Digital Biomarkers

Review Article

Digit Biomark 2025;9:52–66
DOI: 10.1159/000543899

Received: May 28, 2024
Accepted: January 26, 2025
Published online: February 3, 2025

Interpretation of Change in Novel Digital Measures: A Statistical Review and Tutorial

Andrew Trigg[a] Bohdana Ratitch[b] Frank Kruesmann[c]
Madhurima Majumder[d] Andrejus Parfionovas[d] Ulrike Krahn[c]

[a]Medical Affairs Statistics, Bayer plc, Reading, Berkshire, UK; [b]Statistics and Data Insights, Bayer Inc., Mississauga, ON, Canada; [c]Statistics and Data Insights, Bayer AG, Wuppertal, Germany; [d]Statistics and Data Insights, Bayer US LLC, Whippany, NJ, USA

Keywords
Digital health technology · Clinical validation · Interpretation · Minimal clinically important difference · MID

Abstract
Background: Novel clinical measures assessed by a digital health technology tool require thresholds to interpret change over time, such as the minimal clinically important difference. Establishing such thresholds is a key component of clinical validation, facilitating understanding of relevant treatment effects. ***Summary:*** Many of the approaches to derive interpretative thresholds for patient-reported outcomes can be applied to digital clinical measures. We present theoretical background to the use of interpretative thresholds, including the distinction between thresholds based on perceived importance versus measurement error, and thresholds for group- versus individual-level interpretations. We then review methods to estimate such thresholds, including anchor-based approaches. We illustrate the methods using data on cough frequency counts as measured by a wearable device in a clinical trial. ***Key Messages:*** This paper provides an overview of statistical methodologies to estimate thresholds for the interpretation of change.

Introduction

Validation assesses the extent to which a new clinical measure evaluated by a digital health technology (DHT) tool is suitable for use in clinical research and patient care [1–3]. A DHT can be broadly defined as *"a system that uses computing platforms, connectivity, software, and/or sensors for health care and related uses"* [2]; however, in this paper we focus on DHTs used to measure health outcomes in clinical research such as electronic clinical outcome assessments (COAs) and sensor-based digital biomarkers. The V3 framework splits validation into verification, analytical validation, and clinical validation phases [1]; here, we focus on the last step of clinical validation. Ratitch et al. [4] identified four "critical elements" of clinical validation: reliability, associations with other clinical measures, responsiveness, and the minimal clinically important difference (MCID). A previous paper provided an overview of analyses for assessing reliability [5]; this paper is designed to follow on from this, presenting statistical methods related to the MCID.

The MCID was originally defined 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*

karger@karger.com
www.karger.com/dib

Correspondence to:
Andrew Trigg, andrew.trigg@bayer.com

management" [6]. While this original definition focuses on the care of an individual patient, researchers also require thresholds to discern clinically important differences between groups of patients. Additionally, researchers may be interested in what constitutes a moderate or large change, beyond a minimal threshold. Therefore, this paper generally discusses interpretative thresholds for clinical measure values arising from DHT tools, rather than a strict focus on the MCID as originally defined.

The objective of this paper is to summarize methodologies to estimate thresholds for the interpretation of clinical measure values. The paper assumes familiarity with basic statistical concepts (e.g., correlation, logistic regression) and provides a review/tutorial on specific methods to estimate interpretative thresholds that build on these basic concepts. Most of the relevant literature on interpretability stems from patient-reported outcomes (PROs) and other COAs. While these methods are generally applicable to wider DHTs, we highlight instances where this may not be the case. Two recent articles discuss interpretative thresholds in the context of DHTs, albeit without a focus on statistical methods [7, 8]. The article by McCarthy et al. [7] provides a particularly useful list of example studies estimating thresholds.

Theoretical background to the MCID and other thresholds is provided. We then review statistical methods to estimate these thresholds, including the type of data required to implement such techniques. These methods are then demonstrated using data from a cough-recording device administered within a clinical trial.

Theoretical Background

The following sections review the distinction between thresholds based on perceived importance versus measurement error, interpretations at a group- or individual-level, and magnitude of thresholds beyond minimal. A note on terminology is provided in the online supplementary material (for all online suppl. material, see https://doi.org/10.1159/000543899).

The focus of this paper is the comparison of clinical measure values within/between individuals or groups of participants, where all values are observed estimates. Other types of interpretative guideline, such as cross-sectional severity/diagnostic/prognostic cut-offs [9–11], Patient Acceptable Symptom State [12], or normative values [13] are not described further in this review given they compare against a fixed external value rather than an observed estimate. Similarly, the use of a single clinical measure value to identify an event of interest after start of treatment (without pre-post-comparison of the value) is not covered within this review, but has been discussed previously [4].

Perceived Importance versus Measurement Error

Thresholds for interpreting clinical measure values can be based on the perceived importance of changes (often from the patients' perspective) or based on measurement error. Thresholds based on perceived minimal importance are aligned with the concept of the MCID as originally defined. Measurement error is the inverse of reliability, expressed on the same metric as the clinical measure, where changes within the range of measurement error may not represent a true change in the construct of interest. Therefore, thresholds can be set beyond the expected variation in clinical measure values due to measurement error; however, these do not reflect perceived importance. Additionally, thresholds based on perceived importance are more relevant for regulatory authorities [7].

An alternative approach to interpretation is based on standardized effect sizes (SESs), e.g., 0.2, 0.5, or 0.8 standard deviation units taken to represent a "small," "medium," or "large" effect, respectively [14–16]. SESs have previously been grouped with thresholds based on measurement error under the term "distribution-based indices" [17]; however, the SESs do not concern measurement error so we view these as theoretically distinct. Given the arbitrariness of what constitutes a "small," "medium," or "large" effect size, we recommend focusing on thresholds for perceived importance and/or measurement error and therefore do not describe SESs further.

Group-Level and Individual-Level Thresholds

Most interpretative thresholds based on perceived importance can be categorized according to the level of interpretation, type of comparison, and magnitude [18] as per Table 1. In clinical trials where a DHT tool is used to measure health outcomes before and after treatment, it is common to express treatment effect in terms of the between-group difference in mean change from baseline. A group-level threshold for the difference in change over time (case 1A and 2C) is then required to judge clinical importance. Alternatively, one may wish to judge whether an individual patient has improved to a clinically important degree (individual-level change over time, case 1B and 2B). In the context of clinical trials, this can support responder analyses where the proportion of patients with changes meeting or exceeding a threshold are compared

Table 1. Categorization of thresholds based on perceived importance

Dimension	Options
1: Level of interpretation	A: Group B: Individual
2: Type of comparison	A: Difference (cross-sectional) B: Change over time[a] C: Difference in change over time
3: Magnitude	A: Minimal B: Larger than minimal (e.g., moderate or large)

[a]May be further split according to improvement/worsening. Table adapted from Trigg et al. [18].

[19, 20]. Example applications for the remaining cases are as follows:

- Case 1A and 2B: A group-level change over time threshold to ascertain whether a single treatment arm improved by a clinically significant amount.
- Case 1A and 2A: A group-level cross-sectional difference threshold could be used to interpret whether two treatment arms are sufficiently similar in their baseline characteristics, especially useful to judge in observational or "real-world" studies of comparative effectiveness.
- Case 1B and 2C: The difference in change over time between individuals would be evaluated in trials with a "win ratio" endpoint [21].
- Case 1B and 2A: Individual cross-sectional difference thresholds could be used as calipers for propensity score matching in observational studies [22].

It is unlikely that thresholds derived at the group level are transferable to interpretations at an individual level [23–26]; however, they may be a useful starting point in light of no other information [27]. Notably, in the original paper defining MCID, the authors noted it could be used for interpretation *"both in individuals and in groups of patients participating in controlled trials,"* conflating individual- and group-level change which continues to this day [6, 28]. Therefore, we recommend specifying the intended interpretation when estimating and applying any threshold as per the options in Table 1 (e.g., "minimal within-individual change over time").

Statistical Methods for Estimating Thresholds

For the remainder of this article, we generally refer to the interpretation of changes over time for simplicity, but make note where a cross-sectional difference is of specific relevance. We consider the statistical methods to be generally applicable to any continuous digital clinical measure, regardless of its origin or technological underpinnings. The methods should also be appropriate for ordinal measures with a high number of categories (e.g., >7) [29, 30]. Ordinal measures with fewer categories (and binary measures) will carry the inherent assumption that differences between categories are meaningful. To aid understanding, a glossary is provided in Table 2, including definitions and explanations of key terms.

Study Design

Thresholds to interpret clinical measure values should be estimated using data from studies with eligibility criteria reflecting the target population for which the thresholds will be applied in future, especially in terms of diagnosis and disease severity. Longitudinal data on the clinical measure value of interest are required for a thorough estimation of thresholds to interpret change over time. The data should also ideally include potential anchor measures, administered at the times relevant to the endpoint that will be constructed based on the DHT of interest.

An anchor measure is an external criterion reflecting known changes in the same concept of interest (COI) as the DHT tool. The most used anchor is the patient global impression of change (PGI-C), also called transition rating (see online suppl. Table S1). The PGI-C directly asks about perceived change, but is subject to "present state bias" where the respondent's rating is more reflective of their current health than true change over time [32, 33]. A solution to avoid present state bias is to administer a patient global impression of severity (PGI-S) and calculate the change in this measure as the criterion for change. However, change in PGI-S is not as direct as a measurement of the perceived importance of change, where the importance of a 1-category change may require further expert judgment or qualitative investigation [33,

Table 2. Glossary of key terms

Term	Definition/explanation
Anchor measure	An external criterion reflecting known changes or differences in a particular COI
DHT	*"A system that uses computing platforms, connectivity, software, and/or sensors for health care and related uses"* [2]
Discriminant analysis	A nonparametric anchor-based method to display the densities of clinical measure values split by anchor groups. The intersection point of these densities is taken as the threshold estimate
eCDF plot	In the context of anchor-based analyses, a plot with change in clinical measure value on the x-axis and the cumulative proportion of subjects on the y-axis, presented by anchor group. The "empirical" description implies no smoothing is applied
Measurement error	Variability in measurements due to sources other than true variability in the COI. Interpretative thresholds can be defined in terms of the magnitude of change expected to exceed measurement error
MCID	Originally defined 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]. It has since been used as a general term for interpretative thresholds based on patient-perceived importance
MDC	Used to express the variability of measurement error around individual changes in clinical measure values. Set at different magnitudes based on critical values, e.g., 95% or 68%
Point biserial correlation coefficient	A correlation coefficient suitable for associations between a continuous and binary variable
Polyserial correlation coefficient	A correlation coefficient suitable for associations between a continuous and ordinal variable, where the ordinal variable is assumed to be derived from an underlying continuous variable
Predictive modeling method	An anchor-based method based on a logistic regression model with dichotomized anchor measure as the outcome and clinical measure value change as the predictor
PDF plot	In the context of anchor-based analyses, a plot with change in clinical measure value on the x-axis and the density of subjects on the y-axis, presented by anchor group. Smoothing is commonly applied (e.g., kernel smoothing)
ROC analysis	A method to estimate a threshold that optimally discriminates between two groups. In the context of anchor-based analyses, these are commonly "improved" and "not improved" groups
Reliability	The proportion of variance in a measure attributable to true variance in the underlying concept being measured
Sensitivity	In the context of anchor-based analyses for improvement thresholds, the proportion of anchor-based "improved" patients with clinical measure value changes at or exceeding a threshold
Specificity	In the context of anchor-based analyses for improvement thresholds, the proportion of anchor-based "not improved" patients with clinical measure value changes below a threshold
SEM	The standard deviation of measurement error around an individual clinical measure value at a single point in time [31]

34]. Additionally, due to the ordinal scaling of a PGI-S, a 1-category change from "severe" to "moderate" may not reflect the same magnitude of change as a 1-category change from "moderate" to "mild." An alternate patient-reported anchor is to use an existing PRO measuring the same COI, with a pre-established threshold for interpreting change [17].

A limitation with the above anchors in the context of DHTs is that patient-reported health may not be a suitable criterion for objectively measured attributes. For example, correlations between patient-reported sleep quality and actigraphy-measured parameters can be low [35]. One option is to use clinician-reported versions of the PGI-C and PGI-S, but this may still fail to capture objectively measured attributes. Another is to look into the data for so-called clinical anchors based on other measurements such as disease staging, laboratory values, adverse events, or hospitalizations [26, 36, 37].

Nevertheless, several studies have used patient-reported anchor measures to estimate thresholds for objective sensor-based measures [38–40].

If no suitable anchor measure is included in the data, it is still possible to derive thresholds based on measurement error (albeit not targeting perceived importance). An example of this scenario is provided by Demeyer et al. [41] where none of the attempted anchors were sufficient to derive thresholds for physical activity in patients with chronic obstructive pulmonary disease. Alternate study designs relevant for perceived importance are available based on qualitative methods, such as patient interviews, to derive thresholds [36, 42]. Judgment by an expert panel (e.g., of clinicians) is also possible, albeit less patient-centric [43]. Another design relevant to multi-item PROs is bookmarking, described elsewhere [44, 45].

Anchor-Based Methods

Anchor-based methods are currently the most widely recommended, including the FDA and other health authorities [23, 46–48]. As the name suggests, they are reliant on the inclusion of anchor measures. The central idea is that changes on the DHT clinical measure value of interest can be linked to corresponding changes on the anchor measure that have a known perceived importance.

The first step is to evaluate the credibility of the anchor measure to define change in the target COI. While previous recommendations for PROs state an anchor should be easier to interpret than the value of interest [46], we do not think this is necessary for clinical anchors as described above so may not be as relevant for DHTs. An important factor supporting credibility is that the anchor measure is sufficiently associated with the clinical measure value of interest. This is typically evaluated by a correlation coefficient between the anchor measure and change in the clinical measure value, where coefficients of at least 0.3–0.5 have been suggested as cut-offs [49, 50]. A polyserial correlation coefficient is suitable to measure the association between an ordinal anchor and continuous DHT value change, but others may be used depending on the distribution of each measure. Additionally, variability in the anchor (i.e., sufficient number of patients experiencing change) is a necessary requirement.

Once the credibility of an anchor is determined to be sufficient, anchor-based estimates can be obtained. Previous studies and reviews have considered different anchor-based methods to be equivalent in terms of the target estimate, but this is not the case and specific methods are suitable for specific types of interpretation (Table 1). This is best illustrated in the context of a minimal within-individual change threshold, which theoretically should correspond to the location of a threshold on the change in clinical measure value that optimally discriminates between those who experience at least a minimal improvement versus those who do not [51]. Figure 1 shows the theoretical distribution of change in clinical measure value, split by PGI-C response in the top panel and merging these to form binary "improved" or "not improved" groups in the bottom panel. Three values are labeled on the plot: (1) the mean change within patients who are "a little better"; (2) the difference in mean change between patients who are "a little better" and "no change"; (3) the location of the threshold discriminating those at least minimally improved versus not improved.

A simple anchor-based approach is to calculate the mean change in DHT value within subgroups defined by the anchor measure. The mean change within the group of patients experiencing a minimal improvement on the anchor (i.e., those selecting "a little better") is the estimate of the threshold. As shown in Figure 1, this estimate does not effectively target the threshold separating improved versus not improved patients (necessary for interpretation of within-individual change). The mean change method has been proposed as more suited to the interpretation of within-group change over time [52, 53]; however, this assumed suitability has not been empirically tested to date. The mean change within the "much better" group would estimate a higher magnitude threshold beyond minimal. Changes within worsened groups could be used to find specific thresholds for worsening, as in practice the absolute values of thresholds for improvement and worsening can differ [54].

Similar to the mean change, the difference in mean change between those experiencing minimal improvement and no change anchor groups can be calculated. This provides an estimate which has been proposed as suitable for interpreting between-group differences in mean change over time [52, 55]. A linear regression model with a binary anchor as a predictor and change in clinical measure value as the outcome can be used to achieve the same estimate, but with the opportunity to control for other covariates [56], or incorporate repeated measurements [53]. These approaches for between-group differences carry an implicit assumption where the threshold is based on the difference between a group where everyone has (minimally) improved, versus a group where no one has. This rather strong assumption may lead to estimates higher than what is needed in scenarios where a minority of patients on a treatment arm are expected to improve or there is a large placebo effect; therefore, further research in this area is warranted.

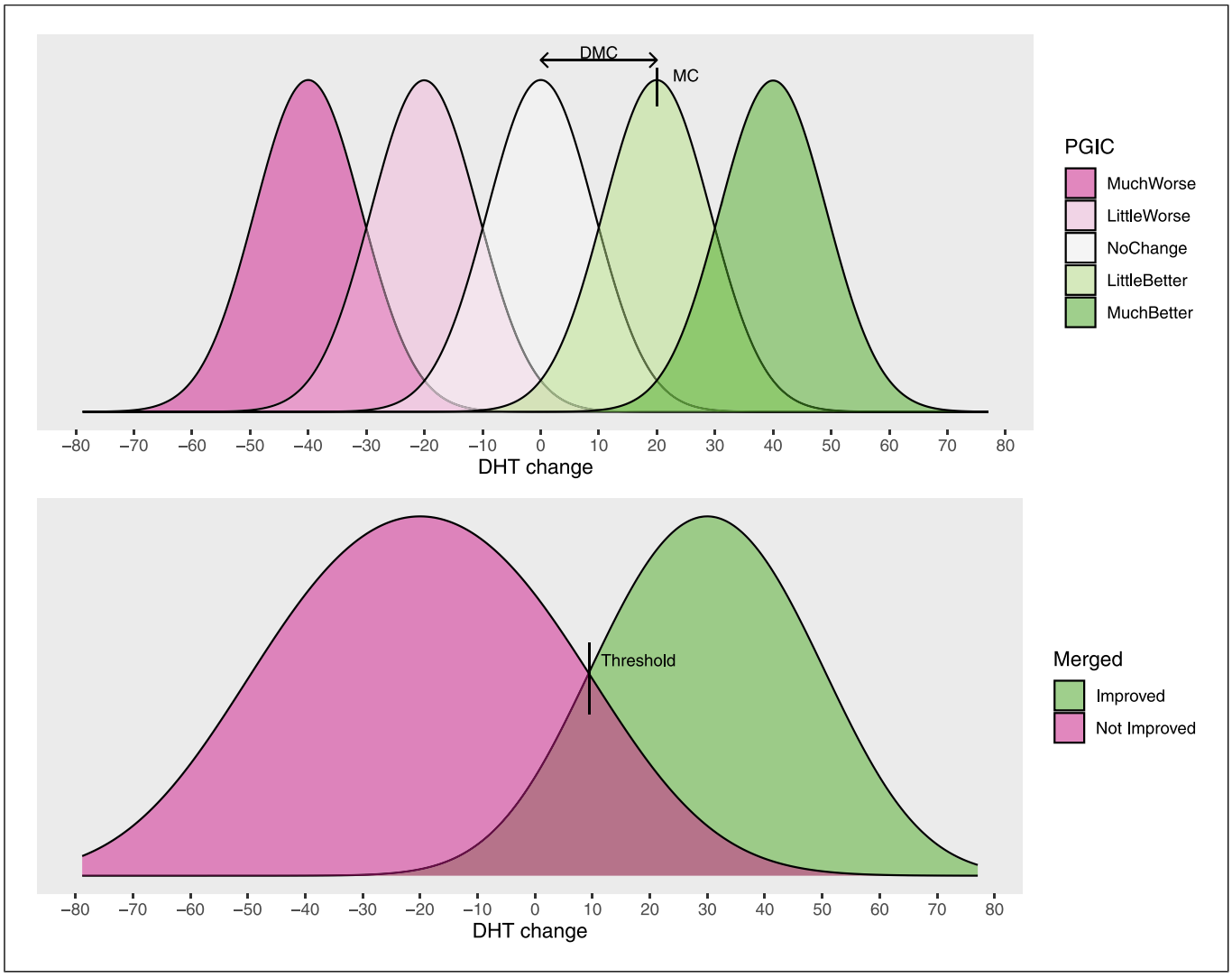

Fig. 1. Graphical representation of anchor-based estimates. DHT, digital health technology; DMC, the difference in mean change between patients who are "a little better" and "no change"; MC, mean change within patients who are "a little better"; PGI-C, patient global impression of change.

Additionally, it has been proposed that the importance of between-group differences can only be judged in light of the overall benefit-risk ratio [57].

The subsequent anchor-based methods described all target the location of the threshold denoted in the lower panel of Figure 1 and are thus more aligned with thresholds for within-individual change [58]. A common approach is the use of receiver operating characteristic (ROC) analysis, where the anchor measure is merged into binary "improved" and "not improved" categories [51, 59]. A central idea behind ROC analysis is that each possible threshold for change on the clinical measure value can also be used to classify patients into categories. The performance of a specific threshold versus the anchor-based classification is summarized in terms of sensitivity (proportion of anchor-based improved patients with clinical measure value changes at or exceeding the threshold) and specificity (proportion of anchor-based not improved patients with clinical measure value changes below the threshold). A ROC curve can be plotted, where the optimal threshold (based on maximizing sensitivity and specificity) can be identified through various formulae [60].

An alternate threshold-based approach, based on a logistic regression model, is referred to as the predictive modeling method [61]. The logistic regression model

comprises a dichotomized anchor variable (improved vs. not improved) as the outcome, and clinical measure value change as the predictor:

$$logit\left(p_i\right) = C + \beta_1 X_i$$

where p_i is the modeled probability of being in the improved anchor category for an individual i, X_i is the clinical measure value change for an individual, C is an intercept, and β_1 is a fixed effect. Terluin et al. [61] demonstrated how the model coefficients can be applied in the following formula to estimate the optimal threshold:

$$Threshold = (I - C)/\beta_1$$

where

$$I = \ln\left(\frac{prop_{improved}}{1 - prop_{improved}}\right)$$

$Prop_{improved}$ is the observed proportion of participants in the "improved" category as defined above. A disadvantage of ROC and the above predictive modeling method is that estimates of the threshold can be biased when the proportion of improved patients is far from 50% [62]. An adjustment formula exists to correct for this [62] as follows:

$$Threshold_{adjusted} = Threshold \pm (0.090 + 0.103 \times |Cor|) \times SD_{change} \times I$$

The above formula is slightly altered from the original to flexibly account for the directionality of scores. If higher scores (and positive changes) equate to better health, the ± becomes a minus symbol and the threshold is reduced when the proportion of improved patients is >50%. If lower scores (and negative changes) equate to better health, the ± becomes a plus symbol and the threshold is increased (i.e., becomes less negative) when the proportion of improved patients is >50%. An absolute point biserial correlation coefficient (*Cor* in the above formula) is used to allow the directionality of clinical measure and anchor to differ. Alternatively to this adjustment method, another adjustment based on the reliability of the anchor [63] or thresholds estimated by latent variable models [64, 65], are also robust to the proportion of improved patients, but calculation is only possible for multi-item electronic COAs rather than all types of DHT. The predictive modeling methods have been shown to outperform ROC analysis in terms of bias and precision [61, 62].

The above methods are all parametric. Nonparametric discriminant analysis can also estimate thresholds, which calculate the densities of change in clinical measure value by anchor group using normal kernels with unequal bandwidth [66, 67]. The threshold at which the kernel densities intersect is taken as the estimate. Notably, while the original paper describing this method [66] framed it in terms of a between-group difference in change over time (i.e., case 1A and 2C in Table 1), subsequent applications have targeted within-individual change over time thresholds [67], which we agree with. While rules of thumb for bandwidth selection are a good starting point [68], we recommend altering the bandwidth manually and visually inspecting the densities.

Two additional descriptive plots can be informative, where their use has been largely driven by inclusion in FDA guidance [23, 46]. The empirical cumulative distribution function (eCDF) plot displays a continuous clinical measure value change on the x-axis and the cumulative percentage of participants experiencing that change on the y-axis. The use of "empirical" denotes the absence of any smoothing. An eCDF curve is plotted for each anchor group so the separation of the curves can be visually compared across a range of potential thresholds. A suitable threshold for minimal within-patient change would have a cumulative percentage >50% in the minimal improvement group, but <50% in the stable group. Probability density functions (PDFs) also plot continuous clinical measure value change on the x-axis, but the (noncumulative) percentage of participants is on the y-axis, plotted as a kernel-smoothed PDF. As such, PDF plots are closely related to discriminant analysis.

Measurement Error

As stated above, thresholds based on measurement error are not appropriate to define interpretative thresholds regarding the *perceived importance* of change, but estimate the degree of measurement error around individual clinical measure values or changes [69]. These benchmarks are useful for understanding the likelihood that observed changes are within the measurement error of the instrument.

The standard error of measurement (SEM) represents the standard deviation of measurement error around an individual clinical measure value [31]. As described in a previous tutorial paper on reliability [5], two versions of the SEM exist based on consistency or absolute agreement definitions. While $SEM_{consistency}$ concerns the rank ordering of individuals, $SEM_{agreement}$ accounts for the agreement of specific clinical measure values and is stricter where $SEM_{agreement} \geq SEM_{consistency}$. However, calculating $SEM_{agreement}$ requires estimates of the individual variance components which are rarely presented in

publications, so commonly researchers will rely on the following formula for $SEM_{consistency}$:

$$SEM_{consistency} = SD \times \sqrt{1 - reliability}$$

where SD is the standard deviation of the clinical measure value (commonly from the baseline visit). The reliability coefficient is most commonly in the form of an intraclass correlation coefficient or internal consistency, which can be taken from an existing publication or possibly estimated in the same dataset as per recommendations in [5]. Alternatives exist that allow reliability, and thus SEM, to vary over the range of observed values [70].

The SEM concerns individual clinical measure values at a point in time. Thresholds for an individual change over time need to recognize that the measurement error is inflated beyond the SEM, due to two measurements being involved [31]. This leads to another important statistic, the minimal detectable change (MDC, also referred to as smallest real difference or coefficient of repeatability):

$$MDC = z \times \sqrt{2} \times SEM$$

where z is the critical value from a normal distribution (e.g., 1.96 for 95%, 1.645 for 90%), and $\sqrt{2}$ is the inflation factor for measurement error at each timepoint. The MDC_{95} (with $z = 1.96$) and MDC_{90} (with $z = 1.645$) are commonly calculated, but any significance level can be set [71]. For example, the MDC_{68} (with $z = 0.994$) can be thought of as the standard deviation of measurement error around an individual's change score. While the MDC is designed for individual-level change, this may be corrected for change at the group level by dividing by the square root of sample size [72]:

$$MDC_{group} = MDC \big/ \sqrt{n}$$

Triangulation

The application of multiple methods and anchors to estimate thresholds will almost always result in a range of values for each threshold [73]. Triangulation is the process of bringing all the estimates together with the aim of converging on a single value or small range of values. Note that estimates for within-group change, between-group difference, and within-individual change should be triangulated separately [25]. Furthermore, we recommend triangulating thresholds for perceived importance separately to thresholds based on measurement error, given they have different conceptual targets.

It is recommended to plot the different estimates with 95% confidence intervals [73], where confidence intervals for thresholds are commonly derived through bootstrapping. Specific to anchor-based analyses, the proximity of anchor to the target DHT tool should also be considered, with more importance to estimates generated from more closely linked concepts and highly correlated pairings. To this end, a correlation-weighted average of anchor-based estimates can be calculated [25], where estimates are weighted by the observed correlations between anchor and change in clinical measure value as follows:

$$M_{weighted} = \frac{\sum_{j=1}^{n} |r_j||x_j|}{\sum_{j=1}^{n} |r_j|}$$

where $|x|$ denotes each [absolute] estimate and $|r|$ denotes the [absolute] correlation coefficient of each anchor-scale combination, for each j of n total estimates. The correlations should be z-transformed prior to their use in the weighted average, and a confidence interval formula is provided [25].

Sample Size Requirements to Estimate Thresholds

Literature on sample size requirements to estimate thresholds is limited, likely in part due to most of these analyses being conducted on data from an existing cohort or clinical trial where sample size was calculated for another purpose. One general guideline for anchor-based estimates, developed with PROs in mind, is based on the precision around estimates [50]. Presented in the form of a credibility instrument to judge published thresholds, 95% CI widths ≤10% of the threshold estimate indicate a threshold is "definitely" precise, or widths ≤25% indicate precision "to a great extent." In the same paper, the authors suggest a minimum sample size of 150 patients to indicate precision "to a great extent" (as an alternate criterion when information on precision is not available). These general guidelines are of limited value given the relationship between precision and sample size differs according to the specific method used. Additionally, methods such as the mean change within a "minimally improved" anchor group only use a subset of the full data, which should be considered.

Ultimately, when working with existing data, we recommend estimating thresholds even when sample size falls short of the above general rules. However, in such scenarios it may be wise to present a range of plausible thresholds rather than a single value, so that sensitivity analyses incorporating this uncertainty can be conducted in future.

Application of Thresholds to Assess Treatment Effect

Once a threshold has been estimated, this can be applied to assess the magnitude of treatment effects (in terms of perceived importance or measurement error).

For group-level thresholds, the estimated between-group treatment difference can simply be numerically compared to the threshold. For within-individual thresholds, it is a two-step process to (1) classify patients as a responder versus not; and (2) compare the proportion of responders between treatment arms. The difference in proportion of responders can be evaluated descriptively or statistically through approaches such as binomial test of proportions, Cochran-Mantel-Haenszel test, or logistic regression. Alternatively, when a range of plausible within-individual thresholds exists, CDF curves can be plotted by treatment arm.

Of note, for a continuous clinical measure, there is a loss of statistical power when dichotomizing this for a within-individual responder analysis [20, 74]. For this reason, it is recommended that clinical trials retain the continuous values for statistical hypothesis testing, with responder-based comparisons as supplementary descriptive analyses [19].

Applied Example

As described earlier, a variety of thresholds are possible based on the type and magnitude of interpretation. The subsequent example focuses on individual-level thresholds for change over time, in terms of minimal perceived change and measurement error.

Data Source

We use data from the randomized placebo-controlled PAGANINI study (NCT04562155) in patients with refractory and/or unexplained chronic cough where the change in 24-h cough count from baseline was assessed [75]. Cough counts were recorded by the VitaloJAK device [76], where participants wore this during six 24-h periods throughout the study: at screening, the day before start of intervention (baseline), and at weeks 2, 4, 8, and 12. The purpose of this example was not to formally validate or to advocate for use of any specific measure of cough count in the refractory chronic cough population. This example is purely illustrative with respect to statistical analyses to estimate interpretative thresholds, and for that purpose we focus on a measure of total cough count within 24 h, focusing on percentage change from baseline to week 4.

In addition to the VitaloJAK device, several PROs were assessed at similar times. A PGI-C asked participants to rate the "overall change in your cough since you started taking the study medication," with responses "very much better," "much better," "a little better," "no change," "a little worse," "much worse," and "very much worse." A PGI-S asked participants to rate "the severity of your cough today," with responses "no cough," "very mild," "mild," "moderate," "severe," "very severe." The Leicester Cough Questionnaire (LCQ) is a 19-item PRO measure that asks about the impact of chronic cough on various aspects of participants' lives using a recall period of 2 weeks [77]. An LCQ total score ranges from 3 to 21 where higher scores indicate better health. The Cough Severity VAS is a single item asking participants to indicate the severity of their cough using a vertically oriented line ordered from 0 (no cough) to 100 (extremely severe cough) [78]. The above PROs were all used as anchor measures, with the following definitions of a minimal improvement: PGI-C "very much better," "much better," or "a little better"; PGI-S 1-category improvement; LCQ total score change ≥1.7 [79]; VAS change ≤−20 mm [78].

Analyses and Results

Of the 310 participants enrolled and randomized in the PAGANINI study, 308 had at least one cough recording so were eligible for this analysis (demographic characteristics reported in online suppl. Table S2). Participants had a mean (SD) baseline 24-h cough count of 27.69 (28.41). The majority of participants improved on each anchor, where all were sufficiently correlated with the observed percentage change in 24-h cough frequency to warrant their further use (Table 3).

ROC curves were plotted for each anchor, to help identify thresholds that optimally discriminate between patients considered "improved" (at least a minimal improvement on the anchor) and "not improved" (stable or worsened on the anchor). The optimal threshold was chosen using the "sum of squares" method [60]: the one which minimized $([1 - \text{sensitivity}]^2 + [1 - \text{specificity}]^2)$. The optimal thresholds for each anchor are presented in Table 4, and the ROC curves are presented in online supplementary Figures S1–S4.

To implement the predictive modeling method, logistic regression models were fitted with each dichotomized anchor ("improved" vs. "not improved") as the outcome and percentage change in cough count as the predictor. When the proportion of improvement was greater than 50% (PGI-C anchor only; see Table 3), the adjusted logistic regression threshold was shrunk closer to zero; when the proportion was less than 50%, the adjusted threshold became more negative (Table 4, adjustment formula presented in statistical methods section). Thresholds estimated through discriminant analysis are also displayed in Table 4, with the plot for PGI-C in Figure 2 (plots for other anchors in online suppl. Fig. S5–S7). An eCDF plot helps demonstrate

Table 3. Anchor measure proportion improved and correlations, percentage change from baseline to week 4

Anchor measure	*N*	Proportion improved	Correlation with percentage change in 24-h cough frequency	*z*-transformed correlation
PGI-C	262	64.1%	0.448 (polyserial)	0.482
PGI-S	263	43.3%	0.372 (polyserial)	0.391
LCQ total score	263	43.0%	−0.409 (Spearman's)	−0.434
Cough severity VAS	248	22.6%	0.370 (Spearman's)	0.388

Correlations were based on the original anchor measure values rather than the dichotomized versions, where polyserial coefficients were applied for the ordinal PGI-C and PGI-S. Fisher z transformation defined as $z = 0.5 \times \ln((1 + r)/(1 - r))$.

Table 4. Threshold estimates for within-patient change, percentage change from baseline to week 4 in 24-h cough frequency

Anchor measure	ROC threshold	Logistic regression threshold (unadjusted)	Logistic regression threshold (adjusted)	Discriminant analysis threshold
PGI-C	−33.07	−21.60	−18.14	−27.80
PGI-S	−37.29	−27.19	−28.77	−25.15
LCQ total score	−37.10	−28.03	−29.75	−34.05
Cough severity VAS	−50.43	−37.39	−44.26	−46.55

that, for the PGI-C anchor, thresholds between −16 and −35 successfully exclude the majority of patients indicating "no change," but capture the majority of patients indicating "a little better" or "much better" (online suppl. Fig. S8, S9).

As shown in Table 4, threshold estimates vary by method and anchor. The unadjusted logistic regression thresholds can be disregarded, given the adjusted approach should provide more accurate values. The ROC-based estimates are consistently the largest in magnitude, followed by discriminant analysis thresholds for three of the anchors. Estimates from the VAS anchor are notably higher than the other anchors, nevertheless are still considered valid given the conceptual appropriateness of the VAS.

The varied estimates were triangulated using a correlation-weighted average of anchor-based estimates. While the ROC-based estimates were calculated for the purposes of the worked example, they were disregarded from the triangulation given this method is outperformed by adjusted logistic regression in simulations [61, 62]. Given the eight estimates from adjusted logistic regression and discriminant analysis in Table 4, and the z-transformed correlations in Table 3, the correlation-weighted average was calculated as follows:

$$M_{\text{weighted}} = \frac{\begin{array}{c}(-18.14\times0.482)+(-28.77\times0.391)\\+(-29.75\times0.434)+(-44.26\times0.388)+\\(-27.80\times0.482)+(-25.15\times0.391)\\+(-34.05\times0.434)+(-46.55\times0.388)\end{array}}{(0.482+0.391+0.434+0.388+0.482+0.391+0.434+0.388)} = -31.31$$

Yielding a triangulated threshold of −31.31 with 95% confidence intervals −24.77 and −37.85. The $SEM_{agreement}$ and $SEM_{consistency}$ for 24-h cough frequency were the same value of 7.14, as the variance component due to visit was practically zero (if it was nonzero, the $SEM_{agreement}$ would have been higher). The MDC_{90} was 16.61, and MDC_{68} was 10.04; however, this value is still on the scale of absolute values of cough frequency rather than percentage change. To calculate MDC_{90} on a percentage change scale, the following approach [80] was employed:

$$MDC_{\%\ change} = \frac{MDC}{Baseline\_Mean} \times 100 = \frac{16.61}{27.69} \times 100 = 59.98\%$$

The corresponding MDC_{68} on a percentage scale is 36.26%. The MDC_{90} for percentage change is considerably larger than the anchor-based thresholds, whereas the MDC_{68} for percentage change is within

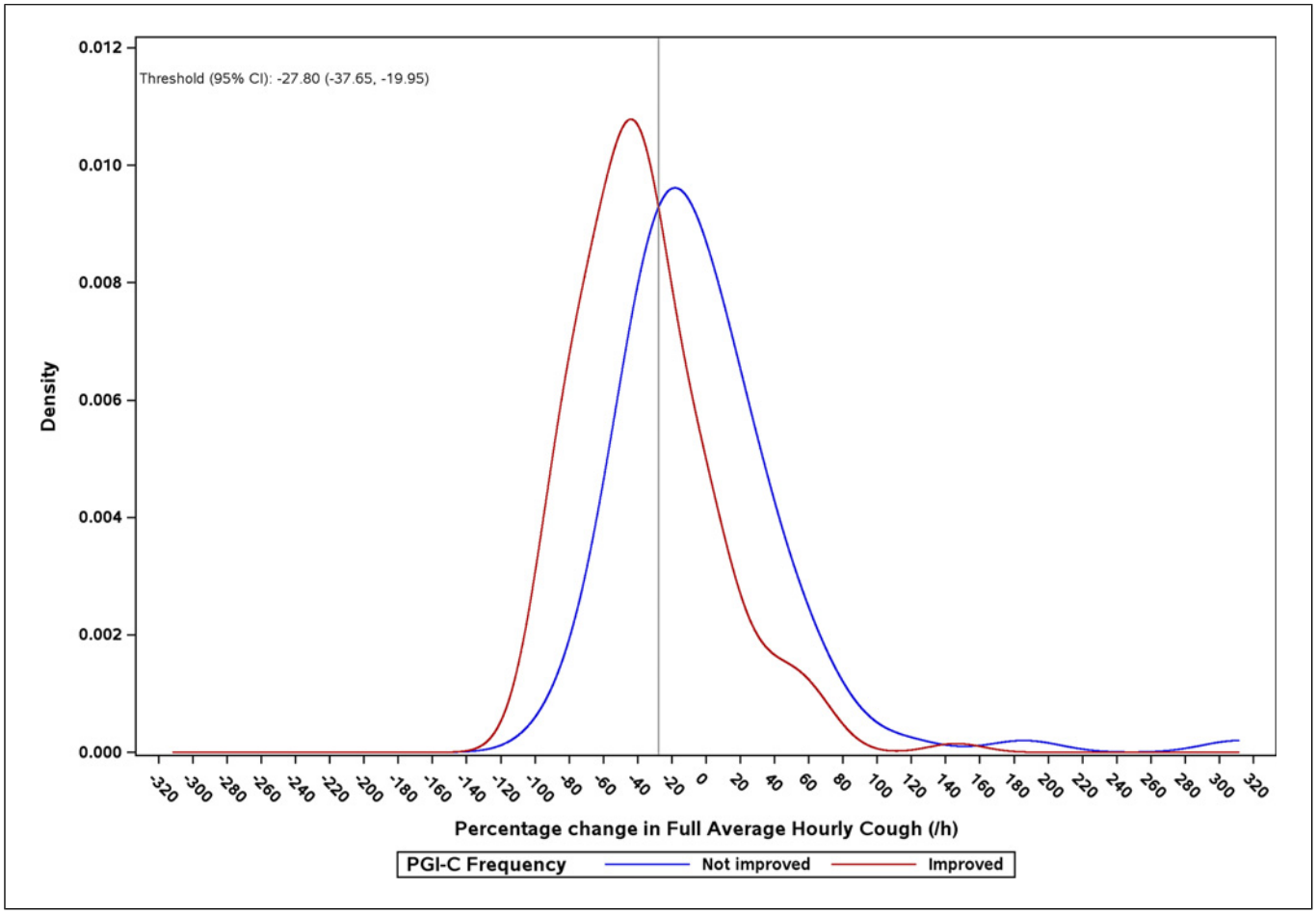

Fig. 2. Discriminant analysis for PGI-C anchor, percentage change from baseline to week 4 in 24-h cough frequency.

the 95% confidence intervals of the triangulated threshold. This scenario presents the following question: do I need to increase the triangulated anchor-based threshold to surpass the MDC values? We would again stress that anchor-based thresholds for perceived importance have a different conceptual target to thresholds based on measurement error, so should exist separately. While ideally the MDC will be smaller than the triangulated anchor-based threshold, the necessity of this is context-specific. For individual patient management where observing an improvement beyond the threshold could prompt a decision to come off treatment, extra caution may be warranted to ensure very few patients are wrongly assumed as improved due to measurement error. On the other hand, in a clinical trial aiming to compare the proportion of responders in each treatment arm, measurement error will be averaged out when calculating the proportion so there is not a strict requirement to surpass the MDC. It is important to remember that measurement error around a change score is bidirectional: some truly stable patients will appear improved due to measurement error added to their score change, while some truly improved patients will appear stable due to measurement error subtracted from their score change.

Conclusion

Several methods to estimate thresholds to interpret change over time have been presented, including a worked example focusing on individual-level change. This aims to equip researchers with the knowledge required to conduct such analyses in practice. With that in mind, as part of the clinical validation of a novel digital measure, interpretative thresholds should be estimated with the following recommendations in mind:

- Follow the V3 validation framework [1], by ensuring prior evidence exists to support verification and analytical validation.
- Be mindful of the desired level of interpretation, type of comparison, and magnitude so that the target threshold is well defined.
- To assess thresholds for interpreting change, design a study or use an existing data source where
  - The sample adequately reflects the intended context of use for application of threshold (i.e., same diagnosis and severity level)
  - Longitudinal data on the target DHT value have been collected, in a setting and timeframe where change over time can be expected in a considerable proportion of subjects (e.g., an interventional study)
  - One or more plausible anchor measures are administered at times aligned with the target DHT
- When calculating anchor-based thresholds
  - Ideally use multiple anchor measures, with a clear conceptual overlap with the measure of interest, and demonstrate correlation at least 0.3 in magnitude
  - For within-individual change, thresholds derived through logistic regression (adjusted for the proportion of improved subjects) and discriminant analysis are recommended as a minimum
  - Triangulate the various anchor-based thresholds using a correlation-weighted average
- Thresholds based on measurement error should be calculated and presented, but should not inform the triangulation of thresholds based on perceived importance.

We acknowledge that some DHT tools may have updates to the underlying algorithms used to derive clinical measures, including machine learning and artificial intelligence-based algorithms that continuously update. In such cases, we recommend periodical reevaluation of thresholds to assess their stability. In most of these cases, the same COI is being targeted so we expect interpretative thresholds to remain stable.

Despite our best efforts to provide a thorough overview of the topic, some more advanced methodological considerations have been omitted. One consideration requiring further research is that the methods described to estimate thresholds for within-individual change are all reliant on between-subject variability. The assumption that results based on between-subject variability are transferable to the subject level is called ergodicity [81], but is largely untested in this context. The use of within-subject standard deviations for measurement error statistics has been proposed, although many repeated measurements are necessary to yield accurate estimates [24]. This may therefore be an option for continuous or daily measurements. In addition, during the period when this manuscript was drafted, the FDA released a draft guidance covering interpretative thresholds [82], but we chose to reserve commenting or summarizing this until the guidance is finalized to avoid the information becoming quickly outdated.

This review mainly concerns thresholds based on perceived importance. However, in the context of DHT prognostic biomarkers (or surrogate endpoints), it could be the case that early imperceptible changes could predict changes that would become noticeably important later. In such cases, the need for a threshold based on perceived importance is questionable, and whichever threshold provided the highest predictive accuracy is likely suitable.

This paper summarizes statistical methods for interpreting change to be applied during clinical validation of novel digital measures. We hope this improves understanding of such methods and the quality of proposed interpretative thresholds in the field.

Conflict of Interest Statement

All authors are salaried employees of Bayer.

Funding Sources

The authors have no funding sources to disclose.

Author Contributions

A.T.: conceptualization, methodology, software, and writing – original draft. B.R.: conceptualization, methodology, software, and writing – review and editing. F.K., M.M., and A.P.: software and writing – review and editing. U.K.: data curation and writing – review and editing.

References

1 Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). Npj Digit Med. 2020;3:55–15. https://doi.org/10.1038/s41746-020-0260-4

2 FDA. Digital health technologies for remote data acquisition in clinical investigations guidance for industry, investigators, and other stakeholders; 2023.

3 FDA. Framework for the use of DHTs in drug and biological product development; 2023.
4 Ratitch B, Rodriguez-Chavez IR, Dabral A, Fontanari A, Vega J, Onorati F, et al. Considerations for analyzing and interpreting data from biometric monitoring technologies in clinical trials. DIB. 2022;6(3):83–97. https://doi.org/10.1159/000525897
5 Ratitch B, Trigg A, Majumder M, Vlajnic V, Rethemeier N, Nkulikiyinka R. Clinical validation of novel digital measures: statistical methods for reliability evaluation. Digit Biomark. 2023;7(1):74–91. https://doi.org/10.1159/000531054
6 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15. https://doi.org/10.1016/0197-2456(89)90005-6
7 Mc Carthy M, Burrows K, Griffiths P, Black PM, Demanuele C, Karlsson N, et al. From meaningful outcomes to meaningful change thresholds: a path to progress for establishing digital endpoints. Ther Innov Regul Sci. 2023;57(4):629–45. https://doi.org/10.1007/s43441-023-00502-8
8 Byrom B, Breedon P, Tulkki-Wilke R, Platko J. Meaningful change: defining the interpretability of changes in endpoints derived from interactive and mHealth technologies in healthcare and clinical research. J Rehabil Assist Technol Eng. 2020;7:2055668319892778. https://doi.org/10.1177/2055668319892778
9 Natale V, Plazzi G, Martoni M. Actigraphy in the assessment of insomnia: a quantitative approach. Sleep. 2009;32(6):767–71. https://doi.org/10.1093/sleep/32.6.767
10 Rhudy M, Dreisbach S, Moran M, Ruggiero M, Veerabhadrappa P. Cut points of the Actigraph GT9X for moderate and vigorous intensity physical activity at four different wear locations. J Sports Sci. 2020;38(5):503–10. https://doi.org/10.1080/02640414.2019.1707956
11 Zaturenskaya M, A Jason L, Torres-Harding S, W Tryon W. Subgrouping in chronic fatigue syndrome based on actigraphy and illness severity. Open Biol J. 2009;2(1):20–6. https://doi.org/10.2174/1874196700902010020
12 Maksymowych WP, Richardson R, Mallon C, van der Heijde D, Boonen A. Evaluation and validation of the patient acceptable symptom state (PASS) in patients with ankylosing spondylitis. Arthritis Rheum. 2007;57(1):133–9. https://doi.org/10.1002/art.22469
13 Sahlberg L, Lapinleimu H, Elovainio M, Rönnlund H, Virtanen I. Normative values for sleep parameters in pre-schoolers using actigraphy. Clin Neurophysiol. 2018;129(9):1964–70. https://doi.org/10.1016/j.clinph.2018.06.027
14 Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Routledge; 1988.
15 Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? Expert Rev Pharmacoecon Outcomes Res. 2004;4(5):515–23. https://doi.org/10.1586/14737167.4.5.515
16 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. 2002;77(4):371–83. https://doi.org/10.4065/77.4.371
17 Mouelhi Y, Jouve E, Castelli C, Gentile S. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. Health Qual Life Outcomes. 2020;18(1):136. https://doi.org/10.1186/s12955-020-01344-w
18 Trigg A, Lenderking WR, Boehnke JR. Introduction to the special section: "Methodologies and considerations for meaningful change.". Qual Life Res. 2023;32(5):1223–30. https://doi.org/10.1007/s11136-023-03413-1
19 Collister D, Bangdiwala S, Walsh M, Mian R, Lee SF, Furukawa TA, et al. Patient reported outcome measures in clinical trials should be initially analyzed as continuous outcomes for statistical significance and responder analyses should be reserved as secondary analyses. J Clin Epidemiol. 2021;134:95–102. https://doi.org/10.1016/j.jclinepi.2021.01.026
20 Snapinn SM, Jiang Q. Responder analyses and the assessment of a clinically relevant treatment effect. Trials. 2007;8:31. https://doi.org/10.1186/1745-6215-8-31
21 Voors AA, Angermann CE, Teerlink JR, Collins SP, Kosiborod M, Biegus J, et al. The SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure: a multinational randomized trial. Nat Med. 2022;28(3):568–74. https://doi.org/10.1038/s41591-021-01659-1
22 Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. Am J Epidemiol. 2014;179(2):226–35. https://doi.org/10.1093/aje/kwt212
23 FDA. Public workshop on patient-focused drug development: guidance 4 – incorporating clinical outcome assessments into endpoints for regulatory decision making. FDA; 2019. Available from: https://www.fda.gov/drugs/development-approval-process-drugs/public-workshop-patient-focused-drug-development-guidance-4-incorporating-clinical-outcome (accessed March 30, 2023).
24 Hays RD, Peipert JD. Between-group minimally important change versus individual treatment responders. Qual Life Res. 2021;30(10):2765–72. https://doi.org/10.1007/s11136-021-02897-z
25 Trigg A, Griffiths P. Triangulation of multiple meaningful change thresholds for patient-reported outcome scores. Qual Life Res. 2021;30(10):2755–64. https://doi.org/10.1007/s11136-021-02957-4
26 Musoro ZJ, Hamel J-F, Ediebah DE, Cocks K, King MT, Groenvold M, et al. Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-analysis protocol. BMJ Open. 2018;8(1):e019117. https://doi.org/10.1136/bmjopen-2017-019117
27 King MT, Dueck AC, Revicki DA. Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? Med Care. 2019;57 Suppl 1(Suppl 5 1):S38–45. https://doi.org/10.1097/MLR.0000000000001111
28 Coon CD, Cappelleri JC. Interpreting change in scores on patient-reported outcome instruments. Ther Innov Regul Sci. 2016;50(1):22–9. https://doi.org/10.1177/2168479015622667
29 Bollen KA, Barb KH. Pearson's R and coarsely categorized measures. Am Sociol Rev. 1981;46(2):232–9. https://doi.org/10.2307/2094981
30 Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods. 2012;17(3):354–73. https://doi.org/10.1037/a0029315
31 Harvill LM. Standard error of measurement. Educ Meas. 1991;10(2):33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x
32 Terluin B, Griffiths P, Trigg A, Terwee CB, Bjorner JB. Present state bias in transition ratings was accurately estimated in simulated and real data. J Clin Epidemiol. 2022;143:128–36. https://doi.org/10.1016/j.jclinepi.2021.12.024
33 Wyrwich KW, Norman GR. The challenges inherent with anchor-based approaches to the interpretation of important change in clinical outcome assessments. Qual Life Res. 2023;32(5):1239–46. https://doi.org/10.1007/s11136-022-03297-7
34 Kelly K, Cardellino A, Shah M, Hanlon J, Taiyari S, Roborel de Climens A, et al. PCR37 understanding symptoms and meaningful change in patient global impression scales for patients with advanced/metastatic non-small cell lung cancer (adv/met NSCLC): a qualitative research study. Value in Health. 2022;25(12):S397. https://doi.org/10.1016/j.jval.2022.09.1972
35 Aili K, Åström-Paulsson S, Stoetzer U, Svartengren M, Hillert L. Reliability of actigraphy and subjective sleep measurements in adults: the design of sleep assessments. J Clin Sleep Med. 2017;13(1):39–47. https://doi.org/10.5664/jcsm.6384
36 Sully K, Trigg A, Bonner N, Moreno-Koehler A, Trennery C, Shah N, et al. Estimation of minimally important differences and responder definitions for EORTC QLQ-MY20 scores in multiple myeloma patients. Eur J Haematol. 2019;103(5):500–9. https://doi.org/10.1111/ejh.13316

37 Teylan M, Kantorowski A, Homsy D, Kadri R, Richardson C, Moy M. Physical activity in COPD: minimal clinically important difference for medical events. Chron Respir Dis. 2019;16:1479973118816424. https://doi.org/10.1177/1479973118816424
38 Polgar O, Patel S, Walsh JA, Barker RE, Clarke SF, Man WD-C, et al. Minimal clinically important difference for daily pedometer step count in COPD. ERJ Open Res. 2021;7(1):00823-2020-2020. https://doi.org/10.1183/23120541.00823-2020
39 Chen H-L, Lin K-C, Hsieh Y-W, Wu C-Y, Liing R-J, Chen C-L. A study of predictive validity, responsiveness, and minimal clinically important difference of arm accelerometer in real-world activity of patients with chronic stroke. Clin Rehabil. 2018;32(1):75–83. https://doi.org/10.1177/0269215517712042
40 Shoemaker MJ, Curtis AB, Vangsnes E, Dickinson MG. Clinically meaningful change estimates for the six-minute walk test and daily activity in individuals with chronic heart failure. Cardiopulm Phys Ther J. 2013;24(3):21–9. https://doi.org/10.1097/01823246-201324030-00004
41 Demeyer H, Burtin C, Hornikx M, Camillo CA, Van Remoortel H, Langer D, et al. The minimal important difference in physical activity in patients with COPD. PLoS One. 2016;11(4):e0154587. https://doi.org/10.1371/journal.pone.0154587
42 Staunton H, Willgoss T, Nelsen L, Burbridge C, Sully K, Rofail D, et al. An overview of using qualitative techniques to explore and define estimates of clinically important change on clinical outcome assessments. J Patient Rep Outcomes. 2019;3(1):16. https://doi.org/10.1186/s41687-019-0100-y
43 Cocks K, King MT, Velikova G, de Castro G, Martyn St-James M, Fayers PM, et al. Evidence-based guidelines for interpreting change scores for the European organisation for the research and treatment of cancer quality of life Questionnaire core 30. Eur J Cancer. 2012;48(11):1713–21. https://doi.org/10.1016/j.ejca.2012.02.059
44 Bingham CO III, Butanis AL, Orbai AM, Jones M, Ruffing V, Lyddiatt A, et al. Patients and clinicians define symptom levels and meaningful change for PROMIS pain interference and fatigue in RA using bookmarking. Rheumatology. 2021;60(9):4306–14. https://doi.org/10.1093/rheumatology/keab014
45 Thissen D, Liu Y, Magnus B, Quinn H, Gipson DS, Dampier C, et al. Estimating minimally important difference (MID) in PROMIS pediatric measures using the scale-judgment method. Qual Life Res. 2016;25(1):13–23. https://doi.org/10.1007/s11136-015-1058-8
46 FDA. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims; 2009.
47 IQWIG. General methods - version 6.1. General methods; 2022.
48 EMA. Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man; 2016.
49 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102–9. https://doi.org/10.1016/j.jclinepi.2007.03.012
50 Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020;369:m1714. https://doi.org/10.1136/bmj.m1714
51 Terwee CB, Peipert JD, Chapman R, Lai J-S, Terluin B, Cella D, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res. 2021;30(10):2729–54. https://doi.org/10.1007/s11136-021-02925-y
52 Sabah SA, Alvand A, Beard DJ, Price AJ. Minimal important changes and differences were estimated for Oxford hip and knee scores following primary and revision arthroplasty. J Clin Epidemiol. 2022;143:159–68. https://doi.org/10.1016/j.jclinepi.2021.12.016
53 Dirven L, Musoro JZ, Coens C, Reijneveld JC, Taphoorn MJB, Boele FW, et al. Establishing anchor-based minimally important differences for the EORTC QLQ-C30 in glioma patients. Neuro Oncol. 2021;23(8):1327–36. https://doi.org/10.1093/neuonc/noab037
54 Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res. 2002;11(3):207–21. https://doi.org/10.1023/a:1015276414526
55 Bell ML, Dhillon HM, Bray VJ, Vardy JL. Important differences and meaningful changes for the functional assessment of cancer therapy-cognitive function (FACT-Cog). J Patient Rep Outcomes. 2018;2(1):48. https://doi.org/10.1186/s41687-018-0071-4
56 Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol. 2017;82:128–36. https://doi.org/10.1016/j.jclinepi.2016.11.016
57 Dworkin RH, Turk DC, McDermott MP, Peirce-Sandner S, Burke LB, Cowan P, et al. Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. Pain. 2009;146(3):238–44. https://doi.org/10.1016/j.pain.2009.08.019
58 Vanier A, Sébille V, Blanchin M, Hardouin J-B. The minimal perceived change: a formal model of the responder definition according to the patient's meaning of change for patient-reported outcome data analysis and interpretation. BMC Med Res Methodol. 2021;21(1):128. https://doi.org/10.1186/s12874-021-01307-9
59 Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. J Clin Epidemiol. 2000;53(3):279–84. https://doi.org/10.1016/S0895-4356(99)00140-7
60 Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. Theoretical considerations and an example application of change in health status. PLoS One. 2014;9(12):e114468. https://doi.org/10.1371/journal.pone.0114468
61 Terluin B, Eekhout I, Terwee CB, de Vet HCW. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. J Clin Epidemiol. 2015;68(12):1388–96. https://doi.org/10.1016/j.jclinepi.2015.03.015
62 Terluin B, Eekhout I, Terwee CB. The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. J Clin Epidemiol. 2017;83:90–100. https://doi.org/10.1016/j.jclinepi.2016.12.015
63 Terluin B, Eekhout I, Terwee CB. Improved adjusted minimal important change took reliability of transition ratings into account. J Clin Epidemiol. 2022;148:48–53. https://doi.org/10.1016/j.jclinepi.2022.04.018
64 Bjorner JB, Terluin B, Trigg A, Hu J, Brady KJS, Griffiths P. Establishing thresholds for meaningful within-individual change using longitudinal item response theory. Qual Life Res. 2023;32(5):1267–76. https://doi.org/10.1007/s11136-022-03172-5
65 Terluin B, Trigg A, Fromy P, Schuller W, Terwee CB, Bjorner JB. Estimating anchor-based minimal important change using longitudinal confirmatory factor analysis. Qual Life Res. 2024;33(4):963–73. https://doi.org/10.1007/s11136-023-03577-w
66 Gerlinger C, Schmelter T. Determining the non-inferiority margin for patient reported outcomes. Pharm Stat. 2011;10(5):410–3. https://doi.org/10.1002/pst.507
67 Gerlinger C, Gude K, Hiemeyer F, Schmelter T, Schäfers M. An empirically validated responder definition for the reduction of moderate to severe hot flushes in postmenopausal women. Menopause. 2012;19(7):799–803. https://doi.org/10.1097/gme.0b013e31823de8ba

68 Jones MC, Marron JS, Sheather SJ. A brief survey of bandwidth selection for density estimation. J Am Stat Assoc. 1996;91(433):401–7. https://doi.org/10.1080/01621459.1996.10476701
69 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. 2006;4:54. https://doi.org/10.1186/1477-7525-4-54
70 Lee MK, Peipert JD, Cella D, Yost KJ, Eton DT, Novotny PJ, et al. Identifying meaningful change on PROMIS short forms in cancer patients: a comparison of item response theory and classic test theory frameworks. Qual Life Res. 2023;32(5):1355–67. https://doi.org/10.1007/s11136-022-03255-3
71 Peipert JD, Hays RD, Cella D. Likely change indexes improve estimates of individual change on patient-reported outcomes. Qual Life Res. 2023;32(5):1341–52. https://doi.org/10.1007/s11136-022-03200-4
72 van Kampen DA, Willems WJ, van Beers LWAH, Castelein RM, Scholtes VAB, Terwee CB. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). J Orthop Surg Res. 2013;8:40. https://doi.org/10.1186/1749-799X-8-40
73 Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes. 2006;4:70. https://doi.org/10.1186/1477-7525-4-70
74 Senn S. Disappointing dichotomies. Pharm Stat. 2003;2(4):239–40. https://doi.org/10.1002/pst.90
75 Dicpinigaitis PV, Morice AH, Smith JA, Sher MR, Vaezi M, Guilleminault L, et al. Efficacy and safety of eliapixant in refractory chronic cough: the randomized, placebo-controlled phase 2b PAGANINI study. Lung. 2023;201(3):255–66. https://doi.org/10.1007/s00408-023-00621-x
76 Smith JA, Holt K, Dockry R, Sen S, Sheppard K, Turner P, et al. Performance of a digital signal processing algorithm for the accurate quantification of cough frequency. Eur Respir J. 2021;58(2):2004271. https://doi.org/10.1183/13993003.04271-2020
77 Birring SS, Prudon B, Carr AJ, Singh SJ, Morgan MDL, Pavord ID. Development of a symptom specific health status measure for patients with chronic cough: leicester Cough Questionnaire (LCQ). Thorax. 2003;58(4):339–43. https://doi.org/10.1136/thorax.58.4.339
78 Martin Nguyen A, Bacci ED, Vernon M, Birring SS, Rosa CL, Muccino D, et al. Validation of a visual analog scale for assessing cough severity in patients with chronic cough. Ther Adv Respir Dis. 2021;15:17534666211049743. https://doi.org/10.1177/17534666211049743
79 Nguyen AM, Schelfhout J, Muccino D, Bacci ED, La Rosa C, Vernon M, et al. Leicester Cough Questionnaire validation and clinically important thresholds for change in refractory or unexplained chronic cough. Ther Adv Respir Dis. 2022;16:17534666221099737. https://doi.org/10.1177/17534666221099737
80 Salvatore MD, Colacino AM, Hess ME, Todd SW, Saunders NW. Concurrent validity and minimum detectable change of Senior Fitness Test components: instrumented vs. manual assessment. Phys Ther Rehabil. 2017;4(1):13. https://doi.org/10.7243/2055-2386-4-13
81 Molenaar PCM, Campbell CG. The new person-specific paradigm in psychology. Curr Dir Psychol Sci. 2009;18(2):112–7. https://doi.org/10.1111/j.1467-8721.2009.01619.x
82 FDA. Patient-focused drug development: Incorporating clinical outcome assessments into endpoints for regulatory decision-making. Draft Guidance; 2023.