Interventional Pain Medicine 1 (2022) 100123

Contents lists available at ScienceDirect

Interventional Pain Medicine

journal homepage: www.journals.elsevier.com/interventional-pain-medicine

Calibrating effect-size for studies of pain treatment

Nikolai Bogduk

The University of Newcastle, Newcastle, PO Box 431, East Maitland, NSW, 2323, Australia

ARTICLE INFO

Keywords:
Effect size
Effectiveness
Pain treatment

ABSTRACT

Effect-size is a statistic often used in studies of pain treatment. Its attraction is that numerical values can be translated into adjectives, such as small, medium, and large, that describe the change in scores for outcome measures. However, these adjectives can be misused to imply that the effectiveness of the treatment is the equivalent of small, medium, or large. Consideration of examples reveals the fallacy of this practice. Treatments with little to no effectiveness can produce effect-sizes described as medium or large. Reciprocally, treatments with good but imperfect effectiveness generate effect-sizes described as very large or even huge. Effect-size was developed specifically to describe the magnitude of statistical changes, but has little proportional bearing on effectiveness of treatment. When evaluating treatments, readers should not be swayed by descriptors of effect-sizes. Instead, they should consult categorical data on success rates, upon which they can base their decision as to how well the treatment works.

1. Introduction

Effect-size is a statistical expression that can be used to grade changes in scores from before to after an intervention. Although there are other formulations the one most commonly used in the pain literature is Cohen's d [1] or some variant of it. Specifically, the effect-size (d) is the dividend between the difference in mean scores and their pooled standard deviation, i.e.

$$d = \frac{\mu_0 - \mu_1}{\sqrt{\frac{(n_0-1)s_0^2+(n_1-1)s_1^2}{n_0+n_1-2}}}$$

where μ_0 and μ_1 are the initial and subsequent mean scores, s_0 and s_1 are their respective standard deviations, and n_0 and n_1 are the respective sample sizes.

One of the attractions of using this effect-size is that it comes with verbal translations, by which a seemingly complex statistic can be converted to a meaningful adjective (Table 1). Thus, for example, when a study generates an effect-size of 0.80, the authors can report that the effect-size is large.

In the pain literature, these descriptors have served authors well. The effect-sizes encountered in studies of pain treatment have ranged between 0.0 and 1.0, whereupon authors can describe them as small, medium or large. However, the risk arises that authors might imply, or readers might infer, that the grade of the effect-size reflects the effectiveness of the treatment. Thus, a treatment with a "large" effect-size might be construed as having "large" effectiveness. In some other contexts this inference might be legitimate, but as we shall see, it is not legitimate for pain treatment.

What is not revealed in the pain literature is that effect-sizes are not restricted to a range between 0.0 and 1.0 and, therefore, their translations are not limited to small, medium, and large. There is an extended scale (Table 2) that encompasses "huge" effect-sizes.

In order that readers of pain studies accurately understand the meaning of effect-sizes, it is worthwhile to calibrate effect-sizes in the context of pain treatment. This essay provides that calibration, and offers a guideline to the interpretation of studies that report effect-sizes in pain treatment.

2. Calibration

Consider the data in Table 3. They show a significant improvement in pain scores after an intervention, with a resultant effect-size of 0.50. That effect-size rates as "medium". The question that arises is whether or not this descriptor indicates a treatment with "medium" effectiveness, or some equivalent to that. That question is answered by Fig. 1.

Inspection of Fig. 1 reveals that there has been an obvious improvement of the pain scores, but the shift is small. Only a handful of patients have reduced their scores to 4, but the remainder occupy the same distribution as before treatment. It would be generous to describe the effectiveness of this treatment as "good" or even "medium". It might be "good" that there has been an improvement of scores, but for clinical

E-mail address: nbogduk@bigpond.net.au.

https://doi.org/10.1016/j.inpm.2022.100123
Received 13 June 2022; Accepted 16 June 2022

Table 1
A list of the descriptors that apply to effect-sizes of various magnitude [1].

EFFECT-SIZE	DESCRIPTOR
0.8	Large
0.5	Medium
0.2	Small

Table 2
An extended list of the descriptors that apply to effect-sizes of various magnitude [2].

EFFECT-SIZE	DESCRIPTOR
2.0	Huge
1.2	Very Large
0.8	Large
0.5	Medium
0.2	Small

Table 3
The mean pain scores, and standard deviations, before and after an intervention, and the result effect-size. The p-value applies to a two-sample *t*-test.

	Pain Score		p-value	Effect-size
	mean	sd		
Before	7.1	1.3	0.02	0.50
After	6.4	1.5		

purposes the improvement has been little to none. Most patients would still remain eligible for enrolment in a study of another treatment.

Table 4 shows the summary data of another example. The improvement in pain scores is greater than it was in Table 1. The effect-size is 0.88, which is rated as “large”. The temptation to rate the intervention as having a “good” effect is challenged by Fig. 2.

Fig. 2 shows an obvious, significant improvement in pain scores, but the improvements are little different from those seen in Fig. 1. A handful of patients have reduced their scores to 4, and are joined by a few whose scores have become 5 and 6. Nonetheless, the majority of patients still have scores that remain in the range of scores before intervention. The latter defies rating the effectiveness of this intervention as “good”. The treatment works, but does not have a powerful effect. Indeed, the difference between the mean scores, before and after treatment, is substantially less than minimal clinically important change for all forms of spinal pain.

Table 4
The mean pain scores, and standard deviations, before and after an intervention, and the resultant effect-size. The p-value applies to a two-sample *t*-test.

	Pain Score		p-value	Effect-size
	mean	sd		
Before	7.1	1.3	0.00	0.88
After	6.0	1.2		

Fig. 3 illustrates the outcomes of a more effective intervention. Those outcomes are representative of what might be expected from a treatment such as transforaminal injection of steroids for lumbar radicular pain. Not all patients benefit, but some 54% do [3]. Of those, some get 50% relief of pain or greater, and others get complete relief. If we use these data as a reference point for a treatment whose effectiveness might be considered “good”, we can determine what the effect-size should be for a “good” treatment.

Table 5 shows the summary data of Fig. 3. The pain scores have improved significantly, but in this instance the effect-size is 1.2. This amounts to a very large effect-size (Table 2). Yet the success rate of lumbar transforaminal injection of steroids (56%) is not very large, and would be regarded by some as only moderate. In essence, the descriptor from the effect-size rather overstates how effective the treatment is.

Fig. 4 shows yet another example. It depicts the pattern of outcomes such as those that might be expected of cervical medial branch radiofrequency neurotomy, in which up to 70% of patients obtain complete relief of neck pain [4].

Table 6 shows the summary statistics for the data plotted in Fig. 4. Mean pain scores have reduced by more than half, and the effect-size is 2.1. This amounts to an effect-size that is beyond “huge” (Table 2). So, were we to adopt the language of effect-sizes, cervical radiofrequency neurotomy would be regarded as hugely successful. Yet it is not; some 30% of patients fail to benefit.

These examples show that, in studies of pain treatment, medium or large effect-sizes reflect only minimal or small changes in pain, and

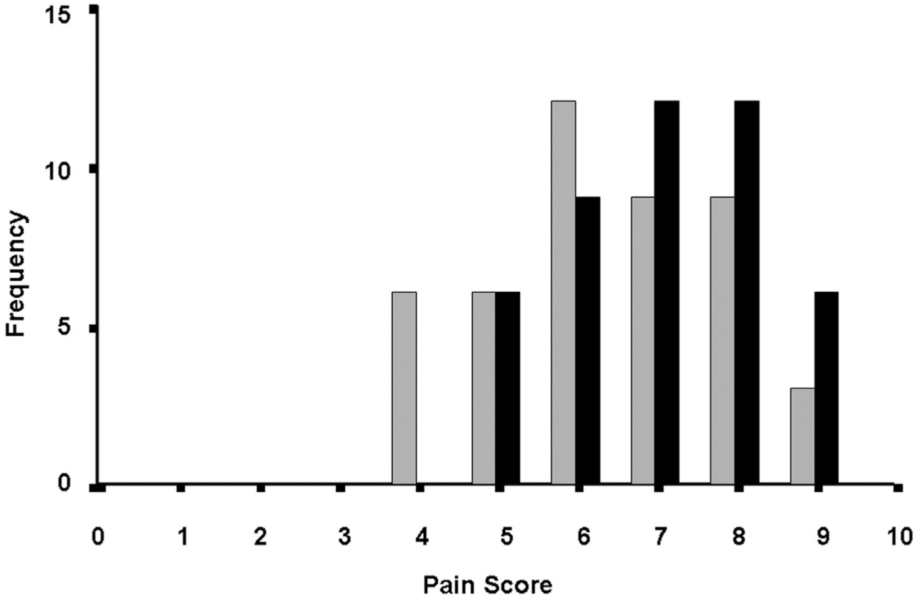

Fig. 1. The raw data summarised in Table 3 plotted in a histogram. Black bars are the scores before intervention. Grey bars are the scores after intervention. The effect-size of the change in scores is 0.50, which is rated as “medium”.

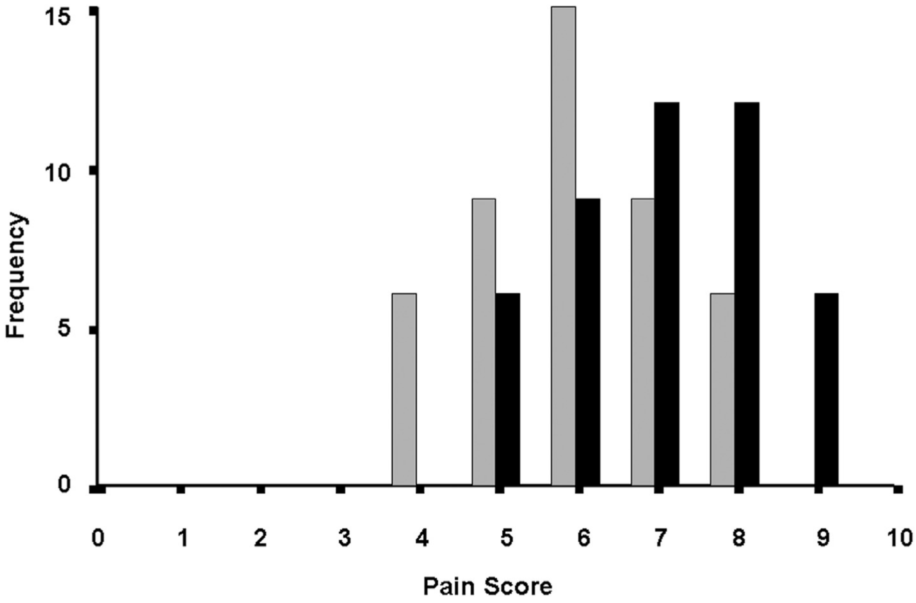

Fig. 2. The raw data summarised in Table 4 plotted in a histogram. Black bars are the scores before intervention. Grey bars are the scores after intervention. The effect-size of the change in scores is 0.88, which is rated as "large".

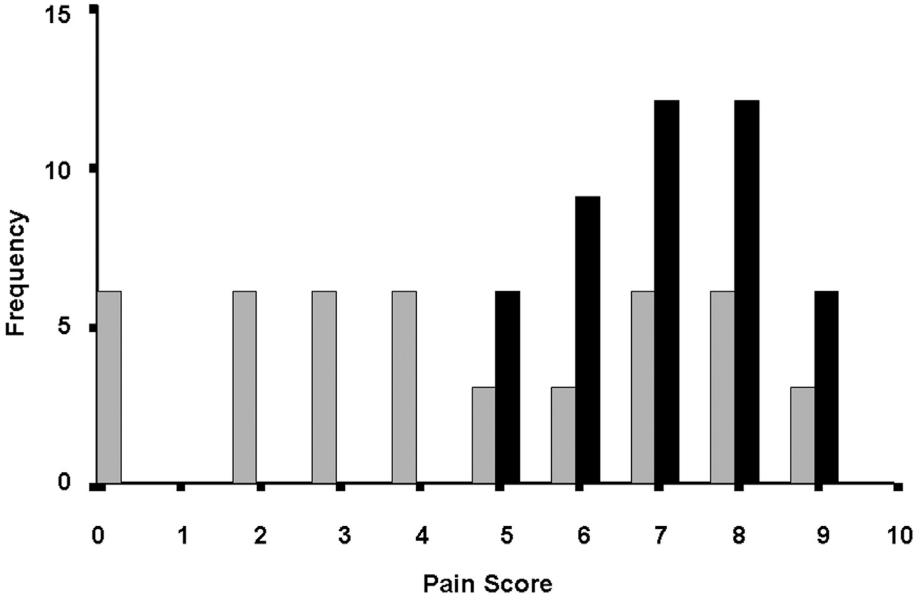

Fig. 3. The distribution of pain scores before and after an intervention after which over 50% of patients reduce their pain by 50% relief of pain or completely. Black bars are the scores before intervention. Grey bars are the scores after intervention.

Table 5
The mean pain scores and standard deviations, and the resultant effect-size, of the data illustrated in Fig. 3. The p-value applies to a two-sample *t*-test.

	Pain Score		p-value	Effect-size
	mean	sd		
Before	7.1	1.3	0.00	1.2
After	4.5	2.8		

indicate only a minimally effective treatment. Conversely, interventions that produce good clinical outcomes, but in only 50%–70% of patients, produce effect-sizes that are rated very large, huge, or beyond huge.

This dissonance arises because the descriptors for effect-sizes were never intended to apply to the intervention; they apply only to the behaviour of the numbers in the data. An effect-size is large if and because the change in scores is visibly obvious, but this does not require the change to be large. Small changes can create a large effect-size if they are consistent across all subjects. Large changes that greatly reduce or eliminate pain create huge effect-sizes.

3. Context

In some contexts, interventions that create small changes with large effect-sizes might be considered effective. In athletics, reducing times for a sprint by only 1 s might be regarded as highly worthwhile; and a training program that produced such changes across a team of athletes might be considered very successful and effective. In education, a program that improved the performance of each pupil in a class by 10% might be rated as highly successful.

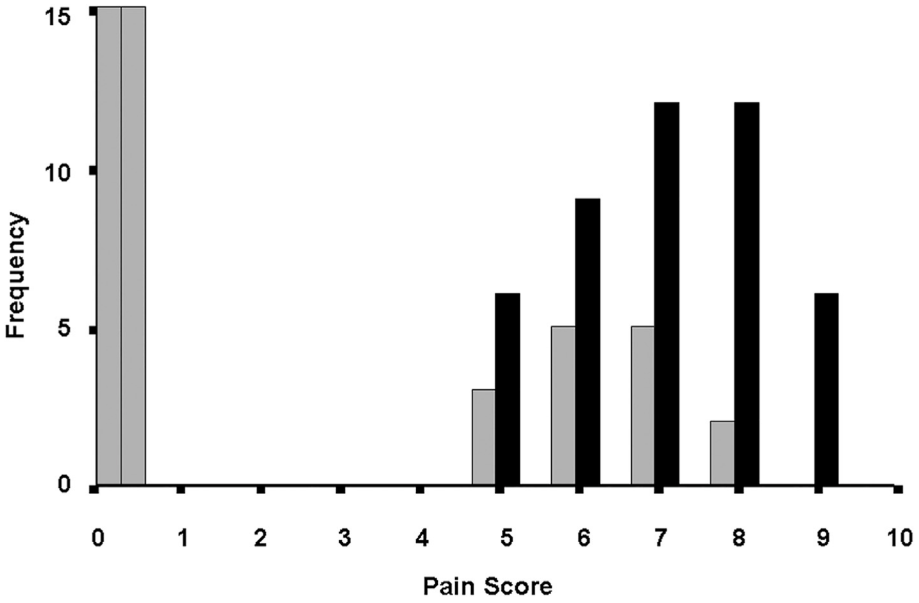

Fig. 4. The distribution of pain scores before and after an intervention. Black bars are the scores before intervention. Grey bars are the scores after intervention. The treatment is not universally effective, but nearly 67% of patients reduce their pain to zero.

Table 6
The mean pain scores and standard deviations, and the resultant effect-size, of the data illustrated in Fig. 4. The p-value applies to a two-sample *t*-test.

	Pain Score		p-value	Effect-size
	mean	sd		
Before	7.1	1.3	0.00	2.1
After	2.4	3.4		

The difference in the treatment of pain is that small changes in scores are almost inconsequential. Such changes do little or nothing to change the burden of illness. They do not change the state of the patients. A patient whose pain score improves from 7 to 6 is still a patient with significant, persisting pain. For changes in pain scores to be clinically significant they have to be large: at least by 50% or more.

4. Technicalities

The algebra of effect-size presumes a normal (binomial) distribution of data, for mean scores and standard deviations apply. Consequently, calculations of effect-size are valid for circumstances in which scores retain a normal distribution after changing, as in the examples illustrated in Figs. 1 and 2.

Problems arise when scores after an intervention do not remain normally distributed. Some treatments of pain create bimodal distributions, of patients who do not respond and patients who respond very well (Figs. 3 and 4). In such cases, the calculations for effect-size can still be applied mechanically, but it not legitimate to do so. The mean score of the entire group after treatment is not a valid representation of what are actually two, distinctly different subgroups. Consequently, the numerical magnitude of the effect-size for the entire group may not be valid, and descriptors such as very large or huge may not be accurate.

However, no formulation for effect-size has been designed to accommodate distributions, arising after an intervention, that are not normal. No lexicon of accurate descriptors for the effect-size in such situations has been developed.

For present purposes, however, this is not a critical issue. The cardinal value of calibrating effect-sizes in pain treatment is not to validate the descriptors for very large and ostensibly huge effects. Rather, the purpose is to show that the effect-sizes commonly encountered in studies of pain treatment, rated as medium or large, do not equate to treatments that have medium or large effectiveness. Both numerically and verbally, such effect-sizes are dwarfed by those produced by moderately successful treatments let alone highly successful ones. Consequently, the field of pain treatment needs more treatments with very large or huge effect-sizes, not more treatments with small, medium, or large effect-sizes.

5. Discussion

Readers of studies of pain treatment should exercise insight when authors report effect-sizes. In the first instance, they should understand that the effect-size pertains to the behaviour of the data, and does not apply to the intervention. In the second instance, they should understand that the verbal descriptors of effect-size likewise apply to the statistical appearance of the data, and not to the treatment. Thirdly, readers should understand that the available descriptors extend beyond large to very large and huge. Against this background, readers should realize that effective treatments, even with modest success rates, generate very large or even huge effect-sizes. Reciprocally, treatments with only medium or large effect-sizes are likely to be only minimally effective in clinical terms.

Readers should not rely on verbal descriptors when drawing conclusions about a treatment. If they want to see if the treatment is effective they should ignore the effect-size. Instead, they should call for a look at the distribution of the data, from which they can decide for themselves if the effectiveness of the treatment actually looks good, very good, or excellent.

Disclosures

The author was not funded to produce this article, and has no conflicts of interest with its contents.

References

[1] Cohen J. Statistical power analysis for the behavioral sciences. second ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
[2] Sawilowsky S. New effect size rules of thumb. J Mod Appl Stat Methods 2009;8:467–74.
[3] Ghahreman A, Ferch R, Bogduk N. The efficacy of transforaminal injection of steroids for the treatment of lumbar radicular pain. Pain Med 2010;11:1149–68.
[4] MacVicar J, Borowczyk JM, MacVicar AM, et al. Cervical medial branch radiofrequency neurotomy in New Zealand. Pain Med 2012;13:647–54.