Identifying the "incredible"! Part 2: Spot the difference – a rigorous risk of bias assessment can alter the main findings of a systematic review

Education Review

Fionn Büttner[1], Marinus Winters[2], Eamonn Delahunt[1, 3], Roy Elbers[4], Carolina Bryne Lura[2], Karim M Khan[5], Adam Weir[6, 7, 8], Clare L. Ardern[9, 10, 11]

1 School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland
2 Research Unit for General Practice in Aalborg, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark
3 Institute for Sport and Health, University College Dublin, Dublin, Ireland
4 Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
5 Centre for Hip Health and Mobility, University of British Columbia, Vancouver, British Columbia, Canada
6 Aspetar Orthopaedic and Sports Medicine Hospital, Doha, Qatar
7 Department of Orthopaedics, Erasmus MC University Medical Center for Groin Injuries, Rotterdam, The Netherlands
8 Sport medicine and exercise clinic Haarlem (SBK), Haarlem, The Netherlands
9 Division of Physiotherapy, Linköping University, Linköping, Sweden
10 School of Allied Health, La Trobe University, Melbourne, Australia
11 Division of Physiotherapy, Karolinska Institute, Stockholm, Sweden

Correspondence to:
Fionn Cléirigh Büttner
School of Public Health, Physiotherapy and Sports Science
University College Dublin
Belfield
Dublin 4
Ireland
E: fionn.cleirigh-buttner@ucdconnect.ie

Word count: 3405

Keywords: review [MeSH], meta-analysis [MeSH], bias [MeSH]

INTRODUCTION

Systematic reviews are a valuable tool to inform healthcare decision-making.(1,2) While a single randomised controlled trial (RCT) is insufficient to definitively guide healthcare decisions, a systematic review synthesising multiple RCTs can overcome this limitation. The results of rigorous systematic reviews possess wide-ranging applicability to numerous stakeholders within the evidence-based medicine 'ecosystem'. Clinicians consult systematic reviews to inform their clinical decisions.(3) Researchers rely on systematic reviews to identify knowledge gaps in existing literature.(4) Health policymakers use systematic review evidence to inform practice guidelines and legislation.(5,6) Journal editors often prioritize systematic reviews for their impact on readership attention and journal metrics.(7) Finally, patients are empowered by systematic reviews that assess the beneficial and harmful patient-important outcomes of available management strategies.(8) Systematic review authors have an important responsibility to ensure their findings provide the most credible results possible.

The biomedical literature expands by twenty-two systematic reviews daily,(9) with no evidence that production is waning. More systematic reviews are desirable if they identify and inform important research questions that improve patient care.(10) However, production of this magnitude is problematic when systematic reviews offer "extensive redundancy, little value, misleading claims, and/or vested interests."(11) As we outlined in Part 1, bias is a systematic deviation from the truth in the results of a research study, due to limitations in study design, conduct, or analysis.(2) Deviations may either over- or under-estimate a study's true findings depending of the type and magnitude of bias. As the results of a systematic review are only as valid as the studies it includes, pooling biased studies can compromise the credibility of systematic review findings when no assessment, or a poor assessment, of risk of bias is performed.(3,12)

Inadequate study design, conduct, or analysis, devalue the credibility of biomedical research and the competence of clinicians to care for patients.(13) Comprehensively assessing risk of bias instead of

study quality – using a domain-based risk of bias assessment instead of a quality scale or checklist – is *the* best method to avoid overlooking biased evidence that propagates misleading systematic review findings. This two-part education primer focuses on critical assessments as a source of misleading systematic review conclusions. In part one, we introduced risk of bias as the perceived risk that the results of a research study deviate from the truth. In part two, we:

1. Evaluate the prevalence of and methods used to critically assess study findings included in systematic reviews published in *BJSM*.
2. Perform a risk of bias assessment on a sample of RCTs in a systematic review, to compare risk of bias assessment findings with study quality assessment findings.
3. Illustrate the impact that different assessment tools have on risk of bias assessment findings, and ultimately, systematic review findings.
4. Provide recommendations to systematic review authors who are planning a risk of bias assessment.

EMPIRICAL EVIDENCE FROM SPORT AND EXERCISE MEDICINE

Sport and Exercise Medicine (SEM) research is vulnerable to bias across many empirical study designs.(14,15) The characteristics of risk of bias assessments in SEM systematic reviews are unknown. Evaluating how risk of bias assessments are conducted is necessary to determine whether risk of bias is adequately assessed across SEM research. We performed a cross-sectional study to determine the methods used to critically appraise the credibility of original study findings in systematic reviews published in *BJSM*.

Methods

We searched the *BJSM* journal archive (http://bjsm.bmj.com/content/by/year) on May 11th, 2017 to identify systematic reviews published since January 1st, 2016. Systematic reviews performing a descriptive synthesis or meta-analysis, and published in *BJSM* between 1st January 2016 and 10th May 2017, were eligible for inclusion.

Two authors (FCB & ED) independently screened titles and abstracts to identify systematic reviews. Review article types that did not systematically identify and select eligible studies, and synthesise relevant study content (e.g., narrative/critical reviews, PEDro syntheses, consensus statements and practice guidelines) were excluded. A third author (CLA or MW) arbitrated disagreements. Descriptive characteristics of each systematic review were independently abstracted by two authors (FCB & ED) using a pre-defined data extraction template (Table 1).

Table 1. Critical assessment tool and method variables extracted from *BJSM* systematic reviews

#	Outcome variable extracted	Levels of each outcome variable
1	Critical assessment tool implemented	Yes, no, unclear.
2	Critical assessment tool name	Cochrane Risk of Bias Tool, PEDro, etc.
3	Critical assessment tool type	**Standard tool:** Systematic review authors used a published tool. **Adapted tool:** Systematic review authors modified a published tool for their use. **Custom tool:** Systematic review authors created a new tool to critically appraise included studies.
4	Critical assessment tool format	**Scale:** Multiple tool items were assigned a quantitative value and combined to generate a single numeric summary score. **Checklist:** Tool items possessed a descriptive, discrete outcome (e.g., "yes", "no", "maybe") without an assigned numeric value. **Domain-based tool:** Separate assessments were made for different risk of bias domains, with subjective judgements of "low risk", "unclear", and "high risk" determined by established criteria.
5	Ranking system prevalence and type, if applicable	**Summary score:** Tool items were summed to generate a single numeric value. **Threshold summary score**: Tool items were summed to generate a single numeric value with arbitrary cut-off values used to classify studies as being at 'low', 'moderate'/ 'unclear', or 'high' quality/risk of bias.
6	Risk of bias assessment method	**Domain-based assessment** **Separate assessments for different subjective and objective outcome types**
7	Method used to incorporate critical assessment findings into systematic review findings	Sensitivity analysis Meta-regression Exclude studies at high or unclear risk of bias (or moderate or low quality) from the synthesis Narrative discussion (with minimal evidence for incorporation deemed acceptable) No attempt to incorporate risk of bias assessment findings into systematic review findings

One author (FCB) performed pilot data extraction on a subsample of systematic reviews prior to full data extraction. Two authors (FCB & ED) independently extracted data from each included

systematic review. A third author (CLA or MW) arbitrated disagreements. Data are presented as absolute frequencies (n) and as a proportion (%) of the sample of systematic reviews.

Results

We included 66 systematic reviews.

- Sixty-five (99%) systematic reviews reported a critical assessment of included studies. Table 2 lists the characteristics of critical assessment tools used. Critical assessment tools used in systematic reviews are included in Supplementary Table 2.
- Forty-two (65%) systematic reviews used a standard tool, 14 (21%) used an adapted tool, and 9 (14%) used a custom tool.
- Eighteen (28%) systematic reviews used a checklist to critically assess included studies, 38 (58%) used a scale, and 9 (14%) used a domain-based assessment tool.
- Fifty (77%) systematic reviews used a ranking system that ranked studies based on critical assessment findings. Nineteen (38% of 50) systematic reviews ranked studies based on a summary score of methodological study quality. Thirty-one (62% of 50) ranked studies using a threshold summary score to classify studies in categories of “high”, “moderate”, or “low quality”.
- Of sixty-five systematic reviews that performed a critical assessment of included studies, 11 (17%) performed a risk of bias assessment for separate outcomes. Three (5%) performed a domain-based risk of bias assessment for separate outcomes; 10 (15%) performed domain-level risk of bias assessments but not for separate outcomes. Forty-seven (72%) systematic revies presented risk of bias assessment findings for each tool item but not for individual risk of bias domains or for separate outcomes.
- Two systematic reviews (3%) performed a meta-regression to examine the quantitative influence of risk of bias or study quality of each included study on individual study effect size. Ten (15%) performed a sensitivity analysis to compare the effect estimates of studies at ‘high’ and ‘unclear’ risk of bias to all included studies. One systematic review (2%) excluded

studies at 'high' and 'unclear' risk of bias from the synthesis to restrict their evidence synthesis to studies at low risk of bias. Forty-three (66%) narratively discussed risk of bias assessment findings in the context of systematic review findings, and 9 systematic reviews (14%) did not incorporate risk of bias assessment findings into review findings.

Table 2. Prevalence & methodological features of critical assessment performed in systematic reviews published in BJSM.

Data items	Reviews k (%)
Critical assessment performed in systematic review	65 (99%)
Critical assessment tool type	
Standard tool	42 (65%)
Adapted tool	14 (21%)
Custom tool	9 (14%)
Critical assessment tool format	
Checklist	18 (28%)
Scale	38 (58%)
Domain-based	9 (14%)
Ranking system	50 (77%)
Simple summary score	19 (38%)
Cut-off threshold score	31 (62%)
Separate critical assessments for each outcome	
Yes	11 (17%)
No	54 (83%)
Domain-based assessment across outcomes	
Yes	10 (15%)
No	55 (85%)
Critical assessment for each tool item	
Yes	47 (72%)
No	18 (28%)
Methods to incorporate risk of bias assessment findings into review findings	
Meta-regression (quantitative)	2 (3%)
Sensitivity analysis (quantitative)	10 (15%)
Excluded studies at high or unclear risk of bias (or moderate or low quality) (qualitative)	1 (2%)
Qualitative description (i.e. narrative discussion)	43 (66%)
No attempt to incorporate risk of bias assessment findings into review findings	9 (14%)

CRITICAL RE-ASSESSMENT USING A RISK OF BIAS ASSESSMENT INSTEAD OF AN ASSESSMENT OF STUDY QUALITY

Features of risk of bias assessments including the terminology, assessment tool and method, and incorporation method used, influence risk of bias assessment findings. In this section, we present a worked example where we re-assess the risk of bias of RCTs in one *BJSM* systematic review.(16) By

assessing the risk of bias in RCTs, we illustrate the impact that different critical assessment tools can have on risk of bias assessment findings, and ultimately, systematic review findings.

Methods of risk of bias assessment

We used pre-specified inclusion criteria to identify a systematic review (from our original group of 65) that included intervention studies and intended to assess risk of bias, but evaluated features other than risk of bias (Eligibility criteria – supplementary file 1). Our eligibility criteria identified a systematic review (16) that investigated the efficacy of therapeutic interventions to improve patient-reported function in participants with chronic ankle instability (CAI). This systematic review included 17 original research studies: 11 RCTs, 1 cohort study, 2 case-control studies, and 3 case-series. For the purpose of this education review, we re-assessed only the 11 RCTs included in this systematic review.(16) An author of the current study (ED) contacted a member of the systematic review team to articulate the current study's aims and to obtain the review authors' permission to perform a risk of bias assessment of RCTs included in their systematic review. The lead and supervising authors of this systematic review(16) also provide a commentary on the current study's findings in the context of their systematic review findings **(**see Kosik & Gribble).

We assessed the risk of bias in outcomes of patient-reported function (activities of daily living (ADL) and sports subscales) in eleven RCTs(16) using the Cochrane Risk of Bias tool 2 (RoB2). RoB2 is the revised, second edition of the Cochrane Risk of Bias tool for RCTs.(17,18) RoB2 is an outcome-focused, domain-based tool that assesses the risk of bias of outcomes in individually-randomized, parallel-group trials, randomized cross-over trials, and cluster-randomized controlled trials.(18)

RoB2 features five risk of bias domains for individually-randomized, parallel-group trials:

1. Bias arising from the randomization process;
2. Bias due to deviations from intended interventions;
3. Bias due to missing outcome data;
4. Bias in measurement of the outcome,

5. Bias in selection of the reported results.

Responses to the RoB2 signaling questions are mapped using a decision algorithm to determine each risk of bias domain judgement.(18) Finally, an overall risk of bias judgement is made for each assessed outcome, in each trial, based on the domain-level assessment. A study outcome is judged as 'low risk' when all domains are judged to be at low risk of bias. A study outcome is determined to have 'some concerns' of bias when one or more domains are judged to be at 'some concerns' of bias. A study outcome is judged as 'high risk' of bias when at least one domain is judged to be at 'high risk' of bias, or when multiple domains have 'some concerns'. Multiple domains at 'some concerns' of bias raise the likelihood of distortion of the treatment effect (i.e., a high overall risk of bias). Strong meta-epidemiological evidence supports the content of risk of bias domains included in RoB2.(19)

Two independent assessors (FCB & ED) re-assessed the risk of bias of the eleven RCTs using the revised version of RoB2 from October 9th, 2018 (http://www.riskofbias.info).(18) Assessors resolved initial disagreement via discussion. A third, independent assessor (MW & RE) arbitrated any persisting disagreements. Assessor were not blind to the findings of the Downs & Black quality assessment performed by Kosik et al.(16)

Results of study quality assessment and risk of bias assessments

Kosik et al.(16) performed an intended risk of bias assessment using the Downs and Black Checklist. The Downs and Black Checklist is a quality assessment scale developed to evaluate the methodological quality of randomized and non-randomized trials.(20) The checklist comprises 27 items across 4 subscales including: completeness of reporting (9 items); internal validity (13 items); precision (1 item), and; external validity (3 items). The Downs and Black Checklist assigns numeric values to item responses ('Yes' = 1; 'No' = 0; 'Unable to determine' = 0). Each item score is summed; higher scale summary scores indicate superior study quality. The Downs and Black Checklist considers the methodological quality of each study irrespective of the number and type of outcomes assessed within each study.

Risk of bias reassessment: Bias arising from the randomization process

Eight trials (73%) had outcomes at ‘some concerns’ of bias arising from the randomisation process. This was due to improper or unclear methods of sequence generation, allocation concealment, or due to imbalances in group baseline demographic characteristics. Outcomes in 2 RCTs (18%) had ‘high’ risk of bias arising from the randomisation process. One RCT (9%) had an outcome at ‘low’ risk of bias.

Risk of bias reassessment: Bias due to deviations from intended interventions

Seven RCTs (64%) had outcomes at ‘high’ risk of bias due to deviations from intended interventions with specific interest in lack of adherence to the intervention. Two trials (18%) had ‘some concerns’ of bias and two trials (18%) had ‘low’ risk of bias.

Risk of bias reassessment: Bias due to missing outcome data

Two RCTs (18%) had outcomes at ‘high’ risk of bias due to missing outcome data. Nine trials (82%) had outcomes that were at ‘low’ risk of bias due to little, or no, missing outcome data.

Risk of bias reassessment: Bias in measurement of the outcome

Nine RCTs (73%) had outcomes at ‘high’ risk of bias, predominantly due to a lack of participant blinding. Trial participants were outcome assessors due the use of patient-reported outcome measures. Two trials (18%) were at ‘low’ risk of bias and 1 trial (9%) was at ‘some concerns’ of bias in measurement of the outcome.

Risk of bias reassessment: Bias in selection of the reported results

One RCT (9%) was at ‘high’ risk of bias in selection of the reported results. Ten trials (91%) had outcomes at ‘some concerns’ of bias. ‘Some concerns’ of bias were due to no available pre-specified trial protocol or analysis plan, and the possibility for many statistical analyses, other than that reported in the results of each trial, to be performed.

Risk of bias reassessment: Overall risk of bias for patient-reported function

All RCTs (k = 11; 100%) were at 'high' overall risk of bias for all intervention comparisons and follow-up assessment time-points. RCTs were at 'high' overall risk of bias due to the presence of at least one risk of bias domain at 'high' risk of bias or multiple risk of bias domains at 'some concerns' of bias (Table 3).(21–31).

Study quality

The assessment of study quality (using the Downs and Black checklist) produced a mean scale summary score of 21 out of 31 (minimum-maximum = 11-24) across eleven RCTs.(16) Eight RCTs were judged by Kosik et al. as high quality and 3 RCTs were judged by Kosik et al. as moderate quality.

Summary

Using the Downs and Black checklist, the majority of included RCTs (8/11) were judged to be high-quality trials. Kosik et al. interpreted study quality assessment findings to provide moderate-to-high quality evidence for therapeutic interventions improving patient-reported function in individuals with chronic ankle instability (CAI). Using ROB2 on the same sample of RCTs, all trials were judged to be at 'high' overall risk of bias. Our interpretation is that the current evidence for therapeutic interventions in individuals with CAI is likely prone to bias-limiting conclusions.

Table 3. Overall and domain-level risk of bias judgements for each review outcome, intervention comparisons, and follow-up assessments timepoints.

Study	Outcome	Intervention comparison	Follow-up assessment time-points	Bias arising due to the randomization process	Bias due to deviation from intended interventions (adhering to the intervention)	Missing outcome data	Bias due to measurement of the outcome	Bias due to selection of the reported result	OVERALL RISK OF BIAS JUDGEMENT
Wright et al., 2017.	**FAAM-ADL FAAM-Sport**	Wobble board balance training vs. ankle strengthen resistance tubing	Day intervention ceased	Some Concerns	Some Concerns	Low Risk	Some Concerns	Some Concerns	**High Risk**
Donovan et al., 2016.	**FAAM-ADL FAAM-Sport**	Exercise rehabilitation using destabilization device vs. exercise rehabilitation using no destabilization device	48-hour follow-up	Some Concerns	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
McKeon & Wikstrom 2016	**FAAM-ADL FAAM-Sport**	Ankle joint mobilization vs. plantar massage vs. triceps surae stretching vs. control	72-hour follow-up	Low Risk	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
			4-week follow-up	Low Risk	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
Cruz-Diaz et al., 2015. DR.	**CAIT**	Group 1: Intervention Group 2: Sham intervention Group 3: Control group	3-week follow-up	Some Concerns	High Risk	Low Risk	Low Risk	Some Concerns	**High Risk**
			6-month follow-up	Some Concerns	High Risk	Low Risk	Low Risk	Some Concerns	**High Risk**
Cruz-Diaz et al., 2015. IJSM.	**CAIT**	Comprehensive exercise program vs. general lower body strength program	Day intervention ceased	Some Concerns	Low Risk	Low Risk	High Risk	Some Concerns	**High Risk**
Lubbe et al., 2015.	**FADI-ADL**	Manual therapy + rehabilitation vs. rehabilitation alone	Day intervention ceased	Some Concerns	High Risk	Low Risk	High Risk	High Risk	**High Risk**
Salom-Moreno et al., 2015.	**FAAM-ADL FAAM-Sport**	TrP-DN + strength/ proprioceptive training vs. strength/proprioceptive training	4-week follow-up	Some Concerns	Some Concerns	Low Risk	High Risk	Some Concerns	**High Risk**
Collins et al., 2014.	**FAAM-ADL FAAM-Sport**	Strain counter-strain vs. sham strain counter-strain	Day intervention ceased	High Risk	Low Risk	Low Risk	Low Risk	Some Concerns	**High Risk**
Beazell et al., 2012.	**FAAM-Sport**	Group 1: Proximal tibio-fibular manipulation vs. Group 2: Distal tibio-fibular manipulation vs. Group 3: Control group – no intervention	Intervention Day 1	Some Concerns	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
			Day 7 of intervention	Some Concerns	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
			Day 14 of intervention	Some Concerns	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
			Day 21 of intervention	Some Concerns	High Risk	Low Risk	High Risk	Some Concerns	**High Risk**
Schaefer & Sandrey (2012).	**FAAM-ADL FAAM-Sport**	GISTM + balance training vs. Sham GISTM + balance training vs. balance training	Day intervention ceased	High Risk	High Risk	High Risk	High Risk	Some Concerns	**High Risk**
McKeon (2008)	**FAAM-ADL FAAM-Sport**	Balance training vs. control group	4-week follow-up	Some Concerns	High Risk	High Risk	High Risk	Some Concerns	**High Risk**

Abbreviations: IJSM = International Journal of Sports Medicine; DR = Disability & Rehabilitation; FAAM-ADL = Foot & Ankle Ability Measure – Activities of Daily Living Subscale; FAAM-Sport = Foot & Ankle Ability Measure – Sports Subscale; CAIT = Cumberland Ankle Instability Tool; TrP-DN = TriggerPoint – Dry Needling; GISTM = Graston Instrument-assisted Soft Tissue Mobilization

IMPLICATIONS & RECOMMENDATIONS

We found a high prevalence of critical assessments among systematic reviews published in *BJSM*. Risk of bias assessments were infrequently performed in systematic reviews published in *BJSM*. Many systematic reviews assessed study quality instead of risk of bias, which impairs accurate inferences about the credibility of study outcomes.(2,32,33)

Performing an assessment of study quality instead of a risk of bias assessment may lead to invalid systematic review findings, conclusions, and recommendations. The quality assessment of 11 RCTs in our worked example produced a mean summary score of 20.5 out of 31 (66%; minimum-maximum = 11–24), indicating moderate-to-high quality evidence.(16) Kosik and colleagues did not categorize scale summary scores into equal tertiles.(16) If categorized into equal tertiles of low (1-10), moderate (11-20), and high quality (21-31), 8 RCTs were of high quality, 3 RCTs were of moderate quality, and no RCT was judged to be of low quality (Table 4). However, using a risk of bias assessment, all RCTs were judged to be at ‘high’ overall risk of bias using RoB2 (Table 4).

Domains should dominate risk of bias assessments

Despite empirical evidence to support domain-based risk of bias assessment tools, approximately one-third of SEM systematic reviews included in our sample did not use an empirically-supported tool. Only 14% performed a domain-based risk of bias assessment while most used scales or checklists. This estimate of 14% in our sample of SEM systematic reviews is lower than in biomedicine where 41% of non-Cochrane systematic reviews used a domain-level risk of bias assessment tool.(9)

Scale summary scores often lack meaning and omit valuable information about specific study limitations in individual studies.(34,35) For example, summary scores presented in the systematic review example ranged from 11/31 to 24/31 (Table 4).(16) Reporting only one numeric value

provides insufficient detail to highlight specific limitations in trial design, conduct, or analysis that threaten trial validity.(36,37)

Using RoB2, in every RCT we identified at least one domain at 'high' risk of bias or multiple domains at 'some concerns' of bias for each assessed outcome. Due to available evidence on the relationships between various risk of bias domains and distorted effect estimates,(9,38,39) the use of an established domain-based tool enables a deeper understanding about the likely direction and size of trial effect estimates associated with bias. Our risk of bias assessment findings highlight that despite high trial quality,(16) all RCTs were at 'high' overall risk of bias. This renders the current evidence of therapeutic interventions for CAI prone to bias-limiting conclusions about patient-reported function.

Focus on outcomes, not trials

Systematic review authors should perform separate risk of bias assessments for subjective and objective outcomes rather than assess all review outcomes simultaneously using the same risk of bias assessment.(2) Study limitations can distort separate outcomes differently,(9,40) necessitating separate risk of bias assessments for different outcome types. For example, a subjective outcome, such as self-reported pain, is more likely to be over-estimated when a patient is aware of their allocation to a specific intervention group, (due to lack of patient blinding) than if they were not aware of their group allocation.(9) Conversely, a patient's awareness of their allocation to a therapeutic intervention group is less likely to influence an objective outcome such as re-injury.

Only half of systematic reviews across biomedicine, and 17% in our sample, performed separate risk of bias assessments for subjective and objective outcomes.(41) Kosik et al. (16) reported patient-reported function (activities of daily living and sporting activities). Due to the subjective nature of patient-reported outcomes, the effect of study limitations due to deviation from intended interventions (e.g., lack of patient blinding) may overestimate differences between intervention and control groups.

Table 4. A tabular comparison of trial quality, using Down & Black Quality Checklist, and risk of bias, using RoB 2.0.

Lead-author (Year)	Outcome measure	Downs & Black Quality Checklist		Cochrane RoB Tool 2.0
		Summary Score	Quality Category	Overall Risk of Bias Judgement
Wright et al., 2017.	FAAM-ADL FAAM-Sport CAIT (ADL)	**22**	High quality	High Risk
Donovan et al., 2016.	FAAM-ADL FAAM-Sport	**22**	High quality	High Risk
McKeon & Wikstrom 2016	FAAM-ADL FAAM-Sport	**23**	High quality	High Risk
Cruz-Diaz et al., 2015. DR.	CAIT (ADL)	**21**	High quality	High Risk
Cruz-Diaz et al., 2015. IJSM.	CAIT (ADL)	**21**	High quality	High Risk
Lubbe et al., 2015.	FADI-ADL	**23**	High quality	High Risk
Salom-Moreno et al., 2015.	FAAM-ADL FAAM-Sport	**22**	High quality	High Risk
Collins et al., 2014.	FAAM-ADL FAAM-Sport	**11**	Moderate quality	High Risk
Beazell et al., 2012.	FAAM-Sport	**24**	High quality	High Risk
Schaefer & Sandrey (2012)	FAAM-ADL FAAM-Sport	**20**	Moderate quality	High Risk
McKeon (2008)	FADI-ADL FADI-Sport	**17**	Moderate quality	High Risk
OVERALL BODY OF EVIDENCE	**Patient-reported function**	**Mean = 21 Min – Max = 11-24**	**HIGH QUALITY**	**HIGH RISK**

Abbreviations: IJSM = International Journal of Sports Medicine; DR = Disability & Rehabilitation; FAAM-ADL = Foot & Ankle Ability Measure – Activities of Daily Living Subscale; FAAM-Sport = Foot & Ankle Ability Measure – Sports Subscale; CAIT = Cumberland Ankle Instability Tool.
Cut-off scores to differentiate quality categories: Low quality = 1-10; Moderate quality = 11-20; High quality = 21-31.

Incorporating risk of bias assessment findings into systematic review findings

Incorporating risk of bias assessment findings into systematic review findings allows an interpretation of over- or under-estimated study outcomes, to avoid misleading conclusions. High numbers of studies at 'some concerns' or 'high' risk of bias necessitate a more cautious interpretation of review findings. Two thirds of systematic reviews in our sample presented a qualitative description of risk of bias assessment findings. However, these systematic reviews did not estimate the likely impact of bias on systematic review outcomes.(19) Approximately one in five systematic reviews used quantitative methods to adjust the review effect estimate based on the risk of bias present in included studies (Table 2).(2,41–44)

Table 5– Incorporating risk of bias assessment findings into review findings

Method	Explanation
Quantitatively incorporating risk of bias assessment findings when undertaking a meta-analysis	Sensitivity analyses can evaluate whether effect estimates change if only studies at 'low' risk of bias are included in meta-analysis compared to all included studies.(34) Meta-regression is a secondary analysis that assesses the quantitative influence of overall risk of bias judgements (e.g., 'low', 'some concerns', or 'high') of studies on the meta-analytic effect estimates.(43)
Qualitative description (descriptive information when there is no meta-analysis)	A detailed discussion can speculate about the extent to which included studies, at 'some concerns' or at 'high' risk of bias, may under-estimate or over-estimate review findings compared to studies at 'low' risk of bias.(19) Excluding studies at a pre-specified risk of bias judgement from the evidence synthesis offers an addition to a simple descriptive synthesis. For example, systematic review authors might decide to include only studies at 'low' risk in specific bias domains. By excluding studies at 'some concerns' and 'high' risk of bias from the synthesis, review authors may plausibly reduce the extent to which review findings are exposed to biased study-level data. Another approach to account for risk of bias is to visually present study findings according to risk of bias assessment findings.

How believable is the review effect estimate if bias is present?

In the RCTs descriptively synthesized by Kosik et al.,(16) small, non-significant between-group differences could represent an under-estimate due to bias that under-represents that true magnitude of an intervention's effect. The true difference between groups may actually be larger, but we cannot be certain because of high risk of bias across many domains. Therapeutic interventions may be more or less effective than concluded in the systematic review.(16) Applying an evidence-grading tool to

'high' overall risk of bias judgments in this systematic review lowers the quality of evidence and strength of recommendation for balance training and multimodal treatment (Strength of Recommendation Taxonomy (SORT) used by Kosik et al.(16)).

In the presence of inarguable demonstrations that substantiate why "most published research findings are false",(45) researchers must be able to identify study outcomes at 'some concerns' or 'high risk' of bias. To accomplish this, researchers need to perform separate domain-based risk of bias assessments (for separate subjective and objective outcome) that evaluate potential threats to a study's internal validity. In table 6, we present recommendations for systematic review authors and editorial team members to inform the conduct of a valid risk of bias assessment. These recommendations reflect a blend of information from contemporary texts in evidence synthesis methods, peer-reviewed meta-research, editorial and evidence synthesis experience, and the findings of our methodological study.

Table 6 – Guidance checklist for systematic review authors, peer-reviewers, and editorial team members when performing and interpreting risk of bias assessments:

1) Assess risk of bias:
Avoid assessing study quality in place of risk of bias. Use a rigorously-developed, study design-specific risk of bias tool such as RoB2 (for randomized intervention trials), ROBINS-I (for non-randomized interventions studies), ROBIS (for systematic reviews), PROBAST (for prediction modelling studies), QUIPS-II (for prognostic studies), or QUADAS-II (for diagnostic accuracy studies).
2) Use a risk of bias assessment tool in its original form:
Do not modify risk of bias tools by adding new items or omitting existing items that might deem to be relevant or irrelevant to the assessment of risk of bias. Modifying a standard risk of bias assessment tool can negatively impact on the sensitivity of the risk of bias assessment, by potentially including an item that does not address risk of bias. Systematic review authors should not develop and/or use their own critical assessment tool to assess bias.
3) Evaluate and present risk of bias using a domain-level risk of bias assessment:
Domain-level risk of bias assessments consider the effect of risk of bias domains on study outcomes rather than one summary score. Individual risk of bias domains contribute differently to the extent that study limitations may distort study results.
4) Avoid using scales and checklists to assess risk of bias:
Scales and checklists frequently generate summary scores by using one numeric value, which omits information about diverse sources of bias. Judging the risk of bias for a particular outcome based on individual domains is recommended because several domains at 'some concerns' or 'high' risk of bias raise the suspicion that an effect estimate may be distorted.
5) Avoid cut-off thresholds that categorise studies based on risk of bias or study quality:
Cut-off thresholds that dichotomise or categorise ordinal scales of study quality into nominal groups (of 'high', 'moderate', or 'low' study quality) omits valuable information about methodological differences between studies. There is no evidence to support the numeric ranking of studies based on overall study quality, particularly when independent items contribute towards the overall quality score of a study.
5) Assess risk of bias separately for different outcomes:

Subjective outcomes, such as pain, are more strongly influenced by participants' and outcome assessors' awareness of methodological phenomena such as knowledge of group assignment. An objective outcome, such as death, is more resistant to the influence of bias.

6) Incorporate risk of bias assessment findings into systematic review findings using quantitative or qualitative (descriptive) methods:

Use quantitative methods, such as sensitivity analyses or meta-regression, when applicable in meta-analyses to investigate the impact of bias on meta-analysis effect estimates. In a systematic review without meta-analysis, visually present study findings according to risk of bias assessment findings or provide an informative discussion to speculate about influence of risk of bias on the credibility of research findings.

7) Provide justification, based on the available risk of bias assessment criteria, to support each risk of bias judgement:

Supplementary File 2 provides the final judgements allocated by assessors to each signalling question of RoB2 in our risk of bias re-assessment of eleven RCTs included in the systematic review by Kosik et al.**(**16)

Summary

Consistent, valid, and trustworthy assessments of risk of bias in RCTs are essential to judge the credibility of a body of evidence. Nearly all systematic reviews published in *BJSM* between January 2016 and May 2017 included a critical assessment tool. However, few systematic reviews used domain-level risk of bias assessment tools or performed domain-level risk of bias assessments. Outcome types were rarely considered separately in risk of bias assessments. Quantitative methods were infrequently used to incorporate risk of bias assessment findings into systematic review findings. We identified discrepancies between risk of bias and study quality assessment findings in our re-assessment of a previous systematic review. Using different critical assessment tools generated different assessment findings, which affected inferences about the credibility of review findings. Risk of bias assessment tools must be correctly selected and administered, and findings interpreted and incorporated to inform the extent to which assessment findings likely impact a body of systematically-reviewed evidence.

References

1. Mulrow CD. The medical review article: state of the science. Ann Intern Med. 1987;106:485–8.

2. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Wiley; 2008. 672 p.

3. Ardern CL, Dupont G, Impellizzeri FM, et al. Unravelling confusion in sports medicine and sports science practice: a systematic approach to using the best of research and practice-based evidence to make a quality decision. Br J Sports Med. 2017;bjsports-2016-097239.

4. Cooper NJ, Jones DR, Sutton AJ. The use of systematic reviews when designing studies. Clin Trials. 2005;2:260–4.

5. Mulrow CD. Systematic Reviews: Rationale for systematic reviews. BMJ. 1994;309:597–9.

6. Andrews JC, Schünemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation—determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66:726–35.

7. Patsopoulos NA, Analatos AA, Ioannidis JPA. Relative Citation Impact of Various Study Designs in the Health Sciences. JAMA. 2005;293:2362–6.

8. Hewlett S, Wit M de, Richards P, et al. Patients and professionals as research partners: Challenges, practicalities, and benefits. Arthritis Care Res. 2006;55:676–80.

9. Page MJ, Shamseer L, Altman DG, et al. Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. PLOS Med. 2016;13:e1002028.

10. Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of healthcare interventions? BMJ Evid-Based Med. 2003;8:100–1.

11. Ioannidis JP a. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q. 2016;94:485–514.

12. Weir A, Rabia S, Ardern C. Trusting systematic reviews and meta-analyses: all that glitters is not gold! Br J Sports Med. 2016;50:1100–1.

13. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet. 2009; 374:86–9.

14. Bandholm T, Henriksen M, Thorborg K. Slow down to strengthen sport and exercise medicine research. Br J Sports Med. 2017;51:1453–1453.

15. Grant HM, Tjoumakaris FP, Maltenfort MG, et al. Levels of Evidence in the Clinical Sports Medicine Literature: Are We Getting Better Over Time? Am J Sports Med. 2014;42:1738–42.

16. Kosik KB, McCann RS, Terada M, et al. Therapeutic interventions for improving self-reported function in patients with chronic ankle instability: a systematic review. Br J Sports Med. 2017;51:105–12.

17. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.

18. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.

19. Page MJ, Higgins JPT, Clayton G, et al. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLOS ONE. 2016;11:e0159267.

20. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health. 1998;52:377–84.

21. Wright CJ, Linens SW, Cain MS. A Randomized Controlled Trial Comparing Rehabilitation Efficacy in Chronic Ankle Instability. J Sport Rehabil. 2017;26:238–49.

22. Beazell JR, Grindstaff TL, Sauer LR, et al. Effects of a Proximal or Distal Tibiofibular Joint Manipulation on Ankle Range of Motion and Functional Outcomes in Individuals With Chronic Ankle Instability. J Orthop Sports Phys Ther. 2012;42:125–34.

23. Donovan L, Hart JM, Saliba SA, et al. Rehabilitation for Chronic Ankle Instability With or Without Destabilization Devices: A Randomized Controlled Trial. J Athl Train. 2016;51:233–51.

24. Cruz-Díaz D, Vega RL, Osuna-Pérez MC, et al. Effects of joint mobilization on chronic ankle instability: a randomized controlled trial. Disabil Rehabil. 2015;37:601–10.

25. Cruz-Diaz D, Lomas-Vega R, Osuna-Pérez MC, et al. Effects of 6 Weeks of Balance Training on Chronic Ankle Instability in Athletes: A Randomized Controlled Trial. Int J Sports Med. 2015;36:754–60.

26. Collins CK, Masaracchio M, Cleland JA. The effectiveness of strain counterstrain in the treatment of patients with chronic ankle instability: A randomized clinical trial. J Man Manip Ther. 2014;22:119–28.

27. Lubbe D, Lakhani E, Brantingham JW, et al. Manipulative Therapy and Rehabilitation for Recurrent Ankle Sprain With Functional Instability: A Short-Term, Assessor-Blind, Parallel-Group Randomized Trial. J Manipulative Physiol Ther. 2015;38:22–34.

28. Salom-Moreno J, Ayuso-Casado B, Tamaral-Costa B, et al. Trigger Point Dry Needling and Proprioceptive Exercises for the Management of Chronic Ankle Instability: A Randomized Clinical Trial. Evidence-Based Complementary and Alternative Medicine. 2015 [cited 2018 Oct 17]. Available from: https://www.hindawi.com/journals/ecam/2015/790209/

29. Schaefer JL, Sandrey MA. Effects of a 4-Week Dynamic-Balance-Training Program Supplemented with Graston Instrument-Assisted Soft-Tissue Mobilization for Chronic Ankle Instability. J Sport Rehabil. 2012;21:313–26.

30. McKeon PO, Wikstrom EA. Sensory-Targeted Ankle Rehabilitation Strategies for Chronic Ankle Instability. Med Sci Sports Exerc. 2016;48:776–84.

31. McKeon PO, Ingersoll CD, Kerrigan DC, et al. Balance training improves function and postural control in those with chronic ankle instability. Med Sci Sports Exerc. 2008;40:1810–9.

32. Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials. 1995;16:62–73.

33. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? The Lancet. 1998;352:609–13.

34. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. BMJ. 2001;323:42–6.

35. Jüni P, Witschi A, Bloch R, et al. The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. JAMA. 1999;282:1054–60.

36. da Costa BR, Hilfiker R, Egger M. PEDro's bias: summary quality scores should not be used in meta-analysis. J Clin Epidemiol. 2013;66:75–7.

37. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostat Oxf Engl. 2001;2:463–71.

38. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012;157:429–38.

39. Savović J, Turner RM, Mawdsley D, et al. Association Between Risk-of-Bias Assessments and Results of Randomized Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study. Am J Epidemiol. 2018;187:1113–22.

40. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336:601–5.

41. Katikireddi SV, Egan M, Petticrew M. How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. J Epidemiol Community Health. 2015;69:189–95.

42. Detsky AS, Naylor CD, O'Rourke K, et al. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol. 1992;45:255–65.

43. Schmid CH, Stark PC, Berlin JA, et al. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol. 2004;57:683–97.

44. Hopewell S, Boutron I, Altman DG, et al. Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study. BMJ Open [Internet]. 2013 [cited 2018 Jan 25];3(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3753473/

45. Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Med. 2005;2:e124.