Interventional Pain Medicine 1 (2022) 100126

Contents lists available at ScienceDirect

Interventional Pain Medicine

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

On comparing groups in studies of pain treatment

Nikolai Bogduk

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

ARTICLE INFO

Keywords:
Success rate
Confidence intervals
Survival
Number needed to treat

ABSTRACT

Studies of pain treatment involve comparing groups. In observational studies scores for outcome variables in the one group are compared before and after treatment. In controlled trials, scores are compared between groups undergoing different treatments. Statistics based on group scores might show that a statistically significant change has occurred but they do not reveal how well a treatment works. That is revealed by categorical data that show the proportion of patients that benefitted by how much after treatment. Those proportions are qualified by the 95% confidence interval of the proportion. In observational studies the magnitude of the success rate, and its 95% confidence interval, is enough to indicate how well the treatment worked. In controlled trials, success rates can be compared to determine how more often is one treatment successful than another. Statistical significance of the difference is established if the confidence intervals of the respective success rates do not overlap, or if the 95% confidence interval of the difference between success rates does not include zero. Pragmatic trials compare one treatment with another, but the comparison does not show if either treatment works well. Each arm of a pragmatic trial constitutes an observational study and the data in each ach show how well each treatment works. Explanatory trials investigate the extent to which the observed success rate is provided by responses to non-specific effects of treatment. The extent to which active treatment is more often effective than is sham treatment can be demonstrated by comparing the success rates of the two treatments, by comparing their survival curves, and by calculating the attributable effect and the number needed to treat of the active treatment.

1. Introduction

One way or another, studies of pain treatment involve comparing the outcomes of groups of patients. In observational studies, the one group is compared before and after treatment. In controlled trials, the groups in each arm are compared both within and between the arms. For readers of such studies, two pertinent issues arise: how best to compare groups, and why.

2. Methods for comparing

The methods most commonly used for comparing groups in studies of pain are statistical tests of group data. Some tests are relatively simple; others are more complex. Simple tests compare the state of one group, before and after treatment, or the states different groups, each before and after treatment. Commonly used, simple tests are the two-sample *t*-test when the data have a normal or near normal distribution, and the Mann-Whitney test or the *U* test for data that have a flat or irregular distribution. More complex tests, such as the Analysis of Variance (ANOVA) simultaneously test for differences between groups and differences over time, such as before and after treatment.

The limitation of such statistical tests is that they tell us only about the numerical behaviour of scores for measures of outcome. They tell us if the scores, on average, improved or not. However, by focussing on the group rather than on individual patients, these tests do not tell us if every patient benefitted, or if only some patients benefitted, if all patients benefitted to the same degree, or if some patients benefitted more than others. Although group statistics might tell us what the mean benefit was, they do not tell us how many patients achieved that degree of benefit, how many patients achieved less, and how many achieved greater benefit.

Although group data provide a global impression about whether or not a treatment works they do not provide more meaningful insights into the effects of treatment. They do not tell us the success rates of the treatment, for different definitions or grades of success. They do not tell the physician or their patients the chances of achieving a successful outcome. However, that information is provided by categorical data.

For any outcome measure, categorical data can report how many patients achieved different scores, and how many patients achieved different changes in scores, together with the proportions and cumulative

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

https://doi.org/10.1016/j.inpm.2022.100126
Received 12 June 2022; Accepted 16 June 2022

proportions who did so (Table 1). Cumulative proportions are the proportions of patients who achieved at least the improvement depicted in the corresponding row, i.e. the depicted improvement plus greater improvements.

For different types of outcomes, such as pain, disability, distress, and reduction in use of analgesics, the column for "scores" can be adapted to suit the scale used by the instrument used to measure the outcome, such as 0–10, 0–100, 0–30, or Likert scales of 1–4 or 1–5. Likewise, the column for percentage improvements can be adapted simply by dividing the change in outcome by the baseline score.

If categorical data are used to compare a single group before and after treatment, the question changes from "did the group improve" to "how often was the treatment successful (for various degrees of improvement)".

Without resorting to any statistical tests, data displayed in the format of Table 1 are informative. That format shows the distribution of scores before and after treatment. It shows how many patients achieved various grades of improvement. A reader can see that after treatment few patients retained high scores, and many achieved low scores. They can also see that many patients achieved high grades of improvement; few patients achieved low grades of improvement; and fortunately none were worse after treatment.

Regardless of whether the scores before and after treatment are significantly different or not statistically, the data displayed in the format of Table 2 tell the reader what they could expect of the treatment were they to adopt it. The display also allows the reader to perform a sensitivity analysis. That means that the reader can scan through the table of data and first, pick an outcome in which they are interested, and read the success rate that they could expect to encounter; secondly, they can then scan up or down to see what the success rates are for other grades of outcome. The notion of "sensitivity" means that the reader can scan through the table to find the lowest or highest success rate that might be attractive to them, or the most or least demanding grade of outcome in which they might be interested.

For example, the raw data of Table 2 tell the reader that they could expect 3% of their patients to achieve 50% improvement, but 70% of their patients could expect to achieve 50% improvement *or better*. Some 60% of their patients could expect to achieve at least 70% improvement, but only 30% could expect complete improvement.

These figures, however, are only estimates of the true success rate. The accuracy of any estimate is governed by the sample-size in any study. Small studies provide only crude estimates, because just one or two patients can skew the success rate downwards or upwards. Larger studies are likely to be more accurate because they constitute a better representation of the general population, and are at lesser risk of having their estimate of the success rate being skewed by the odd patient with "rogue" outcomes that are excessively poor or excessively good.

In as much as a success rate is mathematically a proportion, the statistical test for correcting for sample-size is the 95% confidence intervals (CI) of a proportion. There are more precise means of calculating the 95% CIs, but for general purposes an approximate value serves well enough, and is given by:

$$p^* = p \pm 1.96\sqrt{\frac{p(p-1)}{n}}$$

where p is the observed proportion or success rate, n is the sample-size, 1.96 is the coefficient used to produce 95% confidence, and p* is the upper limit or lower limit of the confidence interval once the ± operation is applied.

Table 2 provides the 95% confidence intervals of the several proportions reported in Table 1. It shows that the estimated success rate (70%) for achieving at least 50% improvement has a 95% confidence interval of 54%–86%. In the first instance, that means that the true success rate of this treatment, for that particular outcome, lies somewhere in the range between 54% and 86%. In the second instance, it informs the reader that, were they to adopt this treatment, they could expect to encounter a success rate, not necessarily of 70%, but maybe as low as 54% or as high as 86%.

Similarly, for achieving 70% improvement or better, the reader should expect a success rate, not of 60%, but anywhere from as low as 43% to as high as 77%. For achieving complete improvement, the chances are somewhere between 14% and 46%.

Categorical data, showing success rates for a range of possible outcomes allow the reader to decide if any of the success rates and their respective outcomes are sufficiently attractive to warrant adopting the treatment. The reader can also decide if the 95% confidence intervals are too wide to be convincing, and that the reader should wait for a more robust study to provide a stronger estimate with tighter confidence intervals.

For interpreting the results of an observational study, classical statistical tests based on group data are essentially irrelevant and uninformative. They might show that there has been a change in scores, and if that change is statistically significant; but they do not show how well the treatment works. In contrast, categorical data show how often the treatment works and to what degrees.

Table 3 shows an example. A two-sample *t*-test shows that the scores after treatment are significantly different, with p = 0.03. The categorical data, however, show that no patient improved by more than 30%; one-third of patients had no improvement, and two-thirds improved by only 10% or 20%. The low p value for the group data belies how weak this treatment is.

Table 1
A table of categorical outcome data for a hypothetical treatment for pain.

Outcome Measure			Improvement			
Score	Before	After		Number	Proportion	Cumulative Proportion
100	2		**100%**	9	0.30	0.30
90	6	1	**90**	0	0.00	0.30
80	5	8	**80**	6	0.20	0.50
70	6	2	**70**	3	0.10	0.60
60	6	1	**60**	2	0.07	0.67
50	5	3	**50**	1	0.03	0.70
40		0	**40**	3	0.10	0.80
30		3	**30**	1	0.03	0.83
20		4	**20**	1	0.03	0.86
10		6	**10**	1	0.03	0.89
0		9	**0**	3	0.10	0.99
			Worse	0	0.00	0.99

Table 2
The success rates and 95% confidence intervals for various grades of outcome, following a hypothetical treatment for pain.

Improvement			
	Number	Proportion	Cumulative Proportion
100%	9	0.30 (0.14–0.46)	0.30 (0.14–0.46)
90	0	0.00	0.30 (0.14–0.46)
80	6	0.20 (0.06–0.34)	0.50 (0.32–0.68)
70	3	0.10 (0.05–0.15)	0.60 (0.43–0.77)
60	2	0.07 (0.00–0.79)	0.67 (0.50–0.84)
50	1	0.03 (0.00–0.16)	0.70 (0.54–0.86)
40	3	0.10 (0.00–0.21)	0.80 (0.66–0.94)
30	1	0.03 (0.00–0.16)	0.83 (0.70–0.96)
20	1	0.03 (0.00–0.16)	0.86 (0.74–0.98)
10	1	0.03 (0.00–0.16)	0.89 (0.78–1.00)
0	3	0.10 (0.00–0.21)	0.99 (0.95–1.00)
Worse	0	0.00	

2.1. Controlled trials

When two different treatments are compared, the question being

Table 3
A table of categorical outcome data for a hypothetical treatment for pain.

Outcome Measure			Improvement			
Score	Before	After		Number	Proportion	Cumulative Proportion
100	2		**100%**	9	0.00	0.00
90	6	3	**90**	0	0.00	0.00
80	5	3	**80**	6	0.00	0.00
70	6	9	**70**	3	0.00	0.00
60	6	5	**60**	2	0.00	0.00
50	5	7	**50**	1	0.00	0.00
40		3	**40**	3	0.00	0.00
30		3	**30**	1	0.00	0.00
20		4	**20**	1	0.30	0.30
10		6	**10**	1	0.37	0.67
0		9	**0**	3	0.33	1.00
			Worse	0	0.00	

asked should change from "do the groups differ?" to "how more often did one treatment achieve successful outcomes than the other?". Here too, categorical data provide the answers. Moreover, 95% confidence intervals provide the statistical test for significant difference. As an example, Table 4 provides the categorical data comparing the outcomes of two treatments.

In the first instance, the data of Table 4 show that fewer patients had less than 50% improvement after treatment A than after treatment B. For this category of outcome, the 95% confidence intervals do not overlap. So, the difference between success rates is statistically significant. This difference implies that there should be a reciprocal difference in the success rates for treatment A achieving greater than 50% improvement; but this is not immediately evident in Table 4.

Although more patients achieved 50–75%, 75–100%, and 100% improvement, after treatment A, the confidence intervals of the respective success rates overlap for these categories of outcome, which denies a statistically significant difference. However, the success rates for achieving at least 50% and at least 75% improvement are significantly greater after treatment A than after treatment B.

This example serves to illustrate some properties of confidence intervals. When approximate 95% confidence intervals do not overlap they serve perfectly well to establish (rule in) a statistically significant difference between two success rates. However, approximate values of the confidence intervals lack discrimination when two sets of confidence contact one another or overlap slightly. They do not rule out a statistically significant difference. In such cases, greater resolution is provided by a related test.

Instead calculating and comparing the difference between the confidence intervals of two proportions, one can calculate the 95% confidence interval of the difference between two proportions. The language in this sentence may not be immediately clear, because it involves so many words that sound alike. The essence is that "the difference between two confidence intervals" is not the same as "the confidence intervals of the difference between two proportions". This is analogous to "the difference between the respective means of two groups" is not the same as "the mean value of individual differences between two groups".

The 95% confidence interval of the difference between two proportions involves a somewhat complicated calculation, it but provides a higher resolution test of the significance of any difference. If the 95% confidence interval of the difference is consistently above or consistently below zero, then one can be 95% confident that the difference is statistically significant ($p < 0.05$). Several calculators for the 95% confidence interval of the difference between two proportions are available on the Internet, e.g. Confidence Interval for the Difference in Proportions Calculator,

https://www.statology.org.

Table 5 provides the 95% confidence intervals of the differences between success rates shown in Table 4. For achieving less than 50% improvements, the confidence interval is entirely less than zero, which means that the success rate for treatment A is significantly less than that of treatment B for achieving this outcome. For achieving 50–75% improvement, the confidence interval ranges from below zero to above zero, which means that there is no significant difference between the success rates for achieving this outcome; the true difference could be negative, favouring Treatment B; it could be zero favouring neither treatment; or it could be positive, favouring treatment A.

For all other grades of outcome the confidence intervals are completely greater than zero, which means that the success rates for Treatment A are significantly greater than those of treatment B, for achieving 75–100% improvement or complete improvement, as well as for at least 50% improvement and for at least 75% improvement.

Given these data, the reader can conclude that treatment A significantly less often fails to achieve less than 50% improvement, and significantly more often achieves 75%–100% improvement, 100% improvement, at least 50% improvement, and at least 75% improvement. The only question that the reader has to ask of themselves is if a failure rate of 23% but a success rate of 44% for complete improvement plus another 23% for 75% improvement makes the treatment worthwhile for their purposes.

Group data and their p values do not provide such insights. Only categorical data do so.

Table 5
The proportions of patients, and the 95% confidence intervals of those proportions, who achieved the outcomes listed after either treatment A or treatment B, and the 95% confidence intervals of the differences between those proportions.

Improvement	Proportion		Difference		Statistical
	A	B	A - B	95% Confidence Interval	Significance
<50%	0.23	0.44	- 0.21	(- 0.32) – (- 0.09)	A < B
50–75%	0.09	0.11	- 0.02	(-0.09 – (+0.05)	no
75–100%	0.23	0.13	+0.09	(+0.01) – (+0.19)	A > B
100%	0.44	0.32	+0.12	(+0.01) – (+0.24)	A > B
>50%	0.79	0.55	+0.24	(+0.10) – (+0.32)	A > B
>75%	0.68	0.45	+0.23	(+0.11) – (+0.35)	A > B

Table 4
The numbers of patients, the proportions of patients, and the 95% confidence intervals of those proportions, who achieved the outcomes listed after either treatment A or treatment B.

Improvement	Number		Proportion		95% Confidence Interval		Significant Difference
	A	B	A	B	A	B	
<50%	43	46	0.23	0.44	0.17–0.29	0.34–0.54	A < B
50–75%	16	11	0.09	0.11	0.09–0.13	0.03–0.17	no
75–100%	43	13	0.23	0.13	0.17–0.29	0.07–0.19	no
100%	81	33	0.44	0.32	0.37–0.51	0.23–0.41	no
Cumulative Data							
>50%	140	57	0.76	0.55	0.70–0.82	0.45–0.65	A > B
>75%	124	46	0.68	0.45	0.61–0.75	0.35–0.55	A > B

3. Why compare?

The cardinal reason for studying any treatment for pain is to determine if it works. This can be done by observational studies, pragmatic controlled trials, or explanatory trials. In that regard, however, it is somewhat enigmatic why systematic reviews and practice guidelines eschew observational studies when making their decisions.

3.1. Observational studies

Ostensibly, the reasons for ignoring observational studies are that they are too often of poor quality, and that good methods are to be found only in controlled trials. However, guidelines are available for making observational studies rigorous [1]. They key features of these guidelines are summarised in Table 6. There is no fundamental reason why a good observational study should be ignored when looking for evidence that a treatment works. Indeed, the pioneer of evidence-based medicine maintained that all evidence, from observational studies as well as controlled trials, should be considered [2].

3.2. Pragmatic trials

Pragmatic trials compare two different treatments. However, each arm of the trial constitutes no more than a good observational study. Comparing the outcomes of the two treatments does not determine if either of them works. That is achieved by the observational study of each arm. Comparing the two treatments serves only to test if one treatment is better than the other. To no small extent this can serve a political, rather than a scientific, objective. Authors can claim that "my treatment" works better than "yours".

A particular problem with most pragmatic trials to date is that they use group data ostensibly to prove that one treatment works better than the other. However, this is illusory. As explained above, statistical tests on group data might show a difference in outcome but they do not show how well a treatment works. For meaningful comparisons, categorical data should be used, to show how often either treatment works, for various grades of outcome. Under those conditions, the research questions changes from "are the outcomes statistically different?" to "how often is one treatment more successful than the other?". Other problems apply to pragmatic trials.

Both treatments might work, and their outcomes might be significantly different statistically but not necessarily clinically. Both treatments, might be ineffective, but with one being significantly less so. In that event, the comparison of the treatments does not prove that the other treatment works, because the observational data in each arm show that neither treatment works well.

It might be claimed that pragmatic trials are superior to observational data because patients are randomized to each treatment, but that can be a specious virtue. That virtue applies only if the two treatments are competitive, and patients can expect that either is the "real" treatment. The virtue of randomization, however, evaporates if the control treatment is obviously inferior, such as returning to usual care, or being told to wait for 6 months to become eligible for the real treatment. Being relegated in this way invites patients to experience a nocebo effect, causing them to rate their experience with their treatment worse than they otherwise might. Any differences in outcome are, therefore, due not to the superiority of the real treatment but to the amplified inferiority of the control treatment.

On the other hand, if one treatment in a pragmatic controlled trial is advertised as a "new" treatment, and the control treatment is obviously not the new treatment, the new treatment can incur what is known as a halo effect. The halo effect encourages patients to expect good outcomes from the new treatment, as if it is a godsend, and thereby invites placebo responses. These effects exaggerate the outcomes of that arm of the trial.

Halo effects are the greatest concern about observational studies. Although observational studies might show that a treatment works, they cannot show the extent to which the outcomes are influenced by non-specific or placebo effects. Resolving those is the role of explanatory trials.

Table 6
Cardinal features of a good observational study (1).

DIAGNOSIS	**ASSESSMENT**
Defined	Minimal loss to follow-up
Valid	Follow-up long enough
PATIENTS	Valid outcome measures
Prospective	For pain, disability, other health care
All consecutive	Independent assessor
Typical sample	Real-time assessment
Stratified for duration, comorbidity	Outcomes quantified
CO-INTERVENTIONS	**ANALYSIS**
None or	Worst-case, intention to treat
Stratified for co-interventions	Sample size sufficiently large
TREATMENT	Success rates provided for various grades of outcome
Defined	
Uniformly applied	Stratified for duration, comorbidity

3.3. Explanatory trials

In an explanatory trial, patients are randomized either to the active treatment or a sham treatment. For the trial to be valid, the sham treatment and its delivery must be identical to the active treatment save for the omission of what is purported to be the active ingredient or component of the active treatment. Thus, in an explanatory trial of a drug, the sham treatment would consist of prescribing and consuming a pill that was the same size, shape, and colour as the active drug, but which did not contain any active agent. In a trial of injection of steroids, the injection would be performed entirely in the same way, but in the sham treatment an inactive agent would be injected instead of the steroid. In a trial of radiofrequency neurotomy, the procedure would be performed in the same way for the same duration, but current would not be passed to generate a lesion.

Under these conditions, the patients cannot tell to which treatment they have been randomized, thereby eliminating, or at least distributing equally between both arms of the trial, any halo effects or nocebo effects. Randomization ensures that any patients more likely or less likely to respond are equally distributed between both treatments.

The arm with the active treatment constitutes an observational study, and establishes how well the active treatment works. The sham treatment constitutes an observational study of how well providing the non-specific components of treatment works. Comparing the two sets of outcomes establishes the extent to which the outcomes of active treatment exceed those of non-specific effects.

In the first instance, the difference between the two treatments can be measured by comparing their success rates for any and all grades of all relevant outcome measures. If the 95% confidence intervals of the respective success rates do not overlap, or if the 95% confidence intervals of differences between success rates are consistently non-zero, the conclusion can be drawn that the active treatment achieves a successful outcome significantly more often than does sham treatment.

3.3.1. Survival analysis

Another method is to compare survival curves, using Kaplan-Meier regression analysis [3] (Fig. 1). Calculation of the statistic to compare the two curves is somewhat complex. It involves comparing the survival rates using chi-squared analysis, but integrating any differences over time. In essence, the test measures the probability that the area between the two curves could have arisen by chance alone, given the sample-sizes of the study. In that regard, survival analysis is a high-resolution method for dealing with small studies, such as may arise for rare conditions. Although at any given point in time survival rates may not be significant statistically because of small sample-sizes, the test accords credit to

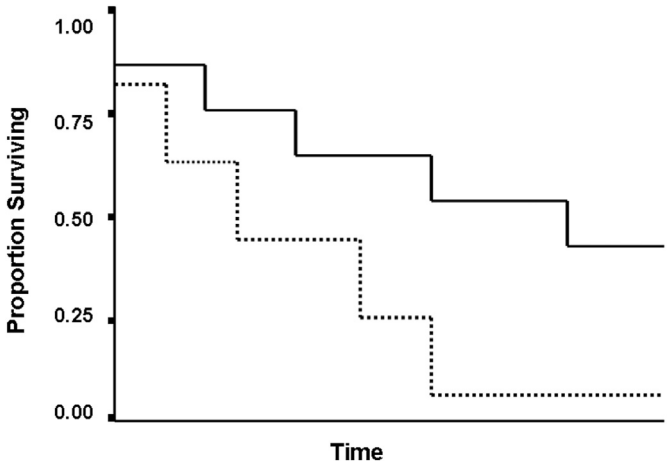

Fig. 1. Survival curves for comparing the success rates of two treatments. The Proportion Surviving measures the proportion of patients with a successful outcome from treatment, which evolves over time.

differences that persist or increase over time. When applied to studies of pain treatment, "survival" is conceptually replaced by "success rate".

If the survival curves prove to be significantly different. It can be concluded that the treatment with the greater survival of success rate is more effective than the other. Whether that difference is clinically worthwhile depends on how the reader values the observed success rate over time.

3.3.2. Number needed to treat

Another statistic can be applied to the outcomes of an explanatory trial in order to show, in an elegant, succinct, manner how well a treatment works [4]. The calculation of this statistic is easy, but explaining its idiom can be awkward.

Table 7 shows that one can list the respective success rate of an index treatment (SR_{IT}) and the success rate of a sham treatment (SR_{SH}). The difference between these success rates can be called the Attributable Effect (AE) of the index treatment, meaning how much of the observed success rate can be attributed to the active component of the index treatment as opposed what is provided to the success rate by non-specific effects.

The example shown in Table 7 states that the observed success rate of the index treatment was 60% (but is expressed as a proportion (0.60) in order to suit the later calculations). The success rate of sham treatment was 40%. The Attributable Effect is 0.60–0.40 = 0.20. This means that, on average, for every 100 patients treated, 40 respond to the index treatment because of non-specific effects, while 20 respond explicitly because of the specific, active component of the treatment.

Such numbers can be used to generate a succinct statistic. It is called the number need to treat (NNT) [4]. Conceptually, this statistic is derived in the following manner.

Let there be a treatment that has a success rate of 60% and an Attributable Effect of 20%. If 100 patients are treated, 60 would have a successful outcome, but only 20 of these outcomes are due to the specific effect of the treatment. In order to appreciate how well the treatment works we need to discount the successes generated by non-specific effects. The true merit of the treatment lies in how many patients respond explicitly to the active ingredient of the treatment, which is what the Attributable Effect tells us. So, we can say:

100 treated	yields	20 successes.

The question can be asked: how many patients do I need to treat in order that, on average, I would get one specific success? This could be derived by trial and error, error, e.g.

100 treated	yields	20 successes.
50 treated	yields	10 successes.
10 treated	yields	2 successes.
5 treated	yields	1 success.

This result means that you need to treat five patients before you can claim that one responded because of the specific effects of the treatment (as opposed to its non-specific effects).

A generic expression can be derived in the following manner.

If N patients treated	[(AE).N] respond specifically to the treatment

How many patients do I need to treat (NNT) for [(AE).N] to equal 1?
(AE). N = 1
NNT = 1/AE.

So, NNT is the reciprocal of the Attributable Effect. For example, if AE = 0.10, NNT = 10; if AE = 0.20, NNT = 5; if AE = 0.30, NNT = 3.3; if AE = 0.40, NNT = 2.5. Since it is not possible to treat 0.3 or 0.5 of a patient, the convention for NNT is to round up any decimal values. So, NNT = 3.3 becomes NNT = 4, and NNT = 2.5 becomes NNT = 3. (Exceptions to this convention are tolerated when reviews determine the "representative" NNT that is the mean value based on a number of different studies.)

Large NNTs indicate that the treatment either has a low success rate or has a success rate with a large component of placebo effects. For treatments with low NNT their success rates are less due to placebo effects. So, a single statistic can provide immediate insight into how well a treatment works, and the extent to which its apparent success is due to placebo effects. Table 8 list some representative values of NNT for commonly used treatments for pain.

Given that post-operative analgesics are considered a "good" treatment, or at least an acceptable one, and given that most analgesics have an NNT between 3 and 5, it could be concluded that an NNT of 3–5 could serve as a reference value for a "good" treatment. Reciprocally, NNTs greater than 5 suggest that the treatment is either weak or heavily relies on placebo effects. However, when interpreting NNTs readers should take care not to misunderstand or misrepresent them. NNTs cannot be cited in isolation. NNTs must be qualified.

Foremost, an NNT needs to be qualified for the definition of success to which it pertains. Two treatments might have an NNT of 4, but for one treatment that number might pertain to achieving 50% relief of pain, whereas for the other treatment the number might pertain to complete relief of pain. So, while both are equivalent in achieving success, one of them achieves a higher grade of success. Table 7 shows some examples of this.

NNTs may also need to be qualified by time. Success rates of treatment may deteriorate over time, and outcomes after active treatment and control treatment may deteriorate at different rates. Consequently the NNT may be different at different times of follow-up. For example, in the case of cervical radiofrequency neurotomy, the NNT for achieving complete relief of pain is 3.0 at three weeks, but becomes 2.4 at two months, and 2.0 at seven months, because the effects of sham treatment deteriorate faster than those of active treatment. So, although the long-term

Table 7
The relationship between success rates in an explanatory trial, and the Attributable Effect of the index treatment.

Treatment	Success Rate	Attributable Effect
Index	SR_{IT}	$SR_{IT} - SR_{SH}$
Sham	SR_{SH}	
Example		
Index	0.60	0.20
Sham	0.40	

Table 8
A list of representaive numbers needed to treat (NNT) for treatments commonly used for pain.

Treatment	NNT	Source
Post-operative Analgesics (50% relief of pain)	1.5–5.7	[5]
Codeine 60 mg (50% relief of pain)	17	[5]
Gabapentin (50% relief of pain)	6–8	[6]
Caudal epidural steroids (marked improvement at 12W)	9	[7]
Lumbar epidural steroids (improvement)	10	[8]
Lumbar transforaminal injection of steroids (>50% relief)	3	[9]
Cervical RFN (complete relief at 3 weeks)	3	[10]
(complete relief at 7 months)	2	

success rate of cervical radiofrequency neurotomy might be only about 60%, the NNT of 2 assures readers that little of that lasting success can be attributed to placebo effects.

4. Discussion

In his teaching, the late Dr Charles Aprill [11] regularly used the expression: "Think about it". This instruction lies at the heart of this essay.

When a study provides you with group data and a p value of >0.05, is that enough to convince you that the treatment works? Technically it does, but only well enough to effect a statistical difference. Do mean scores and their p value tell you if the treatment works well enough for you to use it? Think about it.

As a consumer of medical information, is a p value enough to convince you, or would you prefer to know how often a treatment works, in what respects does it work, by how much, and for how long? Instead of p values, would you prefer to know the success rates of the treatment, and their 95% confidence intervals?

If you want the latter, come to demand it of publications on pain treatment. Reject or ignore publications that give you only group data and p values. Seek out those publications that provide you with transparent, comprehensive categorical data. Base your decision about treatments only on categorical data.

The same precepts apply to randomised controlled trials. Is it enough to see $p < 0.05$ between group scores, or would you prefer to see how often is one treatment more successful than another, and by how much; and an NNT that tells you that you are not fostering a treatment most of whose effect comes from how you deliver it instead of its purported active ingredient?

There is an aphorism that claims that most physicians know more about a buying a used car than they do about assessing trials of pain treatment. When buying a car they know to check the logbook, check the fluids, test the breaks, look for rust, look for leaks, and test the suspension, *inter alia*, lest they buy a lemon. Having read this essay, are you better equipped to handle medical information, or will you continue to settle for group data, and accept a lemon?

Disclosures

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

References

[1] Bogduk N, Kennedy DJ, Vorobeychik Y, Engel A. Guidelines for composing and assessing a paper on treatment of pain. Pain Med 2017;18:2096–104.
[2] Cochrane A. Effectiveness and efficiency. Cambridge: Cambridge University Press; 1977. p. 21–30.
[3] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
[4] Cook RJ, Sackett DL. TThe number needed to treat: a clinically useful measure of treatment effect.he number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452–4.
[5] Moore A, Edwards J, Barden J, McQuay H. Bandolier's little book of pain. Oxford: Oxford University Press; 2003.
[6] Moore RA, Wiffen PJ, Derry S, et al. Gabapentin for chronic neuropathic pain and fibromyalgia in adults. Cochrane Database Syst Rev 2014;v4. CD007938.
[7] Nandi J, Chowdhery A. A randomized controlled clinical trial to determine the effectiveness of caudal epidural steroid injection in lumbosacral sciatica. J Clin Diagn Res 2017;11:RC04–8.
[8] Oliveira_Cb, Maher_Cg, Ferreira_Ml, et al. Epidural corticosteroid injections for lumbosacral radicular pain. Cochrane Database Syst Rev 2020;(4). https://doi.org/10.1002/14651858.CD013577. CD013577.
[9] 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.
[10] Lord SM, Barnsley L, Wallis BJ, McDonald GJ, Bogduk N. Percutaneous radiofrequency neurotomy for chronic cervical zygapophysial-joint pain. N Engl J Med 1996;335:1721–6.
[11] Bogduk N, Maus T. In memoriam Dr. Charles N. Aprill, MD. Spine 2021;46:E800–1.