97078_Revision_Auto_Edited.docx

WORD COUNT 3666
TIME SUBMITTED 23-OCT-2024 06:51PM
PAPER ID 112508656

Name of Journal: *World Journal of Methodology*

Manuscript NO: 97078

Manuscript Type: MINIREVIEWS

Variations in Quantifying Patient Reported Outcome Measures to Estimate Treatment Effect

Abstract

In the practice of healthcare, patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs) are used as an attempt to observe the changes in complex clinical situations. They guide us in making decisions based on the evidence regarding patient care by recording the change in outcomes for a particular treatment to a given condition and finally to understand whether a patient will benefit from a particular treatment and to quantify the treatment effect. For any PROM to be usable in healthcare, we need it to be reliable, encapsulating the points of interest with the potential to detect any real change. Using structured outcome measures routinely in clinical practice helps the physician to understand the functional limitation of a patient which would otherwise not be clear in an office interview and this will allow the physician and patient to have a meaningful conversation as well as a customized plan for each patient. Having mentioned the rationale and the benefits of PROMs, understanding the quantification process is crucial before embarking on management decisions. A better interpretation of change needs to identify the treatment effect based on clinical relevance for a given condition. There are a multiple set of measurement indices to serve this effect and most of them are used interchangeably without clear demarcation on their differences. This editorial details the various quantification metrics used to evaluate the treatment effect using PROMs, their limitations and the scope of usage and implementation in clinical practice.

Key Words: Patient Reported Outcome Measures; Treatment Effect; Minimal Clinical Important Difference; Patient Accepted Symptom State; Minimum Detectable Change; Orthopedics

Core Tip: In healthcare, patient-reported outcomes and patient-reported outcome measures (PROMs) help track changes in complex clinical situations. They provide evidence-based guidance for patient care by showing how a treatment affects a specific condition and if the patient benefits from it. For PROMs to be useful, they must be

reliable and able to detect real changes. Regular use of structured outcome measures helps doctors understand a patient's limitations better than just an office interview. This allows for meaningful discussions and personalized treatment plans. Understanding how to measure treatment effects with PROMs is crucial, as there are many different metrics, often used interchangeably. This editorial explains these metrics, their limitations, and their practical use in healthcare.

INTRODUCTION

In the practice of healthcare, patient reported outcomes (PROs) and patient reported outcome measures (PROMs) are used as an attempt to observe the changes in complex clinical situations. They guide us in making decision based on the evidence regarding patient care by recording the change in outcomes for a particular treatment to a given condition and finally to understand whether a patient will benefit from a particular treatment and to quantify the treatment effect[1]. According to the United States Food and Drug Administration (USFDA), PRO is any report coming directly from patients about a health condition and its treatment[2]. For any PROM to be usable in healthcare, we need it to be reliable, encapsulating the points of interest with a potential to detect any real change[3]. Using structured outcome measures routinely in clinical practice helps the physician to understand the functional limitation of a patient which would otherwise be not clear in an office interview and this will allow the physician and patient to have a meaningful conversation as well as a customized plan for each patient. The importance of serially and routinely measuring outcomes is stressed by Codman, the father of modern-day outcome assessment[4] and the rationale behind collecting PROs are as follows: Better communication aid that also makes the decision making process shared between patient and the provider; subjective assessment of health status and identify the treatment lacunae; quantify the loss of function; distinguish between problems due to physical, emotional and social reasons; identify adverse effects of treatment methods; estimate disease progression and treatment response; aid in change of treatment methods; prognostication of disease course and treatment outcomes. [5–7]

Having mentioning the rationale and the benefits of PROMs, understanding the quantification process is crucial before embarking on management decisions. Traditionally, statistical methods are used to see the difference before and after an intervention. However, statistical significance may not relate to clinical improvement[8]. A better interpretation of change needs to identify the treatment effect based on clinical relevance for a given condition. Thresholds to measure the clinical relevance or significance of change can be of three types[8]. First, the minimum difference to understand the clinical relevance below which it cannot be distinguished from the random error; second, the difference between the scores pre and post intervention which can be perceived as good or bad by the patient and finally, the difference that is perceived as clinically relevant or meaningful. There are several metrics to serve this purpose that were brought down for this purpose as shown in Table 1[8] and some of the most used are discussed in this review.

In a step towards understanding and standardizing the PROMs, Jaeschke and colleagues[9] have described the concept called minimal clinically important difference (MCID) to aid in interpreting the questionnaire scores. Following MCID, other metrics to assess the patient's perception of treatment effect were developed for interpreting PROMs as rightly called as "The Alphabet soup" by Tashjian[4] include patient acceptable symptomatic state (PASS), substantial clinical benefit (SCB) and maximal outcome improvement (MOI). This review analyses these commonly used evaluation metrics of PROMs to aid in better comprehension and implementation of these measures in clinical practice.

MCID

According to Jaeschke *et al*, MCID is defined as "the smallest difference in score in the *domain* of interest which *patients perceive* as beneficial and which would mandate, in the

absence of troublesome side effects and excessive cost, a change in the patient's *management.*" In this definition three important things to note include patient perception; absence of excessive cost and troublesome side effects; and mandating change in management[10]. As discussed in the Table 1, there are multiple terms similar to MCID but they are different in their own definitions. MCID can be estimated for an outcome measure by various methods[11,12]. We discuss the commonly employed methods to estimate MCID for a given PROM. Consensus method, anchor-based method, distribution-based method and a combination of anchor-based and distribution-based method[13].

Consensus method: In this method of assessment of MCID, an expert panel is convened, discussed and consensus is reached on the proposed MCID for the outcome of interest. The main problem in this method is that, patients' perspective is not taken into consideration.

Anchor-based method: Outcome scores are compared with an independent, external face valid criterion called "Anchor" to determine MCID of the particular outcome in question. Generally, anchor is a reliable and valid questionnaire for which patients respond based on their perspective[13]. Transition question, Patient global impression change (PGIC) or Patient global assessment (PGA) of treatment effectiveness are some of the examples of the anchor questions. The threshold used to calculate MCID of an anchor question is *minimally improved (it can be minimally deteriorated as well).* Gum *et al* have estimated MCID for back and leg pain for the NRS 0 – 10 scale using *somewhat better* as threshold and concluded that a decrease in the score of ≥ 3.1 were considered as minimally improved[14]. Anchor based method takes into account, patients' perspective for which these metrics were designed in the first place, unlike purely statistical approaches. . Second, this method cannot be used in conditions where most patients get better and the ones who are unchanged are minimal. Third, anchor questions don't take into consideration the variability in sample. Finally, the idea of MCID will only help in understanding whether there is an understandable improvement but not if that improvement is of any meaningfulness to the patient.

Distribution-based methods: Paradoxically, this method uses statistical means to measure MCID. One of such statistical means is to use standard error measurement (SEM) as it reflects the lack of precision in the measurement. So, any value below SEM cannot be MCID as this doesn't show any real change. Alternatively, the minimal detectable change (MDC) is calculated, which by definition is the smallest change that can be detected beyond the measurement error. In this method, MCID is considered as the upper value of 95% confidence interval (CI) of the average score in non-responders for a specific intervention. Usually MCID is shown to be on average similar to either 1 SEM or half of the standard deviation[19]. Gum *et al* and Carreon *et al* assessed MCID in lumbar fusion surgeries[14,20]. This method is also not without limitations. First, one can define MCID only based on the hope that change in the score is not due to the measurement error. Second, patient perspective is not accounted in this method.

Combination of distribution- and anchor-based methods: In this approach an anchor question is used to differentiate between the responders and non-responders and then use MDC to calculate the MCID as described above and using the upper value of the 95%CI in non-responders as MCID. Using this method, Parker *et al* determined the MCID of VAS (0-10) was 2.1 for neck pain but on validation by receiver-operator characteristics (ROC) analysis, the cut-off was found to be 4.1. However, for arm pain both the methods, MCID and ROC, resulted in 4.1 and 4.0 as cut-off points respectively. This method of assessing is much more complex compared to the previously described methods. However, it is advantageous compared to the pure anchor-based method being vulnerable to the sample variability.

Although MCID is useful to compare the efficacy of treatment in clinical trials and determine the efficacy of treatment in individual patients to inform treatment effect, there are some notable pitfalls. It includes the variability in the metric based on the quality of the data used, method used, anchor type, definition of improvement, population demographics, and their perception of symptoms and functional limitations. Further the weaknesses of using MCID involves the lack of universal fixed attribute that can be used across different patient populations. There is no consensus on the method

to calculate leading to extreme variability. In the study published by Ostelo *et al,* they found that depending on the method used to calculate MCID, Owestry disability score for low back pain, on a scale of 0-100 varied between 2 and 8.6[21]. Taking this data, Wright *et al* explain how this can be disastrous in clinical practice[3].

There is no single value for MCID for any specific outcome measure as it can be influence by type of patient population, method used to estimate. Further, MCID if reported as single point estimate rather than a confidence interval, it can be problematic as it can risk misclassifying the outcomes in patients as not improved even when they did. Rossi *et al.*[22], concluded that the calculation of MCID seem not as important as it seems. They reported MCID "a low bar" and recommended scientific studies to not only provide MCID but also mention PASS and SCB to be meaningful to both the scientific community and more importantly to the patient for whom a meaningful improvement makes sense than a minimum one as illustrated in Figure 1.

SCB

When assessing the clinical outcome of a patient, one does not reach a floor value like MCID but one would expect to reach a substantial clinical improvement, which being the substantial clinical benefit (SCB), a concept introduced by Glassman[23]. SCB is the minimum amount of change in a PROM that allows a patient to feel "sufficiently better" or "substantially better" after treatment. Generally, if MCID is considered the lower limit of treatment effect and SCB is the upper limit of any meaningful outcomes of a treatment.

SCB is measured based on anchor method as detailed for MCID. The commonly used anchor question is "Compared to the first evaluation, how is your physical condition now?" This question is usually answered using a Likert scale response[24]. Statistical analysis is done using various techniques to determine the SCB, but the most commonly used method is the ROC curve analysis. SCB values are determined for every particular PROM and for every condition distinctively.

Depending on the type of anchor questions that are used, there can be an issue of recall bias in calculating SCB as it is with MCID. Hubbard *et al* have used 2 anchor questions instead of one. While the first question was used to find out the improvement in the physical function since the first visit, the other question was used to assess SCB[24]. Similarly, Glassman *et al* have used 5 satisfaction statements in their study[23] to standardize the SCB determined. Although being very important, this metric SCB, is scant among the published literature[25]. However, Wellington *et al.* have shown that when patients were divided into people from different geographical locations or time, there was a high degree of variability in their SCB thresholds for total shoulder arthroplasty[26]. Hence, SCB also suffers variability as that of MCID based on population characteristics.

PASS

Unlike MCID which attempts to compare the pre-intervention and post-intervention scores for a given condition, PASS is a cross sectional evaluation of how the patient feels at a given point in time[27]. It is the magnitude of result that would make the patient feel fulfilled[4]. Several studies in the literature have proposed that overall improvement in the health status of patients is consider on of the crucial factors irrespective of the intervention for a given condition [28–30]. PASS is a holistic satisfaction score of the present health status of the patient and not just related to the symptoms of a particular disease or intervention[31].

PASS cut off point is estimated using anchor question which has a binary response of "Yes" or "No". One of the most common question that is used is ''Taking into account your level of pain and also your functional impairment, if you were to remain for the next few months as you are today, would you consider that your current state is satisfactory?''[32]. Along with anchor question, a 75th percentile method and ROC methods are used to reach a cut-off value for specific PROM for a given condition[33,34]. Considering MCID being called "a low bar" by Rossi *et al.,* PASS is

called as "an ambitious target for disease management" by Maksymowych *et al.*, thereby making it an interesting and something to look forward to in PROMs[22,35].

Figure 1 Illustration of the various quantification indices in the context of numerical pain rating scale for knee osteoarthritis.

Discussion

One of the most important thing that researchers and clinicians need to remember is that these metrics are not universal, they are usually specific for the condition that they are calculated for as well as the outcome measure that is used to calculate the change[36,37]. But, it is interesting to see that MCID is not different with the treatment used when the same condition is assessed with the same outcome measure[13,38]. Katz *et al.* in their review considered "*improvement is improvement regardless of what produced it*". On contrary to Farrar *et al.* who demonstrated on pain intensity – numerical rating scale (PI-NRS) a 11-point pain measurement instrument (0-10) similar MCID among a host of conditions such as osteoarthritis, painful diabetic neuropathy, low back pain andso. However, Stauffer *et al.* demonstrated that when different version of VAS (0-100mm) was used, MCID differed among different disease states such as knee, hip osteoarthritis and back pain[38,39].

Disease severity is another variable which influences these metrics. Patients with lower preoperative scores were easier to achieve MCID and SCB whereas those with higher scores were better off to reach a PASS[13,40]. Patients with a severe disease state have more room to reach the clinically important or significantly better state but those with a low severity have not enough room. So, in patients with low severity, it is difficult to define MCID or SCB as they don't have enough room to become better. Hence, one can consider patients with low health status or a severe disease with significant functional limitations to have a higher chance of achieving MCID or SCB after an intervention as the room to improvement is much higher but the chances of reaching PASS threshold remains unpredictable[40,41].

Use of these metrics namely MCID, SCB and PASS is not practical in a day to day clinical practice as each individual patient may not perceive the change in their health status in a similar manner. For example, if PASS threshold on 11-point PI-NRS is 3 for a specific condition and the patient in the consultation room with PI-NRS 2.5 might still not able to accept the present condition as satisfied[4]. However, Goh *et al.* described the PASS thresholds for multiple PROMs following unicompartmental knee arthroplasty, and recommended these thresholds of PASS as the target to treat the condition in future studies[27].

As already described, patients with higher functional scores or less severe disease status pre-intervention may not be able to reach MCID or SCB owing to the lack of enough room for improvement. In this regard, Berglund *et al* have come out with a new metric called maximal outcome improvement (MOI)[42]. MOI is a threshold for an outcome measure normalized to the maximum possible outcome for each patient who considers to have achieved a satisfactory result. Tashjian in his editorial commentary claimed MOI as the threshold which can be used at an individual patient level in day to day practice[4]. MCID, SCB and PASS are more meaningful when discussing the outcome of a group of patients than an individual patient. These metrics can also be used to assess the sample size, power of a study *etc.*, as well in the statistical aspect of the research which usually takes into consideration only the numbers, but with these metrics, we are introducing the aspect of patient perception of change or satisfaction.

Comparison of these metrics among different studies remains extremely difficult as there is a lack of consensus on their assessment methods[43,44]. Depending on the type of method that is used to assess the threshold, the value of these metrics can and will be different, making it a priority for the researchers to come to a consensus on their estimation methods. The type of anchor, number of questions and responses to be used, and identification of responses that are chosen as no, minimal or substantial change need to be ascertained since they have a significant impact on the evaluation metric that is calculated. Considering the amount of variability, achieving universal threshold for the PROMs does not seem to be in the horizon as of now. One of the ways in which this

variability could be managed is to define a range of measurement as threshold for these evaluation metrics rather than a single cut-off value[45]. Further, standardization of the methods to estimate these threshold ranges need to be developed to aid in universal acceptability and ease of use in both research as well as in day to day clinical practice[46,47]. Table 2 gives the list of common orthopaedic PROMs for hip, knee and shoulder ailments and their MCID and PASS cut-off values.

ASES, American Shoulder and Elbow Surgeons Score; DASH, Disabilities of the Arm, Shoulder, and Hand; FAAM, Functional Ankle Ability Measure; HAGOS, The Copenhagen Hip and Groin Outcome Score; HOOS, hip dysfunction and osteoarthritis outcome score; HOS, hip outcome score; iHOT-12, International Hip Outcome Tool-12; iHOT-33, International Hip Outcome Tool-33; IKDC-SKF, International Knee Documentation Committee Subjective Knee Form; KOOS, knee injury and osteoarthritis outcome score; MCID, minimal clinically important difference; MDC, minimal detectable change; mHHS, modified Harris Hip score; MIC, minimal important change; MOXFQ, Manchester-Oxford Foot Questionnaire; NAHS, nonarthritic hip score; NR, not yet reported in the literature; PASS, patient acceptable symptom state; PRWE, patient-rated wrist evaluation; SDC, smallest detectable change; SPADI, Shoulder Pain and Disability Index; SSS, sport-specific subscore; SST, simple shoulder test; WOMAC, Western Ontario and McMaster Universities Arthritis Index; WOOS, Western Ontario Osteoarthritis of the Shoulder Index; WORC, Western Ontario Rotator Cuff Index; WOSI, Western Ontario Shoulder Instability Index.

The findings of this study call for a unified approach in quantifying the patient reported outcomes and its treatment effect measure for a given condition for the benefit of the readers and researchers. The concept of core outcome dataset (COD) is being developed to emphasize this concept.[78,79] However, they were not put into action as a standard practice due to the lack of necessary reporting guidelines. Authors suggest journals to facilitate the necessary COD for a given condition as a necessary publishing requirement. Although not possible for all study methods, studies of higher clinical impact such as randomized controlled trials should be mandated towards the same.

Having tried to implement the COD concept and looking at its impracticality, the concept of minimum core outcome dataset (mCOD) is now in development for various clinical conditions. The impracticality of the idea lies in the regional differences in the context of outcome measures utilized. The outcome measures and their treatment effect noted to be relevant in one part of the world may not be relevant to the other and making them mandatory only makes them impractical. Hence, the concept of mCOD is in vogue to account for the regional, economic and cultural variations in outcome measurements.[80] Hence, the authors suggest that clinicians move towards a standard mCOD for the condition with a standardized measure of treatment effect to make the reported results meaningful to the readers and researchers in the present and future.

CONCLUSION

There is substantial variability in the estimation of treatment effect through indices such as MCID, SCB or PASS for a given intervention and patient population which prevents their generalizability. Hence, researchers and clinicians must exercise caution while utilizing these indices to their patient population to estimate the treatment effect for any given intervention. The author suggests utilization of mCOD for outcome selection and their recommended estimation of treatment effect for the given conditions to establish a standardized reporting method beneficial to global readers and researchers.

97078_Revision_Auto_Edited.docx

ORIGINALITY REPORT

5%
SIMILARITY INDEX

PRIMARY SOURCES

1	Nathaniel P Katz, Florence C Paillard, Evan Ekman. "Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions", Journal of Orthopaedic Surgery and Research, 2015 Crossref	50 words — 1%
2	c.coek.info Internet	41 words — 1%
3	www.proqolid.org Internet	28 words — 1%
4	Robert Tashjian. "Editorial Commentary: The Alphabet Soup of Understanding Clinical Shoulder Research: MCID (Minimal Clinically Important Difference), PASS (Patient Acceptable Symptomatic State), SCB (Substantial Clinical Benefit), and Now . . . MOI (Maximal Outcome Improvement)", Arthroscopy: The Journal of Arthroscopic & Related Surgery, 2020 Crossref	18 words — < 1%
5	bmchealthservres.biomedcentral.com Internet	15 words — < 1%
6	link.springer.com Internet	14 words — < 1%
7	pure.uva.nl Internet	

14 words — < 1%

8 www.slideshare.net
Internet

12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS