0.50- and 0.75-quantiles have been used to qualitatively label effect sizes as "small", "medium", and "large" across multiple areas including tendinopathy management [8, 27-29]. To apply this approach with symmetric comparative effect sizes, the small, medium and large thresholds were defined by the middle 25% (0.375 to 0.625-quantile), the middle 50% (0.25 to 0.75-quantile) and the middle 75% of the distribution (0.125 to 0.875-quantile), respectively (Figure 1).

Figure 1: Schematics illustrating differences in a non-symmetric comparative effect size distribution (left) and a symmetric comparative effect size distribution centred on zero (right) with small, medium, and large thresholds defined.



Non-symmetric plot (left) provides a schematic used for comparison of exercise therapy versus non-active control. In the schematic, most of the distribution exceeds zero (favouring exercise therapy) and regions (small/medium/large) are defined from the left. Symmetric plot (right) provides a schematic used for comparison of exercise therapy versus exercise therapy, and exercise therapy versus non-exercise interventions. The comparative distribution (right) is centred on zero with small effects closest to zero in either direction, and medium and large effects located further from the centre. Q: Quantile. SMD: Standardized mean difference effect size.

Comparative effect sizes and their sampling variance were calculated using group mean and standard deviation values reported at baseline and at any subsequent time-point. Pairwise comparative standardised mean differences ($\mathrm{SMD}_{AB_{\mathrm{pre}}}$) of an intervention "A" and "B", and their sampling variances σ^2 were calculated using the following formulae [30]:

$$\mathrm{SMD}_{AB_{\mathrm{pre}}} = \left(1 - \frac{3}{4(n_A + n_B - 2) - 1}\right)\left(\frac{\left(\bar{x}_{A_{Post}} - \bar{x}_{A_{Baseline}}\right) - \left(\bar{x}_{B_{Post}} - \bar{x}_{B_{Baseline}}\right)}{Sd_{AB_{Pre}}}\right)$$

where n_A and n_B are the number of participants in intervention A and B. The first term in the equation comprises a small-study bias term $c(n_A + n_B - 2)$, where $c(n_A + n_B - 2) = 1 - \frac{3}{4(n_A+n_B-2)-1}$, and $Sd_{AB_{Pre}}$ is the baseline pooled standard deviation where $Sd_{AB_{Pre}} = \sqrt{\frac{(n_A-1)Sd_{A_{Pre}}+(n_B-1)Sd_{B_{Pre}}}{n_A+n_B-2}}$.

$$\sigma^2\left(\mathrm{SMD}_{AB_{\mathrm{pre}}}\right) = 2c(n_A + n_B - 2)^2(1-\rho)\left(\frac{n_A + n_B}{n_A n_B}\right)\left(\frac{n_A + n_B - 2}{n_A n_B}\right)\left(1 + \frac{\mathrm{SMD}^2_{AB_{\mathrm{pre}}}}{2(1-\rho)\left(\frac{n_A + n_B}{n_A n_B}\right)}\right) - \mathrm{SMD}^2_{AB_{\mathrm{pre}}}$$

where ρ is the correlation between repeated measures.

The empirically obtained effect sizes were modelled using a three-level Bayesian mixed effects meta-analytic model. The three levels included the between study (level 3), the outcome (level 2) and the within study sampling variance (level 1). The application of a meta-analytic model enabled sharing of information across studies to better estimate model parameters and accounted for dependencies within the data due to most studies providing more than one data point (based on reporting multiple outcomes and/or multiple time points following baseline), and multiple studies including more than two trial arms such that multiple pairwise calculations could be made. To account for uncertainty in σ^2 due to non-reporting of correlations between baseline and follow-ups, the values were allowed to vary and were estimated by including an informative Gaussian prior with correlation values centred on 0.7 and ranging from approximately 0.5 to 0.9 [29].

The parameters obtained from the meta-analysis models were then used to calculate small, medium, and large threshold values for the pooled data and for each outcome domain separately. For exercise vs. non-active controls this was achieved by generating posterior predictions from each meta-analysis model and calculating the 0.25-, 0.5-, and 0.75-quantiles. Posterior predictions used the posterior sample for the model parameters to simulate new data. For symmetric distributions modelling across exercise therapies and between exercise therapies and non-exercise interventions, each analysis included all relevant studies and data points with 100 bootstrap samples comprising a +1/-1 random allocation for their pairwise effect sizes. The same coefficient that was obtained with 50/50 probability in each bootstrap sample was applied to all effect sizes obtained from a single study to maintain associations. For each set of posterior predictions across the bootstrap samples, the 0.625-quantile/|0.375|-quantile, 0.75-quantile/|0.25|-quantile, and

0.875-quantile/|0.125|-quantile values were obtained to quantify small, medium, and large thresholds, respectively. Median values and credible intervals (CrI) were reported to express estimates and their uncertainty. Where more than 50 effect sizes were obtained within a single outcome domain, a sub-set analysis was conducted to quantify the outcome domain specific distribution.

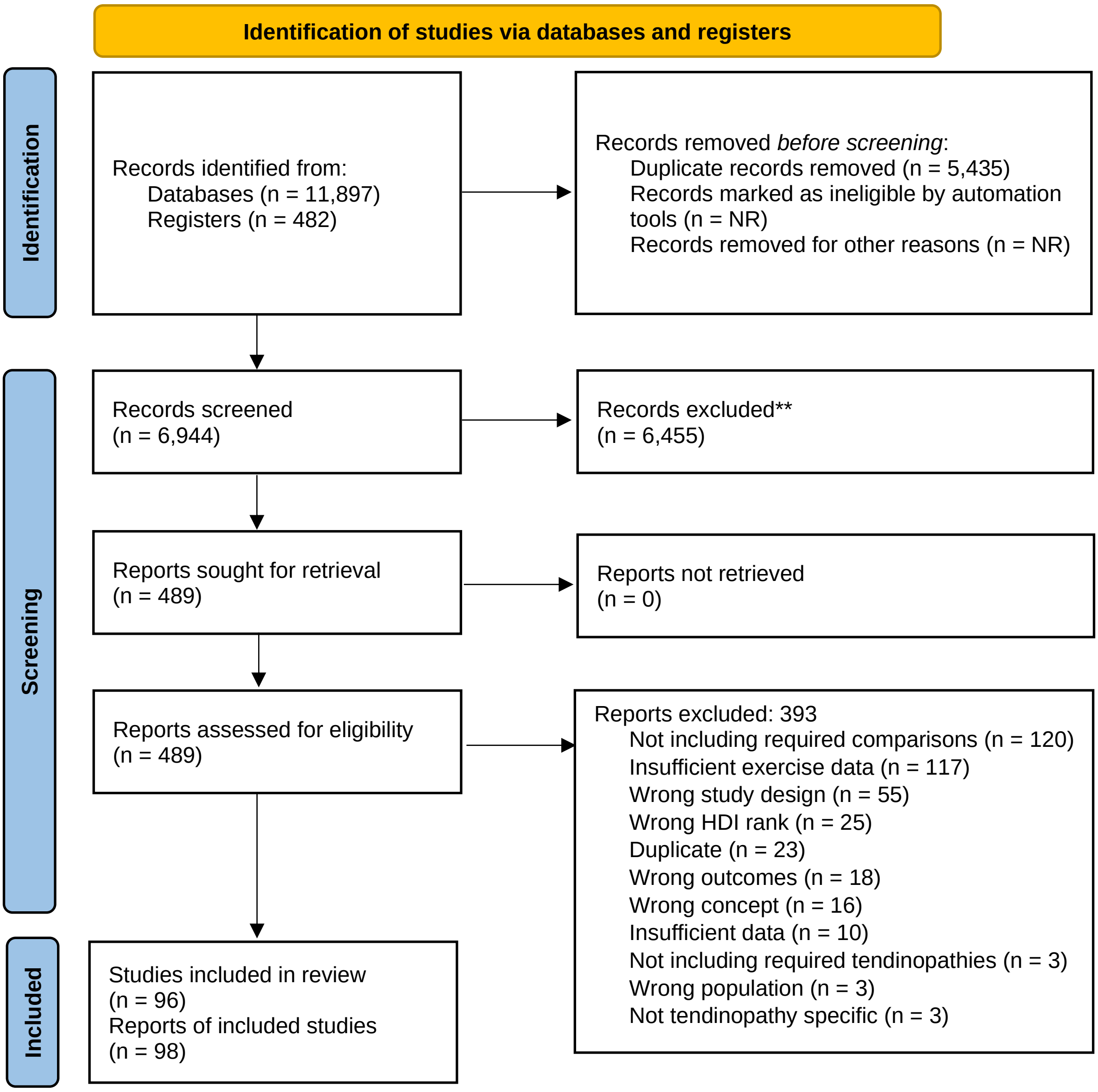
Default weakly informative Student-t and half Student-t priors with 3 degrees of freedom were used for all intercept and variance parameters, respectively [31]. Outlier values were identified by adjusting the empirical distribution by a Tukey g-and-h distribution and obtaining the 0.0035- and 0.9965-quantiles, with values beyond these points removed prior to further analysis [32]. Meta-analyses were performed using the R wrapper package brms interfaced with Stan to perform sampling [33]. Convergence of parameter estimates were obtained for all models with Gelman-Rubin R-hat values below 1.1 [34].

3.0 Results

3.1 Descriptions of data

A flow diagram illustrating study selection with reasons for exclusions is presented in Figure 2. Data to investigate comparative effect sizes were obtained from 96 studies (Supplementary file 6). Of these 96 studies, 40 investigated RCRSP, 23 investigated Achilles tendinopathy, 16 investigated lateral elbow tendinopathy, 14 investigated patellar tendinopathy, and 3 investigated gluteal tendinopathy. A total of 63 studies provided comparisons between different exercise therapies, 26 between exercise therapies and non-exercise interventions, and 12 between exercise and non-active controls. In total, data from 130 pairwise comparisons were obtained (85 between exercise therapies, 33 between exercise therapies and non-exercise interventions, and 12 between exercise and non-active controls). A total of 1075 pairwise effect sizes were obtained (pain: 381/35%, disability: 296/28%, physical function capacity: 206/19%, ROM: 70/7%, QoL: 62/6%, function: 60/6%). Sample sizes across the comparisons ranged from 5 to 134, with median equal to 21 [IQR: 13-30]. Measurement duration relative to baseline ranged from 1 to 260 weeks, with median equal to 10 [IQR: 5-16] weeks.

Figure 2: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) chart illustrating study selection.


Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 11,897)
Registers (n = 482)
Records removed before screening:
Duplicate records removed (n = 5,435)
Records marked as ineligible by automation tools (n = NR)
Records removed for other reasons (n = NR)
Records screened
(n = 6,944)
Records excluded**
(n = 6,455)
Reports sought for retrieval
(n = 489)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 489)
Reports excluded: 393
Not including required comparisons (n = 120)
Insufficient exercise data (n = 117)
Wrong study design (n = 55)
Wrong HDI rank (n = 25)
Duplicate (n = 23)
Wrong outcomes (n = 18)
Wrong concept (n = 16)
Insufficient data (n = 10)
Not including required tendinopathies (n = 3)
Wrong population (n = 3)
Not tendinopathy specific (n = 3)
Studies included in review
(n = 96)
Reports of included studies
(n = 98)

Review of the dominant treatment classes in each of the 85 pairwise exercise therapy comparisons, showed 51 (60%) were between resistance, 18 (21%) between resistance and flexibility, 6 (7%) between resistance and proprioception, 4 (5%) between proprioception, 3 (4%) between flexibility, 1 (1%) between flexibility and proprioception, 1 (1%) between resistance and vibration, 1 (1%) between proprioception and vibration. Review of the 33 exercise therapy and non-exercise intervention comparisons showed the most common non-exercise intervention was electrotherapy (9/27%), followed by injection (8/24%), biomechanical modifications (8/24%), manual therapy (4/12%), and surgery (4/12%).

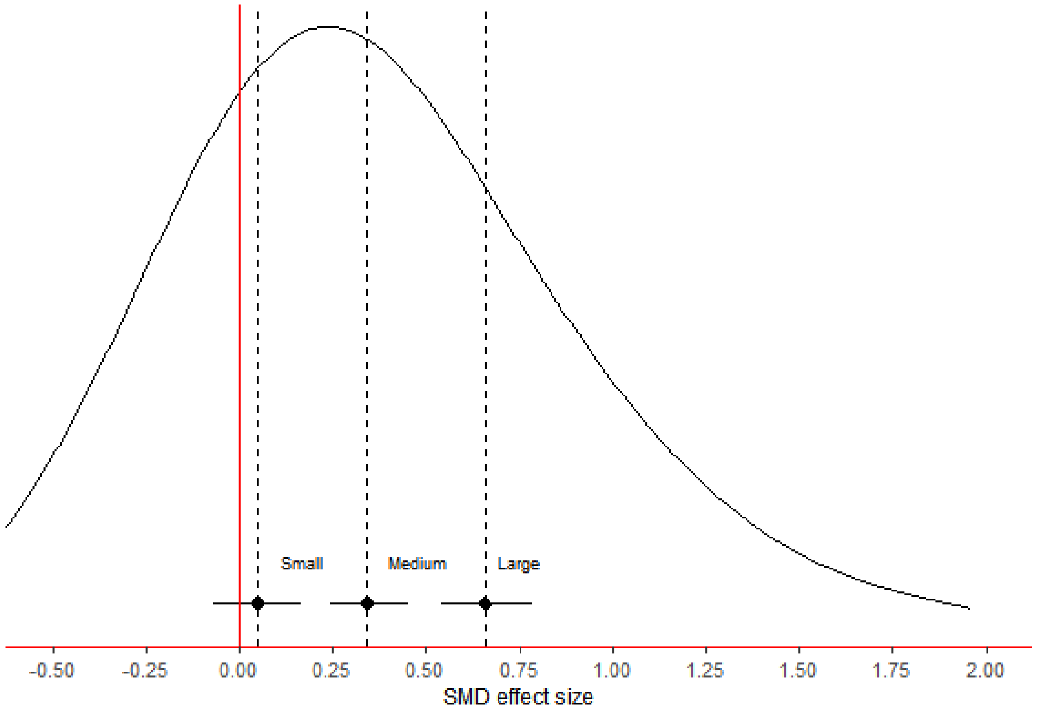
3.2 Risk of bias results

Summary risk of bias assessments for individual studies are presented in supplementary file 7. Risk of bias was highest for blinding of participants (high risk: 41%, unclear risk: 30%) and other bias (high risk: 49%, unclear risk: 11%), where the most common causes for concern included potentially important baseline imbalances and differences in fidelity across the trial arms.

3.3 Exercise and non-active comparative effect sizes

Following removal of outliers, a total of 98 effect sizes were obtained from 12 studies and 12 pairwise comparisons. With positive effect sizes favouring exercise therapy (Figure 3), the results estimated the thresholds as small: 0.05 [95%CrI: -0.06 to 0.17], medium: 0.34 [95%CrI: 0.24 to 0.45], and large 0.66 [95%CrI: 0.54 to 0.79]. Model details including breakdowns of outcome domains and tendinopathies presented in supplementary file 8.

Figure 3: Empirical distribution and modelled comparative effect size thresholds comparing exercise and non-active controls.

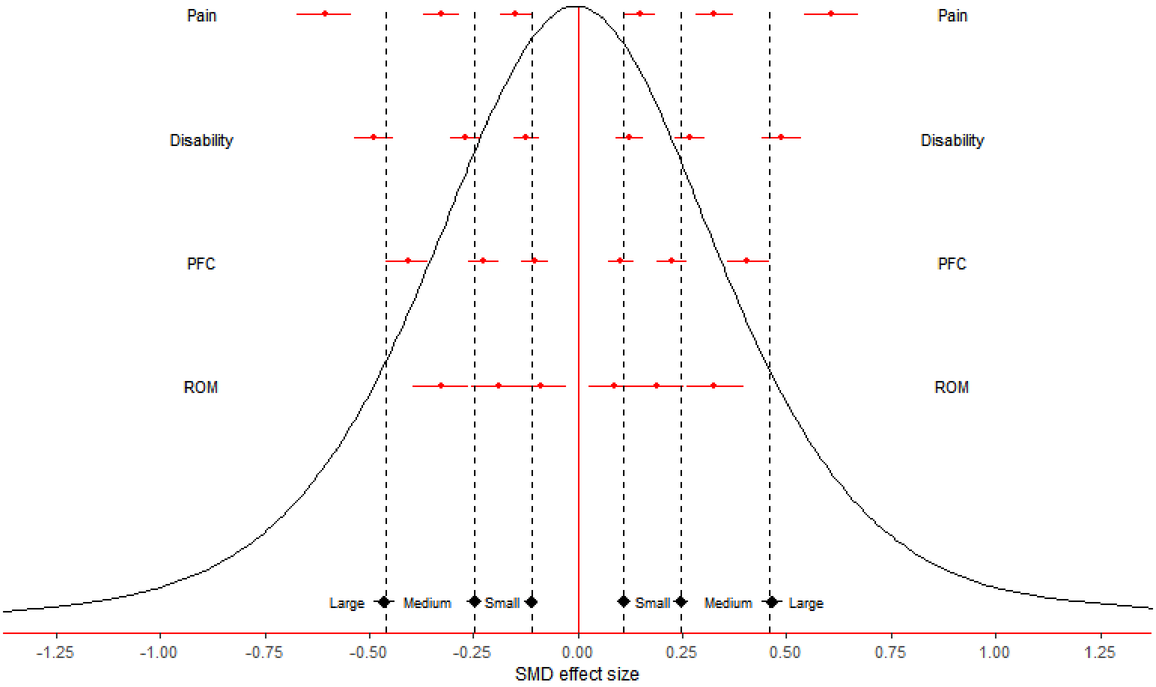


Black curve is a density plot of the directly calculated comparative effect size values across all outcomes. Small, medium, and large thresholds represent the 0.25-, 0.50-, and 0.75-quantiles of predicted draws. Black diamonds represent threshold values based on all outcomes and include 95% credible intervals. Positive values favour exercise therapies. SMD: Standardised mean difference.

3.4 Comparative effect sizes between exercise therapies

Following removal of outliers, an initial analysis was conducted pooling data across all outcome domains comprising 636 effect sizes from 61 studies and 82 pairwise comparisons. The results estimated the thresholds as small: 0.11 [95%CrI: 0.09 to 0.13], medium: 0.25 [95%CrI: 0.23 to 0.27], and large 0.46 [95%CrI: 0.44 to 0.49]. Sufficient data were obtained to estimate domain specific thresholds for pain, disability, ROM and physical function capacity. Figure 4 illustrates that differences were obtained, primarily for medium and large thresholds where the greatest values were obtained for pain, and lowest values obtained for ROM. Model details for all meta-analyses including breakdowns of outcome domains and tendinopathies presented in supplementary file 8.

Figure 4: Empirical bootstrapped distribution and modelled comparative effect size thresholds across exercise therapies for different outcome domains.



Black curve is a density plot of the directly calculated and boot-strapped empirical comparative effect size values across all outcomes. Small, medium, and large thresholds represent the 0.375-/0.625-, 0.25-/0.75-, and 0.125-/0.875-quantiles of predicted draws. Black diamonds represent threshold values based on all outcomes. Redpoint ranges illustrate the outcome domain specific estimates and their uncertainty through the median value (circle) and 95% credible interval. PFC: Physical function capacity; ROM: Range of motion of shoulder.

3.5 Comparative effect size between exercise therapies and non-exercise interventions

Following removal of outliers, an initial analysis was conducted pooling data across all outcome domains comprising 254 effect sizes from 23 studies and 29 pairwise comparisons. The results estimated the thresholds as small: 0.17 [95%CrI: 0.13 to 0.21], medium: 0.37 [95%CrI: 0.33 to 0.41], and large 0.70 [95%CrI: 0.64 to 0.75]. Sufficient data were obtained to estimate domain specific thresholds for pain, disability, and physical function capacity, with model details for all meta-analyses including breakdowns of outcome domains and tendinopathies presented in supplementary file 8.

4.0 Discussion

The results of this systematic review with meta-analysis provided several novel insights into comparative effect size distributions across different contexts of interest in the study of tendinopathy management. Comparison with non-active controls showed most of the distribution favoured exercise with additional improvements obtained beyond natural healing processes and expectancy effects. Comparisons across different exercise therapies showed that half of the comparative effect sizes are expected to be low in magnitude with absolute value standardised mean differences less than 0.25. Sub-analyses identified differences across outcome domains, particularly for large thresholds where greater effect sizes were obtained for subjective measures including pain and disability, compared with more objective measures of physical function capacity and ROM at the shoulder. Small, medium, and large thresholds were all increased for comparisons between exercise therapies and non-exercise interventions, likely reflecting greater heterogeneity of treatment comparisons. These results have implications for interpreting clinical effectiveness of different therapies and conducting future RCT's, with the need to recruit large sample sizes when comparing similar therapies.

The results of this systematic review follow on from a previous review synthesising single-arm non-comparative effect sizes obtained with exercise therapies for tendinopathy [8]. Results from the previous review identified high magnitude effect size distributions across most outcomes, with small (0.25-quantile) medium (0.5-quantile) and large (0.75) effect sizes thresholds of approximately 0.35, 0.75 and 1.2, respectively. Similar values were also obtained across different tendinopathies [8]. In the present review, the small, medium, and large comparative effect size thresholds obtained between exercise and non-active controls were equal to approximately 0.05, 0.35, and 0.65, respectively. The difference in results between reviews shows that whilst a substantive portion of improvements in outcomes used to evaluate tendinopathy interventions may be due to natural healing processes and expectancy effects, exercise therapies are likely to create additional improvements that may be meaningful. The inclusion of comparisons of exercise therapy against non-active controls including a natural history group (such as wait-

and-see) or sham or placebo intervention provide clinicians with a clear baseline and ability to evaluate the overall effectiveness of exercise therapy. A previous meta-analysis conducted by Steuri et al. [35] investigating the effectiveness of non-surgical interventions for the management of shoulder impingement also reported substantive comparative effect sizes favouring exercise relative to placebo or no-treatment groups for outcomes measuring pain (SMD = 0.94) and function (SMD =0.57). The findings of Steuri et al. [35] are in line with those presented here and in reviews evaluating exercise therapies for Achilles tendinopathy but with fewer included studies [36,37]. It is plausible that different comparative effect sizes would be obtained when controlling with natural history, placebo, or sham groups; however, due to the limited number of studies that included these controls, data were combined to facilitate modelling of the overall distribution rather than simply the mean which is typically focussed on.

Reflecting clinicians' interest in optimising exercise therapy, many of the RCTs included in the present review comprised relatively minor adjustments to the same exercise such as comparing concentric and eccentric actions [11-13], or the same exercise with factors such as the setting manipulated [17]. With increased knowledge and speciality, it is understandable that clinicians seek continued refinement with additional improvement potentially being meaningful in some cases. Most of the comparative effect size distribution between exercise therapies was less than ~0.25, with around a quarter of the distribution less than ~0.10. It is important for clinicians to be aware of these findings and that many exercise therapies that match general recommendations are likely to produce similar changes on average. This knowledge can be used to consider more fully the costs and benefits of implementing different exercise therapies. Similarly, for researchers interested in comparing relatively minor adjustments to an exercise therapy, they should be aware that the comparative effect size is likely to be low which will have implications on the sample size required to adequately power the study.

The meta-analyses conducted in this review also showed that comparative effect size distributions were influenced by the outcome domain measured. The finding that effect sizes are likely to be different across outcome domains frequently assessed in tendinopathy has been shown previously. In Swinton et al, [8] differences were primarily obtained between domains measured subjectively and objectively, with the small effect size threshold for subjectively measured outcomes (disability, pain, and function) between the medium (~0.40) and large (~0.65) effect size thresholds for the objectively measured outcomes (physical function capacity and ROM). The difference between outcome domains was not as large in the present review using comparative effect sizes. A high degree of overlap was obtained for the small effect size threshold, with some overlap also obtained for the medium effect size threshold (Figure 4). In contrast, substantial differences were obtained for the large effect size threshold between for example pain (~0.60) and ROM at the shoulder (~0.35). This difference likely reflects the much greater non-comparative effect sizes that can be obtained with outcome domains measured subjectively, and that some percentage of studies (e.g. ≤25%) will compare very different therapies where one is substantially more effective than the other.

Comparisons between exercise therapies and non-exercise interventions provides greater scope to develop interventions that are both outwardly different and encompass greater differences in comparative effectiveness. In the present review, data were obtained from 23 studies including 29 pairwise comparisons of exercise therapies and non-exercise interventions. The most common non-exercise interventions were electrotherapy (8 studies), biomechanical modifications (8 studies), and injection (7 studies). Each of these non-exercise therapies were aimed at different mechanisms of action to assist the healing process ranging from injection therapy, often corticosteroids, targeted primarily at pain relief to electrotherapy such as shock wave therapy to promote cell proliferation and enhance collagen synthesis required for tendon healing. Most shock wave therapy were administered at a frequency of 8 to 12 Hz and 2000 pulses per session, with a pressure between 2.5 and 4.0 bar [38,39]. Depending on the type of therapy, frequency in dose also varies. Exercise protocols in the studies tend to have higher frequency compared to non-exercise therapies. For instance, in one RCT, patients randomised to the eccentric exercise group were asked to complete sets of repetitions twice a day, 7 days per week compared to 3 sessions of shock wave therapy at weekly intervals in the experimental group [40]. These inherent differences between trial groups may result in more discerning effect sizes. Across all outcomes, the small, medium, and large comparative effect size thresholds between exercise therapies and non-exercise interventions were approximately 0.20, 0.40 and 0.70. These values are lower in magnitude, but somewhat close to Cohens original guidelines [41] that are frequently used to interpret effect sizes in many disciplines and areas, including tendinopathy.

One of the main implications from the present review includes sample sizes required for future studies. Of the 96 included studies, samples sizes in each of the trial arms ranged from 5 to 134, with a median of 21 [IQR: 13-30] participants. Whilst statistical power depends on a range of factors including the specific statistical test, number of data points, the underlying structure of the data, and hypotheses conducted (i.e. disjunction, conjunction and individual), general sample size requirements can be illustrated. For simplicity, testing the null hypothesis of zero population average treatment with a two-tailed independent t-test (α=0.05, and power (1-β) = 0.80) and the small, medium, and large effect thresholds identified in the present review returns group sizes of 1299, 253, and 76, respectively, for between exercise therapy comparisons [42]. Using the larger thresholds for exercise versus non-exercise comparisons returns reduced group sizes of 545, 116 and 34 [42].

Alternative approaches to determining a priori sample sizes can be used. One alternative is for researchers to specify a smallest effect size of interest using anchor-based methods reflecting what patients would consider a meaningful improvement [43]. If this effect size can be determined, power calculations can be performed, and subsequent sample sizes selected. The effect sizes determined in the present review represent a distributional-based method for describing change and can be combined with anchor-based approaches to provide greater context for both clinicians and researchers. Where for example, the smallest effect sizes of interest are much greater than that achieved with the majority of treatment comparisons (e.g. exercise vs exercise, or exercise vs non-exercise), clinicians may choose to remain with the current therapy and researchers may choose not to expend resources conducting a clinical trial. In contrast, where the smallest effect size of interest aligns or is smaller than what is often achieved in research, a change in practice or conducting a study to investigate further may be more appropriate. The results obtained in the present review, however, do suggest that much of the research conducted when comparing tendinopathy interventions is underpowered, particularly where relatively similar interventions are compared. Further research investigating the statistical approaches, the number of tests commonly performed and how conclusions are established based on hypothesis test construction in tendinopathy is warranted.

There are several limitations that should be considered when interpreting and evaluating the results and findings from the present study. Most of the meta-analyses were conducted with pooled data obtained across a range of tendinopathies, outcome domains, outcome tools, and measurement durations. Therefore, in some contexts the results obtained may not accurately represent the underlying true effect size distribution. In addition, the precision of estimates may be influenced by the reliance of the review on published data that may include a bias towards larger values. Attempts were made to include results from unpublished studies; however, none met the inclusion criteria. Additionally, whilst standardised mean differences facilitate pooling of different outcomes across studies to provide more data to estimate effect size distributions, these values are not as clinically interpretable as absolute mean differences.

5.0 Conclusion

This systematic review with meta-analysis provides evidence that the interpretation of comparative effect sizes in tendinopathy trials should vary across multiple contexts, with implications for both clinicians and researchers. Given exercise is generally accepted as the most appropriate and first line conservative management in tendinopathy, there is interest in determining which exercise therapy is the most effective. Given the advancement in this area and development of some standard practices, the research presented here highlights that clinicians and researchers should be aware that slight modifications are likely to have limited difference on average. It is recommended that researchers should not use Cohens general benchmarks [41] when planning or interpreting studies as these are likely to be too large in most cases. Instead, the values presented here should be used as a guide. Given there is still likely to be an interest in comparing therapies with relatively small adjustments, researchers may need to consider different analysis approaches. Considering the relatively low effect sizes that are likely to be generated, methods such as high frequency data collection combined with appropriate statistical analyses should be considered.

6.0 References

1) Fu FH, Wang JH-C, Rothrauff BB. BMJ Best Practice Tendinopathy. 2019. Available from: https://bestpractice.bmj.com/topics/en-gb/582 [Accessed 20th July 2019].

2) Ackermann PW, Renström P. Tendinopathy in sport. Sports health. 2012;4(3):193-201. Doi:10.1177/1941738112440957.

3) Hopkins C, Fu SC, Chua E, Hu X, Rolf C, Mattila VM, Qin L, Yung PS, Chan KM. Critical review on the socio-economic impact of tendinopathy. Asia-Pacific journal of sports medicine, arthroscopy, rehabilitation and technology. 2016;4:9-20. Doi:10.1016/j.asmart.2016.01.002.

4) Scott A, Squier K, Alfredson H, Bahr R, Cook JL, Coombes B, de Vos RJ, Fu SN, Grimaldi A, Lewis JS, Maffulli N. Icon 2019: international scientific tendinopathy symposium consensus: clinical terminology. British journal of sports medicine. 202;54(5):260-2. Doi:10.1136/bjsports-2019-100885.

5) Abat F, Alfredson H, Cucciarini M, Madry H, Marmott A, Mouton C, et al. Current trends in tendinopathy: consensus of the ESSKA basic science committee. Part I: biology, biomechanics, anatomy and an exercise-based approach. J Exp Orthop. 2017; 4:18. Doi:10.1186/s40634-017-0092-6.

6) Cooper K, Alexander L, Brandie D, Brown VT, Greig L, Harrison I, MacLean C, Mitchell L, Morrissey D, Moss RA, Parkinson E, Pavlova AV, Shim J, Swinton PA (2023). Exercise therapy for tendinopathy: a mixed methods evidence synthesis exploring feasibility, acceptability and effectiveness. Health Technology Assessment. 2023 (in press). Doi:10.3310/TAHK7102.

7) Shim J, Pavlova AV, Moss RA, MacLean C, Brandie D, Mitchell L, Greig L, Parkinson E, Brown VT, Morrissey D, Alexander L. Patient Ratings in Exercise Therapy for the Management of Tendinopathy: A Systematic Review With Meta-analysis. Physiotherapy. 2023 (in press). Doi:10.1016/j.physio.2023.05.002.

8) Swinton PA, Shim JS, Pavlova AV, Moss R, Maclean C, Brandie D, Mitchell L, Greig L, Parkinson E, Brown VT, Morrissey D. What are small, medium and large effect sizes for exercise treatments of

tendinopathy? A systematic review and meta-analysis. BMJ Open Sport & Exercise Medicine. 2023;9(1):e001389. Doi:10.1136/bmjsem-2022-001389.

9) Johannsen F, Olesen JL, Øhlenschläger TF, Lundgaard-Nielsen M, Cullum CK, Jakobsen AS, Rathleff MS, Magnusson PS, Kjær M. Effect of ultrasonography-guided corticosteroid injection vs placebo added to exercise therapy for Achilles tendinopathy: a randomized clinical trial. JAMA network open. 2022;5(7):e2219661. Doi:10.1001/jamanetworkopen.2022.19661.

10) Bussin E, Cairns B, Gerschman T, Fredericson M, Bovard J, Scott A. Topical diclofenac vs placebo for the treatment of chronic Achilles tendinopathy: A randomized controlled clinical trial. Plos one. 2021;16(3):e0247663. Doi:10.1371/journal.pone.0247663.

11) Murphy M, Travers M, Chivers P, Debenham J, Docking S, Rio E, Gibson W. Is heavy eccentric calf training our best option for mid-portion Achilles tendinopathy? A systematic review and meta-analysis. Journal of Science and Medicine in Sport. 2018;21:S83. DOI:10.1016/j.jsams.2018.09.190.

12) Lim HY, Wong SH. Effects of isometric, eccentric, or heavy slow resistance exercises on pain and function in individuals with patellar tendinopathy: A systematic review. Physiotherapy Research International. 2018;23(4):e1721. Doi: 10.1002/pri.1721.

13) Chen Z, Baker NA. Effectiveness of eccentric strengthening in the treatment of lateral elbow tendinopathy: A systematic review with meta-analysis. Journal of Hand Therapy. 2021;34(1):18-28. Doi:10.1016/j.jht.2020.02.002.

14) Meyer A, Tumilty S, Baxter GD. Eccentric exercise protocols for chronic non-insertional Achilles tendinopathy: how much is enough? Scandinavian journal of medicine & science in sports. 2009;19(5):609-15. Doi:10.1111/j.1600-0838.2009.00981.x.

15) Young JL, Rhon DI, de Zoete RM, Cleland JA, Snodgrass SJ. The influence of dosing on effect size of exercise therapy for musculoskeletal foot and ankle disorders: a systematic review. Brazilian Journal of Physical Therapy. 2018;22(1):20-32. Doi: 10.1016/j.bjpt.2017.10.001.

16) Doiron-Cadrin P, Lafrance S, Saulnier M, Cournoyer É, Roy JS, Dyer JO, Frémont P, Dionne C, MacDermid JC, Tousignant M, Rochette A. Shoulder rotator cuff disorders: a systematic review of clinical practice guidelines and semantic analyses of recommendations. Archives of physical medicine and rehabilitation. 2020;101(7):1233-42. Doi:10.1016/j.apmr.2019.12.017.

17) Gutierrez-Espinoza H, Araya-Quintanilla F, Cereceda-Muriel C, Alvarez-Bueno C, Martinez-Vizcaino V, Cavero-Redondo I. Effect of supervised physiotherapy versus home exercise program in patients with subacromial impingement syndrome: a systematic review and meta-analysis. Physical Therapy in Sport. 2020;41:34-42. Doi:10.1016/j.ptsp.2019.11.003.

18) Thompson BT, Schoenfeld D. Usual care as the control group in clinical trials of nonpharmacologic interventions. Proceedings of the American Thoracic Society. 2007;4(7):577-82. Doi:10.1513/pats.200706-072JK.

19) Vicenzino B, de Vos RJ, Alfredson H, Bahr R, Cook JL, Coombes BK, Fu SN, Silbernagel KG, Grimaldi A, Lewis JS, Maffulli N. ICON 2019—International Scientific Tendinopathy Symposium Consensus: There are nine core health-related domains for tendinopathy (CORE DOMAINS): Delphi study of healthcare professionals and patients. British journal of sports medicine. 2020;54(8):444-51. Doi:10.1136/bjsports-2019-100894.

20) Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355. Doi:10.1136/bmj.i4919.

21) Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JA. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. Bmj. 2011;343. Doi:10.1136/bmj.d5928.

22: Program U, Nations Development. Human Development Reports. New York: United Nations. 2019. Available from: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI [Accessed 15th July 2019].

23) Alfredson H, Pietilä T, Jonsson P, Lorentzon R. Heavy-load eccentric calf muscle training for the treatment of chronic Achilles tendinosis. The American journal of sports medicine. 1998;26(3):360-6. Doi:10.1177/03635465980260030301.

24) Challoumas D, Millar NL. Risk of bias in systematic reviews of tendinopathy management: Are we comparing apples with oranges?. Translational Sports Medicine. 2021;4(1):21-37. Doi:10.1002/tsm2.196.

25) Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association. 2016;23(1):193-201. Doi:10.1093/jamia/ocv044.

26) Zampieri FG, Casey JD, Shankar-Hari M, Harrell Jr FE, Harhay MO. Using Bayesian methods to augment the interpretation of critical care trials. An overview of theory and example reanalysis of the alveolar recruitment for acute respiratory distress syndrome trial. American journal of respiratory and critical care medicine. 2021;203(5):543-52. Doi: 10.1164/rccm.202006-2381CP.

27) Brydges CR. Effect size guidelines, sample size calculations, and statistical power in gerontology. Innovation in aging. 2019 Aug;3(4):igz036. Doi:10.1093/geroni/igz036.

28) Gignac GE, Szodorai ET. Effect size guidelines for individual differences researchers. Personality and individual differences. 2016 Nov 1;102:74-8. Doi:10.1016/j.paid.2016.06.069.

29) Swinton PA, Burgess K, Hall A, Greig L, Psyllas J, Aspe R, Maughan P, Murphy A. Interpreting magnitude of change in strength and conditioning: Effect size selection, threshold values and Bayesian updating. Journal of sports sciences. 2022;40(18):2047-54. Doi:10.1080/02640414.2022.2128548.

30) Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organizational research methods. 2008 Apr;11(2):364-86. Doi:10.1177/1094428106291059.

31) Gelman A. Prior distributions for variance parameters in hierarchical models. International Society for Bayesian Analysis. 2006;3(1):515-533.

32) Verardi V, Vermandele C. Univariate and multivariate outlier identification for skewed or heavy-tailed distributions. The Stata Journal. 2018;18(3):517-32. Doi: 10.1177/1536867X18018003.

33) Bürkner PC. brms: An R package for Bayesian multilevel models using Stan. Journal of statistical software. 2017;80:1-28. Doi:10.18637/jss.v080.i01.

34) Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis: Taylor & Francis; 2014.

35) Steuri R, Sattelmayer M, Elsig S, Kolly C, Tal A, Taeymans J, Hilfiker R. Effectiveness of conservative interventions including exercise, manual therapy and medical management in adults with shoulder impingement: a systematic review and meta-analysis of RCTs. British journal of sports medicine. 2017;51(18):1340-7. Doi:10.1136/bjsports-2016-096515.

36) Murphy MC, Travers MJ, Chivers P, Debenham JR, Docking SI, Rio EK, Gibson W. Efficacy of heavy eccentric calf training for treating mid-portion Achilles tendinopathy: a systematic review and meta-analysis. Br J Sports Med. 2019;53(17):1070-1077. Doi:10.1136/bjsports-2018-099934. 37) Van Der Vlist AC, Winters M, Weir A, Ardern CL, Welton NJ, Caldwell DM, Verhaar JA, De Vos RJ. Which treatment is most effective for patients with Achilles tendinopathy? A living systematic review with network meta-analysis of 29 randomised controlled trials. British journal of sports medicine. 2021;55(5):249-56. Doi:10.1136/bjsports-2019-101872.

38) Rompe JD, Segal NA, Cacchio A, Furia JP, Morral A, Maffulli N. Home training, local corticosteroid injection, or radial shock wave therapy for greater trochanter pain syndrome. The American journal of sports medicine. 2009;37(10):1981-90. Doi:10.1177/0363546509334374.

39) Engebretsen K, Grotle M, Bautz-Holter E, Sandvik L, Juel NG, Ekeberg OM, Brox JI. Radial extracorporeal shockwave treatment compared with supervised exercises in patients with subacromial pain syndrome: single blind randomised study. Bmj. 2009;339. Doi:10.1136/bmj.b3360.

40) Rompe JD, Nafe B, Furia JP, Maffulli N. Eccentric loading, shock-wave treatment, or a wait-and-see policy for tendinopathy of the main body of tendo Achillis: a randomized controlled trial. The American journal of sports medicine. 2007;35(3):374-83. Doi:10.1177/0363546506295940.

41) Cohen J. New York, NY: Routledge Academic; 1988. Statistical Power Analysis for the Behavioral Sciences.

42) Faul F, Erdfelder E, Lang AG, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91. Doi:10.3758/bf03193146.

43) Anvari F, Lakens D. Using anchor-based methods to determine the smallest effect size of interest. Journal of Experimental Social Psychology. 2021 Sep 1;96:104159. Doi:10.1016/j.jesp.2021.104159.