Systematic reviews do not (yet) represent the 'gold standard' of evidence: a position paper

Running head: systematic reviews are problematical

RA Moore [1,4], E Fisher [2,4], C Eccleston [2,3,4]

[1] Newton Ferrers, Plymouth, United Kingdom.
[2] Centre for Pain Research, University of Bath, Bath, United Kingdom.
[3] Department of Clinical and Health Psychology, Ghent University, Ghent, Belgium.
[4] Cochrane Pain, Palliative, and Supportive Care Review Groups, Oxford University Hospitals, Oxford, United Kingdom;

Corresponding author:
Andrew Moore,
Appledore, 84 Court Road, Newton Ferrers, Plymouth PL8 1DA UK.
Tel: +44 1753873532;
andrew.moore@omkltd.org

Category: Position Paper

Funding: No funding was received for this work.

COI: Authors report no potential conflicts of interest other than current (EF) or past (RAM, CE) editorship roles at Cochrane Pain, Palliative, and Supportive Care review group.

Significance: Most systematic reviews are badly done and based on trials that are themselves inadequate: only about 3 in 100 has both adequate methods and is clinically useful. The position paper examines some of the main deficiencies in how evidence assessing the efficacy of interventions for pain is collated and its quality assessed and suggests mechanisms whereby common and serious deficiencies in systematic reviews can be avoided.

Abstract

The low quality of included trials, insufficient rigour in review methodology, ignorance of key pain issues, small size, and over-optimistic judgements about the direction and magnitude of treatment effects all devalue systematic reviews, supposedly the 'gold standard' of evidence. Available evidence indicates that almost all systematic reviews in the published literature contain fatal flaws likely to make their conclusions incorrect and misleading. Only 3 in every 100 systematic reviews are deemed to have adequate methods and be clinically useful. Examples of research waste and questionable ethical standards abound: most trials have little hope of providing useful results, and systematic review of hopeless trials inspires no confidence. We argue that results of most systematic reviews should be dismissed. Forensically critical systematic reviews are essential tools to improve the quality of trials and should be encouraged and protected.

Introduction and background

Systematic reviews are supposed to be the 'gold standard' of evidence-based medicine; numbers are rising rapidly, with an estimated 35,000 likely to be published in 2021. At current rates of growth our journals will soon be publishing more systematic reviews than randomised trials. Driven by issues of data collection during the COVID-19 pandemic, researchers and students have redirected their attention from primary data collection to secondary data analysis, often to be used as part of a thesis for a higher degree.

In pain, most systematic reviews address effectiveness of interventions for treatment of established pain, while some, notably in anaesthetics, evaluate interventions to prevent the occurrence of postoperative pain in the short or long term. These systematic reviews are the subject of this position paper. Systematic review methodology can be used for many different purposes where different rules of evidence can apply.

There are cogent reasons to distrust almost all systematic reviews: many are useless, or worse, misleading. A broad and scathing analysis of systematic reviews dismisses most as flawed or redundant (Ioannidis, 2016). Only 3% – to emphasise, 3 in every 100 systematic reviews – are deemed to have adequate methods and be clinically useful. It is over 25 years since the observation that meta-analyses of low quality produced significantly more positive conclusions (Jadad & McQuay 1996), and the need to improve the quality of systematic reviews has been oft repeated (Eccleston et al, 2010; Moore et al, 2010).

Trust in evidence must be earned through a forensic obsession with bias and quality. All Cochrane Review Groups, including the PaPaS (Pain, Palliative and Supportive Care) Review Group, adopt formal methods for investigating and managing bias, laid out in the Cochrane Handbook (Higgins et al., 2021). The handbook recognises many problems, but inevitably deals in the generalities of bias. There are, however, specific pain-related aspects of research quality that are not covered. These include (but are not limited to):

- the need for patient-reported pain outcomes;
- for initial pain to be of moderate or severe intensity to achieve sensitivity in tests of analgesic efficacy;
- careful consideration of imputation method in chronic, and some acute, pain studies because of its influence on effect size;
- performing sensitivity analysis for the impact of small studies the potential for publication bias.

Omission of any of these can lead to authors drawing incorrect conclusions regarding efficacy and safety. Using the recent IASP task force on cannabis preparations as an example, these key features were unaddressed in over 90% of systematic reviews of cannabis, cannabinoids, and cannabis-based medicines (Moore et al., 2021). Even for 'standard' assessment of risk of bias advocated for by the Cochrane Handbook in the included studies, only 50% properly examined randomisation and blinding. The overview found that only 86% (48/57) of reviews provided critically low or low confidence in the results when judged with AMSTAR-2.

Many of these inadequate systematic reviews concluded that cannabinoids work, in contrast to systematic reviews of higher quality in the Cochrane Library, and the systematic review performed for the IASP Task Force on cannabinoids that used all quality criteria (Fisher et al., 2021). When assessing risk of bias of the 37 included randomised trials (fewer trials than systematic reviews, it should be noted), not a single trial had low risk of bias for every criterion, and 28 of the 37 (75%) had at least one high risk of bias.

This is a common failing of most systematic reviews, not just those about pain. The problem is not simply one of poor-quality systematic reviews, but of poor-quality medical research in general. An updated systematic review of perioperative drug therapy to prevent chronic pain published in 2021 found three times more trials (110 RCTs) with four times more participants (20,000) than the previous update published in 2013. Due to the many different drugs (20), surgeries (14), small size (88% <200 participants), and high risk of bias (96% of trials with at least one high risk of bias), no conclusions could be drawn (Carley et al., 2021). This represents significant research waste and poses important questions about the way ethical approval was sought and given. This is not a lone example: a Cochrane review of psychological therapies (75 trials, 9,400 participants) for chronic pain commented "given a broad mixture of outcome metrics within each domain, and considerable heterogeneity at baseline, we were unable to make any meaningful translation of effect sizes into clinically interpretable changes" (Williams et al., 2020).

The scandal of poor medical research has been a well-known fact for at least a quarter of a century (Altman, 1994), but little seems to be changing (Moore, 2021).

What is important?

Many systematic reviews do little more than regurgitate the results of individual trials or perform summary analyses of outcomes used by trialists. This falls into the trap of making the measured important, rather than evaluating whether individual trials report in some format what is important, especially what is important to people with the lived experience of pain. Rather than some average change, people with pain want their pain reduced by a large extent (Moore et al., 2013). Outcomes at or close to this, or like

substantial or moderate benefit as suggested by IMMPACT (Dworkin et al., 2008), are often to be found but go unreported.

An example can be found in the use of oral morphine for cancer pain (Wiffen et al., 2016). The latest of a series of updates abandoned reporting on a large series of trials with no consistent comparator in favour of evaluating pain reduced to no worse than mild within 14 days of treatment start; the result was a startling and consistent 96%.

Problems with small studies

Small study size has been suggested as one reason why "most published research findings are false" (Ioannidis, 2005), as well as the origin of considerable research waste (Roberts & Ker, 2015). It has been suggested that systematic reviews should use only prospectively registered clinical trials of sufficient quality and size (Roberts et al., 2015). This would dramatically cut down the work involved in systematic review, as over 80% of trials of small size and poor quality would automatically be eliminated from consideration. It would certainly improve our confidence in any results and help eliminate the growing issue of publications later considered as fraudulent (see https://retractionwatch.com/).

There is now increasing recognition that results based on a small number of small underpowered studies are likely to produce incorrect or highly imprecise results. An analysis on the impact of study size in Cochrane reviews has highlighted that if two adequately powered studies are available, omitting all underpowered studies makes little or no difference to the result (Turner et al., 2013). That study also indicated that a large proportion of Cochrane reviews contain ONLY underpowered studies. Small underpowered studies made up the entirety of the evidence in most meta-analyses reported by Cochrane reviews: in 70% of 14,886 meta-analyses, all included studies were underpowered (defined as less than 50% power to detect a 30% relative risk reduction - NNT values of around 5); only 17% of meta-analyses had two or more adequately powered studies. An analysis of Cochrane reviews published by the Cochrane Heart Group showed that of 22 meta-analyses reported to be conclusive by their authors, 12 (55%) contained

insufficient data to detect or rule out a 25% relative treatment effect (AlBalawi et al., 2013).

Small studies are also associated with strong positive bias. For example, a meta-assessment of bias demonstrated that small study size generated significant positive bias in biological and social sciences (Fanelli et al., 2017). An examination of 93 meta-analyses published in leading journals or the Cochrane Library indicated an overestimation in treatment effect of around 48% in studies with fewer than 50 participants, but with significant overestimation even in studies with 100-200 participants (Dechartres et al., 2013). Analysis of 13 meta-analyses of interventions for osteoarthritis indicated that treatment effects were more beneficial in small trials compared to large trials (Nüesch et al., 2010). In six of the 13 meta-analyses, the overall pooled estimate suggested a significant and clinically relevant benefit of treatment, whereas analyses restricted to large trials yielded smaller, and, importantly, non-significant, estimates.

Despite this critique, small studies are important for other reasons. For example, a high-quality crossover trial of only 31 patients demonstrated that 5 patients on amitriptyline and 4 patients on nortriptyline had a good response to that drug but failed to respond to the other despite good blood levels and being pushed to intolerable side effects (Watson et a., 1998). Evidence of the utility of early use of enriched enrollment randomized withdrawal designs in neuropathic pain was demonstrated over a short period in just 100 patients (Hewitt et al., 2011). Meta-analysis of high-quality small studies was useful in establishing dose response of analgesics in acute pain (McQuay & Moore, 2006).

How well does this intervention work?

Size matters not just in terms of direction of effect (does this intervention work?), but even more for the magnitude of effect (how well does this intervention work?) which might need up to 10 times more data (Moore et al., 1998). For systematic reviews to be useful they should describe results in ways that are relevant to professionals and people living with pain. That means avoiding reliance on relative statistical outputs like odds ratios or standardised mean differences and using absolute outputs that are easier to

use and understand (Rose, 1991), such as number needed to treat (Laupacis et al., 1988; Roose et al., 2016), or even converting results to success rates (Moore et al., 2013).

Table 1 about here

Comparing rates of substantial pain relief of two interventions can generate very similar relative statistical outputs (odds ratios, relative risk), but very different absolute outputs (NNT, success rates; Table 1). Standardised mean difference conversion to NNT is possible: when the SMD = 1, the NNT = 2, but as the SMD approaches zero the NNT becomes very large, so SMD of 0.5 becomes an NNT of about 5, and an SMD of 0.25 is equivalent to an NNT of above 15 (Faraone, 2008).

How small is small, and how large is large?

This depends on two main factors: the size of any likely treatment effect, and the practicalities of recruiting patients with the disorder. Treatment effects in acute and chronic pain are known to vary between zero for intravenous immunoglobulins in complex regional pain syndrome (CRPS; Goebel et al., 2017) and 95% for oral opioids in cancer pain (Wiffen et al., 2017), and between these values (Moore et al., 2013). While many pain conditions are relatively common, others are rare, making patient recruitment difficult. Systematic review of complex regional pain syndrome (CRPS) demonstrates that most studies involve very few patients, though at least two studies have recruited around 100 patients (Mbizvo et al., 2015; Goebel et al., 2017). Despite these modest numbers, this systematic review was able to make a relevant observation about the absence of a placebo effect in CRPS.

A recent simulation exercise has suggested that randomization removes random differences between treatment groups when including at least 1,000 participants to exclude bias in effects estimation (Nguyen et al., 2018). Even that is impractical in most pain trials, emphasising the importance of random chance. Researchers have long been concerned about the issue of size (in terms of participants or events (e.g., adequate pain relief)), and the potentially large effects of random chance when these are small (Flather et al., 1997; Pogue et al., 1997; Moore et al., 1998). Flather, for example, suggested a

minimum of 200 events (beneficial or adverse) was needed for any possibility of accurately estimating the extent or causation of those adverse events.

A simulation of 10,000 trials and meta-analyses based on acute pain studies indicated that good quality studies or meta-analyses would necessitate around 400 participants, or 200 per treatment group to provide a confident estimate of effect size using NNT (Moore et al., 1998). That would also agree with around 200 events, where an event was a participant having 'good' pain relief. An unpublished extension of those analyses examined the different situation of chronic pain studies using 100,000 trial simulations (Gavaghan & Moore, unpublished). For most chronic pain, where effect sizes are more modest than acute pain, treatment group sizes in good quality trials or meta-analyses of 1,000 patients or more are needed to have confidence that random effects do not influence a result.

The numbers of participants in comparisons to be reasonably certain of the magnitude of the effect at a level of 90% depends on effect size, numbers, and how certain one needs to be. Table 2 provides an indication for three pain conditions. Where information available is less than this, the GRADE quality of the evidence would be very low, because the likelihood of the result being substantially different is very high. Indeed, in discussing the issue of sample size determination, Lenth makes the point that sample size may not be the main issue and that the real goal is designing a high-quality study (Lenth 2001). The issue of data mining overly large trials for trivial statistical significance is rarely an issue in estimation of efficacy for pain therapy, more important is balancing the judgement of importance between statistical and clinical significance of the results.

Table 2 about here

In most circumstances there are good arguments for excluding any study of group size below 100 participants from systematic reviews. There are cogent reasons for dismissing results from systematic reviews unless obtained from at least 500 participants in total, and only then with tests showing no likely effects of potential publication bias when effect sizes are small (Moore et al., 2008).

The value of Cochrane

For pain, we are fortunate that Cochrane PaPaS has produced 300 high quality systematic reviews and 13 overview reviews, often by going beyond the requirements of Cochrane and by including the additional quality requirements needed for pain. Cochrane review teams typically include a broad range of expertise, such as pain professionals and pain review methodology experts, and increasingly include input from consumers; all of them are needed to ensure high quality.

By contrast, low quality systematic reviews are often conducted by small teams of authors with little or no experience of pain review methods, current knowledge about pain, or including people with pain. Few will appreciate the importance of using relevant domains of measurement and appropriate measurement tools (Smith et al., 2020). Basic errors abound. Some are fortunately caught during peer review, but too many are still published, even in the highest impact journals. The burden for journal editors, peer reviewers, and consumers of selecting the 3% of trustworthy and clinically useful reviews is substantial.

Unfortunately, the bastion of quality represented by Cochrane may be about to fall, due to major funding changes in UK science resulting in the withdrawal of infrastructure funding for all UK Cochrane groups, including PaPaS. In a proposed new model, the existing Cochrane Review Group and Network structure will be replaced by 8-10 larger, multi-topic, interdisciplinary evidence synthesis units, without any promise of building on the foundations of pain evidence (https://www.futurecochrane.space/).

Conclusion and position

There are no easy answers, and no *deus ex machina* to solve the problem. The obvious, if difficult, solution is for international associations like EFIC and IASP to join in a long-term collaboration with Cochrane PaPaS (or any successor). Their aims should include training and education for authors, peer reviewers, editors, and consumers. The products should be high quality systematic reviews and methodological analyses published in Cochrane

and/or other appropriate journals. A goal, perhaps a far-reaching goal, might be the production and maintenance of high-quality evidence as a source for healthcare organisations and governments to use for 'living' recommendations across a portfolio of painful conditions.

Table 3 about here

Table 3 is a distillation of the 16 AMSTAR-2 generic quality items and 9 pain specific items that should be considered when assessing the value of a systematic review of interventions for pain. They build on established work (Jadad et al., 1996; Shea et al., 2017), and might be a useful aide memoir when performing, editing, reviewing, or reading a review, or for attempts to establish agreed standards. Any overlap is the result of different instructions on judging criteria.

The present constellation of circumstances demands that publication of systematic reviews is justifiable only when the highest standards are met. Trials may be few or many, large or small, of good or poor quality, but the process of systematic review should have two essential themes to aid our thinking.

The first theme is to ensure that any conclusions regarding the direction and magnitude of any beneficial or harmful effect is judged by the highest standards currently available. This is not a proscriptive agenda, of stopping publication of all systematic reviews. Many (indeed most) high quality systematic reviews reveal chasms in our knowledge due to inadequate trials or absence of trial evidence. These so-called "empty" reviews, found commonly in Cochrane, are helpful in pointing out what we do not know with any confidence, as with cannabinoids for pain (Fisher et al., 2021).

The second theme is less about the systematic review itself, but about the trials comprising that review, and particularly their nature and their methods. Systematic reviews might (should) be considered primarily about adequacy or inadequacy of our experimental methods. For example, a systematic review of perioperative ketamine demonstrated that only 14 of 86 studies had moderate or severe pain in the control arm and these studies demonstrated a much lower 24-hour pain intensity at rest over control (by 17/100 mm)

compared to when analysing all 86 (4/100 mm) (Brinck et al., 2018). Perioperative ketamine may be useful, but the bulk of the studies lacked sensitivity to show it.

Both these themes should help set any future research agenda based on the firm foundations of good trial methodology. Journal editors should remain interested in systematic reviews. There is nothing intrinsically wrong with the method. However, submission from teams without direct experience of the subject being studied should be discouraged. We will work to encourage all editors of the major pain journals to uphold the highest quality pain specific agreed standards in both conduct and reporting.

Humans are biased toward action and crave the 'certainty' that gives that action credence. Paradoxically, evidence-based medicine is often in the business of declaring 'uncertainty', and so demanding pause and reflection. But pain is a complicated business, and its history is littered with the immiserating consequence of the over-stated, oversimplified, and the over-promised (Bell & Kalso, 2021). It is hard sometimes for readers to hear any truth above the babble of overstated results. Systematic review was embraced to move us away from the biases of single studies. If standards crumble, and PaPaS is destroyed, the tsunami of poor quality and wasteful reviews will amplify the problem rather than fix it. Patients deserve better than this and we need to address this problem before it is too late.

Acknowledgements

The authors are grateful to Professor Luis Garcia-Larrea for his guidance in the development of the manuscript, and to Professor David Gavaghan for help with simulations.

Author Contributions

RAM, EF and CE all contributed to the intellectual development of the ideas and the writing of the manuscript.

References

AlBalawi, Z., McAlister, F. A., Thorlund, K., Wong, M., & Wetterslev, J. (2013). Random error in cardiovascular meta-analyses: how common are false positive and false negative results?. *International journal of cardiology*, *168*, 1102–1107. https://doi.org/10.1016/j.ijcard.2012.11.048

Altman D. G. (1994). The scandal of poor medical research. BMJ (Clinical research ed.), 308, 283–284. https://doi.org/10.1136/bmj.308.6924.283

Bell, R. F., & Kalso, E. A. (2021). Cannabinoids for pain or profit? Pain, 162(Suppl 1), S125–S126. https://doi.org/10.1097/j.pain.0000000000001930

Brinck, E. C., Tiippana, E., Heesen, M., Bell, R. F., Straube, S., Moore, R. A., & Kontinen, V. (2018). Perioperative intravenous ketamine for acute postoperative pain in adults. *The Cochrane database of systematic reviews*, *12*, CD012033. https://doi.org/10.1002/14651858.CD012033.pub4

Carley, M. E., Chaparro, L. E., Choinière, M., Kehlet, H., Moore, R. A., Van Den Kerkhof, E., & Gilron, I. (2021). Pharmacotherapy for the Prevention of Chronic Pain after Surgery in Adults: An Updated Systematic Review and Meta-analysis. *Anesthesiology*, *135*, 304–325. https://doi.org/10.1097/ALN.0000000000003837

Dechartres, A., Trinquart, L., Boutron, I., & Ravaud, P. (2013). Influence of trial sample size on treatment effect estimates: meta-epidemiological study. *BMJ (Clinical research ed.)*, *346*, f2304. https://doi.org/10.1136/bmj.f2304

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Kerns, R. D., Ader, D. N., Brandenburg, N., Burke, L. B., Cella, D., Chandler, J., Cowan, P., Dimitrova, R., Dionne, R., Hertz, S., Jadad, A. R., Katz, N. P., … Zavisic, S. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The journal of pain*, *9*(2), 105–121. https://doi.org/10.1016/j.jpain.2007.09.005

Eccleston, C., Moore, R. A., Derry, S., Bell, R. F., & McQuay, H. (2010). Improving the quality and reporting of systematic reviews. *European journal of pain (London, England)*, *14*(7), 667–669. https://doi.org/10.1016/j.ejpain.2010.05.015

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 3714–3719. https://doi.org/10.1073/pnas.1618569114

Faraone S. V. (2008). Interpreting estimates of treatment effects: implications for managed care. *P & T : a peer-reviewed journal for formulary management*, *33*, 700–711.

Fisher, E., Moore, R. A., Fogarty, A. E., Finn, D. P., Finnerup, N. B., Gilron, I., Haroutounian, S., Krane, E., Rice, A., Rowbotham, M., Wallace, M., & Eccleston, C. (2021). Cannabinoids, cannabis, and cannabis-based medicine for pain management: a systematic review of randomised controlled trials. Pain, 162(Suppl 1), S45–S66. https://doi.org/10.1097/j.pain.0000000000001929

Flather, M. D., Farkouh, M. E., Pogue, J. M., & Yusuf, S. (1997). Strengths and limitations of meta-analysis: larger studies may be more reliable. *Controlled clinical trials*, *18*(6), 568–666. https://doi.org/10.1016/s0197-2456(97)00024-x

Goebel, A., Bisla, J., Carganillo, R., Cole, C., Frank, B., Gupta, R., James, M., Kelly, J., McCabe, C., Milligan, H., Murphy, C., Padfield, N., Phillips, C., Poole, H., Saunders, M., Serpell, M., Shenker, N., Shoukrey, K., Wyatt, L., & Ambler, G. (2017). *A randomised placebo-controlled Phase III multicentre trial: low-dose intravenous immunoglobulin treatment for long-standing complex regional pain syndrome (LIPS trial)*. NIHR Journals Library. DOI: 10.3310/eme04050

Hewitt, D. J., Ho, T. W., Galer, B., Backonja, M., Markovitz, P., Gammaitoni, A., Michelson, D., Bolognese, J., Alon, A., Rosenberg, E., Herman, G., & Wang, H. (2011). Impact of responder definition on the enriched enrollment randomized withdrawal trial design for establishing proof of concept in neuropathic pain. *Pain*, *152*(3), 514–521. https://doi.org/10.1016/j.pain.2010.10.050

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.2 (updated February 2021). Cochrane, 2021. Available from www.training.cochrane.org/handbook.

Ioannidis J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*, e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis J. P. (2016). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank quarterly*, *94*, 485–514. https://doi.org/10.1111/1468-0009.12210

Jadad, A. R., & McQuay, H. J. (1996). Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. *Journal of clinical epidemiology*, *49*(2), 235–243. https://doi.org/10.1016/0895-4356(95)00062-3

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary?. *Controlled clinical trials*, *17*(1), 1–12. https://doi.org/10.1016/0197-2456(95)00134-4

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *The New England journal of medicine*, *318*, 1728–1733. https://doi.org/10.1056/NEJM198806303182605

Lenth RV (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55,* 187-193. https://doi.org/10.1198/000313001317098149

McQuay, H. J., & Moore, R. A. (2007). Dose-response in direct comparisons of different doses of aspirin, ibuprofen and paracetamol (acetaminophen) in analgesic studies. *British journal of clinical pharmacology*, *63*(3), 271–278. https://doi.org/10.1111/j.1365-2125.2006.02723.x

Moore A. (2021). Red for danger in systematic reviews?. *European journal of hospital pharmacy : science and practice*, ejhpharm-2021-003080. Advance online publication. https://doi.org/10.1136/ejhpharm-2021-003080

Moore RA, Barden J, Derry S, McQuay HJ. (2008) Managing potential publication bias. In: McQuay HJ, Kalso E, Moore RA editor(s). Systematic Reviews in Pain Research: Methodology Refined. Seattle: IASP Press:15–23.[ISBN: 978–0–931092–69–5]

Moore, A., Derry, S., Eccleston, C., & Kalso, E. (2013). Expect analgesic failure; pursue analgesic success. *BMJ (Clinical research ed.)*, *346*, f2690. https://doi.org/10.1136/bmj.f2690

Moore, R. A., Gavaghan, D., Tramèr, R. M., Collins, L. S., & McQuay, J. H. (1998). Size is everything--large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain*, *78*, 209–216. https://doi.org/10.1016/S0304-3959(98)00140-7

Moore, A. R., Eccleston, C., Derry, S., Wiffen, P., Bell, R. F., Straube, S., McQuay, H., & ACTINPAIN writing group of the IASP Special Interest

Group (SIG) on Systematic Reviews in Pain Relief and the Cochrane Pain, Palliative and Supportive Care Systematic Review Group editors (2010). "Evidence" in chronic pain--establishing best practice in the reporting of systematic reviews. *Pain*, *150*(3), 386–389. https://doi.org/10.1016/j.pain.2010.05.011

Moore, R. A., Fisher, E., Finn, D. P., Finnerup, N. B., Gilron, I., Haroutounian, S., Krane, E., Rice, A., Rowbotham, M., Wallace, M., & Eccleston, C. (2021). Cannabinoids, cannabis, and cannabis-based medicines for pain management: an overview of systematic reviews. *Pain*, *162*(Suppl 1), S67–S79. https://doi.org/10.1097/j.pain.0000000000001941

Moore, R. A., Straube, S., Aldington, D. (2013). Pain measures and cut-offs - 'no worse than mild pain' as a simple, universal outcome. *Anaesthesia*, *68*(4), 400–412. https://doi.org/10.1111/anae.12148

Mbizvo, G. K., Nolan, S. J., Nurmikko, T. J., & Goebel, A. (2015). Placebo responses in long-standing complex regional pain syndrome: a systematic review and meta-analysis. *The journal of pain*, *16*(2), 99–115. https://doi.org/10.1016/j.jpain.2014.11.008

Nguyen, T. L., Collins, G. S., Lamy, A., Devereaux, P. J., Daurès, J. P., Landais, P., & Le Manach, Y. (2017). Simple randomization did not protect against bias in smaller trials. *Journal of clinical epidemiology*, *84*, 105–113. https://doi.org/10.1016/j.jclinepi.2017.02.010

Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W., Tschannen, B., Altman, D. G., Egger, M., & Jüni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ (Clinical research ed.)*, *341*, c3515. https://doi.org/10.1136/bmj.c3515

Pogue, J. M., & Yusuf, S. (1997). Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled clinical trials*, *18*(6), 580–666. https://doi.org/10.1016/s0197-2456(97)00051-2

Roberts, I., & Ker, K. (2015). How systematic reviews cause research waste. *Lancet (London, England)*, *386*(10003), 1536. https://doi.org/10.1016/S0140-6736(15)00489-4

Roberts, I., Ker, K., Edwards, P., Beecher, D., Manno, D., & Sydenham, E. (2015). The knowledge system underpinning healthcare is not fit for purpose

and must change. *BMJ (Clinical research ed.)*, *350*, h2463. https://doi.org/10.1136/bmj.h2463

Roose, S. P., Rutherford, B. R., Wall, M. M., & Thase, M. E. (2016). Practising evidence-based medicine in an era of high placebo response: number needed to treat reconsidered. *The British journal of psychiatry : the journal of mental science*, *208*, 416–420. https://doi.org/10.1192/bjp.bp.115.163261

Rose G. (1991). Environmental health: problems and prospects. *Journal of the Royal College of Physicians of London*, *25*(1), 48–52.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ (Clinical research ed.)*, *358*, j4008. https://doi.org/10.1136/bmj.j4008

Smith, S. M., Dworkin, R. H., Turk, D. C., McDermott, M. P., Eccleston, C., Farrar, J. T., Rowbotham, M. C., Bhagwagar, Z., Burke, L. B., Cowan, P., Ellenberg, S. S., Evans, S. R., Freeman, R. L., Garrison, L. P., Iyengar, S., Jadad, A., Jensen, M. P., Junor, R., Kamp, C., Katz, N. P., … Wilson, H. D. (2020). Interpretation of chronic pain clinical trial outcomes: IMMPACT recommended considerations. *Pain*, *161*, 2446–2461. https://doi.org/10.1097/j.pain.0000000000001952

Turner, R. M., Bird, S. M., & Higgins, J. P. (2013). The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PloS one*, *8*, e59202. https://doi.org/10.1371/journal.pone.0059202

Watson, C. P., Vernich, L., Chipman, M., & Reed, K. (1998). Nortriptyline versus amitriptyline in postherpetic neuralgia: a randomized trial. *Neurology*, *51*(4), 1166–1171. https://doi.org/10.1212/wnl.51.4.1166

Wiffen, P. J., Wee, B., Derry, S., Bell, R. F., & Moore, R. A. (2017). Opioids for cancer pain - an overview of Cochrane reviews. *The Cochrane database of systematic reviews*, *7*(7), CD012592. https://doi.org/10.1002/14651858.CD012592.pub2

Wiffen, P. J., Wee, B., & Moore, R. A. (2016). Oral morphine for cancer pain. *The Cochrane database of systematic reviews*, *4*(4), CD003868. https://doi.org/10.1002/14651858.CD003868.pub4

Williams, A., Fisher, E., Hearn, L., & Eccleston, C. (2020). Psychological therapies for the management of chronic pain (excluding headache) in adults. *The Cochrane database of systematic reviews*, *8*, CD007407. https://doi.org/10.1002/14651858.CD007407.pub4

Table 1: Outputs from two hypothetical trials

	Success (%) with:		Odds Ratio	Relative risk	Number needed to treat	Success rate
	Treatment	Placebo				
Treatment A	10	5	201	2.0	20	5%
Treatment B	50	25	2.9	2.0	4	30%

Table 2: Examples of numbers of participants in comparisons to be reasonably certain of the magnitude of the effect at a level of 90%.

Pain condition and therapy	Percent with at least 50% pain intensity reduction		NNT	Boundary 90% certainty that true value is ± this value	Total number of patients needed for 90% confidence
	Active	Placebo			
Acute postoperative pain Ibuprofen 400 mg	54	14	2.5	0.5	600
Painful diabetic neuropathy Duloxetine 60/120 mg	48	26	5	1	1000
Fibromyalgia Duloxetine 60/120 mg	28	17	10	5	2000

Table 3: Suggested items to consider when reading a systematic review of efficacy of interventions for pain

	AMSTAR-2 questions	Methodological issue addressed
1	Did the research questions and inclusion criteria for the review include the components of PICO?	Can readers identify patients, disorder, severity, intervention being used (including dose and timing), the comparator group, and the outcomes sought (including timing).
2	Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?	The establishment of a pre-study protocol is intended to avoid data mining and changes in the review process that might introduce a bias.
3	Did the review authors explain their selection of the study designs for inclusion in the review?	This is often avoided or assumed. But a statement about what study design is being sought, and why, and the benefits or difficulties of particular designs is helpful to readers.
4	Did the review authors use a comprehensive literature search strategy?	AMSTAR-2 requires ideally a search for unpublished trials. The actual benefits of this usually time-consuming procedure is debated, but there is limited evidence of much effect, and some evidence of no effect.
5	Did the review authors perform study selection in duplicate?	This guards against mistakes, and against selective data inclusion
6	Did the review authors perform data extraction in duplicate?	This guards against mistakes, and against selective data inclusion
7	Did the review authors provide a list of excluded studies and justify the exclusions?	Always useful, because it provides a background that speaks to the identified potential work and provides readers and researchers with an opportunity to disagree, for instance on inclusion and exclusion criteria.
8	Did the review authors describe the included studies in adequate detail?	Ideally a table providing a description of the main items found in the PICO
9	Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	Assessment of risk of bias is difficult. Most systems (such as Cochrane) are generic and based on (sometimes limited) evidence about the magnitude of bias from a particular risk in a particular situation. Particular circumstances require special attention to those risks of bias that are known to have major effects. In randomised trials for efficacy in pain, randomisation, blinding, imputation method, and small size are known to be associated with very major bias (again in some cases).
10	Did the review authors report on the sources of funding for the studies included in the review?	This is useful, as it can identify circumstances in which studies predominate from a single centre or source of funding.
11	If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	This is difficult to assess and may be situation dependent. The ideal would be to have both an absolute measure of benefit and harm (Risk difference, NNT, NNH to provide an indication of clinical importance, and a relative measure (risk ration, odds ratio, or standardised mean

		difference) to provide an indication of statistical significance.
12	If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	The ideal is for data to be assessed from high-quality studies without significant risk of bias. There are examples where meta-analysis of high-quality studies produces a different (often less effective) result than all studies combined, or lower quality studies. Many systematic reviews are conducted with data sets where most or all individual studies have one or more sources of potential high risk of bias.
13	Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	One of the questions to ask here is also whether the discussion of RoB is appropriate. It depends a lot on how thorough the search for RoB has been, and whether authors are aware of how RoB can impact on efficacy estimates, and assessment of GRADE. It is often inadequate because so many systematic reviews and meta-analyses depend on a few small trials.
14	Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	Heterogeneity is a tricky topic. CLINICAL heterogeneity is where different types of intervention, or its intensity, or types of patients or outcomes are combined. STATISTICAL heterogeneity occurs commonly with small studies because of random chance and is to be expected. Statistical heterogeneity found with large studies that are clinically homogeneous requires investigation.
15	If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	There is no adequate method of determining that publication bias has occurred. What can be done is to evaluate how much null effect data would be required to overturn a result (make it not statistically significant), or (probably better) to reduce the clinical effectiveness below a certain level. What constitutes the level chosen depends on the condition and its impact on people with pain.
16	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	What is being sought here is transparency. Funding or support does not infer bias in itself, but not declaring it would be worrying.
	Pain-specific questions	**Methodological issue addressed**
1	Were studies properly randomised?	Intended to be adjudicated with detailed instructions of the Oxford Quality Score (1996), developed for use with pain studies.
2	Were studies properly double-blind?	Intended to be adjudicated with detailed instructions of the Oxford Quality Score (1996), developed for use with pain studies.
3	Was the diagnostic condition defined?	It is known that the same interventions for pain used at the same intensity can produce different levels of response in different pain conditions. Proper definition and reporting of the pain condition is essential.
4	Was patient-reported pain only stated?	It is known that there are major discrepancies between pain scores elicited by people living with pain and by observers or carers - typically

		in the direction of lower pain scores by observers. The rule for many decades has been that patient reported pain is the rule in trials assessing efficacy.
5	If the primary outcome is pain relief, was a defined minimum pain intensity for study entry defined?	At least moderate pain is required to yield a sensitive assay. Most trials of drug efficacy require at least moderate pain (typically 40% or more of a maximum on any scale); average pain scores range from about 5 to about 8 or more. Many trials of non-pharmacological interventions include people with pain scores that are mild; average pain scores are often below 40% of a maximum, and this limits estimation of efficacy. Sensitivity analyses using only studies with high initial pain score are known to demonstrate different results from those with low pain scores.
6	Was there a sensitivity analysis for small study size?	The bulk of studies in many systematic reviews are small, with fewer than 50 patients in each treatment group. Small study size is often (but not always) associated with higher effect size than larger studies, and this is well documented in pain.
7	Was susceptibility to publication bias assessed?	Methods for calculating the potential amount of null-effect clinical data required to make an effect size clinically irrelevant are used in many pain meta-analyses, and are useful in estimation of GRADE of evidence.
8	Were missing data handled appropriately?	In pain efficacy trials, it is known that last observation carried forward can, notably when adverse event rates are high, produce large positive effects on effect size estimates. It has cogently been argued that LOCF should not be used, but where used, should be regarded as having a potentially high risk of bias.
9	If the primary outcome is other than pain (e.g., disability, return to work) was a clinically relevant minimum value for study entry defined	A clinically relevant status of the primary measure should be identified. Pain outcomes should be subject to a pooled analysis only if trials have appropriate methods for pain.

AMSTAR criteria highlighted in red and shaded indicate those considered to be critically important. Oxford Quality Score detailed instructions are found in Jadad et al., 1966.