



OPEN ACCESS

Why science is less scientific than we think (and what to do about it): The 2022 Gaston Labat Award Lecture

Brian M Ilfeld ^{1,2}

¹Department of Anesthesiology, University of California San Diego, La Jolla, California, USA
²Outcomes Research, Cleveland Clinic, Cleveland, Ohio, USA

Correspondence to

Dr Brian M Ilfeld, Department of Anesthesiology, University of California San Diego, La Jolla, California, USA; bilfeld@health.ucsd.edu

American Society of Regional Anesthesia and Pain Medicine annual meeting, April 2, 2022, Las Vegas, Nevada

Received 14 November 2021
 Accepted 1 December 2021
 Published Online First 1 April 2022

Fun (pain medicine-related) facts: the origin of the English word ‘narcotic’ is from ancient Greek: νάρκω *narkō*, ‘to make numb’, a term used by Hippocrates, Aristotle and Plato.¹ Contemporaries of these medical luminaries treated the pain of operations, headaches, and childbirth with electricity from what they described as *narke* (figure 1), and we now call torpedo fish (‘torpedo’ is itself derived from the Latin *torpere*, ‘to be numb or stunned’). But I digress...



Figure 1 An example of a torpedo fish which can produce electrical current of up to 220 volts.

BACKGROUND

Gout was a significant health issue in Ancient Rome lacking an effective treatment.

METHODS

In the first Century, 20 Roman citizens with severe gout were randomized to touching either a living or dead torpedo fish (following written, informed consent, of course).²

RESULTS

The day following treatment, pain scores of participants who had contact with a living fish ($n=10$) improved an average of 1.6 points, compared with no improvement at all for the control group on a 0–10 Numeric Rating Scale ($p=0.001$).

DISCUSSION

This difference is statistically significant, but is it clinically relevant—should touching a living torpedo fish be offered to all patients with severe gout?

INDIVIDUAL VERSUS GROUP DIFFERENCES

There are two distinct steps in the critical process of applying study results to clinical practice: (1) determining the degree of improvement important

to *individuals* and (2) establishing the degree of improvement between *groups* that is relevant. An example of the former: identifying that a majority of Romans with gout consider a decrease of 2 points on the 0–10 scale to be meaningful. In other words, if we ask any *individual* patient if touching the fish was worth doing, those with at least a 2-point improvement would answer in the affirmative: we have defined the minimum clinically important difference (MCID) as 2 for *individuals*. In our study, of the 10 patients who touched a living torpedo fish, 8 patients had an improvement of 2 points, while remaining two patients experienced no improvement, providing an average of 1.6 for the *group*. The mean improvement for the control group was 0 (no change), and therefore, the difference between the two treatment *groups* was 1.6 points.

Measuring the smallest values *individuals* consider important for patient-reported outcomes such as pain has been well-described and can be relatively straight-forward.³ However, as with our study of Ancient Romans, controlled clinical trials usually produce aggregated results for *groups*, and the smallest difference that is considered important between *groups* is far less well described. Unfortunately, the smallest change that is important to *individual* patients cannot be extrapolated to the evaluation of differences between *groups*.⁴ This is a critical distinction which is frequently unrecognized by investigators, consumers of research, as well grant and manuscript reviewers, often with significant negative consequences.^{5,6} Considering this issue applies to nearly every controlled study which informs patient care, it is not an obscure matter relegated to academic methodological researchers—it is of immense consequence to everyone involved in healthcare delivery.

INDIVIDUAL DIFFERENCES

Multiple methods are available to discern the smallest difference for patient-reported outcomes (eg, pain) that is important to *individual* patients. For example, postoperative patients can rate their pain on a numeric scale, consume an analgesic, and then rate their subsequent pain level as ‘worse’, ‘the same’, ‘slightly better’, ‘much better’ (figure 2). Patients who rated their improvement as ‘much better’ can be combined, and the difference between their preintervention and postintervention scores define a ‘benchmark’ of what improvement is deemed clinically important to each *individual*. The MCID can be significantly influenced by a myriad of patient characteristics such as age,⁷ gender,⁸



© American Society of Regional Anesthesia & Pain Medicine 2022. Re-use permitted under CC BY-NC. No commercial re-use. Published by BMJ.

To cite: Ilfeld BM. *Reg Anesth Pain Med* 2022;**47**:395–400.

body mass index,⁷ culture,⁹ ethnicity,¹⁰ geographic locality,⁹ educational level,¹¹ perceived general health,¹¹ disease status,¹² and socioeconomic status; intervention tolerability, adverse effects, and safety; as well as pain baseline,⁸ duration, intensity, frequency, location, etiology... well, you get the picture.¹³ As just one simplistic example, a small benefit from a treatment that is well tolerated and safe will usually be considered more acceptable than a larger benefit from a treatment that kills 50% of patients. There are a near-infinite number of combinations of characteristics and, therefore, unique *groups* of patients ('populations').

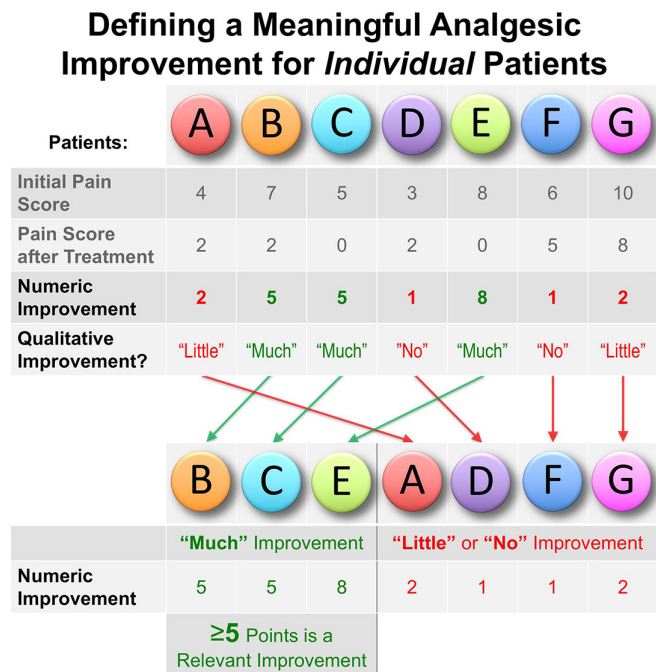


Figure 2 An example of an 'anchor-based' method to determine the smallest improvement in analgesia that *individual* patients consider relevant or important.

Crucially, the smallest important difference for one cannot be assumed valid for others. Nevertheless, it is generally accepted that while we cannot account for all variables, the MCID for specific populations are discernable and function as an estimation of the true MCID for similar *individuals*. For instance, in a widely respected and cited investigation, Farrar *et al* used an anchor-based method to determine that decreases of at least 1.7 points (or ≥28%) on an 11-point pain intensity numerical rating scale identified patients who considered themselves at least 'much improved' after treatment with pregabalin for chronic pain of various etiologies.⁸ There has been a good deal of research involving chronic pain states, resulting in a widely accepted consensus statement involving the analgesic MCID for *individual* patients.³ Regrettably, no similar consensus exists for acute or postoperative pain states.¹⁴ This predicament has lead researchers investigating acute pain to either simply use the MCID from chronic pain states^{3,8}—a practice lacking validation—or rely on the *individual* MCID from the few studies involving acute pain.¹⁵⁻¹⁸

GROUP DIFFERENCES

In contrast, discerning the importance of differences between *group* means requires accounting for clinical and societal

contexts in addition to the smallest important change among *individuals*. Returning to Ancient Rome, of the 10 patients who touched a living torpedo fish, 80% experienced an improvement of 2.0 points on the 0–10 scale, while the remaining two experienced no change (figure 3). Since we identified a 2-point improvement as what each *individual* considers important, 80% of our experimental group considers the treatment beneficial (this is simplified for demonstration purposes—reality is more complex). However, when all of the 10 patients' pain scores are combined for analysis, the *group's* average improvement is only 1.6 points (p=0.001). This is less of an improvement than the 2.0 points required by *individuals* to describe the change as important. When planning the study, if we had defined an 'important difference' between the two treatment *groups* as 2.0 based exclusively on what *individuals* consider important, then we cannot conclude that touching a living torpedo fish is clinically relevant since the difference between our experimental and control *groups* is only 1.6 points.

Smallest Relevant Improvement for *Individuals* versus *Groups*

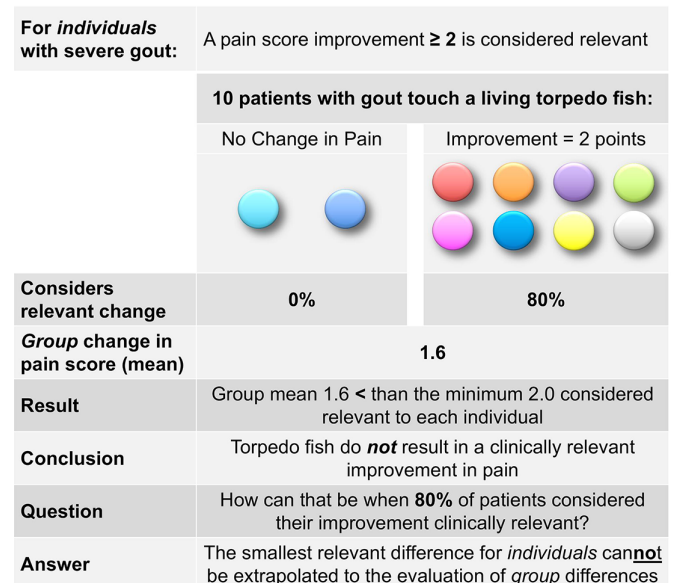


Figure 3 An example of the problem extrapolating the smallest change that is important to *individual* patients to the evaluation of *group* differences. This example is simplified to convey the general concept—reality is (unfortunately) more complex.

Should we now conclude that torpedo fish neuromodulation is an *ineffective* analgesic undeserving of future use? Is this a situation in which we detect a statistically significant difference without concurrent clinical significance? Put another way, did our study reveal an improvement that exists, but is so small that it is clinically irrelevant? That appears improbable since 80% of the treated patients experienced what they consider a clinically relevant effect (2-point or greater improvement), and so one could reasonably and logically conclude that the treatment *is* useful.

The error was to apply the MCID of *individuals* to *group* differences.^{3,4,19} The high successful response rate information is often lost when authors perform aggregated summary statistics between *groups*. In addition, to determine whether the *group*

difference is clinically meaningful, we must take into account many additional variables such as onset speed, treatment effect size magnitude, durability of benefits, convenience, patient adherence, results for secondary efficacy endpoints, safety, tolerability, and cost—to name but a few—relative to other available treatments.^{13–19} If torpedo fish, compared with alternative analgesics of Ancient Rome, provide a rapid analgesic onset and long duration, are plentiful and easily stored, induce no adverse side effects, and are inexpensive, then perhaps we can conclude that the average *group* improvement of 1.6 points is clinically meaningful and should be considered a possible analgesic option (even though *individuals* do not consider less than 2.0 points personally meaningful). In contrast, if torpedo fish require storage in expensive aquariums and frequently induce severe electrical burns, then perhaps the proper conclusion is that the average *group* improvement of 1.6 points does not suggest future use is warranted. And this is a dramatic oversimplification of the issue—completely ignored are issues of a placebo effect, baseline pain score differences, symptom spontaneous resolution, subgroup analysis, and regression to the mean, among others.¹⁹

In general, the smallest relevant improvement for *groups* will be smaller than for *individuals*.²⁰ Consequently, if the MCID for *individuals* is erroneously applied to *groups*, there is a good chance that a potentially useful intervention will be incorrectly declared as clinically unimportant. The use of oral analgesics to treat osteoarthritis pain is a good illustration of this concept. Farrar *et al* previously determined that *individual* patients with chronic pain of multiple etiologies (including osteoarthritis) require an analgesic to improve their pain score by at least 1.7 points to be considered an important change.⁸ However, as noted previously by Dworkin *et al*, when treating osteoarthritis the improvement in average *group* pain scores is less than 1.1 for opioids, non-steroidal anti-inflammatory drugs, paracetamol (acetaminophen), glucosamine, and chondroitin compared with placebo (standardized to an 11-point scale).¹⁹ If the widely respected and cited smallest improvement of 1.7 points deemed important for *individuals* was applied to these studies of *group* differences, we would not have a single oral analgesic considered effective for osteoarthritis!⁸

INVESTIGATOR CONUNDRUM

Determining the smallest improvement deemed important for *groups* is not easily calculated, and there is a lack of consensus regarding appropriate techniques and their application to various scenarios. Unfortunately, investigators cannot simply avoid defining the minimal difference between *groups* since it is usually required for hypothesis testing: in estimating an appropriate sample size, for example. To complicate matters—as if they needed further complication—much of the information required to determine the *group* MCID such as the incidence of adverse events is frequently unknown and requires the trial results to provide the information required to... design the trial in the first place. Which is why authors frequently use the phrase sample size ‘estimation’ or ‘justification’: the sample size must be inferred from imperfect and incomplete information. Regrettably, investigators are frequently asked to pretend they can make a definitive ‘determination’ based on published data. To illustrate the impact misunderstanding these topics can have on clinical research, I offer an example from my own experience.

My coinvestigators and I submitted a grant proposal to investigate the use of treating established post-amputation phantom limb pain with a continuous peripheral nerve block. The MCID for our specific patient population had not been discerned, so

we proposed using Farrar’s finding that decreases of at least 1.7 points on an 11-point pain scale identified *individuals* who considered themselves at least ‘much improved’.⁸ Granted, this was an extrapolation of data from patients with different pain etiologies treated with a different analgesic (pregabalin); but we felt it was the best available data for our purposes. Our estimated sample size was based on an approximation of a *group* difference of 1.1 points, partially informed with the 1.7 MCID for *individuals*. The proposal was rejected, with one reviewer explaining that using 1.1 points as our MCID—when published data demonstrated 1.7 points to be the true MCID for patients with chronic pain—would leave results falling between the two values as uninterpretable.

We subsequently submitted a revised proposal using the reviewer’s preferred smallest meaningful clinical difference of 1.7 points between the treatment and placebo *groups*, and it was funded. On completing the study, we submitted the manuscript reporting that after 4 weeks, average phantom limb pain intensity was a mean of 3.0 in patients given local anesthetic vs 4.5 in those given placebo ($p=0.003$).²¹ The manuscript was rejected with a reviewer explaining that while our primary endpoint was statistically significant, we had prospectively defined the minimal clinically important difference between treatment *groups* as 1.7, and we had found a difference of only 1.5 points; demonstrating that while the intervention did, in fact, decrease phantom limb pain, it failed to do so to a clinically relevant degree.

INVESTIGATORS

As previously noted, investigators have little choice but to define the smallest important difference between treatment *groups* to permit designing, analyzing, and reporting controlled studies. While a full discussion of the possible solutions to this challenge is outside of the scope of the current article, some relatively simple steps warrant mention. Joint hypothesis testing allows for multiple primary end points to be evaluated with a high degree of confidence, so that there are more variables—and information—on which to draw conclusions.²² This is particularly helpful when two outcomes of importance are closely related and drawing conclusions based on just one decreases confidence in the ultimate conclusions. For example, if a new medication improves pain scores yet has no effect on supplemental analgesic consumption, making either of these end points the sole primary outcome measure will result in differing conclusions. However, by prospectively defining both as primary outcomes and requiring one to be superior while the other at least non-inferior, the study results are more easily interpreted with a high degree of confidence.

Another relatively common technique is statistical ‘gate-keeping’: multiple outcomes are combined in sets and these listed in order of importance (figure 4).^{22–23} If at least one outcome in the first (primary) set is found to be statistically significant, the next set of end points are tested (although with a higher ‘bar’ for concluding a difference, depending on the number of variables found significant in the previous group). This continues through each of the gates until the overall type I error is exhausted. Gatekeeping permits statistical comparisons of many variables while retaining a study-wide type I error at a 0.05, increasing confidence in all comparisons and aiding interpretation and generalization of the results. In the phantom limb pain study described above, we prospectively specified a statistical gate-keeping strategy which elevated confidence in various secondary outcomes, the first of which was a scale that demonstrated pain’s interference on emotional and physical functioning was a mean

of 11 for treated participants, vs 28 for the placebo group (lower scores=less interference; $p=0.027$).²¹ Given that the results of multiple published investigations suggest that a mean *group* difference of 2 points on this scale indicates patients who are satisfied or improved with treatment,³ our finding of a difference of 17 points suggested a clinically significant improvement in the experimental over placebo treatments; and the results were subsequently published in different journals.^{21 24}

Example of Joint Hypothesis Testing and Parallel Gatekeeping

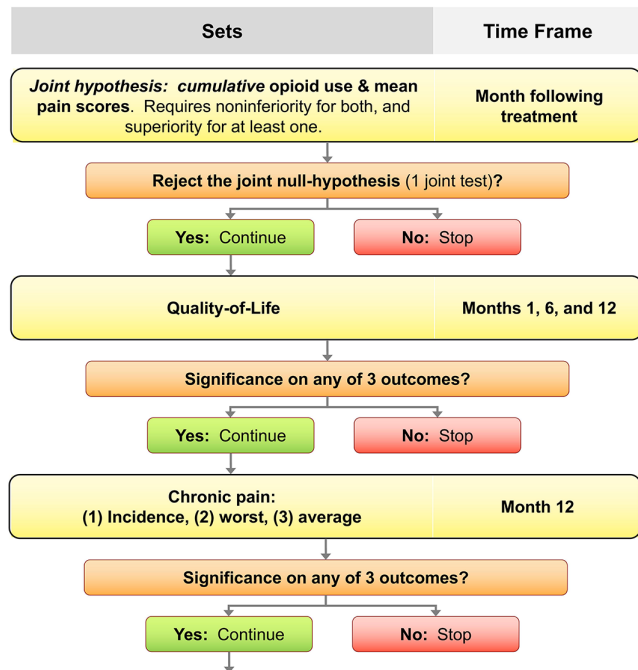


Figure 4 Visual representation of joint hypothesis testing and a parallel gatekeeping procedure. Gatekeeping permits statistical comparisons of many variables while retaining a study-wide type I error at a 0.05, increasing confidence in all comparisons and aiding interpretation and generalization of the results. In this example, the study outcomes are prospectively prioritized into three ordered sets. On study conclusion, statistical testing proceeds through each ‘gate’ to the next set if and only if at least one outcome in the current set reached significance. The significance level for each set will be 0.05 times a cumulative penalty for non-significant results in previous sets (ie, a ‘rejection gain factor’ equal to the cumulative product of the proportion of significant tests across the preceding sets). Within a set, a multiple comparison procedure (Bonferroni correction) is used to control the type I error at the appropriate level, if needed. Adapted from a grant proposal coauthored with Edward Mascha, PhD (Departments of Quantitative Health Sciences and Outcomes Research, Cleveland Clinic, Cleveland, Ohio, USA).

STAKEHOLDERS

While comparing two *group* means facilitates hypothesis testing and thus often preferred by investigators, it also complicates interpretation of the results and extrapolation to differing clinical scenarios. For other stakeholders, including clinicians, patients, administrators, and even other researchers, a more useful measure of treatment effect can be the percentage of each study group which reached various outcomes (related to the ‘number needed to treat’ (NNT)).²⁵ For example, it is probably difficult for *individual* patients with phantom limb pain to evaluate how our reported 1.5-point improvement in group means

applies specifically to them. Potentially more useful information is that 56% of study participants treated with local anesthetic experienced at least a 1.7-point improvement in their pain scores at 4 weeks, compared with only 25% for those in the control group, suggesting that a 6-day continuous peripheral nerve block decreases pain a clinically relevant amount in approximately 31% of treated patients. Increasing the ways in which data is conveyed assists stakeholders in interpreting and applying the study results to their own specific situation/population (figure 5).

Differing Data Presentations

Average phantom pain was a mean (SD) of 3.0 (2.9) in patients given local anesthetic vs 4.5 (2.6) in those given placebo: difference [95% confidence interval] 1.3 [0.4, 2.2]	
Average phantom pain was a median [interquartile] of 3.0 [0,6.0] in patients given local anesthetic vs 5.0 [3.0, 6.6] in the controls.	
Percentage of each treatment group with improvement ≥ 1.7 points	
Local anesthetic: 56%	Placebo: 25%
Percentage of each treatment group with complete resolution of pain	
Local anesthetic: 36%	Placebo: 15%
Percentage of each treatment group with mild pain (score < 3)	
Local anesthetic: 43%	Placebo: 21%

Figure 5 Examples of various presentation formats of the same dataset. Study participants included adult patients with postamputation phantom limb pain who all received a single-injection peripheral nerve block with ropivacaine and a perineural catheter(s). Participants were randomized to receive 6 days of either perineural ropivacaine or normal saline, and the primary outcome measure was the improvement in the average pain score queried 4 weeks following infusion initiation as measured using a 0–10 Numeric Rating Scale. Note that the combination of information improves interpretation of the results, and different formats may be valuable to differing stakeholders such as clinicians, patients, administrators, and other researchers. Adapted from Ilfeld *et al.*²¹ Some results are preliminary and should be used for demonstration purposes only (final analysis currently in preparation).

Importantly, different stakeholders often have differing priorities and therefore frequently draw dissimilar—or even opposite—conclusions from the same investigation. For instance, when presented with these results, a patient with severe phantom limb pain unchanged with any standard treatment may conclude that a 31% improvement represents an excellent alternative analgesic. Conversely, a chronic pain physician may consider these same results as unacceptable for a patient with osteoporosis given the theoretically increased risk of falling with a perineural infusion; and the available alternative of percutaneous peripheral nerve stimulation that does not induce a proprioception, sensory or motor block.²⁶ In contrast, an insurance administrator may conclude the analgesic improvement demonstrates a favorable cost–benefit ratio relative to nerve stimulation and warrants a trial of the perineural infusion.²⁷ **Applicability of study conclusions (positive or negative) to the innumerable populations and scenarios ‘must be conducted on a case-by-case**

basis, and are ideally informed by patients and their significant others, clinicians, researchers, statisticians, and representatives of society at large'.¹⁹

GRANT AND MANUSCRIPT REVIEWERS

There will often be differences in opinion regarding the clinical significance of a *group* difference even for the same population due to divergent priorities among the many consumers of research results, including patients, healthcare providers, administrators, insurance carriers, medication and device manufacturers, journal editors, etc. Although thoughtful sample size estimation based on objective criteria is required for prospective clinical research, it is essential that grant and manuscript reviewers acknowledge that while investigators can provide sound scientific justification of their reasoning, it is not possible to estimate a precise sample size that includes all conceivable variables and will produce results applicable to every plausible population and scenario.

Related to this issue is an appeal to manuscript reviewers and editors to permit some degree of outcomes-based editorializing within the abstract conclusion section. Frequently reviewers require authors to simply restate the results: 'This study found that contact with a living torpedo fish decreased gout-related pain.' Perhaps a more useful conclusion for stakeholders places the results—which they just read in the results section—within some context: 'While contact with a torpedo fish decreased gout-related pain, the significant number of severe electrical burns and poor survival rate when stored in a Pyxis Medstation may limit the use of torpedo fish as a first-line analgesic.'

CONCLUSIONS

It is not my intention to imply that there are no (partial) solutions in assessing the clinical relevance of *group* differences,²⁸ nor to provide a review article of available statistical analysis strategies. Rather, I hope to (1) better-elucidate for readers the smallest improvement that is important to *individual* patients versus the smallest difference considered important between *groups*; (2) draw attention to the erroneous assumption that these two are interchangeable; (3) encourage investigators to use joint hypothesis testing and gatekeeping as well as provide data in multiple formats that will assist stakeholder interpretation and generalization of the results; (4) appeal to reviewers to appreciate the imprecise nature of sample size estimation; and (5) advocate that stakeholders—be they clinicians, patients, administrators, or other researchers—take an active role in applying study results to their own specific situation/population. For readers seeking additional information, I highly recommend outstanding publications by Dworkin *et al*^{3 19 29} and Muñoz-Leyva *et al*,¹⁴ among many others.^{13 28 30–35} Even with beneficial methods such as the NNT and 50% of the SD ($\frac{1}{2}$ SD),²⁸ the nearly infinite combination of diverse populations and clinical scenarios essentially guarantees that there will be a good deal of art in addition to science when basing clinical care and administrative decisions on trial data. A $p < 0.05$ is often viewed as the 'conclusion' of the scientific process, when—in many respects—it is only the beginning.

Acknowledgements The author would like to thank all of his instructors, mentors, and colleagues, who—if listed—would require more pages than this article itself; and without whom this manuscript would never have been conceived and could not have been written.

Contributors BMI conceived and wrote this manuscript. This author is responsible for the overall content as the guarantor and accepts full responsibility for the work and controlled the decision to publish.

Funding The author has not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests The University of California has received funding and product for Dr. Ilfeld's research from Epimed International (Farmers Branch, TX); Avanos (Irvine, CA); Infutronics (Natick, MA); and SPR Therapeutics (Cleveland, OH).

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, an indication of whether changes were made, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Brian M Ilfeld <http://orcid.org/0000-0002-6144-3273>

REFERENCES

- 1 Anstie FE. *Stimulants and narcotics*. Philadelphia: Lindsay and Blakiston, 1865.
- 2 Stillings D. A survey of the history of electrical stimulation for pain to 1900. *Med Instrum* 1975;9:255–9.
- 3 Dworkin RH, Turk DC, Wyrwich KW, *et al*. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain* 2008;9:105–21.
- 4 Cella D, Bullinger M, Scott C, *et al*. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clin Proc* 2002;77:384–92.
- 5 Chin KJ. Erector spinae plane and paravertebral blocks have similar opioid-sparing effects following breast surgery. *Reg Anesth Pain Med* 2020:940–1.
- 6 Ilfeld BM, Donohue MC, Swisher MW. Response to: 'erector spinae plane and paravertebral blocks have similar opioid-sparing effects following breast surgery. *Reg Anesth Pain Med* 2020;45:941–2.
- 7 Clement ND, Bardgett M, Weir D, *et al*. What is the minimum clinically important difference for the WOMAC index after TKA? *Clin Orthop Relat Res* 2018;476:2005–14.
- 8 Farrar JT, Young JP, LaMoreaux L, *et al*. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001;94:149–58.
- 9 Arima H, Carreon LY, Glassman SD, *et al*. Cultural variations in the minimum clinically important difference thresholds for SRS-22R after surgery for adult spinal deformity. *Spine Deform* 2019;7:627–32.
- 10 Osborne JC, Green NS, Smaldone AM. Quality of life of Latino and non-Latino youth with sickle cell disease as reported by parents and youth. *Hisp Health Care Int* 2020;18:224–31.
- 11 Katajapuu N, Heinonen A, Saltychev M. Minimal clinically important difference and minimal detectable change of the world Health organization disability assessment schedule 2.0 (WHODAS 2.0) amongst patients with chronic musculoskeletal pain. *Clin Rehabil* 2020;34:1506–11.
- 12 Quttner AL, Modi AC, Wainwright C, *et al*. Determination of the minimal clinically important difference scores for the cystic fibrosis Questionnaire-Revised respiratory symptom scale in two populations of patients with cystic fibrosis and chronic *Pseudomonas aeruginosa* airway infection. *Chest* 2009;135:1610–8.
- 13 Testa MA. Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. *Med Care* 2000;38:1166–74.
- 14 Muñoz-Leyva F, El-Boghdady K, Chan V. Is the minimal clinically important difference (MCID) in acute pain a good measure of analgesic efficacy in regional anesthesia? *Reg Anesth Pain Med* 2020;45:1000–5.
- 15 Cepeda MS, Africano JM, Polo R, *et al*. What decline in pain intensity is meaningful to patients with acute pain? *Pain* 2003;105:151–7.
- 16 Myles PS, Myles DB, Gallagher W, *et al*. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. *Br J Anaesth* 2017;118:424–9.
- 17 Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. *J Pain* 2003;4:407–14.
- 18 Todd KH, Funk KG, Funk JP, *et al*. Clinical significance of reported changes in pain severity. *Ann Emerg Med* 1996;27:485–9.
- 19 Dworkin RH, Turk DC, McDermott MP, *et al*. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. *Pain* 2009;146:238–44.
- 20 U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome

- measures: use in medical product development to support labeling claims: draft guidance. *Health Qual Life Outcomes* 2006;4:79.
- 21 Ilfeld BM, Khatibi B, Maheshwari K, *et al.* Ambulatory continuous peripheral nerve blocks to treat postamputation phantom limb pain: a multicenter, randomized, quadruple-masked, placebo-controlled clinical trial. *Pain* 2021;162:938–55.
 - 22 Mascha EJ, Turan A. Joint hypothesis testing and gatekeeping procedures for studies with multiple endpoints. *Anesth Analg* 2012;114:1304–17.
 - 23 Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat Med* 2003;22:2387–400.
 - 24 Ilfeld BM, Khatibi B, Maheshwari K, *et al.* Immediate effects of a continuous peripheral nerve block on Postamputation phantom and residual limb pain: secondary outcomes from a multicenter randomized controlled clinical trial. *Anesth Analg* 2021;133:1019–27.
 - 25 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;310:452–4.
 - 26 Gilmore C, Ilfeld B, Rosenow J, *et al.* Percutaneous peripheral nerve stimulation for the treatment of chronic neuropathic postamputation pain: a multicenter, randomized, placebo-controlled trial. *Reg Anesth Pain Med* 2019;44:637–45.
 - 27 Ilfeld BM, Finneran JJ. Cryoneurolysis and percutaneous peripheral nerve stimulation to treat acute pain. *Anesthesiology* 2020;133:1127–49.
 - 28 Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. *COPD* 2005;2:57–62.
 - 29 Dworkin RH, McDermott MP, Raja SN. Preventing chronic postsurgical pain: how much of a difference makes a difference? *Anesthesiology* 2010;112:516–8.
 - 30 Wright A, Hannon J, Hegedus EJ, *et al.* Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther* 2012;20:160–6.
 - 31 Olsen MF, Bjerre E, Hansen MD, *et al.* Pain relief that matters to patients: systematic review of empirical studies assessing the minimum clinically important difference in acute pain. *BMC Med* 2017;15:35.
 - 32 ten Klooster PM, Drossaers-Bakker KW, Taal E, *et al.* Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. *Pain* 2006;121:151–7.
 - 33 Guyatt GH, Osoba D, Wu AW, *et al.* Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–83.
 - 34 Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;40:171–8.
 - 35 Gilron I, Jensen MP. Clinical trial methodology of pain treatment studies: selection and measurement of self-report primary outcomes for efficacy. *Reg Anesth Pain Med* 2011;36:374–81.