Journal of Bodywork & Movement Therapies 23 (2019) 690–694

Contents lists available at ScienceDirect

Journal of Bodywork & Movement Therapies

journal homepage: www.elsevier.com/jbmt

Narrative review

Commentary: Statistical significance and clinical significance - A call to consider patient reported outcome measures, effect size, confidence interval and minimal clinically important difference (MCID)

Michael Fleischmann [a, *], Brett Vaughan [b]

[a] College of Health and Biomedicine, Victoria University, Melbourne, Australia
[b] Department of Medical Education, University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 4 February 2019
Accepted 5 February 2019

Keywords:
Evidence based practice
Statistics
Outcomes
Measurement
Methodology
Clinical practice

ABSTRACT

In healthcare research an intervention may be statistically significant based on quantitative analysis; however, simultaneously it may be relatively insignificant to the health or quality of life of patients affected by a particular condition or disease being treated by the intervention – thus may be interpreted as having low clinical significance. An understanding of statistics is fundamental for evidence informed healthcare. Patient-reported outcome measures (PROMs) direct patients to evaluate aspects of their own health, including quality of life, disability and function. Data obtained from PROMs can be used to demonstrate the impact of healthcare interventions on these elements of a person's quality of life. To interpret outcome measure data for clinical decision making, a clinician must understand the concepts of statistical significance and clinical significance. This commentary explores the concepts of patient reported outcome measures (PROMs), their statistical and clinical significance, and explores their relationship with a practical example for the clinician. Limitations of research that only reports p-values and the need to consider effect size, confidence intervals, and minimal clinically important difference are also discussed. Together, these concepts can assist the clinician to evaluate whether an intervention may be suitable for their clinical practice.

1. Introduction

When reviewing the statistics of research and its interpreted outcomes, a clinician must consider three questions: 1. How reliable are the results? 2. Are the results due to chance? and, 3. Do the results matter to a patient? To answer these questions, an understanding of patient reported outcome measures (PROMs), statistical and clinical significance is needed.

1.1. Patient reported outcome measures (PROMs)

Patient-report outcome measures (PROMs) are used in clinical practice to provide a somewhat objective measurement of patient progress with respect to their management (Cella et al., 2010; Fleischmann and Vaughan, 2018). These measures are particularly valuable for demonstrating improvements in pain levels (Turk et al., 2006), activities of daily living (Cella et al., 2010) and functional activities, to both the patient and third-party payers (e.g. worker's compensation) (Blyth et al., 2003). However, Chiarotto et al. (2016) argue that PROMs lack content validity when measuring physical functioning in low back pain, and lack structural validity as this concept has received limited attention in the literature (Chiarotto et al., 2016, 2018). Notwithstanding, data obtained from PROMs can be used to demonstrate how healthcare interventions may affect various aspects of a person's quality of life and serve as a mechanism to monitor treatment effectiveness (Roach, 2006). Developing an understanding of how best to evaluate PROMs requires an understanding of the concepts of statistical significance and clinical significance to inform clinical decision-making. The purpose of the current commentary is first to explore the basic statistical concepts that can be applied to demonstrate these changes in PROM scores, and then highlights the importance of clinical significance in clinical practice.

* Corresponding author. College of Health & Biomedicine, City Flinders Lane Campus, Victoria University, PO Box 14428, Melbourne, Victoria, 3000, Australia.
E-mail address: michael.fleischmann@vu.edu.au (M. Fleischmann).

https://doi.org/10.1016/j.jbmt.2019.02.009
1360-8592/

2. Statistical significance

In most quantitative experiments, researchers investigate if there is a difference between intervention groups by performing statistical tests and reporting an associated p-value. Box 1 provides an example of how this approach may be applied in a research design.

Once researchers have performed statistical analysis (from Box 1 example) they compare the between group results (p-value) to an *a-priori* alpha level, a parameter designed to determine how likely the result was due to random chance. At this point it is important to consider the definition of a p-value. Ahlbom (2017) clarifies that a p-value is a conditional probability, based on the assumption that H_0 is true (condition) (Ahlbom, 2017). This is the most vital point to understand as a definitive conclusion cannot be drawn from the calculation of p-value alone as to whether H_0 is true or false. Hence, a p-value only measures the strength of evidence against the null hypothesis (H_0) (Dorey, 2010), but does provide a definitive conclusion - albeit it is often reported and/or assumed to be as much.

To further illustrate this point, consider a hypothesis test which yields $p = 0.83$. Probabilistically this demonstrates weaker evidence against the H_0 than a $p = 0.052$. Importantly though, a result of $p = 0.052$ is not statistically significant but could be considered approximate in value to a study with a $p = 0.048$ (which is statistically significant), but may have been influenced by a small sample size. In fact, the H_0 could be false in this case, thus confirming the alternative hypothesis. It is important to understand that even though small p-values demonstrate strong evidence that the H_0 is false, a very large p-value, such as $p = 0.92$, does not provide conclusive evidence that H_0 is true.

In most healthcare research, alpha is set at <0.05, meaning that there is a 5% or less probability ($p < 0.05$) that the result was due to random chance if the test hypothesis (H_1) is true (Cook, 2008; Silva-Ayçaguer et al., 2010). Therefore, if the calculated p-value is less than or equal to the alpha ($p < 0.05$), then the null hypothesis is rejected, and the result is interpreted as statistically significant. If the calculated p-value is greater than ($p > 0.05$), then failure to reject the H_o occurs and the result is statistically non-significant (n.s.) or 'negative'.

The calculation of the p-value increased in popularity since Fisher's publication of an influential book, Statistical Methods for Research Workers (1925) (Fisher, 1925). Statistical significance testing evolved from his idea: suppose we found a difference in the poverty level of those who are malnourished versus those who are not among children under the age of five years. Dahiru (2008) argues that this is indeed a finding, but posits that it could it be a finding due to chance (Dahiru, 2008). Fisher champions the calculation of the p-value because it provides a mathematical estimate of how likely it is that the results of a study were due to chance (Ahlbom, 2017).

Box 1
Sample research design.

Researchers wish to investigate if there is a difference in the effectiveness of two manual therapy techniques for neck pain; one group are to receive muscle energy technique (MET) and one group massage therapy (MT). The researchers would hypothesise that there is a difference in the effectiveness, which is termed the test hypothesis (H_1). Other examples could include investigating effectiveness of a new drug, manual therapy intervention, type of exercise, or other intervention. There is always the possibility that there is no difference between the groups. This lack of a difference is called the null hypothesis (N_O).

2.1. An example of statistical significance in manual therapy research

Consider the scenario of a research group planning to investigate if muscle energy technique (MET) reduces cervical spine pain more so than recommended combined education, mobilisation and supervised exercise (Childs et al., 2008; Tsakitzidis et al., 2013) (Box 2).

However, reporting of statistical significance alone can be misleading. Numerous authors have cautioned the sole reporting of p-values for over 20 years (Anderson and Burnham, 2002; Jacobson and Truax, 1991; Wasserstein and Lazar, 2016) although this is still commonplace in the literature. The p-value supports the concept that the observed differences in the interventions are unlikely to be due to chance. To provide some degree of certainty around the interpretation of the p-value, researchers frequently set confidence intervals (CI) at 95% (Brahman, 1991; Nakagawa and Cuthill, 2007). When putting this into practice, it is important to think that if the same population is sampled on numerous occasions and estimates are made on each occasion, the resulting intervals would remain within the estimates of the true population in 95% of the cases (Brahman, 1991). In other words, the CI provides a numerical boundary, consisting of a lower and upper case limit. This is because the true value of the population mean known as the point estimate, is unknown. So rather than using a point estimate (single value) from the sample, an interval estimate is used: meaning the sample value is used as a midpoint, and a boundary (lower and upper case limit) is provided. However, clinicians ought to note that the existence of a statistically significant difference has no direct mathematical bearing on the interpretation of its clinical significance (Nakagawa and Cuthill, 2007).

3. Clinical significance

3.1. Effect size

Absolute effect size provides an indication of the magnitude of the difference between the averages or mean, between the two interventions in the example described previously (Ferguson,

Box 2
Manual therapy research example

The researchers calculate the appropriate sample size, recruit participants, then divide the pool into two groups; one group receives an MET to C5/6 and, the comparison group receive education, mobilisation and self-care exercises. The dependent variable is the numerical rating scale (NRS) – a validated and reliable measure to rate pain (Farrar et al., 2001; Williamson and Hoggart, 2005). The researchers hypothesise that MET reduces cervical spine pain more than the control group, and state the null hypothesis as there being no difference between the two techniques. They set the alpha *a-priori* at 0.05 (like most manual therapy research) (Nakagawa and Cuthill, 2007). Assuming the characteristics of the group were similar, the researchers apply their intervention, and measure the NRS at the start and at the end - comparing the means scores from the NRS data between the MET and control groups at different time points. They would then calculate the test statistic and p-value, and evaluate whether it is below or above the alpha of 0.05. If the researchers calculated a p-value, for example of 0.02, the authors would report a statistically significant difference between the groups (Markel, 1991).

2009; Nakagawa and Cuthill, 2007). Used to quantify the difference between two groups, it has an advantage over the use of tests of statistical significance alone as it places the emphasis on the size of the difference (Nakagawa and Cuthill, 2007). Multiple authors posit that effect sizes should be reported alongside a p-value to provide an indication as to the clinical significance of the difference between interventions to inform clinicians, ultimately to the benefit of patients (Ferguson, 2009; Halsey et al., 2015; Lakens, 2013).

To establish whether a study has clinical significance, studies which collect quantitative data, such as data from a numerical rating scale (NRS) or visual analogue scale (VAS), may provide very useful information for clinicians to decide whether a treatment effect is large enough to make a difference to a patient (Sullivan and Feinn, 2012). This issue raises questions for the clinician, for example:

- How much of a decrease in pain using NRS is large enough to matter?
- How much improvement in function score on a disability questionnaire is enough to make a treatment worthwhile?
- How many treatments are needed in total to significantly reduce disability?

After all, clinicians want to make a difference.

3.2. Minimum clinically important difference

In 1989, Jaeschke et al. (1989) introduced the concept of minimum clinically important difference (MCID), which defines the smallest amount an outcome must change to be meaningful to patients during the use of PROMs. MCID is considered a patient-centred concept (Beaton et al., 2002; Copay et al., 2007), as it evaluates both the magnitude of the improvement and the importance patients place on the change (see Copay, 2007 for an in-depth review).

In a clinical setting, the MCID score can be utilised to provide parameters for data collected from outcome measures. For example, a clinician seeks to improve lower leg function for a patient. He/she then sets a goal for his/her patient to attain the MCID on the chosen PROM. This serves as an objective, measurable goal that is patient-centred, which can be useful for clinical decision-making. Considering another example, if a patient with chronic neck pain demonstrates no improvement on the Neck Disability Index (NDI) after a defined period of manual therapy treatment, clinical justification for continued care without change may be unwarranted. The MCID score is a key tool for the clinician as it provides another concept with which a clinician can establish alpha, calculate sample size, and assess the therapeutic outcomes to determine the effectiveness of an intervention (Sullivan and Feinn, 2012; Wright et al., 2012).

3.3. Minimum clinically important difference in clinical practice - the dilemma

When assessing the clinical significance of procedures intended to improve outcomes such as pain, the amount of improvement that is important to patients must be determined through use of PROMs. Clinicians must often use their own judgment when deciding how much improvement is clinically important solely based on individual patient feedback, particularly for scenarios where there is no evidence in the literature to provide these values. Researchers may explicitly provide the MCID and this value can be used to guide decision making regarding patient outcomes. However, where this value is not explicitly reported, the effect size and confidence interval can provide guidance (Sullivan and Feinn, 2012). Arguably, this information is not always immediately obvious, as it is usually only presented in the statistical analysis section of the methods part of a manuscript and often sometimes only depicted graphically, which can be confusing.

3.4. Two approaches to MCID

Generally, there are two methodological approaches to MCID: anchor-based and distribution-based (Copay et al., 2007; Jaeschke et al., 1989; Jayadevappa et al., 2012), which are conceptually very different (Barrett et al., 2005; Jaeschke et al., 1989; Norman et al., 2001). Wright et al. have discussed the lack of effectiveness of each approach, including the report that there is no established, standardised method for calculating MCID (Wright et al., 2012). However, many still consider MCID an important concept to determine whether an intervention improves perceived outcomes in patients.

There are various methods within the distribution-based approach, including standard deviation, standard error of measurement, effect size, and others (for a more detailed review: see Cook, 2008). However, it is important to consider the limitation of distribution-based methods; for example Wright et al. (2012) argues that effect sizes are based on standard deviation (SD), which is considered sample specific (Wright et al., 2012). Further, research has shown that distribution methods alone do not provide information regarding the clinical significance of the observed change (Wright et al., 2012).

Contrarily, anchor-based methods do not consider the measurement precision of the instrument of PROMs, but instead base clinical judgment on change, and are thus presumed to be sample-independent (Beaton et al., 2002; Coeytaux et al., 2006; Jaeschke et al., 1989; Norman et al., 2001; Redelmeier et al., 1996). Anchor-based methods are reported to examine the meaning of a degree of change, which is argued to be more patient orientated as it uses a single anchor (Wright et al., 2012). Hence, researchers have called for a combination of distribution and anchor-based methods to be used when determining MCID (de Vet et al., 2007; Swigris et al., 2010; Yost et al., 2011).

There are potential problems in defining a MCID when using the global rating of change (GROC) PROMs (Crosby et al., 2003). Often the problem is associated with patients' inability to understand the construct of improvement (Jaeschke et al., 1989). For example, although patients are asked to report on changes from his or her pre-intervention symptoms, Cook (2008) argues patients often report a current state of health as a comparison against expectations, rather than their own status (Cook, 2008). Further, patients may be subject to recall bias when completing PROMs to evaluate the nature of their prior condition (Beaton et al., 2002), and pain levels retrospectively (Pincus and Newman, 2001). Guyatt et al. (2002) argues that recall bias and patients' report of change in their condition is more directly related to their current health status, rather than the amount of change from a baseline value (Guyatt et al., 2002). Further, baseline severity of symptoms has been reported to effect the outcome of MCID (Copay et al., 2007).

(Cook, 2008) posits that clinicians can expect different MCID findings for the same outcome tool when examined on heterogeneous populations (Cook, 2008). When evaluating an intervention for those with neck pain, Cook (2008) notes that it is important to consider how similar but different conditions might influence MCID. For example, when considering a population with cervical pain as there are possible different biological contributors to pain, such as the presence or absence of radiculopathy (Schwind et al., 2013). Other forms of patient variation that can influence reporting of change include descriptive factors such as age,

socioeconomic status, or education, as shown in a low back pain study (Lauridsen et al., 2006).

It is likely that clinicians will face variable MCIDs when evaluating population characteristics or when using PROMs that were created in an unsound manner. In their works in 1989 about MCID, Jaeschke and colleagues provided two considerations associated with MCID (Jaeschke et al., 1989) which the clinician ought to consider. The first demands that the patient reports the central measure which is in line with the development of present day MCID development and implementation, and the measure of change must be reflective of a patient self-report measure versus a clinical finding or statistically significant change – a component of patient-centred care. The second consideration involves considering whether the findings are significant enough to change patient management. Questions may arise following the evaluation of these two important concepts, including how much patient-reported change is beneficial? How does one decide if the patient reported change is enough or not enough? Who decides if it is enough? A single anchor MCID approach may "make or break" a treatment approach if used in ("make") or out ("break") of proper context and advance clinical reasoning.

4. Conclusion

This commentary has argued that reporting p-values is insufficient in both the research and practice contexts as it provides clinicians with little clinically useful information to measure change in their patients. It is most clinically relevant if data are collected through the use of patient reported outcome measures (PROM), then effect size and MCID are reported so that clinicians can make evidence informed decisions for patient centred care. Clinicians need to consider the statistical significance and clinical significance of research findings that is the effect size and MCID when implementing evidence-based practice, and a patient-centred treatment model.

Statement of competing interests

The authors state that there are no competing interests.

References

Ahlbom, A., 2017. Significance testing: why does it prevail? Eur. J. Epidemiol. 32, 1–2.

Anderson, D.R., Burnham, K.P., 2002. Avoiding pitfalls when using information-theoretic methods. J. Wildl. Manag. 912–918.

Barrett, B., Brown, D., Mundt, M., Brown, R., 2005. Sufficiently important difference: expanding the framework of clinical significance. Med. Decis. Making 25, 250–261.

Beaton, D.E., Boers, M., Wells, G.A., 2002. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr. Opin. Rheumatol. 14, 109–114.

Blyth, F.M., March, L.M., Nicholas, M.K., Cousins, M.J., 2003. Chronic pain, work performance and litigation. Pain 103, 41–47.

Brahman, L.E., 1991. Confidence intervals assess both clinical significance and statistical significance. Ann. Intern. Med. 114, 515–517.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., 2010. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J. Clin. Epidemiol. 63, 1179–1194.

Chiarotto, A., Ostelo, R.W., Boers, M., Terwee, C.B., 2018. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. J. Clin. Epidemiol. 95, 73–93.

Chiarotto, A., Terwee, C.B., Ostelo, R.W., 2016. Choosing the right outcome measurement instruments for patients with low back pain. Best Pract. Res. Clin. Rheumatol. 30, 1003–1020.

Childs, J.D., Cleland, J.A., Elliott, J.M., Teyhen, D.S., Wainner, R.S., Whitman, J.M., Sopky, B.J., Godges, J.J., Flynn, T.W., Delitto, A., 2008. Neck pain: clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the American physical therapy association. J. Orthop. Sports Phys. Ther. 38, A1–A34.

Coeytaux, R.R., Kaufman, J.S., Chao, R., Mann, J.D., Devellis, R.F., 2006. Four methods of estimating the minimal important difference score were compared to establish a clinically significant change in Headache Impact Test. J. Clin. Epidemiol. 59, 374–380.

Cook, C.E., 2008. Clinimetrics corner: the minimal clinically important change score (MCID): a necessary pretense. J. Man. Manip. Ther. 16, 82E–83E.

Copay, A.G., Subach, B.R., Glassman, S.D., Polly Jr., D.W., Schuler, T.C., 2007. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 7, 541–546.

Crosby, R.D., Kolotkin, R.L., Williams, G.R., 2003. Defining clinically meaningful change in health-related quality of life. J. Clin. Epidemiol. 56, 395–407.

Dahiru, T., 2008. P – value, a true test of statistical significance? a cautionary note. Ann. Ib. Postgrad. Med. 6, 21–26.

de Vet, H.C., Ostelo, R.W., Terwee, C.B., van der Roer, N., Knol, D.L., Beckerman, H., Boers, M., Bouter, L.M., 2007. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual. Life Res. 16, 131–142.

Dorey, F., 2010. Brief: the P value: what is it and what does it tell you? Clin. Orthop. Relat. Res. 468, 2297–2298.

Farrar, J.T., Young, J.P., LaMoreaux, L., Werth, J.L., Poole, R.M., 2001. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 94, 149–158.

Ferguson, C.J., 2009. An effect size primer: a guide for clinicians and researchers. Prof. Psychol. Res. Pract. 40, 532.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Genesis Publishing Pvt Ltd.

Fleischmann, M., Vaughan, B., 2018. The challenges and opportunities of using patient reported outcome measures (PROMs) in clinical practice. Int. J. Osteopath. Med. 28, 56–61.

Guyatt, G.H., Osoba, D., Wu, A.W., Wyrwich, K.W., Norman, G.R., 2002. Methods to explain the clinical significance of health status measures. Mayo Clin. Proc. 77, 371–383.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., Drummond, G.B., 2015. The fickle P value generates irreproducible results. Nat. Methods 12, 179–185.

Jacobson, N.S., Truax, P., 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J. Consult. Clin. Psychol. 59, 12.

Jaeschke, R., Singer, J., Guyatt, G.H., 1989. Measurement of health status: ascertaining the minimal clinically important difference. Contr. Clin. Trials 10, 407–415.

Jayadevappa, R., Malkowicz, S.B., Wittink, M., Wein, A.J., Chhatre, S., 2012. Comparison of distribution- and anchor-based approaches to infer changes in health-related quality of life of prostate cancer survivors. Health Serv. Res. 47, 1902–1925.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863.

Lauridsen, H.H., Hartvigsen, J., Manniche, C., Korsholm, L., Grunnet-Nilsson, N., 2006. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Muscoskelet. Disord. 7, 82.

Markel, M.D., 1991. The power of a statistical test what does insignificance mean? Vet. Surg. 20, 209–214.

Nakagawa, S., Cuthill, I.C., 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol. Rev. 82, 591–605.

Norman, G.R., Sridhar, F.G., Guyatt, G.H., Walter, S.D., 2001. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med. Care 39, 1039–1047.

Pincus, T., Newman, S., 2001. Recall bias, pain, depression and cost in back pain patients. Br. J. Clin. Psychol. 40, 143–156.

Redelmeier, D.A., Guyatt, G.H., Goldstein, R.S., 1996. Assessing the minimal important difference in symptoms: a comparison of two techniques. J. Clin. Epidemiol. 49, 1215–1219.

Roach, K.E., 2006. Measurement of health outcomes: reliability, validity and responsiveness. JPO J. Prosthetics Orthot. 18, P8–P12.

Schwind, J., Learman, K., O'Halloran, B., Showalter, C., Cook, C., 2013. 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. 21, 71–78.

Silva-Ayçaguer, L.C., Suárez-Gil, P., Fernández-Somoano, A., 2010. The null hypothesis significance test in health sciences research (1995-2006): statistical analysis and interpretation. BMC Med. Res. Methodol. 10, 44-44.

Sullivan, G.M., Feinn, R., 2012. Using effect size—or why the P value is not enough. J. Grad. Med. Educ. 4, 279–282.

Swigris, J.J., Brown, K.K., Behr, J., du Bois, R.M., King, T.E., Raghu, G., Wamboldt, F.S., 2010. The SF-36 and SGRQ: validity and first look at minimum important differences in IPF. Respir. Med. 104, 296–304.

Tsakitzidis, G., Remmen, R., Dankaerts, W., Van Royen, P., 2013. Non-specific neck pain and evidence-based practice. Eur. Sci. J. ESJ 9.

Turk, D.C., Dworkin, R.H., Burke, L.B., Gershon, R., Rothman, M., Scott, J., Allen, R.R., Atkinson, H.J., Chandler, J., Cleeland, C., 2006. Developing patient-reported outcome measures for pain clinical trials: impact recommendations. Pain 125, 208–215.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's Statement on P-Values: Context, Process, and Purpose. Taylor & Francis.

Williamson, A., Hoggart, B., 2005. Pain: a review of three commonly used pain

rating scales. J. Clin. Nurs. 14, 798–804.
Wright, A., Hannon, J., Hegedus, E.J., Kavchak, A.E., 2012. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). J. Man. Manip. Ther. 20, 160–166.
Yost, K.J., Eton, D.T., Garcia, S.F., Cella, D., 2011. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J. Clin. Epidemiol. 64, 507–516.