EXPERT REVIEWS

A point of minimal important difference (MID): a critique of terminology and methods

Expert Rev. Pharmacoeconomics Outcomes Res. 11(2), 171–184 (2011)

Madeleine T King

Psycho-oncology Co-operative Research Group (PoCoG), School of Psychology, Brennan MacCallum Building (A18), University of Sydney, NSW 2006, Australia
Tel.: +61 290 366 114
Fax: +61 290 365 292
madeleine.king@sydney.edu.au

The minimal important difference (MID) is a phrase with instant appeal in a field struggling to interpret health-related quality of life and other patient-reported outcomes. The terminology can be confusing, with several terms differing only slightly in definition (e.g., minimal clinically important difference, clinically important difference, minimally detectable difference, the subjectively significant difference), and others that seem similar despite having quite different meanings (minimally detectable difference versus minimum detectable change). Often, nuances of definition are of little consequence in the way that these quantities are estimated and used. Four methods are commonly employed to estimate MIDs: patient rating of change (global transition items); clinical anchors; standard error of measurement; and effect size. These are described and critiqued in this article. There is no universal MID, despite the appeal of the notion. Indeed, for a particular patient-reported outcome instrument or scale, the MID is not an immutable characteristic, but may vary by population and context. At both the group and individual level, the MID may depend on the clinical context and decision at hand, the baseline from which the patient starts, and whether they are improving or deteriorating. Specific estimates of MIDs should therefore not be overinterpreted. For a given health-related quality-of-life scale, all available MID estimates (and their confidence intervals) should be considered, amalgamated into general guidelines and applied judiciously to any particular clinical or research context.

Keywords: clinical significance • health-related quality of life • HRQOL • interpretation • MCID • MID • minimal clinically important difference • minimal important difference • patient-reported outcome • PRO

The 'minimal important difference' (MID) is a little phrase with big appeal in a field struggling to interpret health-related quality of life (HRQOL) and other patient-reported outcomes (PROs). It is a deceptively simple term; a nuanced understanding of terminology and methods is needed to avoid oversimplification and misuse as the phrase gains popularity in a field looking for a simple solution to a complex problem.

This article critiques the terminology and methods of the MID, providing a historical context for the various 'how to'-focused papers, which summarize methods and provide recommendations [1–4]. It is presented in six sections, addressing this series of questions: how are various MID-related terms defined and what is their historical sequence? What is the MID used for? Why are HRQOL results difficult to interpret? How is the MID usually determined? How does the MID differ from the smallest statistically detectable difference, and how does it link clinical importance with statistical significance, sample size and power? It concludes by speculating on future directions for the MID in the field of HRQOL and PRO research and practice. The articles selected are not based on a systematic search, but on the author's personal experience, reading and a literature search that grew organically from that.

Evolution of definitions & terminology

Table 1 summarizes the evolution of MID-related definitions and terminology. In 1987, Guyatt *et al.* proposed the minimal clinically important difference (MCID) as the appropriate benchmark of important change against which to assess the responsiveness of an instrument or scale [5]. They did not define the MCID, and acknowledged the difficulty of quantifying it, suggesting that the change induced by an intervention of known efficacy could provide an initial estimate [5]. A total of 2 years later, in perhaps the most influential paper in MID history, the MCID was defined by Jaeschke,

 10.1586/ERP.11.9 ISSN 1473-7167

Table 1. Evolution of key terms and definitions related to the minimal important difference, methods used to operationalize or quantify them, and key distinctions between them.

Study (year)	Term	Abbreviation	Definition	Method used and/or key distinctions	Ref.
Guyatt *et al.* (1987)	Minimal clinically important difference	MCID	MCID not defined, but used definition of responsiveness: 'the ability of evaluative instruments to detect minimal clinically important differences'	Change induced by an intervention of known efficacy	[5]
Jaeschke *et al.* (1989)	Minimal clinically important difference	MCID	The smallest difference that patients perceive as beneficial and that would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management	Global transition item ('how much has your <domain of HRQOL> changed in the past <time period> '), with the threshold based on the change in HRQOL (measured prospectively) in patients who report minimal change (on the global transition item), either for better or for worse	[6]
Osoba *et al.* (1998)	Subjectively significant difference	SSD	The smallest change, either beneficial or deleterious, that is perceptible (discernable) to the subject	As per Jaeschke *et al.* [6], the important distinction is in the definition: meaningfulness is based entirely on the patient's self-assessment of the magnitude of change (note that 'perceptible (discernable)' is similar to the 'detectable' from Normal *et al.* [8]	[9]
Guyatt *et al.* (2002)	Minimal important difference	MID	The smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and that would lead the clinician to consider a change in the patient's management	Methodology is not strictly prescribed; authors suggest corroboration across 'anchor- and distribution-based' methods. Authors note that the MID is the threshold between trivial and small-but-important change. Authors also note that 'subjectively significant' is a conceptually congruent alternative label for 'minimally important'	[1]
Schünemann *et al.* (2005)	Minimal important difference	MID	The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in the patient's management	Methodology is not strictly prescribed, but should be patient-based if possible (and while not specified, the definition implies those patients should be 'fully informed'). If proxies must be used, they should be instructed to focus on what they believe patients consider important (similarly, proxies should be 'fully informed')	[11]
Sloan *et al.* (2002)	Clinical significance		Goes beyond statistical significance to identify whether the statistically significant difference is large enough to have implications for patient care	Anchor- and distribution-based methods as described by Guyatt *et al.* [1] (the methods paper from the Clinical Significance Consensus Meeting Group of the Symposium on the Clinical Significance of Quality-of-Life Measures in Cancer Patients, Mayo Clinic [Rochester, MN, USA])	[66]

HRQOL: Health-related quality of life; R: Reliability of scale; SD: Standard deviation; SEM: Standard error of measurement.

Table 1. Evolution of key terms and definitions related to the minimal important difference, methods used to operationalize or quantify them, and key distinctions between them.

Study (year)	Term	Abbreviation	Definition	Method used and/or key distinctions	Ref.
Norman *et al.* (2003)	Clinically important differences	CID	Differences that are clinically important (as determined by the method of quantification), but not necessarily in any sense minimal	Anchor-based method involving longitudinal follow-up to determine whether subgroups can be identified that have clinically different outcomes, such as rehospitalization, relapse of cancer, Medical Research Council grading or different interventions	[8]
Wyrwich *et al.* (2005)	Clinically significant change		A difference score that is large enough to have an implication for the patient's treatment or care; sometimes corresponds to what a patient might recognize as a MID	Anchor- and distribution-based methods as described by Guyatt *et al.* [1]	[4]
De Vet *et al.* (2006)	Minimally important change	MIC	A change that patients would consider important to reach in their situation, dependent on baseline values or severity of disease, on the type of intervention, and on the duration of the follow-up period	Anchor-based methods are preferred, as they include a definition of what is minimally important	[67]
Norman *et al.* (2003)	Minimally detectable difference	MDD	As per Jaeschke *et al.* [6] – same definition, different term	As per Jaeschke *et al.* [6]. The important distinction is in the terminology: 'clinically important' is dropped in favor of 'detectable' to more accurately reflect the quantification method (i.e., patients who report minimal change on the global transition item)	[8]
Wyrwich *et al.* (1999)	Standard error of measurement	SEM	The standard error in an observed score that obscures the true score	$SEM = SD\sqrt{(1-r)}$ where SD = standard deviation of the sample and R = reliability of the scale A theoretically fixed psychometric property of an instrument or scale Takes into consideration the possibility that some of the observed change may be due to random measurement error	[38]
Beaton *et al.* (2001) and De Vet *et al.* (2006)	Minimum detectable change	MDC	Minimum change (at an individual level) detectable given the measurement error of the instrument (or scale)	$MDC(95\%\,confidence\,level) = 1.96 \times \sqrt{2} \times SEM$ where SEM as above, 1.96 derives from the 95% confidence interval of no change and √2 is included because two measurements are involved in measuring change (e.g., before and after an intervention or clinically significant event)	[14,15]

HRQOL: Health-related quality of life; R: Reliability of scale; SD: Standard deviation; SEM: Standard error of measurement.

Table 1. Evolution of key terms and definitions related to the minimal important difference, methods used to operationalize or quantify them, and key distinctions between them.

Study (year)	Term	Abbreviation	Definition	Method used and/or key distinctions	Ref.
Beckerman *et al.* (2001)	Smallest real difference	SRD	The smallest measurement change, that can be interpreted as a real difference (i.e., beyond zero), considering chance variation or measurement error	$SRD = 1.96 \times \sqrt{2} \times SEM$ (= MDC above)	[68]
Angst *et al.* (2001)	Smallest statistically detectable difference	SDD	The smallest mean change over time (within a group) which is statistically significantly different from zero	For a given sample size of n (number of patients for whom change is measured), two-sided type I error rate (α) and power (1-β, where β = one-sided type II error rate): $SDD = SD(z_\alpha + z_\beta)/\sqrt{(n/2)}$ where z_α and z_β are the values of the standard normal distribution (mean = 0, SD = 1) for α and β, respectively	[29]

HRQOL: Health-related quality of life; R: Reliability of scale; SD: Standard deviation; SEM: Standard error of measurement.

Singer and Guyatt as "the smallest difference 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" [6]. This definition planted the MCID firmly in a shared decision-making context. In 1993, in one of the most widely cited papers on HRQOL interpretation, Lydick and Epstein commended Jaeschke *et al.* on their 'wonderful' definition of the MCID, but noted that they "do not directly suggest an operational method for defining clinical meaningfulness" [7]. Jaeschke *et al.* did, however, allude to the fact that "clinicians who gain experience with a questionnaire develop a sense of the importance of changes seen in their patients' scores". In 2003, Norman *et al.* further noted that: "Nowhere in the operationalization of the MID approach is there a consideration of importance, or of the tradeoff between benefit and side effects or costs … Thus, the criterion may be more appropriately thought of as a minimally detectable difference (MDD)" [8]. Nevertheless, Jaeschke *et al.*'s method has become the standard for determining what people now typically call the MID.

In 1998, Osoba *et al.* used Jaeschke's global transition method, and coined the term subjectively significant difference (SSD), emphasizing the patient's self-assessment of the magnitude of change [9]. They defined the SSD as "the smallest change, either beneficial or deleterious, that is perceptible (discernable) to the subject". Although this term is not widely used, the paper is widely cited in cancer research as the basis of 10-point rule of thumb for the MID for the scales of the EORTC's core HRQOL questionnaire, QLQ-C30 [10].

In 2002, Guyatt *et al.* offered a nuanced definition of the MCID, rebranding it the MID: "the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient's management" [1]. Even though the 'C' for 'clinical' was removed, it is still implicit in this definition. A total of 3 years later, Schünemann and Gyuatt added further qualifications: that patients should be informed; that proxy assessment may be used; and that proxies should also be informed [11].

In a summary of the findings of the 2002 Symposium on Clinical Significance of Quality of Life Measures in Cancer Patients, Wyrwich *et al.* described a clinically significant change in quality of life (QOL) as "a difference score that is large enough to have an implication for the patient's treatment or care" [4]. They noted that it sometimes might correspond to what a patient might recognize as a MID in HRQOL/PRO scores, but that such a change may not lead to a change in treatment or care regimen if it represented an improvement that should not be interrupted or a decline that cannot be prevented by other reasonable alternatives, particularly in advanced-stage disease where palliation is often the focus of cancer treatment [4]. As explored further in the 'Clinical anchors' section later, clinical anchors can be used to determine clinically important differences (CIDs) in HRQOL and PROs, but the extent to which these are MCIDs will depend on the anchor selected, how adjacent groups are defined within that anchor, and the strength of the relationship (conceptually and empirically) between the anchor and the target HRQOL domain.

The MID is becoming the dominant term in this literature, and the usage of this term demonstrates that the differences in terminology, definitions and methods are of little real consequence. For example, Ringash *et al.* reported MIDs from two studies that used exactly the same methods but whose stated aims differed: one was "to estimate the magnitude of difference in QOL that is noticeable to patients" [12], while the other was to "determine what magnitude of change in a patient-reported outcome score is clinically meaningful" [13]. Clearly there is a looseness in the

usage of MID-related terms that readers need to be aware of; Table 1 helps to make explicit the differences and similarities in terms, definitions and methods.

'Minimally detectable': in what sense?

A confusing aspect of the MID-related definitions and terminology arises from the fact that a difference in HRQOL may be 'minimally detectable' in two senses. One is owing to the limits of perception and relates to the use of a global transition item. If we consider only those patients who felt they had got better or worse by the smallest possible increment, then this is the MDD in the sense used by Norman *et al.* [8] (as explained previously). Indeed, Norman *et al.* explain their finding of an apparently universal MID being equivalent to an effect size of 0.5 standard deviation (SD) by reference to psychophysiological evidence that the human limit of cognitive discrimination is approximately one part in seven, which in many empirical circumstances is very close to half a SD. This sense is also captured in Osoba *et al.*'s term SSD, which also relies on the global transition anchor method.

The other sense of 'minimally detectable' is in terms of the limits of measurement – that imposed by measurement error, which relates to the standard error of measurement (SEM), as its name implies. It is in this sense that Beaton *et al.* [14] and de Vet *et al.* [15] use the term 'minimum detectable change' (MDC): the amount of individual-level change that must be observed before it is considered above the bounds of measurement error. In other words, it is the threshold at and above which the individual change observed on a particular scale (with fixed SEM) reflects real change in the underlying (latent) domain of interest. The importance of appreciating the distinction between the MDC and the MID is put succinctly by de Vet *et al.* [16]: "to judge whether the minimally detectable change of a measurement instrument is sufficiently small to detect minimally important changes". Since the SEM is a theoretically fixed psychometric property of an instrument or scale, then so is the MDC (as a function of SEM; Table 1).

Group versus individual differences

An important but sometimes obscure distinction in the MID literature is that of group-level differences versus within-individual changes. This distinction is explored further in the following section. In MID-related literature, the term 'MDD' is typically used in relation to the former, while 'MDC' is used in relation to the latter. If the MDC is larger than the MCID, then the measure is insufficiently precise for individual monitoring. This is of increasing relevance with emerging interest in using PROs and HRQOL scores in monitoring and managing individual patients [17]. However, this issue is immaterial at the group level, where required levels of precision for mean differences (reflected in the standard error of the mean) are provided by adequately powered sample sizes (whether mean differences between groups or mean change within groups). This issue is considered further in the section entitled 'Methods used to determine the MID'.

Uses of the MID: decision-making at the individual & group levels

Health-related quality of life and PRO questionnaires have the potential to play a key role in bringing the patient's voice to evidence-based healthcare. However, to realize this potential, we need to be able to interpret the relevance of PROs in making decisions about treatment. Such decisions are made at both the individual level, when a patient (or their clinician, acting as their agent) chooses among treatment options or decides to cease or reduce treatment, and at the group level, when clinical research is conducted to test the relative effectiveness of treatments, often testing a promising new treatment against current best practice. At both of these levels, we need to know how much of a difference in PRO or HRQOL scores matters. The difficulty is working out to whom it should matter and in what sense it should matter.

Use of the MID in shared decision-making

At the individual level, when managing and monitoring patients in routine care, we need to know how much change in HRQOL is sufficient to warrant a change in treatment, whether starting a new treatment, continuing or stopping a current treatment, or increasing or decreasing the dose. Clearly this will vary across treatment contexts, and will often involve balancing benefits against side effects, inconvenience, financial costs to the patient and other less concrete costs. In the case of a treatment aimed at slowing the progression of Parkinson's disease, if deterioration in mobility is too rapid on the current dose, the decision may be made to increase the dosage, despite side effects. In the case of palliative radiotherapy for bone metastases, it may be that an improvement in pain is needed before treatment is ceased. In the case of adjuvant chemotherapy, it may involve the trade-off of likely survival gains against the HRQOL consequences of the toxicity burden. In most cases, each aspect of benefit, harm or cost will have a threshold beyond which the treatment decision will tip one way or the other. This may be a complex decision, involving the balancing the benefits and downsides of the treatment. How these are balanced will differ from patient to patient. Ideally, each patient's decision will be made at the point that best matches that patient's preferences.

Use of the MID in research

The decision context at the group level is quite different. Typically, a randomized trial is conducted to determine the relative efficacy of two treatment options, with patients randomized to treatment. Such trials provide robust evidence to guide policy at the health service provision level and practice at the individual patient management level. How is the final decision about which is the best treatment made in clinical trials? Analogous to individual-level decisions, it will involve an often complex balancing of benefits and harms. But at the group level, we use hypothesis testing and statistical analysis; the section of this article entitled 'Statistical significance, sample size, power and the smallest statistically significant difference' explores how the MID relates to these. The question still remains: what are the appropriate thresholds to tip a decision one way or another? It seems reasonable to use the average of individual patient's thresholds (MIDs).

Another use of the MIDs at the group level is in responder analysis, and similarly, the presentation of results in terms of proportion of patients that have improved, remained stable or deteriorated. This has been recommended by various influential authors as a means to present group-based results in a way that is more meaningful to clinicians [18–20]. Yost *et al.* have suggested that the upper end of a MID range be used in such cases to account for the higher level of measurement error for an individual change score [21].

The problem of interpreting HRQOL & PRO scores

Across these various uses of the MID, the underlying challenge is meaningful interpretation of patient-reported scores, in particular, the threshold that represents the smallest difference or change that tips a particular treatment decision one way or the other. A recent review found that HRQOL results were rarely interpreted in terms of clinical significance, even in randomized controlled trials reported to a high standard [10]. So while the clinical trials community has accepted the validity and feasibility of HRQOL and PRO assessment, and while such end points are increasingly used in clinical trials, a troubling lack of competence in making sense of the results still persists.

Why is it so hard to interpret HRQOL & PRO data?

Many complex factors confound our understanding of HRQOL and PROs. First, HRQOL is intrinsically a subjective phenomenon, a perception, reliant on self-report. It may mean different things to different people, and therefore defies definition (which is partly why we have moved to the less problematic term 'PRO'). Second, a particular individual's perception of their HRQOL may vary over time as their circumstance and perspective changes. Indeed, the capacity to adapt psychologically to loss of health is a boon to the individual, but the consequent 'response shift' in their self-report of their health and QOL [22,23] is one of the great challenges to interpreting changes in HRQOL data. Third, HRQOL is an umbrella term that covers a wide range of health-related phenomena (or 'constructs'), including physical, social and emotional functioning and a variety of symptoms of disease and side effects of treatment. These are measured by a vast number of questionnaires (or 'instruments'), each of which may contain several domain-specific scales. Each scale is intrinsically different because it includes a unique set of questions (or 'items'). All these scales are somewhat arbitrary in terms of their numeric values because there is no absolute zero or standard scalar increment for phenomena such as pain, fatigue and social function. Fourth, the response options on HRQOL items are ordinal. A fairly typical set of options is: 1 = not at all; 2 = a little; 3 = quite a bit; 4 = very much. These numbers do not have interval properties, that is, the difference between 'not at all' and 'quite a bit' may not be the same as the difference between 'a little' and 'very much', which may not be twice as big as the difference between 'not at all' and 'a little'. This is true for all such self-report scales, including numbered scales anchored by two phrases such as 'none at all' and 'worst imaginable'. Fifth, there can be differences among individuals in the way they use these response scales. For example, a person who has a low threshold for pain or fatigue may rate their level as 'quite a bit', while a more stoic individual may rate the same level as 'a little' pain, and a person's pain threshold may increase with their experience of chronic pain. Sixth, questions are often aggregated into multi-item scales, and the scores from individuals are aggregated into group-level results, often presented as mean HRQOL scores. Each step away from the content of the particular questions in the scale represents a further abstraction. Finally, few people have the requisite understanding of psychometrics or the hands-on experience with HRQOL and PRO assessment methods to confidently interpret the results that arise from specific scales, and there are surprisingly few interpretation manuals available. The generic short-form health survey (SF-36) and the cancer-specific Functional Assessment of Cancer Therapy (FACT-G) provide rare exceptions [24,25].

It is therefore surprising that any sense can be made of HRQOL and PRO data. Yet, despite the odds, when well-developed and validated HRQOL and PRO questionnaires are used, remarkably sensible patterns are apparent in the resultant data. For example, when the QLQ-C30 was used to measure the HRQOL of cancer patients, those with more advanced disease typically reported more symptoms and a poorer QOL across a range of functional domains compared with those with less advanced disease [26], and when the SF-36 was used in a large population sample, the group of people who developed a new long-term health condition on average reported a decline in all but one domain of HRQOL [27]. These patterns inspired confidence that HRQOL data could be interpreted meaningfully. In 1993, Lydick and Epstein provided an insightful review and influential taxonomy of methods for interpreting QOL results [7]. A subset of these now persist as methods used to determine the MID. The following section describes these, and gives some historical perspective on each one.

Methods used to determine the MID

This section describes the four methods that are historically and currently the most commonly used methods for determining the MID, and includes some other less widely used methods. The first and fourth of these described are typically called 'anchor-based' and the second and third are called 'distribution-based', after Lydick and Epstein [7], or are alternatively termed 'externally-referenced' and 'internally-referenced', respectively [28].

Global transition questions

Patient retrospective rating of change using a global transition question (as the 'anchor' or 'external reference') was first reported in 1989 [6], and has become the most commonly used method for determining the MID. The PRO is assessed prospectively at two time points, at the second of which the subject is also asked to think back to the first time point and judge the degree of change in that particular outcome, using a single item that has a series of graded options, often this five-point version: 'much worse', 'a little worse', 'the same/ no change', 'a little better' and 'much better'. For multidomain HRQOL instruments, this is typically performed by domain. So, for example, if the MID for an emotional functioning scale is to be determined, the global change question would ask about the degree of change in emotional

functioning since the previous HRQOL assessment time point, and this would be linked with the prospectively measured change in emotional function.

Typically, the mean change of the groups that differ by 'a little' is taken as the estimate of the MID. Some authors estimate separate MIDs for improvement and deterioration, and adjust the slightly better/worse results by subtracting the mean change that occurs in the 'no-change' group, for example Angst *et al.* [29]. The latter correction, similar to the adjacent category mean difference method of Cella *et al.* [30] and Maringwa *et al.* [31], is not universally accepted. While Hays *et al.* support the practice of comparing the change in HRQOL for individuals that have been deemed to change by a minimal amount with the change observed for those who are deemed to have stayed the same (not changed), they do not support the subtraction of the latter from the former [2]. Rather, they recommend that if the mean change for the no-change group is similar to that of the minimally changed group, then the MID estimate is suspect. However, if the MID change exceeds that of the no-change group, the MID estimate is useful and does not need to be adjusted by the HRQOL change observed in the no-change group. For example, if the minimally important change group is found to have an average change in HRQOL of four points versus two points for the no-change group, then the four points is the estimated MID and two points is not enough to constitute a MID.

Note that a change deemed to be 'a little better/worse' is not explicitly important or significant in any sense, which is why Osoba *et al.* called it the SSD [9]. Such thresholds are certainly relevant to the communication between patient and clinician because they represent the degree of change where patients begin to notice an improvement or decline; clearly anything smaller cannot be relevant to the therapeutic encounter. This method has some limitations. First, because judgements are retrospective, they may be prone to response shift and recall bias [32]. Second, patients' retrospective estimates of change are more highly correlated with their present state than with their change in health state [27,33]; this has been confirmed in cognitive interviews [34]. Third, their validity as measures of change has not been formally evaluated, and fourth, they are single items and so are more prone to measurement error than multi-item scales are. Fifth, when transition scales contain more than the five possible options already listed, the cut-points used to define the MID group are somewhat arbitrary, and it is often assumed that change related to an improvement is the same as that for a decline. For example, four change groups were defined in a 15-point transition scale: trivial (- 1, 0 or 1), minimal (2, 3 or -2, -3), moderate (4, 5 or -4, -5) and large (6, 7 or -6, -7) by Metz *et al.* [35].

Finally, two points should be noted about the results of this method. First, they demonstrate that a lot of variation exists among individuals, as illustrated in figure 1 from Osoba *et al.* [9] and figure 2 of Knox and King [27]. So while the means of the various change groups generally follow the expected trend (largest mean deterioration in HRQOL in the group, which felt very much worse, through to the largest mean improvement in HRQOL in the group, which felt very much better), in each group there are likely to be at least some individuals whose prospectively measured change scores contradict their retrospectively assessed global change, and this may be a significant proportion of patients in the smallest change groups. It is unclear the extent to which this reflects the truth of how these individuals felt versus measurement error and other limitations of this method previously described.

A related issue is that sample sizes in each change group are often quite small (as the total sample size is divided into five or seven change groups), so corresponding mean change scores tend to have large confidence intervals. For example, several of the 95% confidence intervals on the mean change for the two smallest change groups in Osoba *et al.*'s figure 1 include zero. Despite this, the ballpark message from Osoba *et al.*'s article, which now echoes throughout the literature [10], is that a ten-point change is the MID, regardless of clinical context or HRQOL domain. This demonstrates that simple messages resonate more readily in the research literature than complex ones, and thereby become embedded in research practice.

Individual variation is expected in all biological phenomena, and the use of a mean MID in clinical research to calculate sample size and interpret aggregate results is consistent with practice for objective health outcomes. However, when using HRQOL to monitor and manage individual patients, the subjective and multidimensional nature of HRQOL and the personalized trade-offs that patients make mandate that patient's opinions and preferences should be sought and considered if decision-making is to be truly shared with the clinician.

Increasingly, global transition questions are being used in another way to determine MIDs via receiver-operator curve (ROC) analysis, as described in the section on 'Other less commonly used methods'. De Vet *et al.* provide an example, illustrating how the ROC approach can be used to address the question: 'How sure we are that this MID value holds for every patient?' [36].

Standard error of measurement

Another commonly used approximation for the MID is the SEM, a theoretically fixed psychometric property of a HRQOL or PRO scale. Conceptually, the SEM is a measure of the spread of observed scores of a notional individual around their true score, had that patient been repeatedly assessed on the same measurement scale, with no memory and/or response effects and while having the same underlying HRQOL or other target PRO. The estimation of the SEM does not involve a patient or proxy's input about whether a change is minimally important in any sense; it has no 'anchor' or 'external-reference point'. Thus the SEM is not really a method for estimating the MID; it is merely a convenient proxy for the MID, easily calculated from available data or published estimates of between-person SD and scale reliability (r):

$$SEM = SD\sqrt{(1-r)}$$

While some argue that test–retest reliability should be used [37], others make the case that Cronbach's α is suitable for HRQOL-related phenomena, particularly those that are highly fluid even over short time periods [38].

Historically, the SEM was used in the field of 'individual differences' in psychology. A confidence interval was constructed around an individual's observed score using the standard normal-based approximation of 68% confidence within 1 SEM and 95% confidence within SEM (or more accurately 1.96 SEM). This indicated the limits beyond which an observed change was likely to reliably reflect true change, as opposed to being an artefact of measurement error. It could also be used to determine whether an individual whose observed score fell close to a cut-point really fell above or below it. This approach is illustrated for HRQOL measures by McHorney and Tarlov in assessing whether five commonly used instruments are sufficiently reliable for monitoring and managing individual patients [37].

The SEM is now used in the HRQOL field as a convenient criterion for estimating MIDs, following validation of this approach by Wyrwich and colleagues for various measures [38–40]. These validation analyses were based on a small set of studies that compared the SEM with established MID thresholds; three studies suggested that the SEM was about the same size as the MID, while the other three suggested the MID was more than twice as large as the SEM. Wyrwich reconciled these apparent differences by considering the extent of change considered to be minimally clinically important; in the former three studies it was change ratings of 'a little better' or 'somewhat better', while in one of the latter three studies (the only one that used a global transition item as the anchor), the MCID was based on patients who felt 'a good deal better' to 'a very great deal better'. Wyrwich concluded that, in both cases, one SEM was equivalent to the change experienced by patients who felt 'a little/somewhat better' [39]. The growing number of articles reporting the SEM alongside other MID estimates provides the opportunity to further assess the generalizability of this relationship. For example, Turner *et al.* recently reported that one SEM provided a reasonable approximation to anchor-based estimates of the MID for two respiratory questionnaires [41].

Effect size

The effect size (ES) is the most general approach to MID determination and, like the SEM, it has no external reference point or anchor for interpretation. It is a 'signal-to-noise ratio': the mean difference (or change) in HRQOL divided by the variability among individuals (SD). Two ES summary statistics are commonly used to estimate the MID: a fifth and a half of a SD. The convention in the MID literature is to use the between-person SD, typically at baseline, perhaps influenced by Kaziz *et al.* who recommended this in 1989 as an aid to interpretation [42]. This statistic is also called the standardized mean difference and Cohen's D [43] and, although it is just one of many variants of the very general notion of an ES, it is commonly 'the' ES in the HRQOL literature, again probably influenced by the Kaziz paper (note its title) [42]. Further confusion in terminology can arise because of the more general medical use of the term as a synonym for 'intervention effect' or 'effect estimate'.

Historically, the ES provided a solution to the problem of interpretation across numerous scales, as needed in meta-analysis, where the same PRO (such as depression or pain) is measured on different scales. The ES standardizes all scales to a common metric; because the numerator (mean difference) and denominator (SD) are both in the same measurement units (the particular HRQOL scale), their ratio is unit-less or scale-free. This allows PRO or HRQOL effects measured on different scales to be directly compared in terms of the variability among individuals, or 'standard deviation units'. However, like the SEM, it does not answer the question of whether a difference is minimally important in any sense. For example, if a new treatment that shifts the mean HRQOL by, say, half a SD, while the current best treatment shifts it by only a third of a SD, should we update policy and practice to the new intervention?

In order to address this question, Norman *et al.* conducted a systematic review of studies that computed a MID and contained sufficient information to compute an ES [8]. A total of 38 studies yielded 62 ESs with an average ES for the MID of 0.48. They consequently proposed that an ES of 0.5 be adopted as a universal standard MID. In a reanalysis of the same data, Farivar *et al.* reported a somewhat lower average ES (0.42) due to different assumptions, inclusions and exclusions, and noted the wide variation among studies (range: 0.11–2.3) [44]. So while a universal standard MID is appealing in its simplicity, opinion remains divided about the accuracy and utility of such a generalization [44–46].

Many years ago, Cohen proposed operational definitions of small, medium and large ESs for the standardized mean difference of 0.2, 0.5 and 0.8, respectively [43]. As the title of his well-known book (still widely available 41 years after its first publication) suggests, he was motivated by the prevalence of underpowered studies in the social sciences. His guidelines are now widely used in healthcare research, not only to calculate sample sizes suitably powered to test hypotheses (as he intended), but also to interpret results. Interestingly, Cohen described his guidelines as 'arbitrary conventions, recommended for use only when no better basis for estimating the effect size is available' [25,43]. While this caveat generally seems to have been overlooked, some researchers have taken up the challenge of developing evidence-based ESs. King *et al.* used an innovative method combining systematic review of published studies, expert opinion and meta-analysis to address this issue for the widely used cancer-specific HRQOL questionnaire, the FACT-G [47]. For some domain scales, the evidence-based ESs were considerably larger than Cohen's guidelines, in which case use of Cohen's guidelines would lead to overpowered studies and to over-interpretation of the clinical significance of an observed effect. For other domain scales, evidence-based ESs were considerably smaller than Cohen's thresholds; in these cases, use of Cohen's guidelines would lead to underpowered studies and inconclusive results. King *et al.*'s results also revealed variation between cross-sectional and longitudinal results, and between domains of HRQOL. Similar conclusions were reached by Cocks *et al.* who undertook a similar exercise for the EORTC's QLQ-C30 [48]. As the ES is signal-to-noise ratio, such variations may be driven by differences in both the signal detected by the scale (reflected in the mean differences) and the variability among individuals (reflected in the SD).

Relationship of SEM & ES, & limitations for MID estimation

As noted by Wyrwich *et al.* [4], there is a relationship between the SEM and the ES; the higher reliability, the lower the ES needed

to achieve a MID. For example, for a measure with reliability of 0.75, 1 SEM implies an ES of 0.50, while for a measure with higher reliability of 0.96, 1 SEM implies an ES of 0.20. However, as noted by Hays *et al.* [2], neither ESs nor SEM provide information about the size of a difference or change in a measure that is minimally important. Evidence such as that collated by Norman *et al.* [8] and Wyrwich [39] has been used to make the case that simple guidelines generalize across measures, and therefore ESs and SEMs can be used as convenient proxies of MIDs. Hays *et al.* are more circumspect in suggesting that ESs be used to explore the extent to which MID estimates are similar or vary across instruments, and recommend that anchor-based methods should be the primary method of estimating the MID [2], as do Revicki *et al.* [3]. Such anchor-based methods include global transition items and clinical anchors.

Clinical anchors

Another method used to determine the MID, designed to aid interpretation of mean HRQOL results, is to group the HRQOL scores by clinical criteria that clinicians are familiar with, called clinical anchors [7]. This is sometimes called the 'known groups approach', where 'known' is short-hand for 'the clinical status of the groups is known' [49]. Several criteria must be satisfied for this method to work [1–3]. Clinicians should be familiar with the anchor, usually because it is widely used in assessing and/or managing patients. The anchor itself should be interpretable. There should be a theoretical basis for the relationship between the anchor and the relevant HRQOL domain(s), and an empirical correlation of at least 0.30 between the anchor and those HRQOL domain(s). Anchors with these characteristics are often used during validation to test the clinical criterion validity of HRQOL and PRO measures.

A classic anchor in cancer is the ubiquitous clinician-rated Eastern Cooperative Oncology Group Performance Status (Table 2). It is used by clinicians to rate a patient's daily activities of living. It is commonly used in cancer clinical trials as an inclusion criterion and codified in practice guidelines for chemotherapy and surgery on the basis that the patient needs to be well enough to survive these treatments. It is commonly used in validation of cancer-specific HRQOL measures. King's review demonstrated that groups with a worse performance status consistently had worse physical function, role function and cognitive function and more fatigue, nausea and pain, but the emotional and social scores did not follow this pattern, confirming its usefulness as an anchor for developing interpretation guidelines for most, but not all, HRQOL domains [26]. Clinician-rated performance status has been used as an anchor to determine MIDs (and, more generally, CIDs), cross-sectionally and longitudinally (for improvement and deterioration, separately), for various cancer-specific measures in the Functional Assessment of Chronic Illness Therapy suite [50] and, more recently, for the EORTC's QLQ-C30 [31].

By their nature, anchor-based estimates of MIDs are dependent on the choice of anchor and the strength of the relationship between the specific HRQOL domain and the anchor chosen. For example, using change in hemoglobin level as an anchor, Cella *et al.* found larger differences in fatigue and anemia-focused scales than in the total FACT-G score [30]. Furthermore, as these anchors are by their very nature clinically meaningful, anchor-based HRQOL differences are likely to be more meaningful to clinicians and researchers than to patients, thus they are CIDs. The extent to which they are MCIDs depends on the anchor selected and how adjacent groups are defined within that anchor. For example, when hemoglobin level was used as the anchor for the Functional Assessment of Chronic Illness Therapy fatigue and anemia scales, and adjacent groups were defined by trichotomizing hemoglobin level as <8g/dl, 8–9.99 g/dl and 10–11 g/dl, mean differences between the adjacent groups were similarly small and, arguably, each was minimally clinically important [30]. But when performance status (PS) was used as an anchor, and groups were defined as 0, 1 and 2–3 (the latter being combined due to small sample sizes), the mean HRQOL difference between groups PS1 and PS2–3 was two-to-three-times larger than that between groups PS0 and PS1. Arguably, the latter difference was more likely to be an MCID, while the former was a CID but not an MCID. More generally, collapsing anchor categories owing to limitations in sample size (a relatively common practice) may lead to overestimation of the MID.

Other less commonly used methods

Each of the following methods is innovative and provides for the input of various stakeholders to the judgment of what is minimally important. While they provide information-rich results, they are more logistically complex and/or labor intensive than the methods already described, which may explain why they are less commonly used.

Redelmeier *et al.* developed a method for estimating the MID that requires patients to judge themselves in relation to others with the same condition (based on between-patient differences), and found it produced similar results to the global transition method previously described (based on within-patient changes) for the Chronic Respiratory Questionnaire [51]. More recently, Redelmeier collaborated with Ringash *et al.* to apply this to cancer [12,13]. This method avoids the major problems of response shift and compounded measurement error of the global transition method.

Receiver-operator curves, commonly used to determine the ability of a diagnostic test to detect true cases of disease (in turn determined by a gold-standard method), have also been used to determine MIDs. In this context, the HRQOL measure is considered the diagnostic test and a clinical anchor functions as the gold standard. The anchor distinguishes persons who are significantly improved or deteriorated from persons who have not significantly changed. Various cut-points on the HRQOL instrument's scale(s) are used to classify patients as improved or not improved, and the cut-point with the optimal ROC characteristics (sensitivity and specificity) is taken as an estimate of the MID. The ROC approach has been applied in two ways. Originally, the anchor was a clinical criterion, as illustrated by Deyo *et al.* [52]. Increasingly, the anchor is a global transition question, as illustrated by Kvam *et al.* [53] and de Vet *et al.* [16].

Expert opinion has also been used in two ways. In the first, Wyrwich and colleagues triangulated views from expert physicians, patients and the clinicians treating these patients on

Table 2. Eastern Cooperative Oncology Group performance status.

Grade	Description
0	Fully active, able to carry on all predisease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature (e.g., light house work, office work)
2	Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours
3	Capable of only limited selfcare, confined to a bed or chair more than 50% of waking hours
4	Completely disabled. Cannot carry on any selfcare. Totally confined to a bed or chair
5	Dead

From [69], with credit to Eastern Cooperative Oncology Group, Robert Comis MD, Group Chair.

how much change in a HRQOL measure was needed for that change to be considered a trivial, small, moderate or large clinically important improvement or decline in asthma [54], heart disease [55] and chronic obstructive pulmonary disease [56]. The expert panels participated in complex and lengthy consensus processes about what constituted clinically important differences, and devised wording for global transition questions and eligibility criteria for the patient participants, who completed questionnaires (including global transition questions) and participated in interviews bimonthly for 1 year. The clinicians treating these patients completed baseline assessments on each patient's health state and then evaluated the change in each patient's condition at subsequent visits during the next year. In asthma and chronic obstructive pulmonary disease, the patient-perceived estimates were consistent with the results of previous global change-based MIDs but were notably lower than those derived from the expert panel and the managing clinicians. In heart disease, however, they found little consensus and concluded that MID estimates depended largely on the rater's perspective and the method used. The authors nevertheless felt that this approach demonstrated the value of patient and physician perspectives and the need for improved dialogue and understanding in the interpretation and use of HRQOL results.

King *et al.* used expert judgement in another way, combining it with clinical anchors and systematic review. Three clinicians with many years of experience managing cancer patients and using HRQOL outcomes in clinical research each reviewed 71 papers that reported mean scores of the FACT-G, a cancer-specific HRQOL measure. Blinded to the FACT-G results, they considered the various clinical anchors associated with FACT-G mean differences, predicted which dimensions of HRQOL would be affected and whether the effects would be trivial, small, moderate or large. These size classes were defined explicitly in terms of clinical relevance. The experts' judgments were then linked with FACT-G mean differences and inverse-variance weighted mean differences and ESs were calculated for each size class [47,57]. Cocks *et al.* applied a similar method to the QLQ-C30 [48]. In both of these studies, variations in MIDs were found across domains of HRQOL.

Statistical significance, sample size, power & the smallest statistically significant difference

Yet another angle on the MID is its relationship with statistical significance. It is often noted that the MID is informative for calculating sample sizes, as demonstrated, for example, in the study by Cocks *et al.* [48]. Fayers and Machin provide a comprehensive description of sample size calculation for various HRQOL scale types and analysis methods [58]. In simple terms, sample size calculation determines the number of patients required to allow a reasonable chance (power, the complement of the type II error rate) of detecting a predetermined difference (which may be the MID) in the outcome variable at a given level of statistical significance (type I error or false-positive rate).

The smallest (statistically) detectable difference (SDD) is the smallest difference that can be detected as statistically significantly different from zero, given nominated type I and II error rates and fixed sample size. It is a function of these quantities and the SD of scores at baseline, as described in Angst *et al.* (Table 1) [29]. The SDD may be smaller or larger than the MID. If it is larger, then the study is underpowered to detect the MID as statistically significantly different from zero; the confidence interval will include the MID and zero. In the arthritis and rheumatism literature, the SDD of candidate outcome measures is sometimes estimated and compared with the MCID (e.g., see Angst *et al.* [29]).

This is the group-level decision-making research context. Here, the MID (or equivalently MCID) is the smallest difference that will convince clinicians to change their treatment practice or that will convince policy-makers to change their practice guidelines or the treatments they make publicly available on subsidized schedules. If, at the planning stage, sample size has been calculated to detect the MID, then the study is appropriately powered to detect the MID, and when the data are finally in, the interpretation of the results will be straightforward. Problems may arise if sample size is based on other considerations, such as when HRQOL is a secondary outcome and the trial is powered on the primary end point. Overpowered HRQOL comparisons may arise in randomized trials powered for survival end points (which typically require larger samples than HRQOL end points), and in population-based surveys or cohort studies, where large sample sizes are likely. Here, the danger is that very small HRQOL differences (clinically trivial) will be statistically significant.

Thus, statistical significance can only be used to interpret HRQOL results if the sample size was determined *a priori* on the MID. Even then, the clinical significance should be considered and discussed to provide a useful interpretation of the results for readers. However, Cocks *et al.* found that of 82 cancer randomized controlled trials reporting EORTC QLQ-C30, clinical significance was only addressed in 38% of these [10]. Where clinical significance was not addressed, reliance was usually based on

statistical significance. This misuse of statistical significance is compounded in HRQOL studies, where the multidimensional nature of HRQOL leads to multiple hypothesis testing and the associated danger of false-positive findings [58].

Expert commentary & five-year view

The occurrence of 'the MID' and related terms in the HRQOL literature has approximately trebled every 5 years over the past 20 years. Many of the recent studies are determining MIDs, either for the first time for a measure or again in another clinical context, or using MIDs to interpret the clinical significance of mean differences or to determine the proportion of patients with clinically important change. This represents progress in the interpretation of HRQOL and PRO results in general.

Interestingly, the term 'MID' (or any of the related terms in Table 1) was conspicuously absent from the US FDA's final guidance for industry on PRO measures [59]. Instead, the term 'responder definition' was used, defined as 'the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit.' They said that it should be determined empirically, and went on to describe the four most commonly used methods for determining MIDs described in the 'Methods used to determine the MID' section, stipulating that transition questions and clinical anchors should provide the primary evidence, with ES and SEM as supportive evidence, as per recommendations for MID determination of Revicki *et al.* [3]. It is unclear why the authors avoided the term 'MID'. They were certainly talking about something that closely resembles what others might call a MID, although with the added time dimension.

Revicki *et al.*'s stance on recommended methods and the hierarchy of evidence for MIDs [3] was also taken in the consensus statement of the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) [60], which covered MIDs, MCIDs and CIDs. These two guidance publications, both published in 2008, reflect current consensus on methods and therefore probably indicate future trends, at least for the next 5 years: the methods described in the 'Methods' section are likely to remain the most widely used, alone and in combination. Expert opinion may also be enlisted to quantify a range of CIDs beyond the MID, following the methods such as those pioneered by Wyrwich and colleagues [54–56], and King and colleagues [47,57].

In addition, it is likely that a plethora of MID estimates will appear in the coming years. While it may be tempting to adopt a single published MID *prima facie*, we need to be mindful that it is just an estimate, as prone to sampling variation as any other, and influenced by the method used, the patient population, the clinical context and perspective [2,3]. As Ware and Keller sagely observed in 1996, interpretability is not established by a single psychometric maneuver; rather, it develops gradually as a body of evidence accumulates with repeated experience from a variety of perspectives [61]. Simple rules of thumb for interpreting HRQOL measures are appealing, but should be used judiciously. Ringash *et al.* provide a good example of the balance required. After reporting MID estimates for different domains, with 95% confidence intervals, they quite reasonably simplified these to: "One rule of thumb for interpreting a difference in QOL scores is a benchmark of about 10% of the instrument range", adding the caveat, "Patients appear to be more sensitive to favorable differences, so an improvement of 5% may be meaningful" [13]. In summarizing our results for the FACT-G, we heeded the advice of Guyatt *et al.* to avoid misleading oversimplifications [62]. As we believed that interpretation guidelines for HRQOL scales require some flexibility to accommodate different patient groups and clinical circumstances, we summarized our results for each size class and domain as probable ranges. Furthermore, the degree of variation of component estimates within size classes in our meta-analysis highlighted the limitations of individual studies for deriving general interpretation guidelines. In addition, rather than focus on the MID, we accommodated the possibility that in some circumstances, the MID may be of a moderate absolute size, while in others it may be relatively small.

As estimates of MIDs emerge from individual studies, we need to consolidate them into a growing store of knowledge. Who should do this, and how should it be done? Some widely used instruments are managed by large organizations, such as the EORTC's Quality of Life Group and Department (for the QLQ-C30 and its modules) and QualityMetric (for the SF-36 and other measures). These are the obvious entities to take on this responsibility, preferably in the form of continuously updated interpretation guidelines. Individual researchers, or consortia such as IMMPACT [60], may have the interest and means to prepare and update reviews of available evidence about MIDs, and present them with accompanying text that explains to less expert readers the suggested use and caveats of MIDs within broader interpretation guidelines that emphasize the importance of context. Different MID estimates may be graphed to visually depict the range of estimates, with informal weighting and synthesis, as illustrated by Revicki *et al.* [3], or by meta-analysis, as illustrated by King *et al.* [47,57] and Cocks *et al.* [48] While such guidelines will lack the immediate appeal of a general rule-of-thumb, they will encourage a more sophisticated practice in the interpretation of HRQOL data, as recently recommended [2,3].

Part of the sophistication that we as a research community should aspire to is the matching of MIDs to clinical contexts and treatment decisions, as emphasized by various authors [2,3,59,63,64]. In reviewing medical product development to labeling claims, the FDA will "evaluate an instrument's responder definition in the context of each specific clinical trial" [59]. Wyrwich *et al.* provide a good example of context-specific interpretation that firstly involves the determination of thresholds based on treatment satisfaction questions and then the determination of the doses of desvenlafaxine that provide the degree of symptom relief considered important by menopausal women, as defined by the satisfaction thresholds [65]. Yet the MID history demonstrates that simple messages tend to resonate and propagate through the research literature and practice. Do time-poor researchers really want to acknowledge that there is no universal MID, that 'the MID' does not exist? As de Vet *et al.* said, "A balance needs to be struck between the practicality of a single MIC [sic] value and the validity of a range of MIC [sic] values" [16]. The point of minimal important difference is indeed elusive.

Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Key issues

- Several minimal important difference (MID)-related terms differ only slightly in definition: minimal clinically important difference, clinically important difference, minimally detectable difference and the subjectively significant difference. Others appear similar but have quite different meanings: minimally detectable difference versus minimum detectable change. Four main methods are commonly used to estimate MIDs: patient retrospective rating of change using global transition items (patient-change anchor); commonly used clinical scales or classifications (external clinical anchors or 'known groups'); standard error of measurement; and effect size. These are described and critiqued in this article.
- Definitions and methods are summarized in Table 1 of this article. Nuances of definition of the MID-related terms are rarely of any consequence in the way these methods are applied, and the results reported and used.
- There is no global MID, although an effect size of between 0.2 and 0.5 may provide a useful ballpark guideline. For a particular patient-reported outcome (PRO) instrument or scale, the MID is not an immutable characteristic, but may vary by population and context. There is considerable variation in individual-specific MIDs.
- At the group level, the MID may need to be adjusted for the clinical context and decision at hand, whether other benefits or side effects are considered in that decision, the baseline from which the patient starts (relatively well or sick), and whether the patient is improving or deteriorating.
- At the individual level, when used in shared decision-making, the MID should be adjusted to match the patient's preferences.
- Empirical estimates are known to differ with domain-specific scales and by which method is used (particularly with clinical anchors). Therefore, specific estimates of MIDs should not be overinterpreted. For a given PRO scale, all available MID estimates and ranges should be considered and applied judiciously to any particular clinical or research context.
- PRO scales commonly used in research settings may not be reliable enough ($\alpha > 0.9$) to detect MIDs at the individual patient level because they are relatively short (typically between two and five questions in each scale). The advent of computerized adaptive testing may provide a solution if/when this technology becomes widely adopted.
- In clinical research, statistical significance cannot be used to interpret health-related quality of life and PRO results unless the sample size has been based *a priori* on a specified MID. Even then, the clinical significance should be considered and discussed to provide a useful interpretation of the results for readers.
- Future directions: multiple methods should be used to determine not only MIDs but also a wider range of clinically important differences (small, moderate and large effects), with global transition questions and clinical anchors providing primary evidence, and standard error of measurement and effect size as supportive evidence. It would be helpful for researchers if available estimates of MIDs and clinically important differences were consolidated into interpretation guidelines for specific health-related quality of life and other PRO measures, with periodic updates as further evidence emerges.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1 Guyatt GH, Osaba D, Wu AW *et al.* Methods to explain the clinical significance of health status measures. *Mayo Clin. Proc.* 77(4), 371–383 (2002).

2 Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD* 2(1), 63–67 (2005).

3 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* 61(2), 102–109 (2008).

•• **Authoritative summary of recommended methods.**

4 Wyrwich KW, Bullinger M, Aaronson N *et al.* Estimating clinically significant differences in quality of life outcomes. *Qual. Life Res.* 14(2), 285–295 (2005).

5 Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J. Chronic Dis.* 40(2), 171–178 (1987).

6 Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control. Clin. Trials* 10, 407–415 (1989).

7 Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual. Life Res.* 2, 221–226 (1993).

•• **Insightful review and influential taxonomy of methods to interpret health-related quality of life (HRQOL).**

8 Norman GR, Sloan JA, Wyrwich WK. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med. Care* 41(5), 582–592 (2003).

9 Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J. Clin. Oncol.* 16(1), 139–144 (1998).

10 Cocks K, King MT, Velikova G, Fayers PM, Brown JM. Quality, interpretation and presentation of EORTC QLQ-C30 data in randomised controlled trials. *Eur. J. Cancer* 44, 1793–1798 (2008).

• **Review of standard of reporting and interpretation for HRQOL outcomes in randomized controlled trials.**

11 Schünemann HJ, Guyatt GH. Goodbye (M)CID! Hello MID, where do you come from? (Commentary). *Health Serv. Res.* 40(2), 593–597 (2005).

12 Ringash J, Bezjak A, O'Sullivan B, Redelmeier DA. Interpreting differences in quality of life: the FACT-H&N in laryngeal cancer patients. *Qual. Life Res.* 13(4), 725–733 (2004).

13 Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. *Cancer* 110(1), 196–202 (2007).

14 Beaton DE, Bombardier C, Katz JN *et al.* Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J. Rheumatol.* 28(2), 400–405 (2001).

15 de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual. Life Outcomes* 4, 54 (2006).

16 de Vet HC, Ostelo RW, Terwee CB *et al.* Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual. Life Res.* 16(1), 131–142 (2007).

17 Snyder C, Aaronson N. Use of patient-reported outcomes in clinical practice. *Lancet* 374(9687), 369–370 (2009).

18 Fayers PM, Machin D. *Quality of Life: Assessment, Analysis and Interpretation (1st Edition)*. John Wiley & Sons Ltd, NY, USA (2000).

19 Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients? *Qual. Life Res.* 16, 1097–1105 (2007).

20 Osoba D, Bezjak A, Brundage M, Zee B, Tu D, Pater J; Quality of Life Committee of the NCIC CTG. Analysis and interpretation of health-related quality of life data from clinical trials: basic approach of the National Cancer Institute of Canada Clinical Trials Group. *Eur. J. Cancer* 41, 280–287 (2005).

21 Yost KJ, Cella D, Chawla A *et al.* Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. *J. Clin. Epidemiol.* 58(12), 1241–1251 (2005).

22 Norman G. Hi! How are you? Response shift, implicit theories and differing epistemologies. *Qual. Life Res.* 12(3), 239–249 (2003).

23 Schwartz CE, Sprangers MA. *Adaptation to Changing Health: Response Shift in Quality-of-Life Research (1st Edition).* American Psychological Association, Washington, DC, USA (2000).

24 Cella D. *Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System (4th Edition).* Evanston Northwestern Healthcare & Northwestern University, IL, USA (1997).

25 Ware JE Jr. *SF-36 Health Survey: Manual and Interpretation Guide.* The Health Institute, Boston, MA, USA (1993).

26 King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual. Life Res.* 5, 555–567 (1996).

27 Knox S, King MT. Validation and calibration of the SF-36 health transition question against an external criterion of clinical change in health status. *Qual. Life Res.* 18(5), 637–645 (2009).

28 Osoba D, King M. Interpreting QOL in individuals and groups: meaningful differences. In: *Assessing Quality of Life in Clinical Trials: Methods and Practice.* Fayers P, Hays R (Eds). Oxford University Press, Oxford, UK, 243–257 (2005).

29 Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum.* 45(4), 384–391 (2001).

30 Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J. Pain Symptom Manage.* 24(6), 547–561 (2002).

- **Illustrates methods with an insightful discussion.**

31 Maringwa JT, Quinten C, King M *et al.* Minimal important differences for interpreting health-related quality of life scores from the EORTC QLQ-C30 in lung cancer patients participating in randomized controlled trials. *Support. Care Cancer* DOI: 10.1007/s00520–010–1016–1015 (2010) (Epub ahead of print).

32 Ross M. Relation of implicit theories to the construction of personal histories. *Psychological Rev.* 96(2), 341–357 (1989).

33 Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J. Clin. Epidemiol.* 50(8), 869–879 (1997).

34 Wyrwich K, Tardino V. Understanding global transition assessments. *Qual. Life Res.* 15(6), 995–1004 (2006).

35 Metz SM, Wyrwich KW, Babu AN, Kroenke K, Tierney WM, Wolinsky FD. A comparison of traditional and Rasch cut points for assessing clinically important change in health-related quality of life among patients with asthma. *Qual. Life Res.* 15(10), 1639–1649 (2006).

36 de Vet HC, Terluin B, Knol DL *et al.* Three ways to quantify uncertainty in individually applied 'minimally important change' values. *J. Clin. Epidemiol.* 63(1), 37–45 (2010).

37 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual. Life Res.* 4(4), 293–307 (1995).

38 Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J. Clin. Epidemiol.* 52(9), 861–873 (1999).

39 Wyrwich K. Minimal important difference thresholds and the standard error of measurement: is there a connection? *J. Biopharm. Stat.* 14(1), 97–110 (2004).

40 Wyrwich KW, Tierney WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Qual. Life Res.* 11(1), 1–7 (2002).

41 Turner D, Schünemann HJ, Griffith LE *et al.* The minimal detectable change cannot reliably replace the minimal important difference. *J. Clin. Epidemiol.* 63(1), 28–36 (2010).

42 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med. Care* 27(3 Suppl.), S178–S189 (1989).

43 Cohen J. *Statistical Power Analysis for the Behavioural Sciences (2nd Edition).* Lawrence Erlbaum Associates, NJ, USA (1988).

44 Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Rev. Pharmacoeconomics Outcomes Res.* 4(5), 515–523 (2004).

45 Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Med. Care* 41(5), 593–596 (2003).

46 Wright JG. Interpreting health-related quality of life scores: the simple rule of seven may not be so simple. *Med. Care* 41(5), 597–598 (2003).

47 King MT, Stockler MR, Cella DF *et al.* Meta-analysis provides evidence-based effect sizes for a cancer-specific quality of life questionnaire, the FACT-G. *J. Clin. Epidemiol.* 63(3), 270–281 (2010).

• **Provides empirical alternative to Cohen's arbitrary guidelines.**

48 Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer quality of life questionnaire core 30 (EORTC QLQ-C30). *J. Clin. Oncol.* 29(1), 89–96 (2011).

• **Interesting presentation of expert opinion.**

49 Aaronson NK, Cull A, Kaasa S *et al.* The European Organisation for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: an update. In: *Quality of Life and Pharmacoeconomics in Clinical Trials.* Spilker B (Ed.). Lippincott-Raven Publishers, PA, USA, 179–189 (1996).

50 Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval. Health Prof.* 28(2), 172–191 (2005).

51 Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques [see comments]. *J. Clin. Epidemiol.* 49(11), 1215–1219 (1996).

52 Deyo RA, Inui TS, Leininger J, Overman S. Physical and psychosocial function in rheumatoid arthritis. Clinical use of a self-administered health status instrument. *Arch. Intern. Med.* 142(5), 879–882 (1982).

53 Kvam AK, Fayers P, Wisloff F. What changes in health-related quality of life matter to multiple myeloma patients? A prospective study. *Eur. J. Haematol.* 84(4), 345–353 (2010).

54 Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Interpreting quality-of-life data: methods for community consensus in asthma. *Ann. Allergy Asthma Immunol.* 96(6), 826–833 (2006).

• **Exemplifies an innovative and complex methodology and a thorough analysis approach.**

55 Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. *Health Serv. Res.* 42(6 Pt 1), 2257–2274; discussion 2294–2323 (2007).

56 Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Measuring patient and clinician perspectives to evaluate change in health-related quality of life among patients with chronic obstructive pulmonary disease. *J. Gen. Intern. Med.* 22(2), 161–170 (2007).

57 King MT, Cella D, Osoba D *et al.* Meta-analysis provides evidence-based interpretation guidelines for the clinical significance of mean differences for the FACT-G, a cancer-specific quality of life questionnaire. *Patient Reported Outcome Measures* 2010(1), 119–126 (2010).

58 Fayers PM, Machin D. Sample sizes. In: *Quality Of Life: The Assessment, Analysis And Interpretation Of Patient-Reported Outcomes*. Wiley, Chichester, UK, 247–270 (2007).

59 Food and Drug Administration. Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. *Federal Register* 74(235), 65132–65133 (2009).

60 Dworkin RH, Turk DC, McDermott MP *et al.* Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J. Pain Symptom Manage.* 9(2), 105–121 (2008).

61 Ware JE, Keller SD. Interpreting general health measures. In: *Quality of Life and Pharmacoeconomics in Clinical Trials.* Spilker B (Ed.). Lippincott-Raven, NY, USA, 445–460 (1996).

62 Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR; Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin. Proc.* 77(4), 371–383 (2002).

63 Beaton D, Boers M, Wells G. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr. Opin. Rheumatol.* 14(2), 109–114 (2002).

•• **Thorough and thoughtful review.**

64 Hays R, Woolley J. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics* 18, 419–423 (2000).

•• **Insightful analysis of conceptual issues.**

65 Wyrwich KW, Spratt DI, Gass M, Yu H, Bobula JD. Identifying meaningful differences in vasomotor symptoms among menopausal women. *Menopause* 15(4 Pt 1), 698–705 (2008).

66 Sloan JA, Cella D, Frost M, Guyatt GH, Sprangers M, Symonds T; Clinical Significance Consensus Meeting Group. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clin. Proc.* 77(4), 367–370 (2002).

67 de Vet HC, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. *J. Rheumatol.* 33(2), 434; author reply 435 (2006).

68 Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. *Qual. Life Res.* 10(7), 571–578 (2001).

69 Oken MM, Creech, RH, Tormey DC *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am. J. Clin. Oncol.* 5, 649–655 (1982).