Interpreting Effectiveness Evidence in Pain: Short Tour of Contemporary Issues

Neil E. O'Connell, G. Lorimer Moseley, James H. McAuley, Benedict M. Wand and Robert D. Herbert

PHYS THER. Published online April 30, 2015
doi: 10.2522/ptj.20140480

Running Head: Interpreting Evidence in Pain

ProfessionWatch

Interpreting Effectiveness Evidence in Pain: Short Tour of Contemporary Issues

Neil E. O'Connell, G. Lorimer Moseley, James H. McAuley, Benedict M. Wand, Robert D. Herbert

N.E O'Connell, PhD, Health Economics Research Group, Institute for the Environment Health and Societies, Department of Clinical Sciences, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH United Kingdom. Address all correspondence to Dr O'Connell at: neil.oconnell@brunel.ac.uk.

G.L. Moseley, PhD, Sansom Institute for Health Research, University of South Australia, Adelaide, Australia, and Neuroscience Research Australia, Randwick, Australia.

J.H. McAuley, PhD, Neuroscience Research Australia.

B.M. Wand, PhD, School of Physiotherapy, The University of Notre Dame Australia, Fremantle, Australia.

R.D. Herbert, PhD, Neuroscience Research Australia.

[O'Connell NE, Moseley GL, McAuley JH, et al. Interpreting effectiveness evidence in pain: short tour of contemporary issues. *Phys Ther*. 2015;95:xxx–xxx.]

Published Ahead of Print: xxxx

Accepted: April 19, 2015

Submitted: October 24, 2014

Abstract

There is no shortage of treatment approaches offered to sufferers of pain. The maze of options presents patients and clinicians with difficult choices. Key to making those choices is evidence of treatment effectiveness provided by clinical trials and systematic reviews. Recent growth in the number of clinical trials and systematic reviews, of both high and low quality, makes it vital that users of this evidence - clinicians, researchers, patients and policy makers - have the skills and knowledge to critically interpret these studies. In this review we discuss some contemporary issues regarding evidence of effectiveness derived from clinical trials and systematic reviews - issues that we think are critical to understanding the field. We focus on evidence of treatment effectiveness in pain, though many of these issues are relevant to and transferable across the spectrum of evidence-based practice.

Sufferers of pain and the clinicians who help them are faced with a maze of treatment options, each backed by enthusiastic and highly motivated advocates, all of whom lay claim to 'evidence'. Negotiating the treatment maze has never been more difficult. How can patients make an informed choice about their own care and how can clinicians best inform that choice?

Clinical trials remain the best tool for reducing uncertainty about the effects of treatment. The recent growth in the number of clinical trials and systematic reviews, of both high and low quality, makes it vital that clinicians, researchers, patients and policy makers have the skills and knowledge to critically interpret the available evidence. Here we discuss some contemporary issues regarding evidence of effectiveness from clinical trials and systematic reviews in pain - issues that we think are critical to understanding the field.

Clinical trials can be designed to test 'efficacy' (whether an intervention delivers an effect in ideal conditions) or 'effectiveness' (whether an intervention delivers an effect 'in the real world'). In reality, many trials test something that falls somewhere on a continuum between the two[1]. We focus on evidence of 'effectiveness' of treatments for pain, particularly chronic pain. We also examine evidence from the world of pharmacological interventions for pain, to consider what lessons there may be for interpreting non-pharmacological evidence. Many of these issues are also relevant to evidence of 'efficacy' of treatments for pain and are also transferable across the spectrum of evidence-based practice.

Beyond "p". The search for importance.

In effectiveness research, the p-value has long been a critical determinant of whether or not a treatment is thought to work. We contend that the p-value has been a significant barrier to efforts to establish which treatments are truly effective and therefore worthwhile. There has been an implicit acceptance among researchers, and users of evidence, that statistical

significance represents clinical effectiveness. Unfortunately this position suffers from substantial conceptual flaws[2,3]. It is perfectly possible for a treatment effect to be statistically significant and clinically meaningless. Conversely, though perhaps less commonly in the field of pain research, a treatment might provide important benefits to patients but be unjustly ignored because it does not cross this arbitrary statistical threshold.

Recently, over-reliance on the p-value to determine treatment effectiveness has come under further scrutiny[4]. While it is commonly held that a p-value of <0.05 suggests a type I error rate of less than 5%, the actual false discovery rate is dependent upon the prior probability of a treatment having an effect[4]. For example, if we assumed that just 10% of the various interventions that we test might provide a real treatment effect, an alpha-level of <0.05 would actually translate to a false discovery rate of 36% rather than the nominal 5% (for a review of this concept see[4]). This is particularly pertinent for chronic pain trials where we recruit participants who have proven refractory to interventions and hence the prior probability of intervention success is likely to be low. Aside from this problem, inappropriate, or perhaps innovative, statistical analyses can yield supportive looking p values[5].

When assessing the effectiveness of a treatment, the size, precision and subsequent clinical importance of the treatment's effects are of greater importance than whether the apparent effect could have occurred by chance. Most chronic pain patients want a cure for their pain[6] and treatments are routinely promoted and marketed as delivering large benefits quickly. Unfortunately the prospect of such an outcome remains very unlikely. So the question shifts to one of how much improvement would be needed to be meaningful to a patient? This 'minimal clinically important difference' (MCID)[7] or 'smallest worthwhile effect' (SWE)[8] should represent the minimum treatment effect with which patients would be satisfied. Recognising this metric is a step forward – it allows us to classify a treatment as imparting enough of an effect to be of value 'in the real world'.

What exactly constitutes an MCID, or a SWE, on any given outcome measure remains contentious, although various methodological approaches are being applied to the problem (for a review with regards to back pain, see[8]). Remarkably it seems that many of these approaches do not actively consider the patient perspective[8]. It is likely that what a patient would be satisfied with might differ substantially between individuals, patient groups, interventions[9], the point in the care pathway at which the patient arrives, and a range of other possible factors. In addition, people almost certainly have variable thresholds of what level of risk, inconvenience or cost, associated with the intervention they would consider to be prohibitive – an issue that, to our knowledge, has received very little attention. This variability, and the requirement that any potential benefit of an intervention must be weighed against its potential harms (including cost and inconvenience), suggests that the construct of a generic MCID for chronic pain interventions is problematic.

The Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) has offered some provisional benchmarks for important change in chronic pain. These benchmarks are based upon studies that compared pain scores with global impression of change in patients with neuropathic pain[10], arthropathies[11] and pain following spinal cord injury or amputation[12]. According to IMMPACT, a 30% reduction in pain in an individual patient represents the lower threshold for considering an effect to be 'moderately clinically important', and a 50% reduction represents a 'substantially clinically important' change[13]. There are obvious problems with applying cut-offs arbitrarily and across the board. For example, it does not seem reasonable to assume that we might require the same degree of change when agreeing to receive a short educational booklet as we might when agreeing to undergo an invasive surgical procedure. These cut-offs are also sensitive to baseline levels of symptoms. A 30% improvement in a severe intractable pain is probably quite a different proposition to a 30% change in a mildly bothersome ache or twinge.

Rather than focusing on the SWE a number of studies have used the Patient-Centred Outcomes Questionnaire (PCOS) to investigate the threshold of symptom improvement required for people with chronic pain to consider treatment successful[14-16]. These studies suggest that around 54-58% improvement in pain intensity, and 63-68% improvement in pain interference is required for treatment success. But, as with the SWE, we might expect judgements of success to be specific to the intervention and other contextual factors. Zeppieri et al.[16] investigated participants about to start physical therapy, whereas Robinson et al.[14] and O'Brien et al.[15] sampled patients from pain clinics and a Rheumatology department respectively, with no intervention specifically identified. So while these estimates suggest that large changes may be necessary it is not appropriate to assume that these estimates should apply across all interventions.

The elusive 'average' patient and the elusive 'responder'

The IMMPACT benchmarks and the thresholds derived from them reflect within-patient change from baseline. This is an appealing concept because it has a real-world resonance, being the amount of change experienced by an individual undergoing the intervention of interest. Unfortunately, within-patient change from baseline provides a poor measure of the effects of the intervention because it includes the influences of natural recovery, statistical regression, and the non-specific effects associated with clinical contact, including but not limited to 'placebo effects'[17] (although see[18] for an alternative understanding of placebo). Within-patient change in outcome might tell us how much an individual's condition improved but it does not tell us how much of this improvement is due to treatment. In most common randomised trial designs, the only value that can help us estimate the actual effect of the intervention is the average between-group difference after treatment[19]. It is only recently that this important principle has been applied to MCID or SWE research.

Ferreira et al.[9] used the benefit-harm trade off method to try and determine the SWE for physical therapy in people with chronic low back pain based on intervention-control between-group comparison, attempting to capture change due to treatment not simply change over time. Participants were informed that their pain and disability is likely to improve 30% without intervention and were then asked to estimate how much additional improvement would be needed to make the intervention worthwhile. The results of this study suggest that on average people with chronic low back pain would need to experience an additional 20% improvement in pain and disability compared to no treatment to perceive that the effect of physical therapy was worthwhile, that is an overall 50% change.

There are limitations inherent in interpreting the average effects of interventions in clinical trials. The question arises of who, if anyone, experiences the average treatment effect. It has been argued that, in the world of pharmaceutical trials for chronic pain, the response pattern is often bimodally distributed[20-22]. Simply put, some patients do very well with the intervention, some have minimal to no effect and very few experience intermediate (moderate) effects. In this instance, the average effect might be the effect that the fewest participants actually demonstrate[20]. The commonly proposed solution to this problem is to conduct a 'responder analysis', which compares the proportion achieving a clinically important improvement from baseline in the treatment and control groups. It has been proposed that this type of analysis better quantifies individual participant responses to treatment and enables the calculation of easily interpreted measures such as the number needed to treat (NNT). The NNT is the number of people we would need to treat with the intervention instead of the control condition for one more participant to achieve the outcome of interest (often a predefined MCID).

This approach too has important limitations. The term 'responder analysis' is a misnomer and is frequently misunderstood[23]. In this type of analysis, 'responders' are identified by within-

person change from baseline. For many participants in each group we are not really measuring treatment 'response', we are measuring 'good outcome', which, as mentioned above, might be due to natural recovery, non-specific treatment effects and regression to the mean as well as, or instead of, the effects of the intervention. Also, it is possible that some individuals who responded strongly to the intervention might not be counted as responders. If the natural history of a person during the treatment period would have been significant worsening, yet with treatment their condition remains stable, they will be counted as non-responders despite receiving significant benefit from the intervention. So, while the between-group difference in the proportion of participants who experience a good outcome reflects the net increase in the proportion of patients who 'responded' during the treatment period, it does not get any closer to telling us about the effects of intervention on individual people. Methods for distinguishing true 'responders' from those who just improved regardless of the treatment have their own substantial difficulties[24]. Responder analysis for a subjective outcome measured on a continuous scale (e.g. pain severity) may be sensitive to the cut-offs used to define clinical importance, and these cut-offs are often arbitrary. Moreover, because the outcome is measured imperfectly and 'responders' may be frequently misclassified, responder analyses might underestimate true effects[25]. This approach also potentially introduces the problem of only detecting positive change - not negative change. That is, all "non-responders" are considered equal. In reality, the response within this group might vary from mild improvement to severe deterioration. Finally, the dichotomisation of outcomes in responder analyses greatly reduces the precision of estimates of effect.

While the use of responder analysis is growing, currently such data remain scarce, particularly for non-pharmacological interventions[26]. A good case for responder analyses in rehabilitation trials has not been clearly established. The observation that patterns of outcome may be bimodal for some specific interventions is not evidence that they are necessarily

bimodal for others. More importantly, evidence of bimodal outcomes is not evidence of bimodal treatment effects. The belief that responder analysis will demonstrate treatment effects on individuals that are not apparent in other analyses may be unfounded. Data from drug trials in chronic pain, where such analyses are more common, rarely report NNTs below 6 [20.]

Do clinical trials underestimate effectiveness?

It is commonly argued that clinical trials are not fit for the purpose of evaluating physical therapies because they fail to capture the true effects of physical therapy treatments. Such arguments seem common at physical therapy conferences, particularly perhaps from those who find the disappointing results from clinical trials to be at odds with their clinical experience. The most common criticisms are that treatments are inadequately targeted in clinical trials because they are shoe-horned into a one-size-fits-all approach; therapies in clinical trials differ from real-world therapy which is complex, tailored and often multimodal; and the effects of treatment are diluted by the application of single interventions to a complex, heterogeneous group with diverse treatment needs. These criticisms are certainly justified in some, though not all trials. In the field of chronic pain there are additional difficulties presented in establishing meaningful diagnoses. Existing diagnostic labels (for example chronic non-specific low back pain, complex regional pain syndrome, fibromyalgia) often identify heterogeneous cohorts of people who share similar symptom profiles but not necessarily similar disease mechanisms.

The one-size-fits-all criticism is arguably an unfair characterisation of many modern therapy trials. Indeed, in recent years many if not most trials allow the therapist some discretion to tailor their approach to the individual, usually within a specific theoretical framework and often in a way that closely models existing clinical practices. For example in the

'manipulation' arm of the UK back pain exercise and manipulation trial[27], therapists were free to deliver a range of soft tissue, joint and neural manual therapy techniques. In addition the therapist could prescribe various exercises for the spine and hip, provide education on activity and return to work as well as address simple psychological issues[28]. In the recent PROMISE trial of exercise for chronic whiplash[29], therapists were able to tailor multi-modal exercise, manual therapy and cognitive behavioural techniques to the individual patient. And in the SWIFT trial, participants randomised to the physical therapy arm received a combination of individualized education/advice, exercise therapy, and manipulative therapy at the discretion of the treating physiotherapist based on usual practice[30].

There are not yet firmly established robust and widely accepted models for subgrouping patients with chronic pain to facilitate better targeting of treatment. Efforts at subgrouping have largely returned mixed outcomes[31]. Much of this work has focused on the treatment of low back pain, both acute and chronic, for which numerous approaches to subgrouping have been developed and tested. The picture that emerges is one in which positive trials (e.g [32,33]) tend to demonstrate small positive effects on primary outcomes, though these often fail to replicate[34-37] or are currently awaiting independent replication[38]. Subgrouping algorithms are frequently based on retrospective analysis of trial data rather than on prospective tests of predictions based on theoretical frameworks or biological mechanisms. Moreover, some subgroup analyses have been shown to be dependent on the cut-points used to determine MCID[39] and many subgroup analyses conducted within trials have been severely underpowered and poorly reported[40]. Better tailoring or subgrouping of cohorts to treatments may still improve outcomes, but so far the promise of subgrouping remains largely unfulfilled.

A further assertion is that the true effects of an intervention are lost in the cacophony of competing real-world variables, including social and psychological factors, competing therapies, adherence, participation and compliance. This assertion maintains that the signal of effective treatment cannot always be detected in the presence of noise. Again, there may be some truth in this, but the best way around it is to conduct large trials which can provide precise estimates of average treatment effects. Therein lies the challenge facing all health interventions: to demonstrate clear benefit in the chaos of the real world. The 'noise' may be particularly loud in chronic pain, but we should recognise that in both clinical practice and research, interventions cannot be provided in the clinical equivalent of a soundproofed room.

Recently Morley[41] has argued, specifically in the case of cognitive behavioural therapy interventions, for greater integration of 'practice-based evidence' in which data generated from routine clinical practice is afforded greater importance. In this approach clinical outcome data are compared to 'benchmark' effect sizes generated from the treatment and control arms of clinical trials. This allows a degree of control over the effects of natural history and non-specific effects of treatment, though it does not offer the high level of control offered by randomisation. One possible risk associated with this approach is that where effect sizes are sufficiently low, it may encourage the celebration of possibly dubious successes. As such it seems best suited to demonstrating 'proof of concept' of new hypotheses regarding treatment innovation, for subsequent testing in RCTs.

Exaggeration, misreporting and spin.

It is also worth considering the alternative possibility, that clinical trials might generate exaggerated estimates of effectiveness. In the context of clinical trials for physical interventions, treatments are often provided by more experienced clinicians, patients are given more time and greater steps are taken to ensure treatment compliance than would be the

case in routine clinical practice, potentially offering a more effective package of care than might be realised in routine clinical practice. More importantly, the observed effectiveness of a treatment is represented by the true effect of treatment plus the effect of biases that can also positively influence outcome. These biases are often sub-optimally controlled in clinical trials so the observed effect represents both the effectiveness of the intervention and bias. Many readers will be familiar with the conventional risk of bias criteria by which trials are assessed in systematic reviews. Meta-epidemiological evidence shows that these criteria are associated with treatment effect sizes, particularly for subjective outcome measures such as pain[42,43]. Blinding of patients and care providers is often not achieved in trials of physical or psychological treatments[44], and it is notable that trials of physical therapies commonly fall short on a number of other criteria. Though quality is improving[45] it is likely that the effect sizes reported by most clinical trials represent more than just the effects of treatment on patient outcomes.

In clinical trials size matters - small studies increase the risk of false negatives by virtue of their low statistical power, but in clinical trials they tend to also result in false positives and inflated effect sizes[46-48]. There are a number of possible reasons for this phenomenon: small studies may include more homogenous clinical groups for which effects are more consistent[49], and it is easier to deliver high quality interventions in smaller trials[49]. Small trials are also often more loosely controlled and of lower methodological quality. Negative small studies have a tendency not to be published, rendering the available sample of published small trials unrepresentatively positive. Large and significant effects arising from small underpowered studies are at higher risk of being false positives than if they arose from large, well-powered studies[50]. The benefits of meta-analysis do little to correct this problem - even where a pooled estimate includes a large number of participants it may be prone to small study bias if it is dominated by small studies.

Managing loss to follow-up of participants and protocol violation during trials is difficult. Traditionally we look for an intention to treat analysis, in which all participants are analysed by the treatment to which they were allocated, regardless of what follows that allocation. Currently the application and reporting of intention to treat analyses in analgesic trials is inconsistent[51], reflecting a common risk of bias in this field. Common methods for dealing with missing data themselves introduce bias. Evidence from drug trials in chronic pain suggests that the commonly used 'last observation carried forward' approach to imputing data inflates effect sizes[52]. This is often an issue with adverse event withdrawal, where the last observation precedes the adverse event, but might also hold true for other reasons for withdrawal - withdrawal may be associated with worsening symptoms or a realisation that the treatment is not really helping, both of which may occur after the last formal observation. New methods for analysing clinical trials, particularly multiple imputation, may improve estimates of effect in the presence of substantial loss to follow-up[53] though further data is needed to formally evaluate this perspective.

Beyond these threats related to methodology are challenges to the balanced conduct and communication of trials. Selective outcome reporting is considered as a risk of bias on many assessment tools and involves the selective presentation of results that are more positive or statistically significant, and the withholding on negative or non-significant results[54]. This can be achieved through poor practices such as deviating from the trial protocol by switching the primary outcomes in light of the trial results. A recent review of analgesic trials[55] compared records in international trials registers with the final published study reports and found discrepancies between the primary outcomes in 79% of the available data with 30% of trials containing what were defined as "uinambiguous" discrepancies, where a registered primary outcome was either not reported in the published trials or was demoted to a secondary outcome. A similar review of acupuncture trials[56] found inconsistency in the primary

outcomes in 45% of available trials, of which 71% had a discrepancy that favoured a statistically significant "positive" result on the primary outcome.

There is also evidence of a strong positivity bias in the interpretation and presentation of results from clinical trials. Boutron et al.[57] found evidence of 'spin' – presenting an experimental treatment as beneficial - in 40% of statistically negative trials. In rheumatology trials, Mathieu et al.[58] found that 23% of trials had conclusions that were misleading, and that the only predictor of misleading conclusions was a statistically negative result. This pattern is also apparent in the analgesic trial literature. Worryingly, some type of positive spin was identified in at least one part of the abstract of 61% of analgesic trials with statistically non-significant results in their primary analysis; most commonly the placing of undue emphasis on statistically significant results from secondary analyses[59]. It seems that beyond the difficulty of getting negative results published, researchers do not like to accept negative results in the first place. Perhaps this is partially motivated by the 'publish or perish' culture of modern research. Notwithstanding that, it clearly represents a failure of the scientific process in which there is a bias towards one possible answer to the research question. For consumers of research papers, the message is that simply looking to the abstract or conclusions of a trial for the truth carries risk – an issue we have touched on before[60].

Pursue success, expect failure?

Looking across the Cochrane library at reviews of common interventions for chronic pain, and being somewhat selective by avoiding interventions where the evidence suggests no effect at all, reveals that most 'effective' therapies appear to provide only very small, short-term effects on pain or other important patient-centred outcomes such as function, distress and quality of life. We must bear in mind that, particularly for complex interventions, the meta-analyses that produce these estimates contain multiple sources of clinical heterogeneity

that have the potential to impact on effect size; they combine interventions that are often quite different in terms of content and dose[61]; the quality of the intervention is often hard to determine, though of great potential importance[62] ; the theories underpinning the interventions often vary significantly between studies or are not clearly established[63] ; the contextual equivalence of the control group interventions is variable[64] ; adherence levels vary and patients are drawn from diverse sources. All that said, we suggest that, when we do not currently have robust means of identifying a priori those who might respond to treatment, it is the average between-group effects that represent our best estimate of the intervention-specific benefit for any individual.

For drug therapies, treatment response, when it comes, is usually rapid. Moore and colleagues[20] recommend that when we introduce a new therapy we should expect failure, be alert to a lack of treatment response, and switch quickly to another agent if outcomes are poor. Such an approach might maximise the chance of finding an effective option as quickly as possible while minimising the risks of adverse events from drugs that confer no individual benefit, though it makes the potentially tenuous assumption that, without intervention, the patient's symptoms would not have changed substantially.

Could this approach be applied to non-pharmacological interventions? We think so, but to avoid pushing patients through a mill of ineffective therapies, we also think that we should limit the potential options to interventions that possess at least biological plausibility (a foundation stone that can be difficult to find in our field) and rigorous evidence of effectiveness.[65]

Reviewing this evidence can leave one with a somewhat negative impression. We acknowledge that there is a danger here - that such focused attention on rigour and bias can appear hypercritical and unduly negative, and take away whatever desire clinicians and

patients may have had to negotiate the evidence maze. That, however, is our challenge: to be dispassionate, recognise bias and make balanced appraisals of the strength and direction of the evidence, and that must in the end be a positive step. In the words of the physicist Richard Feynman, "For a successful technology, reality must take precedence over public relations, for Nature cannot be fooled"[66].

These issues are not unique to the field of chronic pain research - many of them apply across the range of clinical disciplines. This is important because there are examples from other clinical fields of the development and validation of clearly successful interventions, interrogated by high quality clinical trials and systematic reviews. Such compelling evidence of effectiveness from other, comparably complex fields, offers genuine hope for our own field. For example, we can now be confident that stress urinary continence can be prevented and treated with pelvic floor muscle training[67] and the risk of falling in the elderly can be decreased with exercise programmes[68].

We should reflect that, in chronic pain treatment and research, we all have some sort of vested interest[69]. If we offer assessments of the evidence for our treatments without due diligence regarding bias and limitations we will not serve our patients well. Our patients may be given choice but not the choice they need. By its very nature, clinical research *should* threaten current practice. Acknowledging what does not work, as well as what does (and by how much), is of great value and will force us to innovate. In fair tests[70] if our treatments achieve their goals meaningfully and consistently, then the effect sizes will reflect that truth. An appreciation of how to interpret evidence of effectiveness is a critical skill not just for those engaged with research, but in those who wish to use it in clinical practice.

Altering the natural course of any clinical condition is a difficult and complex challenge. In the words of the epidemiologist Archie Cochrane, after whom the Cochrane Collaboration is

named: "…. one should….be delightfully surprised when any treatment at all is effective, and always assume that a treatment is ineffective unless there is evidence to the contrary"[71]. This has genuine resonance in chronic pain, in which we set ourselves the substantial challenge of changing symptoms in a group defined by the fact that those symptoms have so far proven unchangeable. We suggest that we should always have that perspective in mind, whilst remaining at the ready to be delightfully surprised.

Acknowledgements

Dr O'Connell, Dr Moseley, Dr McAuley, and Dr Wand provided concept/idea/project design. All authors provided writing.

Dr Moseley and Dr Herbert are supported by the National Health and Medical Research Council of Australia.

DOI: 10.2522/ptj.20140480

References

1. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. Criteria for distinguishing effectiveness From efficacy Trials in Systematic Reviews. Agency for Healthcare Research and Quality (US) 2006; Report No: 06-0046
2. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986; 292:746-750.
3. Kline RB (2004) Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research. Washington: American Psychological Association.
4. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p values. *Royal Soc Open Sci* 2014; 1: 140216
5. Simonsohn U, Nelson LD, Simmons JP. P-Curve: A Key to the File-Drawer. *J Exp Psych* 2014; 143:2:534–547
6. Hush, JM, Refshauge K, Sullivan G, DeSouza L, Maher CG, McAuley JH. Recovery: what does this mean to patients with low back pain? *Arthritis Rheum* 2009; 61:124–131.
7. Jaeschke R, Singer J, Guyatt GH Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; 10: 407-415.
8. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *J Clin Epidemiol* 2012;65(3):253-61.
9. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Grotle M, et al. The smallest worthwhile effect of nonsteroidal anti-inflammatory drugs and physiotherapy for chronic low back pain: a benefit-harm trade-off study. *J Clin Epidemiol* 2013; 66(12):1397-404.

10. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001; 94:149-158.
11. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *Eur J Pain* 2004; 283-291.
12. Hanley MA, Jensen MP, Ehde DM, Robinson LR, Cardenas DD, Turner JA, Smith DG: Clinically significant changes in pain intensity ratings in persons with spinal cord injury or amputation. *Clin J Pain* 2006; 22:25-31.
13. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain* 2008; 9(2):105–21.
14. Robinson ME, Brown JL, George SZ, Edwards PS, Atchison JW, Hirsh AT, Waxenberg LB, Wittmer V, Fillingim RB. Multidimensional success criteria and expectations for treatment of chronic pain: the patient perspective. *Pain Med* 2005; 6(5):336-45.
15. O'Brien EM, Staud RM, Hassinger AD, McCulloch RC, Craggs JG, Atchison JW, Price DD, Robinson ME. Patient-centered perspective on treatment outcomes in chronic pain. *Pain Med* 2010; 11(1):6-15
16. Zeppieri G, Lentz TA, Atchison JW, Indelicato PA, Moser MW, Vincent KR, George SZ. Preliminary results of patient-defined success criteria for individuals with musculoskeletal pain in outpatient physical therapy settings. *Arch Phys Med Rehabil* 2012; 93(3):434-40
17. Herbert R, Jamtvedt G, Mead J, Hagen KB. Outcome measures measure outcomes, not effects of intervention. *Aus J Physiother* 2005;51: 3-4

18. Moseley GL. Placebo effect: Reconceptualising placebo. *BMJ* 2008; 336(7653):1086.
19. Rubin DB Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psych* 1974; 66: 688-701
20. Moore A, Derry S, Eccleston C, Kalso E. Expect analgesic failure; pursue analgesic success. *BMJ* 2013;346(7911):f2690
21. Moore RA, Derry S, Simon LS, Emery P. Nonsteroidal anti-inflammatory drugs, gastroprotection, and benefit-risk. *Pain Pract* 2014;14:4:378-95
22. Moore RA, Cai N, Skljarevski V, Tölle TR. Duloxetine use in chronic painful conditions--individual patient data responder analysis. *Eur J Pain* 2014;18:1:67-75
23. Senn S. Individual therapy: new dawn or false dawn. *Drug Inf J* 2001; 35:1479-1494
24. Jafari N, Hearne J, Churilov L Why caution is recommended with post-hoc individual patient matching for estimation of treatment effect in parallel-group randomized controlled trials: The case of acute stroke trials. *Stat Med* 2013; 32:4467-4481.
25. Marschner IC, Emberson J, Irwig L, Walter SD. The number needed to treat (NNT) can be adjusted for bias when the outcome is measured with error. *J Clin Epidemiol* 2004; 57: 1244-1252.
26. Henschke N, Van Enst A, Froud R, Ostelo R. Responder analyses in randomised controlled trials for chronic low back pain: an overview of currently used methods. *Eur Spine J* 2014; 23(4):772-8.
27. UK BEAM Trial Team. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. BMJ 2004; 329(7479):1377.
28. Harvey E, Burton AK, Moffett JK, Breen A; UK BEAM trial team. Spinal manipulation for low-back pain: a treatment package agreed to by the UK

chiropractic, osteopathy and physiotherapy professional associations. *Man Ther* 2003; 8(1):46-51

29. Michaleff ZA, Maher CG, Lin CW, Rebbeck T, Jull G, Latimer J, Connelly L, Sterling M. Comprehensive physiotherapy exercise programme or advice for chronic whiplash (PROMISE): a pragmatic randomised controlled trial. *Lancet* 2014; 384(9938):133-41
30. Hurley DA, Tully MA, Lonsdale C, Boreham CA, van Mechelen W, Daly L, Tynan A, McDonough SM. Supervised walking in comparison with fitness training for chronic back pain in physiotherapy: results of the SWIFT single-blinded randomized controlled trial (ISRCTN17592092). *Pain* 2015; 156(1):131-47
31. Kamper SJ, Maher CG, Hancock MJ, Koes BW, Croft PR, Hay E. Treatment-based subgroups of low back pain: a guide to appraisal of research studies and a summary of current evidence. *Best Pract Res Clin Rheumatol* 2010;24(2):181–91.
32. Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, Konstantinou K, Main CJ, Mason E, Somerville S, Sowden G, Vohora K, Hay EM (2011) Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet* 2011; 378:9802:1560-71.
33. Fritz JM, Delitto A , Erhard RE . Comparison of classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain: a randomized clinical trial *Spine* 2003 ; 28 : 1363–7
34. Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. *Eur Spine J* 2008; 17(7):936-43.

35. Apeldoorn AT, Ostelo RW, van Helvoirt H, Fritz JM, Knol DL, van Tulder MW, de Vet HC. A randomized controlled trial on the effectiveness of a classification-based system for subacute and chronic low back pain. *Spine* 2012; 37(16):1347-56

36. Henry SM, Van Dillen LR, Ouellette-Morton RH, Hitt JR, Lomond KV, DeSarno MJ, Bunn JY. Outcomes are not different for patient-matched versus nonmatched treatment in subjects with chronic recurrent low back pain: a randomized clinical trial. *Spine J* 2014;14(12):2799-810.

37. Dougherty PE, Karuza J, Savino D, Katz P. Evaluation of a modified clinical prediction rule for use with spinal manipulative therapy in patients with chronic low back pain: a randomized clinical trial. *Chiropr Man Therap* 2014; 22(1):41

38. Vibe Fersum K, O'Sullivan P, Skouen JS, Smith A, Kvåle A.Efficacy of classification-based cognitive functional therapy in patients with non-specific chronic low back pain: a randomized controlled trial. *Eur J Pain* 2013; 17:6:916-28.

39. Schwind J, Learman K, O'Halloran B, Showalter C, Cook C. Different minimally important clinical difference (MCID) scores lead to different clinical prediction rules for the Oswestry disability index for the same sample of patients. *J Man Manip Ther* 2013; 21(2):71-8.

40. Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;39(7):618-29.

41. Morley S. Efficacy and effectiveness of cognitive behaviour therapy for chronic pain: Progress and some challenges. *Pain* 2011;152 (3 Suppl):S99-106.

42. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different

interventions and outcomes: Meta-epidemiological study. *BMJ* 2008;336(7644):601–5.

43. Savović J, Jones H, Altman D, Harris R, Jűni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess* 2012;16(35):1-82.
44. Mathieu E, Herbert RD, McGeechan K, Herbert JJ, Barratt A. A theoretical analysis showed that blinding cannot eliminate potential for bias associated with beliefs about allocation in randomised clinical trials. *J Clin Epidemiol* 2014; 67: 667-671.
45. Moseley AM, Elkins MR, Janer-Duncan L, Hush JM. The quality of reports of randomized controlled trials varies between subdisciplines of physiotherapy. *Physiother Can* 2014; 66: 1: 36-43.
46. Moore RA, Gavaghan D, Trame MR, Collins SL, Mcquay HJ. Size is everything – large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain* 1998;78:209–16.
47. Nuesch E, Trelle S, Reichenbach S, Rutjes a. WS, Tschannen B, Altman DG, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ*;341:c3515–c3515.
48. Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. *BMJ* 201;346:f2304.
49. Herbert RD, Bø K Analysis of quality of interventions in systematic reviews. *BMJ* 2005; 331: 507-509
50. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews. Neuroscience 2013; 14(5):365–76.

51. Gewandter JS, McDermott MP, McKeown A, Smith SM, Pawlowski JR, Poli JJ, Rothstein D, Williams MR, Bujanover S, Farrar JT, Gilron I, Katz NP, Rowbotham MC, Turk DC, Dworkin RH. Reporting of intention-to-treat analyses in recent analgesic clinical trials: ACTTION systematic review and recommendations. Pain 2014; 155(12):2714-9.

52. Moore RA, Straube S, Eccleston C, Derry S, Aldington D, Wiffen P, et al. Estimate at your peril: Imputation methods for patient withdrawal can bias efficacy outcomes in chronic pain trials using responder analyses. *Pain* 2012;153(2):265-8.

53. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; 338: b2393.

54. Higgins JPT, Altman DG, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

55. Smith SM, Wang AT, Pereira A, Chang RD, McKeown A, Greene K, Rowbotham MC, Burke LB, Coplan P, Gilron I, Hertz SH, Katz NP, Lin AH, McDermott MP, Papadopoulos EJ, Rappaport BA, Sweeney M, Turk DC, Dworkin RH. Discrepancies between registered and published primary outcome specifications in analgesic trials: ACTTION systematic review and recommendations. Pain 2013; 154:12:2769-74.

56. Su CX, Han M, Ren J, Li WY, Yue SJ, Hao YF, Liu JP. Empirical evidence for outcome reporting bias in randomized clinical trials of acupuncture: comparison of registered records and subsequent publications. Trials. 2015 Jan 27;16(1):28

57. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303(20):2058–64.

58. Mathieu S, Giraudeau B, Soubrier M, Ravaud P. Misleading abstract conclusions in randomized controlled trials in rheumatology: comparison of the abstract conclusions and the results section. *Joint Bone Spine* 2012;79(3):262–7

59. Gewandter JS, McKeown A, McDermott MP, Dworkin JD, Smith SM, Gross RA, Hunsinger M, Lin AH, Rappaport BA, Rice AS, Rowbotham MC, Williams MR, Turk DC, Dworkin RH.Data interpretation in analgesic clinical trials with statistically nonsignificant primary analyses: an ACTTION systematic review. *J Pain* 2015; 16(1):3-10.

60. Harvie D, O'Connell N, Moseley L. Dry needling for myofascial pain. Does the evidence make the grade? 2014;

61. Waterschoot FP, Dijkstra PU, Hollak N, de Vries HJ, Geertzen JH, Reneman MF. Dose or content? Effectiveness of pain rehabilitation programs for patients with chronic low back pain: a systematic review. Pain 2014; 155(1):179-89.

62. Herbert RD1, Bø K. Analysis of quality of interventions in systematic reviews. BMJ 2005; 331(7515):507-9.

63. Williams AC, Eccleston C, Morley S. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2012;11: CD007407.

64. Bishop FL, Fenge-Davies AL, Kirby S, Geraghty AW. Context effects and behaviour change techniques in randomised trials: a systematic review using the example of trials to increase adherence to physical activity in musculoskeletal pain. Psychol Health 2015; 30(1):104-21.

65. Bø K, Herbert R. When and how should new therapies become routine clinical practice? Physiotherapy 2009; 95: 51-57.

66. Feynman R. "Appendix F - Personal observations on the reliability of the Shuttle". *Kennedy Space Center* 1986

67. Dumoulin C, Hay-Smith EJC, Mac Habée-Séguin G. Pelvic floor muscle training versus no treatment, or inactive control treatments, for urinary incontinence in women. *Cochrane Database Syst Re*v 2014;5:CD005654

68. Gillespie LD, Robertson MC, Gillespie WJ, Sherrington C, Gates S, Clemson LM, Lamb SE. Interventions for preventing falls in older people living in the community. *Cochrane Database Syst Rev* 2012; 9: CD007146

69. Moseley L. Finding the love between scientists and clinicians – a response to Dr Butler on noijam. 2013; http://www.bodyinmind.org/finding-the-love-between-scientists-and-clinicians-a-response-to-dr-butler-on-noijam/

70. Evans I, Thornton H, Chalmers I, Glasziou P (2011) Testing Treatments (2nd edn). London: Pinter & Martin. Available open access at http://www.testingtreatments.org/

71. Cochrane AL. Effectiveness and Efficiency. Random Reflections on Health Services. London: Nuffield Provincial Hospitals Trust, 1972. (Reprinted in 1989 in association with the BMJ, Reprinted in 1999 for Nuffield Trust by the Royal Society of Medicine Press, London (ISBN 1-85315-394-X). Available here: http://www.nuffieldtrust.org.uk/publications/effectiveness-and-efficiency-random-reflections-health-services

Physical Therapy
Journal of the American Physical Therapy Association

Interpreting Effectiveness Evidence in Pain: Short Tour of Contemporary Issues

Neil E. O'Connell, G. Lorimer Moseley, James H. McAuley, Benedict M. Wand and Robert D. Herbert

PHYS THER. Published online April 30, 2015
doi: 10.2522/ptj.20140480