

Should this systematic review and meta-analysis change my practice? Part 1: exploring treatment effect and trustworthiness

Mervyn J Travers,^{1,2} Myles Calder Murphy,^{1,3}
James Robert Debenham,¹ Paola Chivers,⁴ Max K Bulsara,⁴
Matthew K Bagg,^{5,6} Thorvaldur Skulli Palsson,⁷ William Gibson¹

INTRODUCTION

As a service to its readership, BJSM publishes educational editorials in order to provide methodological guidance and direction to clinicians and researchers.¹ For example, a recent BJSM invited commentary outlined important methodological aspects of *randomised controlled trials* (RCTs) in order to help readers better understand and interpret findings.² Similarly, a recent BJSM editorial written for clinicians has outlined five essential methodological considerations or 'hacks' to consider when reading a systematic review (SR), while another editorial has outlined common pitfalls for authors conducting a SR.^{3,4} In light of recent suggestions of untrustworthiness of systematic review and meta-analyses (SRMA) findings,⁵⁻⁷ researchers are encouraged to maximise the transparency of their work. Similarly, clinicians and policymakers are encouraged to scrutinise SRMA in order to decide if the findings should influence clinical decisions or policies. The purpose of our two-part educational editorial series is to highlight important methodological aspects of SRMA of the effects of interventions and thus help clinicians

make judicious choices when considering whether to accept the findings. By using a worked example, we hope this resource serves as a guide for clinicians when critically appraising SRMA findings and deciding whether (or not) to accept the reported findings or the interpretation of the authors.

WHAT IS A SYSTEMATIC REVIEW AND META-ANALYSIS AND HOW SHOULD IT INFORM MY PRACTICE?

A SR attempts to collate all empirical evidence that fits prespecified eligibility criteria in order to answer a specific research question.⁸ An interventional SR specifically assesses the benefits and harms of an intervention and should provide an unbiased, robust, transparent and reproducible overview of the effect of an intervention and the quality of the evidence from identified clinical studies.⁹ Should the SR contain sufficient and appropriate quantitative data from the included studies, a meta-analysis (MA) can be conducted to provide a pooled estimate of effect, and to examine variation among individual study effect estimates.⁹ As such, SRMA represent the highest standard of evidence of the effects of interventions and have the capacity to usefully inform clinical decisions, reduce research waste and direct policy making.^{3,5,7,10} However, the flip-side and increasingly worrying prospect is that poorly conducted or redundant SRMAs may mislead clinical decisions, policy makers and the public.^{6,11-15} This can reasonably lead clinicians to question whether the findings of a SRMA can be trusted.¹⁶ Thus, it is vital they have the confidence and ability to assess the quality of SRMA and therefore interpret findings appropriately. We encourage clinicians to appraise the findings before deciding 'Should this SRMA change my practice?'. Central to this decision is an understanding of transparency and reproducibility in SRMA.

The working example below is intended to illustrate these issues.

HOW IS TREATMENT EFFECT DETERMINED?

An interventional SRMA pools the effects from multiple studies to obtain a pooled estimate of the effect of an intervention. Studies that use the same measure for a continuous outcome (e.g. pain intensity during running using a visual analogue scale [VAS]) may be pooled using an effect measure called the 'mean difference' (MD). The pooled effect size from the MA reflects the weighted average across studies of the difference (in pain during running) between groups, observed in each study, at the time point of interest (e.g. 12 weeks postintervention).¹⁷ Studies that use different measures for continuous outcomes may be pooled using an effect measure called the 'standardised mean difference' (SMD). Here, the difference in outcome between groups in each study is standardised to a uniform scale (by dividing this value by the standard deviation (SD) observed in the outcome across both groups). Thus, different measures such as a VAS (usually scored 0–100) and the pain component from the SF-36 questionnaire which is scored from 1 to 6 (1=None; 6=Very severe) can be pooled in a MA. The SMD is often more difficult to interpret than a MD because the scale reflects units of SD in outcome, rather than units of difference in outcome.¹⁷

In our worked example, Miller *et al* (2017) reported a pooled SMD of 0.47 (95% CI 0.22 to 0.72, $p < 0.001$) which suggests a moderate treatment effect in favour of platelet-rich plasma injections (PRP) for tendinopathy. This is illustrated in the forest plot below (figure 1). In this example, negative values favour control injections, while positive values favour PRP. The SMD for each study is indicated by the boxes in the forest plot and is calculated from the mean outcome score, SD and sample size of each arm from each study at the time point of interest. Of note, the included studies used different outcome measures which measured different outcomes. For example, pain intensity (scored on a VAS) and disability (the Victorian Institute of Sport Assessment—Achilles (VISA-A)). This comparison raises a key question; was the use of SMD appropriate?

The varied outcome measures employed by trials included in the example SRMA are listed in figure 2 and raise concerns about what treatment effect is actually being reported. Pain (e.g. VAS) is a

¹School of Physiotherapy, University of Notre Dame Australia, Fremantle, Western Australia, Australia

²School of Physiotherapy and Exercise Science, Curtin University, Perth, Western Australia, Australia

³Sports Science Sports Medicine Department, Western Australian Cricket Association, Perth, Western Australia, Australia

⁴Institute for Health Research, University of Notre Dame Australia, Fremantle, Western Australia, Australia

⁵Pain Research, Education and Management Program, Neuroscience Research Australia (NeuRA), Sydney, New South Wales, Australia

⁶UNSW Prince of Wales Clinical School, Sydney, New South Wales, Australia

⁷Laboratory for Musculoskeletal Pain and Motor Control, Center for Sensory-Motor Interaction (SMI), Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

Correspondence to Dr Mervyn J Travers, School of Physiotherapy, University of Notre Dame Australia, Fremantle, WA 6160, Australia; mervyn.travers@nd.edu.au

Discussion

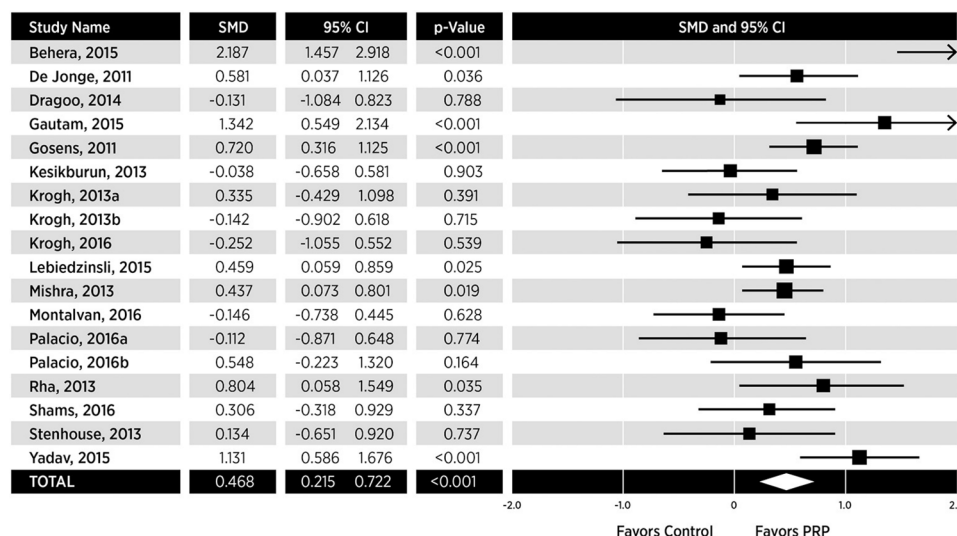


Figure 1 Forest plot of PRP versus control on tendinopathy pain. Reproduced from Miller LE, *et al.* BMJ Open Sport Exerc Med 2017;3:e000237. doi:10.1136/bmjsem-2017-000237 (CC BY-NC 4.0). See: <http://creativecommons.org/licenses/by-nc/4.0/>. PRP, platelet-rich plasma; SMD, standardised mean difference.

different outcome than disability (e.g. VISA-A) as they reflect different domains of the condition.¹⁸ The clinician could be justified in asking ‘how could a result from a pain score at the elbow be compared with a disability scale for Achilles tendinopathy?’. The authors correctly did not use MD, given that the outcome measures differed across trials. However, combining measures of different domains of tendinopathy with the SMD is nevertheless not appropriate here. The SMD is used to standardise effects from different scales that measure the same outcome domain. It may not be used to standardise effects from scales used to measure different outcome domains (pain and disability, in this example). Consequentially, the pooled effect of 0.47 is difficult to interpret in terms of what outcome domain the effect pertains to. It may be argued that the example SRMA would have better informed clinicians had they pooled measures for pain and disability into separate analyses. In this case, use of SMD would have been appropriate had different scales measured the same outcome domain.

INTERPRETING THE TREATMENT EFFECT AND CONSIDERING ITS TRUSTWORTHINESS

Clinicians do not work with SMDs in practice. Translating a pooled SMD into clinically usable terms is aided when the original trial data (between group comparisons) are presented by the authors. Having these data readily available would allow them to gain some understanding of the between group differences that create

the reported SMD. For example, a SMD for the VISA-A of 0.03 (95%CI -0.51 to 0.56) is very difficult to interpret in clinical terms. This means there is no difference between groups at the time point of interest but may be more easily interpreted by knowing that it is derived from a control group mean (SD) 22.4 (17.2) and a PRP group mean (SD) 21.8 (25.9) for the VISA-A.

These trial data were not reported in the example SRMA. Transparent methodological reporting and data sharing facilitate reproducibility. In order to explore transparency and appraise the trustworthiness in our worked example, we attempted to source the outcome data from each individual RCT included in the MA by Miller *et al* (2017).

Using these data, we sought to reproduce the forest plot. Three authors (MJT/MCM/MKB), independently extracted data from the 16 included RCTs. However, eight of the included studies did not explicitly report the required data (mean and SD) for the outcome measure used in the MA. The Cochrane Handbook for Systematic Reviews of Interventions¹⁹ outlines recommended options for obtaining the required data.²⁰

Accordingly, we contacted the corresponding authors of the RCTs that did not report the data required. Only one author replied with the requested data. For three RCTs we calculated the SD from the existing estimate of variance (standard error (SE) or 95% CI) and included these studies in the forest plot. One additional RCT only reported the mean *change* in pain from baseline and

SE of that change so the SE reported at baseline was used to calculate SD. Finally, one RCT did not give an estimate of variance at all, an estimate of the SD was derived from the group means and reported p value.

We were unable to reproduce the SMD for two studies. For one, it was unclear what outcome variable was actually used to derive the SMD as there were multiple possible outcomes that may have been appropriate to use; however, this was not defined. The second study did not report raw scores for the pain outcome used.

Given our difficulties in obtaining the data required to reproduce the forest plot (outlined in detail in online supplementary file 1), we wrote to the corresponding author of the example SRMA to request the data and clarity on how these SMDs were generated. Unfortunately we received no response. Nonetheless, we used the data we extracted or calculated as described in an attempt to reproduce the example SRMA's forest plot. We present this (and reasons for exclusions) in figure 3. The difficulty encountered and our ultimately unsuccessful attempt to exactly reproduce the reported forest plot illustrates that non-transparent reporting seriously inhibits reproducibility and therefore undermines the trustworthiness of SRMA findings. As expected, our results do not match exactly with the example SRMA but we obtained a pooled effect for PRP compared with the comparator of SMD of 0.41 (95% CI 0.11 to 0.70, p=0.007). However, because our results are derived from an incomplete dataset with varied

Table 1 Patient and study design characteristics

Study	Treatment period	Sample size		Female gender (%)		Age (year)		Tendinopathy	Minimum symptom duration (month)	Pain outcome	Follow-up (month)
		PRP	Control	PRP	Control	PRP	Control				
Behera <i>et al</i> ¹³	2011–2011	15	10	80	50	38	37	Lateral epicondylar	3	VAS	12
de Jonge <i>et al</i> ¹⁴	2008–2009	27	27	52	52	49	50	Achilles	2	VISA-A	12
Dragoo <i>et al</i> ¹⁵	2009–2012	A10	13	11	0	28	40	Patellar	1.5*	VAS	6
Gautam <i>et al</i> ¹⁶	2011–2012	15	15	—	—	—	—	Lateral epicondylar	6	VAS	6
Gosens <i>et al</i> ¹⁷	2006–2008	51	49	52	56	47	47	Lateral epicondylar	6	VAS	24†
Kesikburun <i>et al</i> ¹⁸	2011–2011	20	20	65	70	46	51	Rotator cuff	3	VAS	12
Krogh <i>et al</i> ¹⁹	2009–2010	20	20	55	55	48	45	Lateral epicondylar	3	PRTEE	3
Krogh <i>et al</i> ⁴⁰	2009–2010	20	20	55	45	48	44	Lateral epicondylar	3	PRTEE	3
Krogh <i>et al</i> ²⁰	2009–2011	12	12	42	50	47	52	Achilles	6	VAS	3
Lebiedziński <i>et al</i> ²¹	2009–2011	64	56	47	74	47	54	Lateral epicondylar	1.5	DASH	12
Mishra <i>et al</i> ²²	2006–2011	112	113	—	—	48	47	Lateral epicondylar	3	VAS	6
Montalvan <i>et al</i> ²³	2010–2014	25	25	32	32	47	46	Lateral epicondylar	—§	VAS	12
Palacio <i>et al</i> ²⁴	2012–2014	20	20	—	—	47	48	Lateral epicondylar	—	PRTEE	6¶
Palacio <i>et al</i> ²⁴	2012–2014	20	20	—	—	47	46	Lateral epicondylar	—	PRTEE	6¶
Rha <i>et al</i> ²⁵	2010–2011	20	19	55	58	52	54	Rotator cuff	6	SPADI pain	6
Shams <i>et al</i> ²⁶	2013–2015	20	20	50	45	52	50	Rotator cuff	3	ASES	6
Stenhouse <i>et al</i> ²⁷	2010–2012	15	13	47	62	53	48	Lateral epicondylar	6	VAS	6
Yadav <i>et al</i> ²⁸	2012–2014	30	30	67	77	37	37	Lateral epicondylar	1	VAS	3

‘—’ indicates missing data.

*Patients failed to respond to ≥6 weeks physical therapy; total symptom duration not reported.

†Data extracted through 12 months only for meta-analysis per systematic review methods (ie, data extraction at 3, 6 and 12 months); however, total symptom duration not reported.

‡Study includes same PRP group and different control groups. PRP group sample size adjusted in meta-analysis based on number of groups.

§Maximum symptom duration was 3 months.

¶Data extracted through 3 months only for meta-analysis due to implausible reported 6-month outcomes.

ASES, American Shoulder and Elbow Surgeons; DASH, Disabilities of the Arm, Shoulder and Hand score; PRP, platelet-rich plasma; PRTEE, Patient-Rated Tennis Elbow Evaluation; SPADI, Shoulder Pain and Disability Index; VAS, visual analogue scale; VISA-A, Victorian Institute of Sport Assessment-Achilles questionnaire.

Figure 2 Pain outcomes included in example systematic review and meta-analysis. Reproduced from Miller LE, *et al*. BMJ Open Sport Exerc Med 2017;3:e000237. doi:10.1136/bmjsem-2017-000237 (CC BY-NC 4.0). See: <http://creativecommons.org/licenses/by-nc/4.0/>.

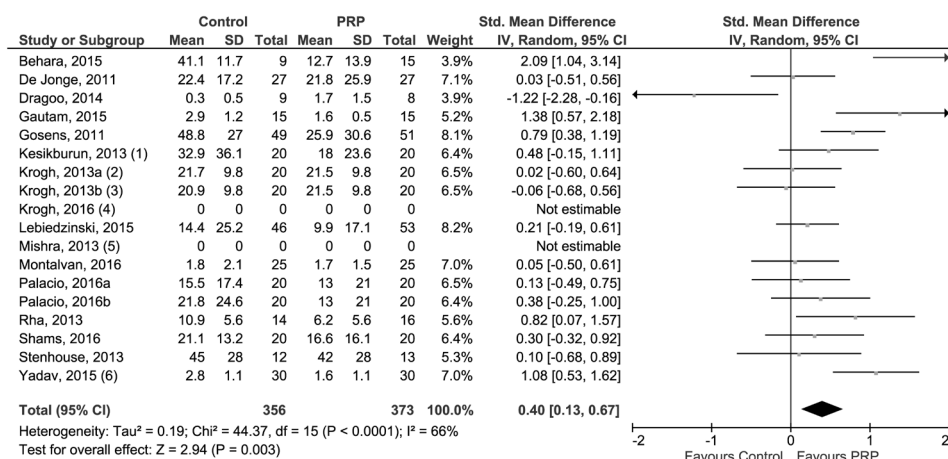
outcomes it cannot be inferred that PRP is effective for tendinopathy.

It is possible that the authors of the example SRMA faced similar difficulties and may have needed to contact RCT authors or impute SD from similar studies. However, this was not reported. Reporting to this level of detail would afford the

reader greater clarity on how data were obtained. This is especially pertinent as we could not directly reproduce a result for six of the SMDs and could not reproduce two of them at all. Furthermore, SRMA authors are encouraged to pre-register their protocol. Transparency, methodological rigour and reproducibility are

enhanced by pre-registering the SRMA protocol in advance, though the example SRMA did not report on their protocol being accessible in any registry (e.g., PROSPERO).

Discussion



Footnotes

- (1) Received additional data from author
- (2) Baseline variance carried forward
- (3) Baseline variance carried forward
- (4) Unclear on pain outcome measure reported
- (5) No raw pain data provided
- (6) Calculated variance from p-value

Figure 3 Our reproduction of PRP versus control on tendinopathy pain with 'Not Estimable' comparisons removed. Review Manager (Computer program) V.5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014. Not Estimable=insufficient data to generate standard mean difference; Total=sample size. PRP, platelet-rich plasma; SMD, standardised mean difference.

HOW SHOULD THIS BE INTERPRETED BY CLINICIANS?

An interventional SRMA should provide an unbiased, robust, transparent and reproducible overview of the effect of an intervention and the quality of the evidence supporting this finding. These reviews

may usefully inform treatment choices for clinicians, but only if the clinician can assess the robustness and trustworthiness of the findings. Our worked example illustrates that there is merit in reading reviews with a sceptical and forensic eye as not all peer-reviewed findings are trustworthy. This example highlighted the importance of carefully considering the pooled estimate of the treatment effect with respect to whether MD or SMD is reported. If the latter, it is necessary to assess what outcome measures have been combined in a MA. Clinicians are well placed to critically judge whether combining varied outcome measures makes sense from a clinical perspective. If the combination of outcome measures does not appear appropriate, then it is reasonable to question whether the pooled estimate is meaningful. Second, when a pooled estimate of effect is provided (MD or SMD), we encourage clinicians to question whether individual trial data reporting appears sufficient to allow for an informed assessment of how the pooled estimate was generated. A lack of transparency through missing data or data simply not being reported renders the results non-reproducible and therefore undermines their trustworthiness.

Acknowledgements Benedict Martin Wand provided advice, direction and oversight to the planning of this project.

Contributors All authors contributed to the inception of the research idea. MJT, MCM and MKB performed the independent data extraction. WG provided oversight to the data extraction and associated disputes. MJT, WG and MCM performed the statistical

analysis with oversight from PC and MB. All authors contributed to drafting / review of manuscript. MKB = Matthew K Bagg; MB = Max K Bulsara

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

© Author(s) (or their employer(s)) 2019. No commercial re-use. See rights and permissions. Published by BMJ.

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/bjsports-2018-099958>).



To cite Travers MJ, Murphy MC, Debenham JR, et al. *Br J Sports Med* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bjsports-2018-099958

Accepted 18 March 2019

Br J Sports Med 2019;0:1–5.
doi:10.1136/bjsports-2018-099958

REFERENCES

- 1 Verhagen E, Stovitz SD, Mansournia MA, et al. *BJSM* educational editorials: methods matter. *Br J Sports Med* 2018;52:1159–60.
- 2 Mansournia MA, Altman DG. Invited commentary: methodological issues in the design and analysis of randomised trials. *Br J Sports Med* 2018;52:553–5.
- 3 Ardern CL. Systematic review hacks for the sports and exercise clinician: five essential methodological elements. *Br J Sports Med* 2016;50:447–9.
- 4 Bero L. Getting the systematic review basics right helps clinical practice: 4 common pitfalls for systematic review authors to avoid. *Br J Sports Med* 2019;53:6–8.
- 5 Annane D, Jaeschke R, Guyatt G. Are systematic reviews and meta-analyses still useful research? Yes. *Intensive Care Med* 2018;44:512–4.

Key messages

- Systematic Review and Meta-analysis (SRMA) of the effects of interventions may usefully inform treatment choices for clinicians, but only if the clinician can assess the robustness and trustworthiness of the findings. Thus, we encourage clinicians to:
 1. Consider the pooled estimate of the treatment effect with respect to whether mean difference or standardised mean difference (SMD) is reported.
 - If SMD is reported, it is important to assess what outcome measures have been combined in a meta-analysis.
 2. Examine whether individual trial data reporting appears sufficient to allow for an informed assessment of how the pooled estimate was generated.
 3. Recognise a lack of transparency and reproducibility through missing data or data simply not being reported undermines the trustworthiness of the results of a SRMA.

- 6 Chevret S, Ferguson ND, Bellomo R. Are systematic reviews and meta-analyses still useful research? No. *Intensive Care Med* 2018;44:515–7.
- 7 Møller MH, Ioannidis JPA, Darmon M. Are systematic reviews and meta-analyses still useful research? We are not sure. *Intensive Care Med* 2018;44:518–20.
- 8 Green S, Higgins JPT, Alderson P, et al. Chapter 1: Introduction. In: Higgins JPT, Green S, eds. *Cochrane Handbook for systematic reviews of interventions version 5.1.0 (updated March 2011)*. The Cochrane Collaboration, 2011.
- 9 Gurevitch J, Koricheva J, Nakagawa S, et al. Meta-analysis and the science of research synthesis. *Nature* 2018;555:175–82.
- 10 Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *J Clin Nurs* 2003;12:77–84.
- 11 Altman DG. The scandal of poor medical research. *BMJ* 1994;308:283–4.
- 12 Ebrahim S, Bance S, Athale A, et al. Meta-analyses with industry involvement are massively published and report no caveats for antidepressants. *J Clin Epidemiol* 2016;70:155–63.
- 13 Ioannidis J. Next-generation systematic reviews: prospective meta-analysis, individual-level data, networks and umbrella reviews. *Br J Sports Med* 2017;51:1456–8.
- 14 Ioannidis JPA. The mass production of redundant, misleading, and Conflicted systematic reviews and meta-analyses. *Milbank Q* 2016;94:485–514.
- 15 Jørgensen AW, Hilden J, Gøtzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. *BMJ* 2006;333.
- 16 Weir A, Rabia S, Arden C. Trusting systematic reviews and meta-analyses: All that glitters is not gold! *Br J Sports Med* 2016;50:1100–1.
- 17 Higgins JPT, Deeks JJ, Deeks JJ, et al. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, eds. *Cochrane Handbook for systematic reviews of interventions version 5.1.0 (updated March 2011)*. The Cochrane Collaboration, 2011.
- 18 Murphy M, Rio E, Debenham J, et al. Evaluating the progress of MID-PORTION Achilles tendinopathy during rehabilitation: a review of outcome measures for self-reported pain and function. *Int J Sports Phys Ther* 2018;13:283–92.
- 19 Higgins JPT, Deeks JJ. Chapter 7: Selecting studies and collecting data. In: Higgins JPT, Green S, eds. *Cochrane Handbook for systematic reviews of interventions version 5.1.0 (updated March 2011)*. The Cochrane Collaboration, 2011.
- 20 Wan X, Wang W, Liu J, et al. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol* 2014;14:135–35.