Doi:10.51224/SRXIV.469 SportRχiv Preprint version 1

Substantial underpowering of research investigating exercise interventions for tendinopathy: A quantitative review of the evidence base.

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.

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; @lynzalexander; @AHPkaycooper

Doi: 10.51224/SRXIV.469

SportRχiv hosted preprint version 1

13/10/2024

Please cite as: Swinton PA, Shim JSC, Pavlova AV, Morrissey D, Alexander L, Cooper K. Substantial underpowering of research investigating exercise interventions for tendinopathy: A quantitative review of the evidence base. 2024. Pre-print available from SportRχiv. https://doi.org/10.51224/SRXIV.469

Doi:10.51224/SRXIV.469 SportRχiv Preprint version 1

Abstract

Background: Tendinopathy is a common musculoskeletal condition, with exercise therapy being the cornerstone of conservative management. An extensive evidence base has compared various exercise therapies and alternative treatments, often using exercise therapy as a control. As in many fields, concerns have emerged of inflated effect sizes leading to underpowered studies and unreliable findings.

Objective: To review sample size determination in experimental trials investigating exercise therapy for tendinopathy management and to assess statistical power given the likely true effect sizes.

Eligibility criteria for selecting studies: A systematic literature search was conducted for trials involving participants diagnosed with rotator cuff, lateral elbow, gluteal, patellar or Achilles tendinopathy comparing exercise and non-exercise interventions.

Review and analysis: Details regarding the development and setting of target effect sizes, statistical models, and calculated sample sizes were extracted. Between intervention standardised mean difference effect sizes were calculated for studies comparing two exercise therapies, or an exercise therapy with a non-exercise intervention. Effect sizes were adjusted for expected overestimations using a shrinkage method based on the signal to noise ratio of observed values. Symmetric distributions centred on zero were then constructed with small, medium, and large thresholds determined. Sensitivity analyses assessed whether adjusted effect size distributions varied by tendinopathy location, outcome domain, and time from baseline measurement.

Results: The review included 126 studies, with 81 reporting a priori power calculations. The median target effect size used for powering studies was 0.85 (IQR: 0.64 to 1.06, range: 0.25 to 1.7), resulting in a median required group sample size of 21 (IQR: 17 to 34, range: 6 to 176). The most common statistical analysis was a t-test applied across group change scores. Shrinkage analysis was applied to 639 effect sizes from

exercise-vs-exercise (EvE) comparisons, and 988 effect sizes from exercise-vs-non-exercise (EvNE) comparisons. The thresholds for small, medium and large effects were equal to 0.04, 0.12, and 0.26 for EvE, and 0.08, 0.21, and 0.50 for EvNE. For medium effect sizes, the typical statistical analysis used suggests required group sizes of 1092 and 357, far exceeding sample sizes reecruited by factors of ~20 to 50. Sensitivity analyses showed distributions were consistent across tendinopathy location, outcome domain, and time from baseline measurement.

Conclusions: The findings highlight a prevalent issue of underpowering in exercise therapy trials for tendinopathy, often due to overestimated effect sizes. This review suggests that future research should incorporate adjusted effect size estimates to ensure adequate power and justify resource allocation. Researchers are encouraged to adopt more realistic effect size thresholds and consider alternative study designs, such as high-frequency data collection and real-world evidence studies, for incremental improvements in treatment. Enhanced methodological approaches and collaboration with clinicians and patients are needed to refine power calculations and improve the quality of tendinopathy research.

Key words: Sample size, Statistical power, Exercise therapy, Applied statistics

Introduction

Tendinopathy is a prevalent musculoskeletal condition affecting athletic, active, and sedentary populations. The condition is characterized by pain, functional impairment, and disability (1,2). Management strategies for tendinopathy include various interventions, with exercise therapy comprised predominantly of resistance exercise considered the cornerstone of conservative treatment (3,4). A recent scoping review revealed a substantial body of research on exercise therapies for tendinopathy, with randomized controlled trials (RCTs) being the most frequently employed study design (5). Notably, over 90% of this research focuses on conditions such as rotator cuff-related shoulder pain (RCRSP), lateral elbow tendinopathy, patellar tendinopathy, and Achilles tendinopathy (5). The most commonly assessed outcome domains include pain, disability, physical functional capacity, and shoulder range of motion (ROM) (5). Exercise therapies are often compared against each other or used as comparators to other treatments, including injections, laser therapy, extracorporeal shockwave therapy, manual therapy, and splinting/taping (5).

Most RCTs investigating exercise therapies adopt a parallel-group superiority design, with the primary goal of estimating average treatment effects between groups to inform clinical recommendations. Exercise therapies provide a cost-effective, non-invasive treatment option, making them ideal comparators in trials assessing the efficacy of more expensive or invasive alternatives such as surgery or injection therapy (6,7). Trials comparing different exercise therapies often focus on subtle modifications—such as concentric versus eccentric muscle actions or the influence of exercise setting—reflecting a process of incremental refinement aimed at identifying the optimal exercise stimulus for maximizing patient outcomes (8–10)

In RCTs, drawing inferences from group comparisons typically relies on statistical inference and null hypothesis testing. When testing the null hypothesis that two treatments produce an equivalent mean response, researchers mitigate the risk of Type I and Type II errors by setting a low threshold for rejecting

the null hypothesis (e.g., $\alpha = 0.05$) and recruiting sufficient sample sizes to ensure that the null hypothesis can be reliably rejected when false. The probability of correctly rejecting a false null hypothesis is referred to as statistical power, which, for a given sample size, depends on the true difference in mean response between interventions. Since the true population difference is unknown, researchers must specify an a priori target difference, or effect size, which the study is designed to detect (11). This target effect size is typically determined using one of two approaches: it may reflect a difference deemed important by stakeholders, such as healthcare professionals or patients, or it may be based on a realistic estimate derived from existing evidence, including meta-analyses of similar studies (11). Some argue that both approaches should be considered when selecting a target effect size (12).

Researchers face challenges in determining an appropriate effect size for powering their studies, regardless of the approach chosen. A large body of research exists in most clinical areas, including tendinopathy, to quantify minimal clinically important differences (MCIDs) for a given outcome. Whist the general conception of an MCID appears to align with target effect sizes from a patient's view of importance, many variations and definitions exist (13,14). In the study of tendinopathy, researchers have tended to quantify MCIDs based on within-patient change following a single intervention (15–18). In contrast, the primary goal of a parallel-group superiority RCT is to identify between-group differences and the effect size should reflect the importance of these differences. To distinguish between these concepts, some researchers use the term "minimal clinically important change" (MCIC) to denote within-group change, reserving "difference" for between-group comparisons (19). In studies comparing an established exercise therapy with a new intervention, the standard therapy may already produce a mean response that exceeds the MCIC. For additional improvements to be considered clinically meaningful, other factors such as cost, acceptability, and resource implications may be considered (11). Despite the relevance of these factors, methods such as health economic or decision-based analyses are seldom employed in the determination of effect sizes for sample size calculations (11,20). Given that most RCTs in tendinopathy use active controls

reflecting standard and effective treatments, powering studies based on within-group changes may inflate expectations of effect size, leading to reduced statistical power and compromised study quality.

When prior RCTs align with the interventions of a planned trial, the observed differences can inform the target effect size for power calculations. Whilst this is the most common approach used in health-related research (11), a naïve review of results from previous studies is likely to overestimate true effect sizes for several reasons (21). Effect sizes from newly discovered true (non-null) differences are often inflated due to the process of claiming "statistical significance," with greater inflation occurring when the statistical power of the research base is low (21,22). Additionally, factors such as researcher degrees of freedom, selective reporting, questionable research practices, and publication bias further inflate effect sizes (21). To address this issue, van Zwet and colleagues (23) developed a method to "shrink" inflated effect sizes reported in RCTs. Incorporating data from more than 20,000 RCTs included in the Cochrane Database of Systematic Reviews, they created a shrinkage estimator based on the joint distributions of reported effect sizes and their standard errors (23). Their analysis indicated that effect size estimates significant at the 5% level typically overestimate the true value by a factor of 1.7. The shrinkage estimator reduced this overestimation by adjusting for the signal-to-noise ratio and providing greater shrinkage for noisier estimates. Their approach showed superior performance in reducing error, bias, and exaggeration of effect sizes in both conditional and unconditional analyses (23). van Zwet et al. (23) recommend that this shrinkage estimator be reported alongside standard estimators in trials to minimize the risk of overestimated effects.

The primary aim of this quantitative analysis is to evaluate the effect sizes reported in prior experimental trials of exercise therapies for tendinopathy and assess the potential for overestimation. By reviewing realistic treatment effects and sample sizes from previous studies, this review seeks to determine the statistical power of the existing research. Trials comparing different exercise therapies will be analyzed

separately from those comparing exercise therapies with other active interventions, as the former may produce smaller effect sizes due to the similarities between treatments. Additionally, this review examines the methods researchers employ for conducting a priori power analyses, focusing on how target effect sizes are selected, the typical values of these effect sizes, and the statistical techniques used to test group differences. Understanding likely effect sizes and the factors influencing them is crucial for both researchers and clinicians to assess the reliability of the current research. Furthermore, this knowledge can guide future studies and highlight instances where additional resources may be necessary.

Methods

This quantitative review was part of a project funded by the National Institute for Health Research (NIHR) and follows on from a systematic review with meta-analysis quantifying non-comparative within-group effect size distributions (5,24). The following sections outline the approaches used to identify, select and extract information from included studies, and the methods used to shrink and describe the effect size distributions obtained.

Inclusion criteria

This quantitative review included research conducted with 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. 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 the online supplementary file. 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. We included studies that featured either a between exercise-only comparison, or between exercise and non-exercise comparison. Non-exercise interventions included injection, electrotherapy, biomechanical modifications, manual therapy or surgery (definitions are presented in the online supplementary file). Trial arms that included both exercise and non-exercise elements were not included. Based on the results of our initial scoping review (5) and subsequent stakeholder workshops, we extracted data from outcomes that assessed four domains: 1) disability; 2) physical function capacity; 3) pain (on loading/activity, over a specified time, or without further specification); and 6) ROM (shoulder joint only). Definitions of each domain and example tools used to measure the outcomes are presented in the online supplementary file. We included RCTs and non-randomized experimental trials that included treatment arms that matched our comparator and other

inclusion criteria. Studies were limited to those conducted in nations defined as very high or high on the Human Development Index (top 62 countries at the time of protocol development) (25).

Search strategy and data extraction

A detailed description of the search strategy used for this quantitative review has been presented previously (5), with search terms for MEDLINE presented in the online supplementary file. The searches were initiated from 1998 as this was the year the heavy load eccentric calf-training protocol by Alfredson et al (26) was published and represents the point at which research of exercise interventions for tendinopathies proliferated. The last date of the search was conducted on 25/03/2022.

Two independent reviewers screened titles and abstracts followed by full-text copies. Data relevant to the calculation of effect sizes and specifics of treatment arms, tendinopathy location, and outcome domain 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 online supplementary file. Each entry was then independently checked. Data relevant to power calculations and statistical models were extracted by a single statistician. For power calculations the alpha, beta, and minimum sample for each group were extracted based on the author text provided. The approach used by authors to generate a target effect size was categorised according to previous criteria (11) including: 1) anchor (focus on clinically important value through MCID or MCIC); 2) distribution (focus on minimum detectable change); 3) health economic (focus on difference relative to cost or other similar factor); 4) opinion-seeking (focus on meaningful difference as established by asking experts); 5) pilot study (use of information from smaller version of trial conducted by the authors); 6) review of evidence (focus on information including difference and population variance estimates from previous full trials); and 7) standardised effect size (use of conventional or topic specific thresholds on scale-invariant measure of effect). Where the authors provided relevant information such as the target difference and a measure of population variance, the magnitude of the target effect size was also extracted. Finally, as the statistical model and related processes also influence statistical

power, information relevant to these domains were also extracted. The primary statistical model used to test for mean differences between groups was categorised as: 1) parametric change omnibus (typified by analysis of variance (ANOVA) applied to change scores across more than two groups, or two or more time points); 2) non-parametric change omnibus (typified by Kruskal-Wallis test to change scores across more than two groups, or two or more time points); 3) parametric change (typified by independent Student's t-test applied to change scores at a single time point between two groups); 4) nonparametric change (typified by Mann-Whitney U test applied to change scores at a single time point between two groups); 5) ANOVA interaction (typified by the group-by-time interaction in a mixed ANOVA); 6) analysis of covariance (typified by a linear regression model of a post-outcome where baseline value and other potential covariances along with group are included as explanatory variables); 7) hierarchical or multilevel (typified by linear mixed models or generalized estimating equations); and 8) multivariate (typified by a multivariate ANOVA combining multiple outcomes in the same model). A record was also made if authors included alpha-adjustments in their analysis to account for multiple tests either over time or outcomes.

Data processing and analyses

Comparative between group effect sizes were calculated for studies comparing at least two trial arms that matched the inclusion criteria. 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 sample 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 (27). 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.

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 (28):

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

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 $Sd_{AB\,Pre} = \sqrt{\frac{(n_A-1)Sd_{A\,Pre}+(n_B-1)Sd_{B\,Pre}}{n_A+n_B-2}}$.

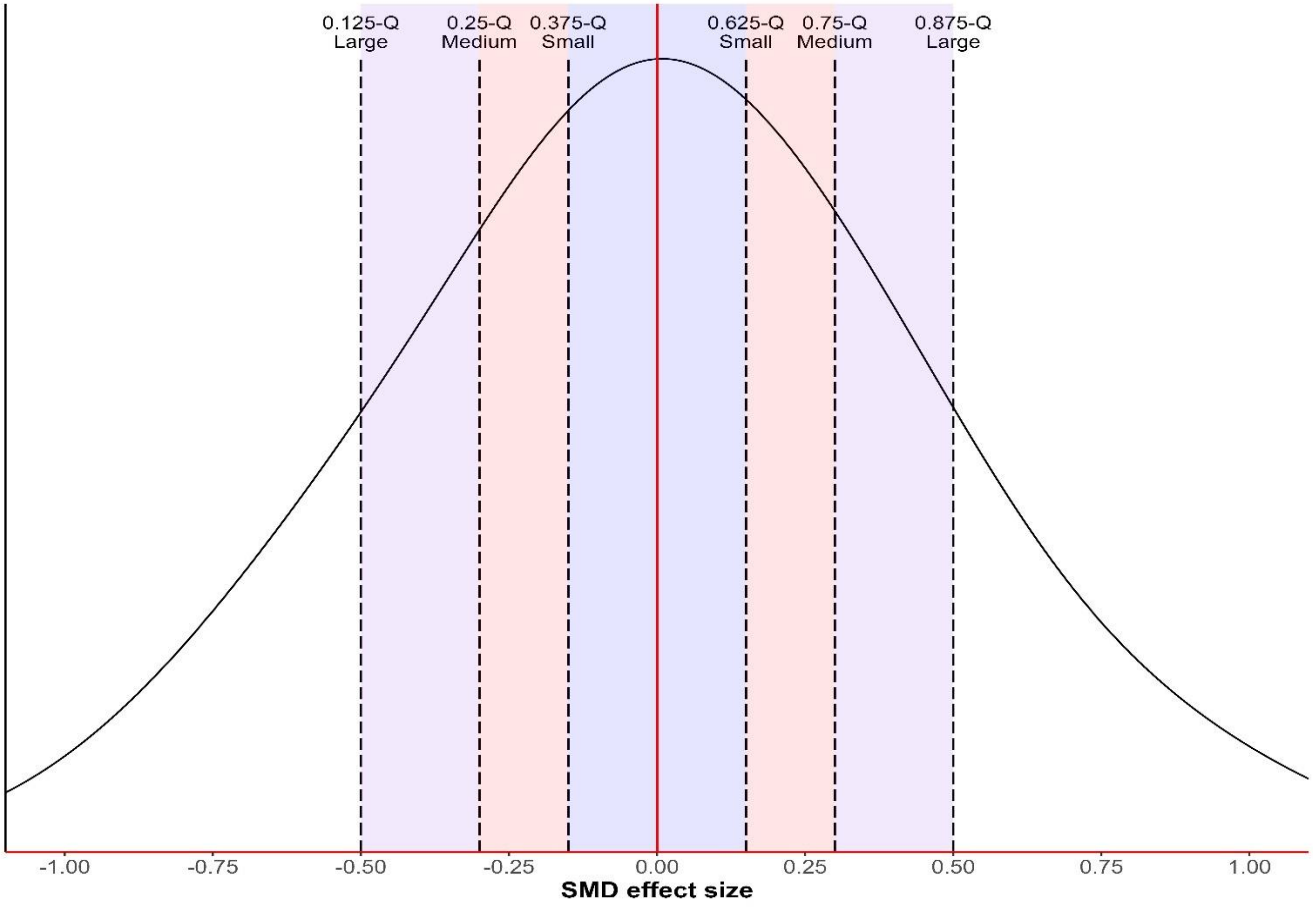
$$\sigma^2\left(\mathrm{SMD}_{AB\mathrm{pre}}\right) = 2c(n_A + n_B - 2)^2(1-\rho)\left(\frac{n_A + n_B}{n_A n_B}\right)\left(\frac{n_A + n_B - 2}{n_A n_B}\right)\left(1 + \frac{\mathrm{SMD}^2_{AB\mathrm{pre}}}{2(1-\rho)\left(\frac{n_A + n_B}{n_A n_B}\right)}\right) - \mathrm{SMD}^2_{AB\mathrm{pre}}$$

where ρ is the correlation between repeated measures and we imputed with the value 0.7 (24).

Shrinkage was applied to each effect size using the method presented by van Zwet and colleagues (23). Briefly, an individual trial was represented by a set of three numbers β, b, s, where β is the treatment effect, b is our standard estimate $\mathrm{SMD}_{AB\mathrm{pre}}$, and s is the standard error $\sigma\left(\mathrm{SMD}_{AB\mathrm{pre}}\right)$. We also define the z-value $z = b/s$ and the signal-to-noise ratio SNR, where $SNR = \beta/s$. Using effect size estimates and standard errors from Cochrane Database of Systematic Reviews, van Zwet and colleagues (23) estimated the marginal distribution of z-values as a mixture of four zero-mean normal distributions returning four variance values (0.61^2,1.42^2,2.16^2,5.64^2) and four mixing proportions (0.32,0.31,0.3,0.07). Since $\beta = s \cdot SNR$, van Zwet and colleagues (23) proposed the shrinkage estimator $\hat{\beta} = s \cdot \hat{\mathrm{E}}(SNR|z)$, where the conditional expectation is calculated on the basis of the distribution being comprised as a mixture of zero-mean normal distributions. For each analysis, the 0.625-quantile/|0.375|-quantile, 0.75-quantile/|0.25|-quantile, and 0.875-quantile/|0.125|-quantile values of the shrunken effect estimates were obtained to quantify small,

medium, and large thresholds, respectively (see Figure 1). Boot strapping was applied in each analysis to quantify uncertainty in the thresholds calculated. Boot strapping was achieved by randomly sampling with replacement effect sizes, then performing a series of random shuffles multiplying all effects sizes obtained within a single treatment comparison by either 1 or -1 and performing the calculations across 10,000 bootstrap samples. The 0.025- and 0.975-quantiles of the bootstrap samples were used to obtain 95% confidence intervals (95%CI). The process was completed separately for exercise vs exercise, and exercise vs non-exercise comparisons. The primary analyses included all effect sizes, with sensitivity analyses then conducted with effect sizes obtained from different tendinopathy locations, different outcome domains, and different time points where the total number of effect sizes was equal to or greater than one hundred.

Figure 1: Schematic showing a hypothetical effect size distribution centred on zero and the small, medium, and large thresholds comprising 25, 50, and 75% of the overall distribution, respectively.



SMD: Standardized mean difference effect size

Results

Descriptions of data

Data from a total of 126 studies comprising RCRSP (57 studies), Achilles (28 studies), lateral elbow (24 studies), patellar (15 studies), and gluteal (2 studies) tendinopathies were included (see supplementary online file). A total of 61 studies provided 130 different pairwise comparisons across different exercise versus exercise comparisons creating a total of 639 comparative effect sizes. The dominant exercise classes across these treatment arms were resistance exercise (74.1%), flexibility (15.7%), proprioception (9.0%), and vibration (1.2%). A total of 73 studies provided 175 different pairwise comparisons across exercise versus non-exercise comparisons creating a total of 988 comparative effect sizes. The dominant non-exercise classes across these treatment arms were electrotherapy (22.8%), injections (22.8%), biomechanical modifications (21.5%), manual therapy (17.7%) and surgery (15.2%).

Power calculations and statistical models

A total of 81 studies (64.3%) reported that they performed a power calculation to inform sample size determination. Within these studies, the most common methods used to obtain a target effect size was review of evidence (29 studies 35.8%) and anchor methods (24 studies 29.6%), followed by use of pilot studies (5 studies 6.2%), standardised effect size (2 studies 2.5%), and opinion seeking (2 studies 2.5%). Eighteen studies (22.2%) did not provide enough information to identify the method used to obtain a target effect size, and 1 study (1.2%) adopted a hybrid approach where the power calculation was initially based on an anchor method but then updated following a pilot study. From these 81 studies, 34 (42.0%) provided sufficient information on the effect size value used. The median effect size was 0.85 (IQR: 0.64 to 1.06) and ranged from 0.25 to 1.7. Seventy-six studies reported the power and alpha value, with 64 studies setting power to 0.8, and values ranging from 0.8 to 0.99. Seventy-three studies set alpha to 0.05, with values ranging from 0.05 to 0.5. Seventy-nine studies reported the required group sample size with a

median value of 21 (IQR: 17 to 34) and range of 6 to 176. This distribution was close to the actual number of participants included across all 126 studies, with a median value of 21 (IQR: 14 to 30) and range of 5 to 178. Similar distributions were also obtained for the actual number of participants included in studies that compared exercise versus exercise (median: 20 (IQR: 12 to 29) and range of 5 to 174), or exercise with non-exercise interventions (median: 23 (IQR: 15 to 33) and range 8 to 178).

Of the 126 studies included, 122 (96.8%) provided sufficient detail of the statistical analyses used, and 119 studies reported performing a statistical test of a group difference null hypothesis. Sixty-five studies reported performing a statistical test on change scores (12 studies starting with parametric omnibus tests, 29 studies stating with parametric pairwise tests, 13 studies starting with nonparametric omnibus tests, and 11 studies stating with nonparametric pairwise tests). Twenty-two studies performed mixed ANOVAs testing for group differences with the time by group interaction. Sixteen studies performed analysis of covariance (ANCOVA), 14 studies performed hierarchical or multilevel models (9 studies linear mixed models and 5 studies generalized estimating equations), 1 study performed a MANOVA, and in 1 study the primary analysis was conducted with Fisher's exact test. Twenty-nine studies reported using an alpha adjustment approach in their hypothesis testing.

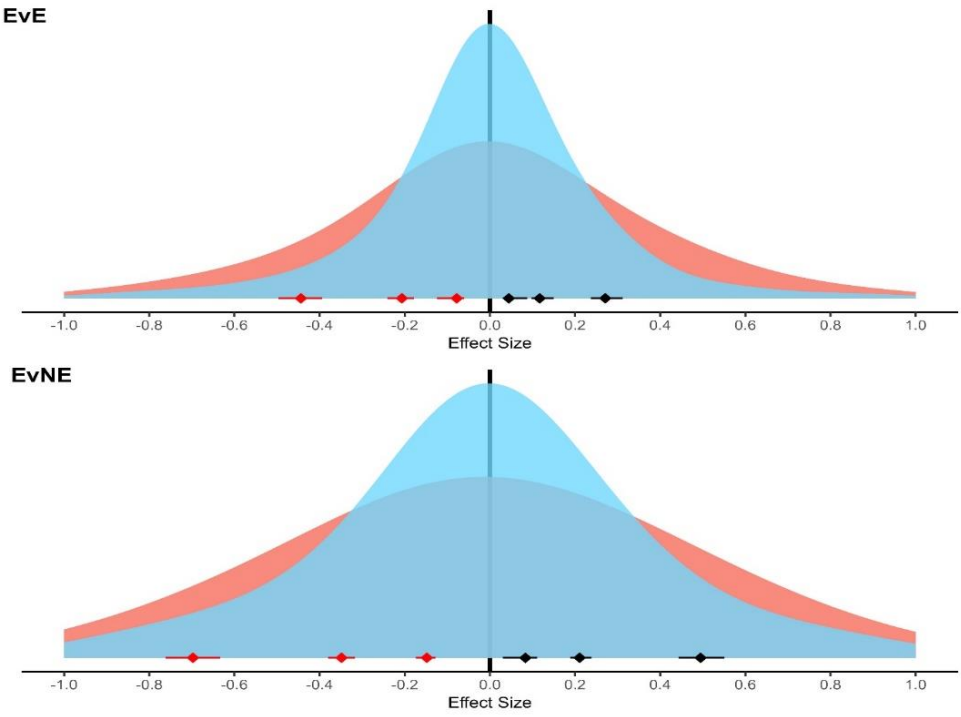
Effect size distributions

Across all 639 effect sizes obtained for comparisons of different exercise therapies, the directly observed values provided small: 0.08 (95%CI: 0.06 to 0.12), medium: 0.20 (95%CI: 0.18 to 0.24), and large 0.44 (95%CI: 0.39 to 0.49) thresholds. Following shrinkage, the thresholds reduced to small: 0.04 (95%CI: 0.03 to 0.08), medium: 0.12 (95%CI: 0.10 to 0.14), and large 0.26 (95%CI: 0.23 to 0.30). A visual comparison of the effect sizes prior to and following shrinkage is presented in Figure 2. Sensitivity analyses across tendinopathy locations and time periods did not identify any substantial changes in threshold estimates

(Table 1). Potential differences were identified across outcome domains, with greater estimates generally obtained for outcomes measuring pain (Table 1).

Greater effect size values were consistently obtained for exercise versus non-exercise comparisons, than were obtained with exercise versus exercise comparisons. Across all 988 exercise versus non-exercise effect sizes, the directly observed values provided small: 0.15 (95%CI: 0.13 to 0.17), medium: 0.35 (95%CI: 0.32 to 0.38), and large 0.70 (95%CI: 0.63 to 0.76) thresholds. Following shrinkage, the thresholds reduced to small: 0.08 (95%CI: 0.07 to 0.11), medium: 0.21 (95%CI: 0.19 to 0.24), and large 0.50 (95%CI: 0.44 to 0.55). A visual comparison of the effect sizes prior to and following shrinkage is presented in Figure 2. Sensitivity analyses across tendinopathy locations and time periods did not identify any substantial changes in threshold estimates (Table 1).

Figure 2: Density plot of between group standardised mean difference effect sizes prior to and post 'shrinkage' for exercise versus exercise (EvE) comparisons, and exercise versus non-exercise (EvNE) comparisons.



Density plot in red illustrates standardised mean different effect sizes prior to shrinking. Density plot in blue illustrates post shrinking. Red diamonds and whiskers illustrate small, medium and large thresholds prior to shrinking and associated 95% confidence intervals. Black diamonds and whiskers illustrate small, medium and large thresholds after shrinking and associated 95% confidence intervals. Labelling of effect sizes as positive and negative is only for formatting.

Table 1: Small, medium and large thresholds obtained for directly observed and shrunken standardized mean difference effect sizes for exercise vs. exercise, and exercise vs. non-exercise comparisons.

Comparison		**Exercise vs. Exercise**				**Exercise vs. Non-exercise**		
	Data	*Small* (95%CI)	*Medium* (95%CI)	*Large* (95%CI)	*Data*	*Small* (95%CI)	*Medium* (95%CI)	*Large* (95%CI)
All Data								
No Shrinkage	Effect sizes: 639 TA comparisons: 130 Studies: 61	0.08 (0.06 to 0.12)	0.20 (0.18 to 0.24)	0.44 (0.39 to 0.49)	Effect sizes: 988 TA comparisons: 175 Studies: 73	0.15 (0.13 to 0.17)	0.35 (0.32 to 0.38)	0.70 (0.63 to 0.76)
Shrinkage		0.04 (0.03 to 0.08)	0.12 (0.10 to 0.14)	0.26 (0.23 to 0.30)		0.08 (0.07 to 0.11)	0.21 (0.19 to 0.24)	0.50 (0.44 to 0.55)
Time (Short)								
No Shrinkage	Effect sizes: 445 TA comparisons: 116 Studies: 55	0.10 (0.08 to 0.13)	0.25 (0.21 to 0.28)	0.48 (0.42 to 0.55)	Effect sizes: 571 TA comparisons: 134 Studies: 56	0.14 (0.12 to 0.17)	0.33 (0.29 to 0.38)	0.67 (0.59 to 0.76)
Shrinkage		0.06 (0.04 to 0.09)	0.14 (0.12 to 0.17)	0.29 (0.25 to 0.34)		0.08 (0.06 to 0.11)	0.20 (0.17 to 0.24)	0.47 (0.41 to 0.55)
Time (Intermediate and long)								
No Shrinkage	Effect sizes: 194 TA comparisons: 58 Studies: 28	0.06 (0.01 to 0.15)	0.13 (0.09 to 0.21)	0.31 (0.24 to 0.41)	Effect sizes: 417 TA comparisons: 101 Studies: 42	0.16 (0.12 to 0.21)	0.37 (0.32 to 0.43)	0.74 (0.64 to 0.87)
Shrinkage		0.05 (0.01 to 0.13)	0.08 (0.05 to 0.17)	0.19 (0.14 to 0.27)		0.09 (0.07 to 0.14)	0.23 (0.19 to 0.28)	0.53 (0.44 to 0.65)
Outcome domain (Pain)								
No Shrinkage	Effect sizes: 222 TA comparisons: 95 Studies: 44	0.13 (0.09 to 0.18)	0.30 (0.24 to 0.37)	0.57 (0.49 to 0.68)	Effect sizes: 325 TA comparisons: 115 Studies: 48	0.18 (0.14 to 0.22)	0.39 (0.34 to 0.46)	0.78 (0.68 to 0.89)
Shrinkage		0.07 (0.05 to 0.12)	0.18 (0.13 to 0.23)	0.36 (0.30 to 0.45)		0.10 (0.08 to 0.13)	0.25 (0.21 to 0.30)	0.57 (0.49 to 0.68)
Outcome domain (Disability)								
No Shrinkage	Effect sizes: 177 TA comparisons: 103	0.09 (0.05 to 0.14)	0.21 (0.16 to 0.27)	0.43 (0.35 to 0.53)	Effect sizes: 317 TA comparisons: 156	0.15 (0.12 to 0.21)	0.37 (0.30 to 0.44)	0.72 (0.62 to 0.86)

Shrinkage	Studies: 48	0.05 (0.03 to 0.10)	0.12 (0.09 to 0.16)	0.25 (0.20 to 0.33)	Studies: 65	0.09 (0.06 to 0.15)	0.22 (0.18 to 0.29)	0.52 (0.43 to 0.64)
Outcome domain (Physical functional capacity)								
No Shrinkage	Effect sizes: 136 TA comparisons: 44 Studies: 21	0.07 (0.02 to 0.15)	0.15 (0.10 to 0.22)	0.33 (0.24 to 0.44)	Effect sizes: 116 TA comparisons: 47 Studies: 21	0.14 (0.08 to 0.25)	0.31 (0.22 to 0.46)	0.64 (0.46 to 0.89)
Shrinkage		0.05 (0.01 to 0.11)	0.09 (0.06 to 0.16)	0.20 (0.14 to 0.28)		0.10 (0.05 to 0.24)	0.20 (0.13 to 0.36)	0.45 (0.28 to 0.67)
Tendinopathy type (Rotator cuff related shoulder pain)								
No Shrinkage	Effect sizes: 310 TA comparisons: 60 Studies: 29	0.08 (0.06 to 0.12)	0.19 (0.15 to 0.23)	0.37 (0.31 to 0.43)	Effect sizes: 593 TA comparisons: 87 Studies: 35	0.15 (0.12 to 0.19)	0.36 (0.31 to 0.40)	0.69 (0.61 to 0.78)
Shrinkage		0.05 (0.03 to 0.08)	0.11 (0.08 to 0.14)	0.22 (0.19 to 0.27)		0.08 (0.06 to 0.12)	0.22 (0.19 to 0.26)	0.50 (0.43 to 0.58)
Tendinopathy type (Achilles)								
No Shrinkage	Effect sizes: 136 TA comparisons: 26 Studies: 11	0.07 (0.02 to 0.17)	0.22 (0.16 to 0.29)	0.47 (0.38 to 0.57)	Effect sizes: 136 TA comparisons: 39 Studies: 17	0.15 (0.09 to 0.21)	0.31 (0.23 to 0.41)	0.60 (0.45 to 0.79)
Shrinkage		0.05 (0.01 to 0.11)	0.12 (0.09 to 0.17)	0.28 (0.22 to 0.35)		0.09 (0.05 to 0.14)	0.19 (0.14 to 0.30)	0.40 (0.29 to 0.54)
Tendinopathy type (Lateral elbow)								
No Shrinkage					Effect sizes: 207 TA comparisons: 37 Studies: 16	0.16 (0.10 to 0.28)	0.37 (0.28 to 0.49)	0.77 (0.61 to 0.96)
Shrinkage						0.11 (0.06 to 0.24)	0.23 (0.16 to 0.37)	0.55 (0.42 to 0.71)

TA: Treatment arms. CI: Confidence interval.

4.0 Discussion

This quantitative review provides several important insights into the evidence base surrounding exercise therapy for tendinopathy management. A key finding is the likely widespread underpowering of experimental trials, raising concerns about the confidence that can be placed in their findings. Underpowering appears to be more pronounced in trials comparing different exercise therapies than in those comparing exercise therapies to non-exercise interventions such as injections or electrotherapy. The analysis indicates that the research base has likely overestimated the effect sizes that should be observed in trials, partly due to inflated effect sizes in published studies and misunderstandings regarding the definition of clinical importance and its role in informing sample size calculations. To address these overestimations, this review applied a shrinkage approach that adjusts effect sizes based on the signal-to-noise ratio from the available data. Future trials investigating exercise therapies for tendinopathy should incorporate these adjusted, or "shrunken," estimates when determining sample size, ensuring that adequate resources are allocated and justified. When comparing similar interventions with only minor modifications, researchers should expect small effect sizes, often approaching zero. In such cases, the traditional superiority parallel pre-post design may not be advisable unless it is deemed appropriate to use resources significantly exceeding those from previous studies. Additionally, this review highlights limitations in the statistical practices of many prior studies, which have further contributed to reduced statistical power.

The effect sizes used to power studies and the corresponding sample sizes reported in the included trials deviate significantly from previously established benchmarks. Rothwell et al. (11) reviewed the sample size practices of 107 RCTs in health research, reporting a median standardized effect size of 0.30 (IQR: 0.28 to 0.38) and a maximum anticipated value of 0.76. In contrast, this review, which included 34 studies with sufficient data for effect size calculation, identified a median effect size of 0.85 (IQR: 0.625 to 1.05), with values ranging from 0.25 to 1.7. The discrepancy between the present findings and those of Rothwell et al. (11) may be partly explained by their focus on RCTs funded by the National Institute of Health Research and published in their Health Technology Assessment Journal, which tend to be of higher quality. Another

contributing factor could be the inconsistent differentiation between within-group change and between-group difference in prior studies. In our previous quantitative review of standardized *change* scores in tendinopathy, we found that effect size distributions varied across outcome domains and differences could be extensive (24). For self-reported measures of pain, disability, and function, small, medium, and large effect size thresholds were approximately 0.6, 1.0, and 1.7, respectively (24). However, the current study demonstrates that across all outcome domains, standardized *difference* scores are considerably smaller, with more realistic values for small, medium, and large effect sizes following shrinkage being 0.05-0.1, 0.1-0.25, and 0.25-0.5, respectively. This discrepancy between standardized change scores and difference scores explains why fewer participants are generally required to detect a meaningful within-group change, while far larger sample sizes are needed to detect differences between interventions. A smaller number of studies have included non-active interventions arms, such as wait-and-see groups, which allow researchers to differentiate between improvements due to natural healing processes and additional improvements from active intervention (29). It should be expected that effect sizes for RCTs comparing active and non-active interventions are greater than those comparing active interventions, but smaller than those reflecting within-group change.

The two most common approaches to determining target effect sizes identified in this review were the use of clinical importance measures and the application of results from previous studies. Frequently, clinical importance was conceptualized in terms of MCID. However, the MCID literature is complex, with inconsistent terminology and a variety of methodological approaches (14). The concept was originally introduced by Jaeschke et al. (30) in 1989, who defined the MCID as "the smallest difference in score in the domain of interest that patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive costs, a change in the patient's management". The standard method for this is to calculate the mean change in the outcome of interest among patients who rated their global change from baseline as "slightly better." This approach is now commonly referred to as the mean change method (31). Following Jaeschke et al (30), the concept of MCID was further developed by Redelmeier

and Lorig (32), who in 1993 adopted an approach where they subtracted the mean change scores from patients that rated their global change as 'about the same', from those that rated their global change as 'slightly better'. This approach is commonly referred to as the mean difference of change method (31) and is likely to better reflect important differences in a between group context relevant to RCTs. Often those that rate their global change as 'about the same' are still expected to show improvements in the outcome of interest, such that the mean difference of change method will return lower values than the mean change method.

Regardless of the method used, the clinical importance measure must be scaled appropriately for use in standard sample size calculations. Typically, researchers employ relatively simple software, such as G*Power 3, for these calculations. To use this software, the clinical importance measure must be standardized—typically by dividing by an estimate of the population standard deviation (33). In the studies included in this review, this estimate was often derived from previous research (16,34,35). If the estimated standard deviation is too small, the target effect size will be inflated, resulting in an underpowered study. Therefore, knowledge of typical standardized effect sizes as presented in this review is crucial to ensuring that the translation of MCID or MCIC into sample size calculations is appropriate and accurate.

We identified that the most popular approach to obtain target effect sizes for the included studies was a review of previous research. This approach, while common, may be subject to various biases. In Rothwell et al's (11) previous review of target effect sizes in health research, by focussing on the Health Technology Assessment Journal studies, the authors not only restricted their analysis to higher quality studies but also mitigated the effects of publication bias, as all NIHR-funded projects are required to publish regardless of results. However, it is important to note that publication bias is not the only factor contributing to inflated effect sizes in research. Other factors, such as selective reporting, p-hacking, and small sample sizes, can also lead to overestimation of effects (36). The use of shrinkage methods, as adopted in the present review, provides an approach to account for this overestimation. The results from the present review show

substantial reductions in effect size distributions, with smaller values obtained for within-exercise comparison, as would be expected from comparisons of inherently similar interventions. Sensitivity analyses did not identify consistent differences in exercise effect size distributions when organised by tendinopathy type, outcome domain, or time. Larger data sets would be required, however, to formally compare distributions.

In addition to the low relative sample sizes, the current review identified that most studies employ basic statistical models that do not make efficient use of data, thereby further reducing statistical power. The most common statistical analyses included parametric (e.g. independent t-test) and non-parametric (e.g. Mann-Whitney U) tests performed on group change scores. The next most common statistical analysis was ANOVA and the time-by-group interaction effect. It is consistently recommended, however, that ANCOVA be preferred over ANOVA and test of change scores due to lower bias, superior precision and greater statistical power (37). Additionally, many studies included alpha adjustment approaches to account for family-wise error rates when conducting multiple tests due to additional time points and secondary outcomes. In all cases, Bonferroni corrections were selected, which are among the most conservative of alpha adjustment approaches, assuming independence between tests that is often unlikely and reducing statistical power to unacceptable levels (38).

The substantial underpowering of the exercise and tendinopathy research base is best demonstrated using the data extracted and the most common statistical test employed. When using the median effect size reported in the a priori power calculations (0.85) and performing a t-test on the change scores, a sample size of 16 per group is required for standard parameters (α=0.05 and statistical power of 0.8). This sample size aligns closely with the median observed in the research, suggesting that many studies may be adequately powered based on their reported target effect sizes. However, when more realistic effect sizes are used the required sample sizes increase substantially. For medium effect sizes of 0.35 and 0.25 based on directly observed values prior to shrinkage, the required sample sizes increase to 130 and 253 per group. After

applying shrinkage, medium effect sizes can be expected to reduce in some instances to ~0.12 (see Table 1), requiring a sample sizes of 1092 per group. Similarly, for small effect sizes of 0.08 and 0.04, the required sample sizes increase to 2454 and 9813 per group. These numbers would fall beyond the practical capabilities of most research, highlighting a significant challenge in detecting small but potentially meaningful effects over time and within an incremental perspective of treatment.

The findings presented in this review are not unique to exercise therapy and tendinopathy; similar issues have been observed in fields such as psychology and sports science, where small effect sizes are of interest but resource constraints often limit the feasibility of increasing sample sizes (39–41). In the context of experimental trials, researchers may need to explore novel approaches, such as high-frequency data collection paired with more advanced statistical techniques (42,43). Another argument is that well-conducted systematic reviews with meta-analyses can help address the low reliability of underpowered studies by pooling their results to generate more precise estimates. However, while research shows most studies included in systematic reviews are underpowered (44), the pooling of heterogenous studies in terms of populations, outcomes and interventions often limits the strength of findings. Additionally, the time required to conduct and synthesize sufficient primary research is considerable, leaving clinicians and researchers with limited high-quality information to guide decision-making in the meantime. An alternative approach to enhance comparative effectiveness research is to supplement RCTs with real-world evidence (RWE) studies (45). RWE research may offer better insights into practical issues with treatments and address concerns about the diversity of participants included in RCTs (46–48). However, conducting informative RWE research poses several challenges, including the collection of high-quality data on a large scale, ensuring reproducibility and replicability, and explaining complex analyses in a way that earns the trust of both clinicians and patients (49,50).

There are several limitations to this quantitative review that should be considered when interpreting the findings. While the analysis includes a relatively large number of effect sizes, these were drawn from a

much smaller pool of studies. The tendency for most exercise and tendinopathy studies to report multiple outcomes across various time points (5) raises concerns about the number of potential type I errors in the research base also. Studies that did not focus exclusively on exercise therapies or that combined exercise with non-exercise therapies were excluded from this analysis. However, it is likely that studies comparing different modifications of combined exercise and non-exercise therapies would also report effect sizes close to zero as both treatments would be expected to be effective. Additionally, it is important to note that the parameters used in the shrinkage approach were based on the signal-to-noise ratio obtained from research across multiple disciplines. Discipline-specific parameters could potentially improve the accuracy of the shrinkage method (23). However, there were insufficient data to adjust the shrinkage parameters specifically for the exercise therapy and tendinopathy literature. Similarly, there was not enough data to formally assess potential differences in effect size distributions across key moderators such as time, outcome domain, and the location of tendinopathy.

In conclusion, experimental trials involving exercise therapy are likely to be significantly underpowered. This underpowering seems to stem from a poor understanding of what constitutes a meaningful difference, as well as reliance on prior data that likely reflect substantial overestimations. Although there is extensive research investigating MCIC in tendinopathy, these values are unlikely to provide meaningful insights when comparing two interventions that are both generally effective. From a service perspective, it may be more appropriate to incorporate factors such as cost, including clinician time, when defining a target effect size. From a patient perspective, measures of acceptability could be more relevant. Significant work is needed in collaboration with clinicians and patients to determine how best to incorporate these different considerations when powering future studies. In the interim, the shrunken effect size distributions presented in this review should inform future research. For traditional parallel pre-post studies, researchers should provide strong justification if applying target effect sizes that deviate from these adjusted values. For those focusing on incremental improvements consistent with the shrunken effect sizes, alternative approaches may be more feasible given the resource demands. These could include high-frequency data

collection or the development of infrastructure for conducting high-quality RWE research. Further methodological research is necessary to assist researchers in adopting these alternative approaches.

Acknowledgements

The authors would like to extent their gratitude to the following colleagues who contributed to the larger NIHR funded research project from which this quantitative review was developed. These colleagues include Leon Greig, Rachel Moss, David Brandie, Laura Mitchell, Eva Parkinson, Victoria Tzortziou Brown, and posthumously, Colin Maclean.

References

1. Millar NL, Silbernagel KG, Thorborg K, Kirwan PD, Galatz LM, Abrams GD, et al. Tendinopathy. Nat Rev Dis Primer. 2021;7(1):1. DOI: https://doi.org/10.1038/s41572-020-00234-1

2. Page MJ, O'Connor DA, Malek M, Haas R, Beaton D, Huang H et al. Patients' experience of shoulder disorders: a systematic review of qualitative studies for the OMERACT Shoulder Core Domain Set. Rheumatol. 2019. DOI: https://doi.org/10.1093/rheumatology/kez046

3. Challoumas D, Crosbie G, O'Neill S, Pedret C, Millar NL. Effectiveness of Exercise Treatments with or without Adjuncts for Common Lower Limb Tendinopathies: A Living Systematic Review and Network Meta-analysis. Sports Med - Open. 2023;9(1):71. DOI: https://doi.org/10.1186/s40798-023-00616-1

4. Karanasios S, Korakakis V, Whiteley R, Vasilogeorgis I, Woodbridge S, Gioftsos G. Exercise interventions in lateral elbow tendinopathy have better outcomes than passive interventions, but the effects are small: a systematic review and meta-analysis of 2123 subjects in 30 trials. Br J Sports Med. 2021;55(9):477–85. DOI: https://doi.org/10.1136/bjsports-2020-102525

5. Cooper K, Alexander L, Brandie D, Brown VT, Greig L, Harrison I, et al. Introduction. In: Exercise therapy for tendinopathy: a mixed-methods evidence synthesis exploring feasibility, acceptability and effectiveness. National Institute for Health and Care Research. 2023;27(24):1-389. DOI: https://doi.org/10.3310/tfws2748

6. Hopewell S, Keene DJ, Marian IR, Dritsaki M, Heine P, Cureton L, et al. Progressive exercise compared with best practice advice, with or without corticosteroid injection, for the treatment of patients with rotator

cuff disorders (GRASP): a multicentre, pragmatic, 2 × 2 factorial, randomised controlled trial. The Lancet. 2021;398(10298):416–28. DOI: https://doi.org/10.1016/S0140-6736(21)00846-1

7. Wilson R, Abbott JH, Mellor R, Grimaldi A, Bennell K, Vicenzino B. Education plus exercise for persistent gluteal tendinopathy improves quality of life and is cost-effective compared with corticosteroid injection and wait and see: economic evaluation of a randomised trial. J Physiother. 2023;69(1):35–41. DOI: https://doi.org/10.1016/j.jphys.2022.11.007

8. Christiansen DH, Hjort J. Group-based exercise, individually supervised exercise and home-based exercise have similar clinical effects and cost-effectiveness in people with subacromial pain: a randomised trial. J Physiother. 2021;67(2):124–31. DOI: https://doi.org/10.1016/j.jphys.2021.02.015

9. Morrey ME, Dean BJF, Carr AJ, Morrey BF. Tendinopathy: Same Disease Different Results—Why? Oper Tech Orthop. 2013;23(2):39–49. DOI: https://doi.org/10.1053/j.oto.2013.06.004

10. Prudêncio DA, Maffulli N, Migliorini F, Serafim TT, Nunes LF, Sanada LS, et al. Eccentric exercise is more effective than other exercises in the treatment of mid-portion Achilles tendinopathy: systematic review and meta-analysis. BMC Sports Sci Med Rehabil. 2023;15:9. DOI: https://doi.org/10.1186/s13102-023-00618-2

11. Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal. Trials. 2018;19(1):544. DOI: https://doi.org/10.1186/s13063-018-2886-y

12. Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinska B, Freedman LS. Sample size calculation for clinical trials: the impact of clinician beliefs. Br J Cancer. 2000;82(1):213–9. DOI: https://doi.org/10.1054%2Fbjoc.1999.0902

13. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? PharmacoEconomics. 2000;18(5):419–23. DOI: https://doi.org/10.2165/00019053-200018050-00001

14. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res. 2011;11(2):171–84. DOI: https://doi.org/10.1586/erp.11.9

15. Boudreau N, Gaudreault N, Roy JS, Bédard S, Balg F. The Addition of Glenohumeral Adductor Coactivation to a Rotator Cuff Exercise Program for Rotator Cuff Tendinopathy: A Single-Blind

Randomized Controlled Trial. J Orthop Sports Phys Ther. 2019;49(3):126–35. DOI: https://doi.org/10.2519/jospt.2019.8240

16. Granviken F, Vasselijen O. Home exercises and supervised exercises are similarly effective for people with subacromial impingement: a randomised trial. J Physiother. 2015;61(3). DOI: https://doi.org/10.1016/j.jphys.2015.05.014

17. Hotta GH, Gomes de Assis Couto A, Cools AM, McQuade KJ, Siriani de Oliveira A. Effects of adding scapular stabilization exercises to a periscapular strengthening exercise program in patients with subacromial pain syndrome: A randomized controlled trial. Musculoskelet Sci Pract. 2020;49:102171. DOI: https://doi.org/10.1016/j.msksp.2020.102171

18. Paavola M, Malmivaara A, Taimela S, Kanto K, Inkinen J, Kalske J, et al. Subacromial decompression versus diagnostic arthroscopy for shoulder impingement: randomised, placebo surgery controlled clinical trial. BMJ. 2018;362:k2860. DOI: https://doi.org/10.1136/bjsports-2020-102216

19. Kamper SJ. Interpreting Outcomes 3—Clinical Meaningfulness: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019;49(9):677–8. DOI: https://doi.org/10.2519/jospt.2019.0705

20. Hollingworth W, McKell-Redwood D, Hampson L, Metcalfe C. Cost-utility analysis conducted alongside randomized controlled trials: are economic end points considered in sample size calculations and does it matter? Clin Trials Lond Engl. 2013;10(1):43–53. DOI: https://doi.org/10.1177/1740774512465358

21. Ioannidis JPA. Why Most Discovered True Associations Are Inflated. Epidemiology. 2008;19(5):640–8. DOI: https://doi.org/10.1097/ede.0b013e31818131e7

22. Sidebotham D, Barlow CJ. The winner's curse: why large effect sizes in discovery trials always get smaller and often disappear completely. Anaesthesia. 2024;79(1):86–90. DOI: https://doi.org/10.1111/anae.16161

23. van Zwet E, Schwab S, Senn S. The statistical properties of RCTs and a proposal for shrinkage. Stat Med. 2021;40(27):6107–17. DOI: https://doi.org/10.1002/sim.9173

24. Swinton PA, Shim JSC, Pavlova AV, Moss R, Maclean C, Brandie D, et al. What are small, medium and large effect sizes for exercise treatments of tendinopathy? A systematic review and meta-analysis. BMJ Open Sport Exerc Med. 2023;9(1):e001389. DOI: https://doi.org/10.1136/bmjsem-2022-001389

25. Nations U. Human Development Index [Internet]. Human Development Reports. United Nations; Available from: https://hdr.undp.org/data-center/human-development-index

26. Alfredson H, Pietilä T, Jonsson P, Lorentzon R. Heavy-load eccentric calf muscle training for the treatment of chronic Achilles tendinosis. Am J Sports Med. 1998;26(3):360–6. DOI: https://doi.org/10.1177/03635465980260030301

27. Zampieri FG, Casey JD, Shankar-Hari M, Harrell 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. Am J Respir Crit Care Med. 2021;203(5):543–52. DOI: https://doi.org/10.1164/rccm.202006-2381cp

28. Morris SB. Estimating effect sizes from pretest-posttest-control group design. Organ Res Methods. 2008;11(2):364–86. DOI: https://doi.org/10.1177/1094428106291

29. Vlist AC van der, Winters M, Weir A, Ardern CL, Welton NJ, Caldwell DM, et al. Which treatment is most effective for patients with Achilles tendinopathy? A living systematic review with network meta-analysis of 29 randomised controlled trials. Br J Sports Med. 2021;55(5):249–56. DOI: https://doi.org/10.1136/bjsports-2019-101872

30. Jaeschke R, Singer J, Guyatt GH.. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4). DOI: https://doi.org/10.1016/0197-2456(89)90005-6

31. Kanto K, Lähdeoja T, Paavola M, Aronen P, Järvinen TLN, Jokihaara J, et al. Minimal important difference and patient acceptable symptom state for pain, Constant-Murley score and Simple Shoulder Test in patients with subacromial pain syndrome. BMC Med Res Methodol. 2021;21(1):45. DOI: https://doi.org/10.1186/s12874-021-01241-w

32. Redelmeier DA, Lorig K. Assessing the Clinical Importance of Symptomatic Improvements: An Illustration in Rheumatology. Arch Intern Med. 1993;153(11):1337–42. DOI: https://doi.org/10.1001/archinte.1993.00410110045008

33. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009;41:1149–60. DOI: https://doi.org/10.3758/brm.41.4.1149

34. Arias-Buría JL, Truyols-Domínguez S, Valero-Alcaide R, Salom-Moreno J, Atín-Arratibel MA, Fernández-de-las-Peñas C. Ultrasound-Guided Percutaneous Electrolysis and Eccentric Exercises for Subacromial Pain Syndrome: A Randomized Clinical Trial. Evid-Based Complement Altern Med ECAM. 2015;2015:315219. DOI: https://doi.org/10.1155/2015/315219

35. Rabusin CL, Menz HB, McClelland JA, Evans AM, Malliaras P, Docking SI, et al. Efficacy of heel lifts versus calf muscle eccentric exercise for mid-portion Achilles tendinopathy (HEALTHY): a randomised trial. Br J Sports Med. 2021 May 1;55(9):486–92. DOI: https://doi.org/10.1186/s13047-019-0325-2

36. Stefan AM, Schönbrodt FD. Big little lies: a compendium and simulation of p-hacking strategies. R Soc Open Sci. 2023;10(2):220346. DOI: https://doi.org/10.1098/rsos.220346

37. Egbewale BE, Lewis M, Sim J. Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. BMC Med Res Methodol. 2014;14(1):49. DOI: https://doi.org/10.1186/1471-2288-14-49

38. Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. Synthese. 2021;199:10969–1000. DOI: https://doi.org/10.1007/s11229-021-03276-4

39. Abt G, Boreham C, Davison G, Jackson R, Nevill A, Wallace E, et al. Power, precision, and sample size estimation in sport and exercise science research. J Sports Sci. 2020;38(17):1933–5. DOI: https://doi.org/10.1080/02640414.2020.1776002

40. Götz FM, Gosling SD, Rentfrow PJ. Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. Perspect Psychol Sci. 2022;17(1):205–15. DOI: https://doi.org/10.1177/1745691620984483

41. Mesquida C, Murphy J, Lakens D, Warne J. Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: potential barriers to replicability. J Sports Sci. 2023;41(16):1507–17. DOI: https://doi.org/10.1080/02640414.2023.2269357

42. Dodge HH, Zhu J, Mattek NC, Austin D, Kornfeld J, Kaye JA. Use of High-Frequency In-Home Monitoring Data May Reduce Sample Sizes Needed in Clinical Trials. PloS One. 2015;10(9):e0138095. DOI: https://doi.org/10.1371/journal.pone.0138095

43. Swinton P. Adequate statistical power in strength and conditioning may be achieved through longer interventions and high frequency outcome measurement. SportRxiv; 2024.DOI: https://doi.org/10.51224/SRXIV.364

44.Turner RM, Bird SM, Higgins JPT. The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. PLOS ONE. 2013;8(3):e59202. DOI: https://doi.org/10.1371/journal.pone.0059202

45.Simon GE, Platt R, Watanabe JH, Bindman AB, John London A, Horberg M, et al. When Can We Rely on Real-World Evidence to Evaluate New Medical Treatments? Clin Pharmacol Ther. 2022;111(1):30–4. DOI: https://doi.org/10.1002/cpt.2253

46.Mondini Trissino da Lodi C, Landini MP, Asunis E, Filardo G. Women Have Tendons… and Tendinopathy: Gender Bias is a "Gender Void" in Sports Medicine with a Lack of Women Data on Patellar Tendinopathy—A Systematic Review. Sports Med - Open. 2022;8:74. DOI: https://doi.org/10.1186/s40798-022-00455-6

47.Schuppisser MV, Mondini Trissino da Lodi C, Albanese J, Candrian C, Filardo G. Achilles tendinopathy research has a gender data gap: A systematic review and meta-analysis. Knee Surg Sports Traumatol Arthrosc. 2024. DOI: https://doi.org/10.1002/ksa.12046

48.Talaski GM, Baumann AN, Salmen N, Curtis DP, Walley KC, Anastasio AT, et al. Socioeconomic Status and Race Are Rarely Reported in Randomized Controlled Trials for Achilles Tendon Pathology in the Top 10 Orthopaedic Journals: A Systematic Review. Foot Ankle Orthop. 2024;9(1). DOI: https://doi.org/10.1177%2F24730114231225454

49.Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol. 2022;22(1):287. DOI: https://doi.org/10.1186/s12874-022-01768-6

50.Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest. 130(2):565–74. DOI: https://doi.org/10.1172/jci129197