


Meaningful change: Defining the interpretability of changes in endpoints derived from interactive and mHealth technologies in healthcare and clinical research

Journal of Rehabilitation and Assistive Technologies Engineering
Volume 7: 1–8
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2055668319892778
journals.sagepub.com/home/jrt


B Byrom¹ , P Breedon², R Tulkki-Wilke³ and JV Platko⁴

Abstract

Immersive, interactive and mHealth technologies are increasingly being used in clinical research, healthcare and rehabilitation solutions. Leveraging technology solutions to derive new and novel clinical outcome measures is important to the ongoing assessment of clinical interventions. While demonstrating statistically significant changes is an important element of intervention assessment, understanding whether changes detected reflect changes of a magnitude that are considered meaningful to patients is equally important. We describe methodologies used to determine meaningful change and recommend that these techniques are routinely included in the development and testing of clinical assessment and rehabilitation technology solutions.

Keywords

Meaningful change, outcome measurement, interpretability, clinical endpoints, statistical analysis (medical)

Date received: 7 January 2019; accepted: 1 November 2019

Introduction

Miniaturisation of sensors and circuitry has given rise to huge proliferation in the development and commercialisation of wearable and sensor-based technologies with application to health and wellness. Much research and development activity has involved the application of interactive and immersive technologies in the areas of healthcare and rehabilitation. Immersive technologies are those that merge the physical world and the digital or simulated world, thereby creating a sense of immersion, such as virtual reality (VR) applications. In healthcare, VR systems, for example, have shown promise in improving outcomes such as muscle balance, dexterity and grip in comparison to traditional rehabilitation in upper limb rehabilitation after spinal cord injury,¹ and to improve the emotional health of cancer patients and decrease associated disease-related psychological symptoms.² Applications using motion-based gaming platform technology, in particular using the Microsoft Kinect[®] (Microsoft Corp., Redmond, Washington, USA), have been reported to provide

therapeutic benefits in areas such as stroke,³ Parkinson's disease,^{4,5} multiple sclerosis⁶ and cerebral palsy.⁷ Mobile Health (mHealth – the use of mobile and wireless devices to improve health outcomes, healthcare services and health research⁸) solutions have, for example, been reported to improve outcomes such as working memory, concentration and interference processing in children suffering from attention-deficit hyperactivity disorder.⁹

In addition to providing the basis for new treatment interventions, interactive technologies are being utilised to develop new ways to measure health outcomes in

¹Product Management, Signant Health, London, UK

²Medical Design Research Group, Nottingham Trent University, Nottingham, UK

³Product Management, Signant Health, Helsinki, Finland

⁴ECO Science, Signant Health, Plymouth Meeting, USA

Corresponding author:

B Byrom, Product Management, Signant Health, London, UK.
Email: bill.byrom@signanthealth.com



clinical research. This enables researchers to track and monitor changes in health status over time or resulting from treatments such as the application of new pharmaceutical products. For example, the Microsoft Kinect has been used to measure performance-based health outcomes such as measures of gait and balance in indications including multiple sclerosis,^{10,11} stroke¹² and Parkinson's disease,⁴ and measures of upper extremity range of motion in adhesive capsulitis¹³ and stroke.¹⁴ mHealth applications, in particular those leveraging the inbuilt sensors and components of modern smartphones, have been used to measure new and novel health outcomes in a variety of indications, speeded by the availability of platforms such as Apple ResearchKit and Google StudyKit. For example, Roche Pharmaceuticals (Basel, Switzerland) has developed an innovative mHealth platform and app that enables the measurement of a number of health outcomes via active performance tests conducted by the patient. These leverage Android smartphone components including the accelerometer, gyroscope and touchscreen.¹⁵ Roche have recently implemented this application to study phonation, tremor, balance, gait and dexterity in Parkinson's disease clinical trials.¹⁶

Over the past two decades, clinical drug development programmes have become increasingly complex and have included a greater volume and variety of assessments and clinical procedures. In particular, there has been increased interest and uptake in the use of new technologies, including gaming platforms and sensor-based solutions to provide richer information on the efficacy and safety of new potential medications. Pfizer (New York, NY), for example, has recently reported positive results from a study evaluating the measures of multitasking and interference processing performance collected while using the Project: EVO videogame (Akili Interactive Labs, Boston, MA) as biomarkers to enable the selection and longitudinal assessment of Alzheimer's patients.¹⁷ Clinical drug developers are also seeking to understand how to leverage other interactive technologies to measure and track intervention effects, including (for example) wearable sensors, motion-based gaming platforms and VR applications.

Healthcare solutions, clinical research and clinical trials rely upon robust and validated methodologies to measure health status and to detect treatment-related changes over time. This enables the efficacy and safety of new interventions to be accurately assessed and measured. In some cases, these evaluations rely upon subjective assessments by the investigator or the patient, such as patient-reported outcome measures (PROMs) and clinician-reported outcomes (ClinROs). While important, in some instances these subjective measures may not be sensitive enough to

detect treatment-related changes and may be unable to conclusively demonstrate treatment effects when they exist. Clinical trials using subjective clinician assessments often must include comprehensive rater training to limit the effects of intra- and inter-rater variability affecting the sensitivity of endpoints measured.¹⁸ For example, a primary reason suggested for the inability to detect differences between treatment groups in a large multinational trial of fluoxetine compared to placebo in geriatric depression was the variability between raters in the subjectively assessed ClinRO used as the primary endpoint.¹⁹ Leveraging mHealth and interactive technologies may provide an opportunity to supplement treatment evaluations with new, novel and objective health outcome measures that may permit increased precision to detect treatment effects and/or enable the measurement of constructs not previously possible. Understanding the clinical relevance of changes detected is of vital importance to understand if an intervention is producing the magnitude of change that will be seen to impact patients. This applies both to the use of interactive technologies to measure intervention effects and the understanding of effects of using interactive technologies when they are intended as an intervention.

In this article, we explore a recommended approach to defining meaningful change in new health outcome measures. This is an element that is largely overlooked in the development and assessment of immersive, interactive and mHealth technologies developed as interventions and technologies to measure health status. We recommend that this becomes a component of rehabilitation and clinical assessment solution development and testing when using immersive, interactive, wearable and mHealth technologies.

Clinical outcomes and endpoints

Clinical outcomes are measurable characteristics influenced by an individual's baseline state or an intervention.²⁰ These might include estimates of functional reaching volumes within immersive game-driven applications, free-living activity measurements using an accelerometer or dexterity measures using an mHealth tapping test on a mobile phone. Outcome assessments are used to define efficacy endpoints when developing a therapy for a disease or condition. Interactive and mHealth technologies have the potential to produce many possible clinical outcome measures, and determining those important in measuring pertinent aspects of health status that are important to the patient or condition studied is essential.

A clinical endpoint is defined as: "*A characteristic or variable that reflects how a patient feels, functions, or survives*".²¹ Endpoint descriptions should also include

a definition of how and when they are measured, how they are calculated from outcome data, rules for missing data and how they should be analysed. For example, an endpoint derived from the outcome data collected using a wearable continuous glucose monitor may be the change from baseline in mean daily time within target range measured over a seven-day interval after 12 weeks of treatment.

In regulatory clinical trials, an endpoint model is required to be defined within the study protocol and regulatory submission materials. This model will detail each study concept of interest, and identify how the endpoints selected relate to each concept, and indicate which endpoints are primary (that the study is powered to adequately assess), secondary and exploratory. Understanding how to interpret changes measured in study endpoints is an important component of the assessment of any intervention, whether a pharmaceutical treatment or a healthcare/rehabilitation solution.

Interpretability

Meaningful change

Meaningful change can be considered to represent the smallest difference in an endpoint measure that would be perceived by patients as beneficial. For many new and novel endpoints derived from the new application of technologies, this understanding of meaningful change thresholds may not already exist. It will be important to assess this if the approach is to be used to measure and monitor changes in health outcomes. In this section on interpretability, we use the illustration of a clinical endpoint measuring activity derived from the output of a wearable step-counting device, as very few published examples exploring meaningful change for other mHealth and interactive technology solutions exist.

Meaningful change is likely to be different between patient populations. For example, in an mHealth intervention measuring and encouraging stepping activity, measures of meaningful change in the number of steps walked per day are likely to be lower in less active patient populations such as those suffering from chronic obstructive respiratory disease (COPD) compared to more active groups, such as type 1 diabetics, where the average number of steps per day has been reported to be 2237 and 8008 steps/day, respectively.²²

While the importance of statistical significance in demonstrating the effects of an intervention is unquestioned, it is also important to recognise that effect sizes detected through statistical tests may be of insufficient magnitude to be considered relevant to the patient. Understanding the threshold for meaningful change (also referred to as clinically relevant change) is

important when interpreting any statistically significant effect sizes detected.

This meaningful change may be represented by the clinically important difference (CID), also called the minimal important difference (MID) or the minimally clinically important difference (MCID), or by the minimal individual change that distinguishes a responder (an individual exhibiting a meaningful improvement) from a non-responder. The CID/MID/MCID represents the minimum change in group means considered clinically relevant, whereas the individual responder definition represents the magnitude of individual change considered clinically relevant.

In general, the responder definition is useful in interpreting the results at an individual patient level, for example, by determining the proportion of patients that achieved the defined responder definition. This is arguably easier to interpret than an effect size resulting from an analysis of group mean differences. However, on account of the loss of information associated with converting a continuous measure to a binary outcome, and the associated loss of statistical power,^{23,24} the analysis of group mean changes in outcome measures derived from interactive and mHealth technologies in clinical trials remains important. Responder analyses typically provide helpful complementary context and interpretation. The analysis of group mean differences will also typically drive clinical intervention study power calculations where it is appropriate to power studies based on this endpoint.

The MCID and responder definition of change provide similar but not identical values: the MCID is defined in terms of differences in mean scores, whereas a responder definition is considered by evaluating individual changes. The responder definition could therefore be larger or smaller than the MCID depending on the degree of change considered. It is recommended that when identifying the amount of change that is meaningful to patients, researchers should aim to estimate both the MCID and one or more individual responder definitions. Both of these provide useful information in terms of both study design and interpretation.

Measuring meaningful change

There are a number of approaches that have been reported with which to demonstrate the clinical relevance of change observed with an outcome measure. These include consensus-based, anchor-based and distribution-based methods, as described below.

Consensus-based methods. Consensus-based methods utilise an expert panel of clinical and domain experts to define a threshold for clinically relevant change in the

specific patient population to be studied. Typically, consensus approaches use techniques such as Delphi methods that operate in an iterative manner. Using Delphi methods, each panel member provides an initial estimate of the MCID along with the rationale guiding their choice. Panel members then review all estimates and explanations, via a blinded summary provided by a facilitator, and are encouraged to revise their estimates based on the responses of other panel members. Iterating this process typically results in a consensus value being reached.

This approach, while helpful in obtaining agreement on the clinical relevance of an endpoint per se, may be less able to determine the true meaning of changes observed. Consensus approaches typically assume that clinicians and health professionals are able to determine the magnitude of change that is important to a patient. The FDA in their draft guidance on patient-focussed drug development state that “Patients are experts in their own experience of their disease or condition”,²⁵ and so ultimately the relevance of changes in health status experienced should be determined by the patient. For these reasons, where possible, consensus methods should not be solely relied upon to estimate the MCID or responder definition for clinical data interpretation.

Distribution-based methods. Distribution-based methods leverage understanding of the distribution of the outcome measure recorded to identify the magnitude of change that would be unlikely to be observed by chance alone. It is common to use more than a single distributional method to obtain a consensus or range for the MCID value. A number of distributional methods exist including the standard error of measurement (SEM), empirical rule effect size, Cohen’s effect size and $0.5 \times$ baseline standard deviation (see Demeyer et al.²⁶ for example, Table 1). It should be noted that distribution-based methods of estimating MCID and responder definition can often be less able to determine

Table 1. MCID estimates for number of steps per day in COPD patients.

Method	MCID calculation	MCID estimate (steps/day)
SEM	$SD_{\text{Baseline}} \times \sqrt{(1 - ICC)}$	599
Empirical rule effect size	$0.08 \times 6 \times SD_{\Delta}$	1029
Cohen’s effect size	$0.5 \times SD_{\Delta}$	1072
$0.5 \times SD$	$0.5 \times SD_{\text{Baseline}}$	1131

SEM: standard error of measurement; MCID: minimally clinically important difference; ICC: intraclass correlation coefficient; SD: standard deviation.

Reproduced from Demeyer et al.²⁶

robust estimates when sample sizes are small or when there is large variability of data at baseline.

Despite their inherent simplicity, distribution-based methods, however, fail to associate statistical changes with whether a truly meaningful change has occurred. Along with other authors (e.g. McLeod et al.²⁷), we agree that distribution-based methods should be considered supportive to anchor-based methods (see below) as opposed to providing primary measures of MCID and responder definitions. However, in some circumstances, anchor-based approaches may not be possible due to the lack of a suitable anchor measure, and it may be necessary to rely more heavily on this approach.

A good example of determining meaningful change using wearable technology to measure physical activity levels is reported by Demeyer et al.²⁶ They used distribution-based methods to estimate the MCID in total daily steps recorded using an accelerometer amongst COPD patients after pulmonary rehabilitation. They reported the MCID in this population as between 600 and 1100 steps/day based on the range of values obtained using a variety of distribution-based methods (Table 1). While these values may be high when compared to the average steps per day achieved amongst COPD patients reported elsewhere, estimated to be 2237 steps/day,²² the clinical importance of improvements of at least 600 steps/day was further demonstrated by reduced risk for hospital re-admission for patients achieving this activity improvement threshold.

Anchor-based methods. Anchor-based methods compare endpoint measures obtained using the new technology to an anchor that is itself interpretable in having known relevance to patients.²⁸ Anchors must be simple and measure a concept that is directly associated with the outcome measure under evaluation. In addition, when designing a study to estimate meaningful change it will be important to select an intervention and time period for which a change in the anchor measure is expected. It is possible that anchors may be objective measures using other related instrumented approaches, or they may be based on subjective assessments made by the patient or clinician. For example, if we wish to determine the meaningful change in an outcome measure of dexterity, collected using a tapping test delivered using an mHealth application on a smartphone, an objective anchor measure may be the count of finger to thumb taps achieved in a 30-s interval during an in-clinic performance test. Subjective anchors, on the other hand, might be a clinician-rating using the 0 to 4 scale for finger tapping assessment in the Unified Parkinson’s Disease Rating Scale, or a global patient impression of change score where the patient is asked to rate any perceived change in bradykinesia

(slowness of movement). To be applicable, meaningful change in the anchor measure must be understood, and changes observed in the anchor measure must be at least moderately associated with changes in the new outcome measure. This association may be assessed by Pearson or Spearman correlations, or more simply via visual inspection of trends. Because associations are typically modest, and anchors may be measuring slightly different concepts, researchers typically include a range of anchor measures to enable a range or consensus value to be derived.

Anchor-based methods can be used to define the MCID for group mean changes or individual responder definitions. As the interpretability of the anchor is known, it is possible to define an anchor value, or range of anchor values, that represent the minimally important group mean change or an individual change indicative of a responder. In each case, by collecting both the new outcome measure and the anchor values in a suitable intervention study in which change is expected, the corresponding MCID and responder definition of the target outcome measure can be estimated.

When determining the MCID using anchor-based methods, it is usual to define a value, or a range of values, for the anchor measure that correspond to the MCID and then calculate the target score that corresponds to that value.²⁸ An example is reported by Motl et al.²⁹ In their study, patients with multiple sclerosis were provided with a wearable accelerometer to measure daily steps over a seven-day period. The study included a number of subjective PROMs as anchor measures including the Multiple Sclerosis Walking Scale (MSWS-12) and the Patient-Determined Disease Steps (PDDS) scale. Meaningful change in both measures is well understood. A 10-point change in the total score of the MSWS-12 instrument and a 1-point change in the PDDS are considered to be the smallest changes deemed meaningful to the patient. Relating changes in steps/day observed to the anchor values collected, the authors reported MCID estimates of 642 and 915 steps/day for the two anchor comparisons, respectively. Figure 1 presents the anchor-based MCID assessment resulting from the MSWS-12 analysis. While a trend is evident, this is a good illustration of the need for multiple anchors to be examined in the same study to enable the MCID to be triangulated – as the between-subject variability in the relationship between the anchor and the new measure is often high, as in this case.

The most common approach to determine an individual responder definition is to classify responders based upon the anchor definition, and then use receiver operating characteristic curves to determine the optimal cut-off point for the target measure to define a responder, based on minimising responder misclassification (see Ward et al.³⁰ and Deyo and Centor³¹ for example).

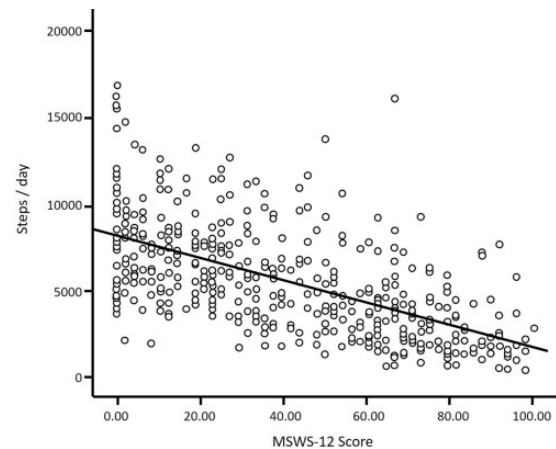


Figure 1. Association between anchor measure (MSWS-12 score) and endpoint derived from wearable device (steps/day). This chart was redrawn from Motl et al.²⁹

It is acknowledged that in some cases suitable anchor measures cannot be determined. For example, despite attempting to derive the MCID for total daily steps in COPD patients from anchors based on the 6-minute walking test distance and Chronic Respiratory Disease Questionnaire (CRDQ) scores, Demeyer et al.²⁶ reported that this was not possible in their evaluation as these measures were found to be only poorly correlated with the total daily steps measurements. In this case, they relied upon distribution-based methods to evaluate the MCID.

Incorporation within product development

Defining meaningful change should be a component of development activities when creating applications using immersive, interactive and mHealth technologies to assess clinical outcomes in clinical research. In addition, defining pertinent clinical outcomes that can be automatically calculated during the use of rehabilitation applications, and their associated values of meaningful change, should be a component of the development and testing of rehabilitation applications. For example, when developing an application to assist patients in the correct and regular conduct of a rehabilitation exercise regimen, technologies such as Microsoft Kinect also enable the accurate tracking of joint coordinates and movements which can be the basis of objective outcome measures defining, for example, joint ranges of motion. These outcome measures can be used to define clinical endpoints that can be tracked over time, and the MCID or responder definition of these new endpoints should be understood. The minimal meaningful change can also be important to understand when defining an acceptance region for equivalence studies, for example comparing

(a) PGI-S	(b) PGI-C
Please rate the severity of your X right now:	Since the start of the study, my overall status is:
1: Not present	1: Very Much Improved
2: Very mild	2: Much Improved
3: Mild	3: Minimally Improved
4: Moderate	4: No Change
5: Moderately severe	5: Minimally Worse
6: Severe	6: Much Worse
7: Extremely severe	7: Very Much Worse

Figure 2. The patient global impression of severity (PGI-S) and change (PGI-C) scales.

the results of a new approach to that of a gold standard or predicate device.

Studies to assess the use of these technologies should include the incorporation of relevant anchor measures. This has the benefit of providing an additional demonstration of intervention effect in addition to providing the means to assess meaningful change in the outcomes derived from the new technology, in the ways described above.

Where it is difficult to determine suitable anchor measures, the use of a patient global impression of severity (PGI-S) or change (PGI-C) scale might be sensible; where a one-point change in PGI-S from baseline would represent a minimal individual change that is meaningful, a score other than “No change” is meaningful on the PGI-C (Figure 2).

This methodology will enable greater understanding of the value and interpretation of new and novel endpoints derived from new technologies and provide interpretation to effect sizes observed during rehabilitation application assessment.

Conclusions

Meaningful change determination is an essential component of clinical endpoint development. mHealth, immersive and interactive technologies offer great potential in the development of novel clinical endpoints that may

provide important insights into the assessment and monitoring of health status. This includes applications developed to measure health status, in addition to those intended as interventions where any resulting longitudinal changes may also be measured while using the technology application. However, the literature contains very few examples of the estimation of meaningful change associated with endpoints derived from mHealth and interactive technologies, making the interpretability of the impact of interventions measured using these technologies problematic. The approaches summarised in this paper, however, are strongly recommended to be implemented alongside application development. This can be facilitated by early planning to include anchor measures within pilot and validation studies to ensure the utility of new clinical technology applications intended to measure clinical change and solutions developed as interventions to generate important changes in health status.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

26. Demeyer H, Burtin C, Hornikx M, et al. The minimal important difference in physical activity in patients with COPD. *PLoS One* 2016; 11: e0154587.
27. McLeod LD, Coon CD, Martin SA, et al. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res* 2011; 11: 163–169.
28. Brožek JL, Guyatt GH and Schünemann HJ. How a well-grounded minimal important difference can enhance transparency of labeling claims and improve interpretation of a patient reported outcome measure. *Health Qual Life Outcomes* 2006; 4: 69–75.
29. Motl RW, Pilutti LA, Learmonth YC, et al. Clinical importance of steps taken per day among persons with multiple sclerosis. *PLoS One* 2013; 8: e73247.
30. Ward MM, Marx AS and Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. *J Clin Epidemiol* 2000; 53: 279–284.
31. Deyo RA and Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis* 1986; 39: 897–906.