Labelling a Patient's Change Status: It's a Confidence Game

Paul W. Stratford, MSc, PT

ABSTRACT

Purpose: The past several decades have seen considerable interest in identifying and applying threshold change values with outcome measures commonly used by physiotherapists. The crucial question of interest to clinicians is, To what extent can valid inferences be drawn from an outcome measure's change or improvement score? To date, typical reporting by researchers includes the presentation of a validity coefficient, often in the form of the area under a receiver operating characteristic curve, and a threshold change or improvement value. A limitation of existing work is that it does not convey the confidence that a clinician can have in a decision based on applying the proposed threshold change value. **Methods:** This knowledge translation article presents three questions, or building blocks, to consider when making a judgment about a patient's change status: (1) to what extent does a measure assess change in the context of interest, (2) what is the threshold change value, and (3) how confident can a clinician be in making the correct decision about a patient's change status when applying the threshold change value? **Results:** This article provides a process for combining clinical expertise with the results from threshold value studies to enhance confidence in clinical decisions about individual patients' change status. **Conclusions:** The article shows how a graph can be used to efficiently translate the results from threshold value studies to convey the chance of making the correct decision about a patient's change status.

Key Words: diagnostic test; outcome assessment; outcome measures; reproducibility of results.

RÉSUMÉ

Objectif : depuis quelques décennies, on s'intéresse beaucoup à déterminer et à appliquer les valeurs seuils de changement pour mesurer les résultats cliniques qu'utilisent couramment les physiothérapeutes. Une question intéresse particulièrement les cliniciens : dans quelle mesure des inférences valides peuvent-elles être tirées d'un score de changement ou d'amélioration d'une mesure? Jusqu'à présent, les chercheurs ont surtout présenté des coefficients de validité, souvent sous forme de surface sous la courbe de la fonction d'efficacité du récepteur, et une valeur de changement ou d'amélioration du seuil. Les travaux existants comportent une limite, celle de ne pas transmettre la confiance qu'un clinicien peut porter à une décision d'après l'application de la valeur seuil de changement proposée. **Méthodologie :** le présent article d'application du savoir s'attarde sur trois questions (ou éléments fondamentaux) à examiner au moment d'évaluer un statut de changement du patient : 1) dans quelle mesure une mesure évalue-t-elle un changement dans le contexte d'intérêt?; 2) quelle est la valeur seuil de changement?; 3) à quel point un clinicien peut-il avoir confiance de prendre la bonne décision sur le statut de changement d'un patient lorsqu'il applique la valeur seuil de changement? **Résultats :** le présent rapport présente un processus pour combiner les compétences cliniques aux résultats des études des valeurs seuils pour accroître la confiance envers les décisions cliniques sur le statut de changement d'un patient donné. **Conclusion :** l'article démontre comment utiliser un graphique pour transmettre avec efficacité les résultats d'études de valeurs seuils pour établir la chance de prendre la bonne décision à l'égard du statut de changement du patient.

Forming measurable goals and assessing whether they have been met are fundamental aspects of clinical practice. Goals can be written in the form of a change score (e.g., to increase knee flexion by ≥ 9 degrees in 2 wk) or in terms of a terminal target value (e.g., to increase knee flexion to 142 degrees in 4 wk). This article focuses on the former goal; it provides suggestions for how researchers can better translate the results of their studies and how clinicians can apply these results to have greater confidence when interpreting a patient's change score.

Historically, the threshold goal value and the expected interval for meeting this value have been based on clinical

experience alone and have been to some extent arbitrary (i.e., not substantiated with evidence). The term *threshold change value* is used to represent the minimum amount of change on a measure beyond which a patient is deemed to have changed. Over the past several decades, numerous studies have been devoted to providing threshold change or improvement values for many outcome measures relevant to physiotherapy practice.¹⁻⁷ The intent of providing this information has been to complement clinical expertise with evidence.

Although the goal of providing this information has been to assist clinical decision making about a patient's

From the School of Rehabilitation Science, McMaster University, Hamilton, Ont.

Correspondence to: Paul Stratford, School of Rehabilitation Science, Institute of Rehabilitation Science, McMaster University, 1400 Main St. West, Hamilton, ON L8S 1C7; stratfor@mcmaster.ca.

Competing Interests: None declared.

Physiotherapy Canada 2019; 71(2):121–129; doi:10.3138/ptc.2018-18

change status, investigators have not formally reported how applying an evidence-based threshold change value translates into level of confidence in interpreting a change value and reporting it in the medical record. Existing studies have typically only examined the validity of a measure in a declared context and reported a single threshold change or improvement value. However, these two requisite pieces of information fall short of offering complete and clinically relevant answers to the question, To what extent can valid inferences be drawn from an outcome measure's change or improvement score? This article introduces three sub-questions that serve as building blocks to assist in answering that question.

METHODS

The three questions, each of which builds on the previous one, are as follows:

1. To what extent does a measure assess change in the context of interest?
2. What is the threshold change value?
3. How confident can a clinician be in making the correct decision about a patient's change status when applying the threshold change value?

The answers to these questions would be provided by researchers and presented in a format that is easily interpreted by clinicians, typically a graph. Clinicians would then use this information to make a decision about a measure's ability to assess change (Question 1), determine whether a patient met the threshold change value (Question 2), and assess the chance of making the correct decision about a patient's change status given that they had applied the threshold value (Question 3).

This article answers each question in turn, accompanied by one or more examples using Roland-Morris Questionnaire (RMQ) data from patients with low back pain.

Background scenario applied to subsequent examples

The RMQ is a 24-item pain-related disability measure. Each item is scored 0 if left blank or 1 if endorsed.⁸ Total RMQ scores can vary from 0 to 24 points, with higher scores representing greater levels of pain-related disability. The scenario's sample consisted of 200 patients with low back pain whose initial RMQ scores ranged from 3 to 23. The sample's mean initial and follow-up scores were 12.8 (SD 5.1) and 7.1 (SD 6.1), respectively. According to an external reference standard, 64 patients remained unchanged and 136 patients improved. Analyses for these data were performed using STATA, version 15.1 (StataCorp, College Station, TX).

RESULTS

Question 1: To what extent does a measure assess change in the context of interest?

This question focuses on the validity of a measured change score: the measure's ability to assess true change in

the characteristic of interest beyond measurement error. Before proceeding, it is important to establish that validity is neither an all-or-none property nor a property of a test or measure. Validity exists to a degree, and it exists in a specific context.⁹ For example, the 6-minute walk test may display moderate correlations with other lower extremity functional status measures in persons with osteoarthritis of the hip or knee, but lower correlations when applied to varsity athletes. This distinction matters because investigators have often made pronouncements such as "Measure X is valid" that ignore the fact that validity is context specific and not an all-or-none property.

Validation designs

Having acknowledged this point, the next step is to review frequently reported validation study designs.

For many outcomes relevant to patients seen by physiotherapists (e.g., pain, functional status, health-related quality of life), no gold or error-free standard exists. Accordingly, three construct validation designs are frequently applied:

Convergent construct validation: The first design is based on the premise that measures believed to be assessing the same characteristic should display moderate to high correlations.¹⁰

Known-groups validation: The second design examines a measure's ability to detect significantly different amounts of change (typically improvement) in groups that, on the basis of theory or existing evidence, should change by different amounts.¹⁰ Test statistics for this design include the *t*-test for independent sample means, *F* statistic, and area under the receiver operating characteristic (ROC) curve. Of these analytics, ROC curve analysis is perhaps used most often because, in addition to providing a validity coefficient (the area under the ROC curve), it can be applied to estimate a threshold improvement value.¹¹

Discriminant validation: The third design is based on a contrast analysis. Specifically, the outcome measure's change scores are correlated with change scores from (1) a measure believed to be assessing a similar characteristic and (2) a measure believed to be assessing a characteristic different from the characteristic of interest.¹⁰ Support for the outcome measure under investigation exists if a substantially higher correlation is obtained with the comparison measure assessing the same characteristic compared with the comparison measure assessing a different characteristic.

Example: To what extent does a measure assess change in the context of interest?

To provide continuity with subsequent examples, this one uses a known-groups validation design. Specifically, a researcher applied an ROC curve analysis to the scenario's data set and generated a validity coefficient in the form of

Figure 1 ROC curve with selected RMQ improvement values (area under ROC curve = 0.9053).

ROC = receiver operating characteristic; RMQ = Roland-Morris Questionnaire.

the area under the ROC curve (see Figure 1).¹¹ The researcher produced the ROC curve by plotting sensitivity against 1 – specificity for all possible change scores, as categorized by an external reference standard. The area under the curve is interpreted as follows: The area can vary from 0 to 1, with an area of 0.5 – represented by the area below the diagonal line – indicating that a measure does no better than chance alone at distinguishing between unimproved and improved patients. The area under this ROC curve is 0.90 (95% CI: 0.86, 0.94). Given the proximity of 0.90 to an area of 1 and the narrow CI, the interpretation is that, in the context of these study patients, the measure displays a high level of validity in distinguishing between unimproved and improved patients.

Question 2: What is the threshold change value?

Paralleling investigations that provide information about the validity of outcome measures have been studies targeting the identification of the measures' threshold change values. The threshold change value is what would be applied when writing a patient's measurable goal. Although reporting a single threshold change value is necessary when doing this, it is restrictive when interpreting the measured change obtained from a patient. When a single threshold change value is applied, it suggests – albeit implicitly – that all change scores, either greater than or less than the threshold value, are interpreted with the same level of confidence. Presumably, the greater the change beyond the identified threshold value, the more confident a clinician can be that a patient

has changed. This limitation is addressed in the Question 3 commentary.

Two study designs and analytic methods have been applied to estimate a threshold change value at the level of an individual patient. One method applies a test–retest reliability design (distribution-based method) and the other uses a known-groups validation design (anchor-based method).⁵

Applying a test–retest reliability design

In brief, the reliability design assembles a group of patients with the condition of interest and applies the outcome measure under investigation at two points in time. A crucial assumption is that a patient's true value for the characteristic has not changed over the reassessment interval. Stated another way, any difference in a patient's results between test and retest is a result of measurement error, where measurement error includes errors owing to inherent variation in the patient, rater, or equipment and interactions among the patient, rater, and environment.¹²

Calculating the standard error of measurement

A second critical assumption, the verification of which is almost always absent from these studies, is that the distribution of errors has a normal distribution, with a mean of zero. This latter assumption is essential because the reliability-obtained threshold change value is based on a normal distribution. Specifically, the analysis calculates the standard error of measurement (SEM), which provides an estimate of measurement error in the same units as the original measurement.¹³

A convenient way of thinking about the SEM is to imagine the distribution of difference scores between test and retest.¹⁴ The SEM is calculated as the standard deviation of the distribution of difference scores divided by the square root of 2. Typically, a multiple of the SEM is used to calculate a quantity known as the minimum detectable change (MDC).¹⁵ To acknowledge that there is error associated with both the test and the retest values (initial and follow-up assessment measurements in the clinical setting), the SEM is multiplied by the square root of 2. Subsequently, this product is multiplied by a *z*-value (standard normal deviate) associated with the confidence level of interest, and for this reason the distribution of difference scores between test and retest must be consistent with a normal distribution. For 90% and 95% confidence levels, the *z*-values are 1.65 and 1.96, respectively.

The interpretation of MDC_{90} , for example, is that 90% of truly unchanged patients would display random fluctuations (measurement errors) within the limits defined by MDC_{90} . An important caveat is that the MDC does not provide the chance to label a patient as changed or unchanged. To make this determination, one needs to know the location and distribution of truly changed (either improved or worsened) patients with respect to the distribution of truly unchanged patients. An elaboration of this concept is provided elsewhere; however, the salient point is that these distributions will overlap.¹⁶

Applying diagnostic test methodology

The deficiency associated with MDC of not being able to estimate the chance of making the correct decision about a patient's change status is overcome using a known-groups design that applies diagnostic test methodology.¹⁶ This design uses a reference standard to classify the change status of patients into two groups. Published reference standards have included global ratings of change by patients and clinicians as well as information from competing measures.⁵⁻⁷

Let us return to our sample of 200 patients, a group composed of 64 unchanged patients and 136 changed (typically improved) patients. For each change or improvement value on the measure under investigation, the researcher created a two-by-two table that compared the reference standard's labelled improvement status – taken to be the truth – with the improvement status had the measure's observed change value been used to classify the patients.¹⁶ Sensitivity and specificity values were calculated for each two-by-two table. *Sensitivity* represents the proportion of patients correctly identified by the measure as improved, divided by all those who truly improved, according to the reference standard. *Specificity* equals the proportion of patients correctly identified by the measure as not improved, divided by all those who truly did not improve, according to the reference standard. Table 1 shows the two-by-two table and accompanying sensitivity

Table 1 Two-by-Two Table for a Threshold Improvement Value of 4 RMQ Points

RMQ result	Reference standard	
	Improved	No change
Improved ≥ 4	109	10
Unimproved < 4	27	54
Total	136	64

Note: Sensitivity: $109 \div 136 = 0.8015$; specificity: $54 \div 64 = 0.8438$; likelihood ratio: $0.8015 \div (1 - 0.8438) = 5.13$.

RMQ = Roland-Morris Questionnaire.

and specificity values for an improvement value of 4 RMQ points – the threshold value that maximizes sensitivity and specificity.

Creating an ROC curve

An ROC curve was then created by plotting sensitivity against $1 - \text{specificity}$ for each observed change or improvement value. Investigators have commonly considered misclassification errors in either direction to be equally important (i.e., declaring a truly unimproved patient as improved and vice versa), and when this qualification is applied, the improvement value associated with the ROC curve point closest to the top left corner of the plot is used to identify the threshold change value.¹¹ An advantage of the ROC curve threshold value estimation method over the MDC approach is that it allows one to estimate the chance of making the correct decision when the threshold change value is applied.¹⁶ This advantage is expanded on in the commentary on Question 3.

Example: What is the threshold change value using the MDC method?

The distribution of difference scores between test and retest for the 64 unchanged patients was consistent with a normal distribution (Shapiro–Wilk $Z = 0.79$, $p = 0.78$). The standard deviation of the difference between test and retest was 2.5 RMQ points. The SEM was calculated to be 1.8 RMQ points, and the MDC_{90} threshold value was estimated to be 4.2 RMQ change points. For interest, the test–retest intraclass correlation coefficient (Shrout and Fleiss type 2, 1) was 0.82 (95% CI: 0.71, 0.90).¹⁷

Example: What is the threshold change value using the ROC curve method?

Recall that the reference standard classified 64 patients as unchanged and 136 patients as improved. The ROC curve analysis produced the result shown in Figure 1. Inspecting the information in this figure reveals that an improvement of 4 RMQ points provided the optimal threshold value, given the qualification that misclassification errors in either direction are deemed to be of equal importance. For reference, Table 1 shows that 4 RMQ points is the threshold value that maximizes both sensitivity and specificity.

Typically, studies applying diagnostic test methodology have reported a single threshold value. A consequence of this practice is that it assigns equal weight to all change values meeting or exceeding the threshold value: patients are declared as having improved with the same level of confidence. The opposite is also true: patients not meeting the threshold value are labelled unimproved with the same level of confidence. It is possible that this unspoken limitation has occurred because investigators have not connected threshold values to the confidence that a clinician can have in applying these values. We address this drawback in the discussion of Question 3.

Question 3: How confident can a clinician be in making the correct decision about a patient's change status when applying the threshold change value?

There are two considerations when answering this question. The first one acknowledges a two-step process, which combines information from clinical research with clinical expertise. The second consideration recognizes that the amount of measured improvement should influence a clinician's confidence in labelling a patient as improved or not: greater improvement should translate into greater confidence. Because an MDC threshold change value does not provide information about the confidence in making a correct decision about a patient's change status, in this section we focus on information obtained from studies that apply diagnostic test methodology.

Calculating the pre-measurement chance of improvement

Quantifying the chance of making the correct decision about a patient's improvement status involves two steps. The first step applies clinical expertise, which may include information from prognostic studies; the second step builds on the first by integrating sensitivity and specificity values specific to the threshold improvement value. Clinical expertise, and information from prognostic studies when it exists, is then applied to assign the pre-measurement chance of improvement.

Applying clinical expertise: The pre-measurement chance of improvement represents the chance of improvement that a clinician assigns to a patient before administering and interpreting the result from the specific outcome measure of interest. It can be implemented at two points in time. The first opportunity exists when writing a measurable goal. For example, at the time of a patient's initial assessment, a clinician might write, "To decrease the patient's RMQ score by 4 points in 2 weeks." We will assume that 4 points is the threshold value, identified from an ROC curve analysis, that maximizes sensitivity and specificity, but where does the 2-week time frame come from?

Often the time frame for meeting a goal value is based on clinical experience and the results from prognostic

studies in the form of growth curves, when available.¹⁸ In the absence of prognostic studies, it is likely that the implicit process involves forming an impression of the expected rate of improvement for a specific patient on the basis of recollections of the rate of improvement typically seen in similar patients. If one literally equates the typical patient with the average patient, this would represent the 50th percentile, or a 50% chance of meeting the goal value in the specified interval.¹⁹ Information provided by growth curves also represents the expected change for the average patient with the prognostic characteristics of interest. Once again, this would equate with a 50% chance of meeting the goal value in the specified interval.

The second opportunity to implement the pre-measurement chance of improvement occurs during ongoing patient visits and is shaped by applying information gained from sources other than the outcome measure of interest (e.g., a patient's comments, clinical observations, and other measurements). For example, a patient may state that she has noticed improvement, her pain has decreased by 2 points (on a numeric rating scale ranging from 0 to 10 points), and her Modified Schöber flexion range has increased. Armed with this information, the clinician may decide to readminister the RMQ and assign a pre-measurement chance of improvement of, for example, 75%.

Integrating the sensitivity and specificity values: The next step is to use the sensitivity and specificity values obtained for the threshold change value to calculate a likelihood ratio: sensitivity ÷ (1 – specificity).²⁰ This likelihood ratio is combined with the pre-measurement chance of improvement to estimate the chance of correctly labelling a patient as improved or not. A step-by-step example of this process is shown in the Appendix.

Using information gain to improve confidence level

However, once this concept is understood, a more practical way of thinking about pre- and post-measurement chances of improvement would be to conceptually partition the percentage chance continuum into a hierarchy of about five categories – for example, *extremely unlikely* (≤15%), *unlikely* (16%–40%), *a toss-up* (41%–60%), *likely* (61%–85%), and *extremely likely* (>85%). The term *information gain* is used to describe the extent to which the post-measurement chance of improvement moves the clinician's confidence level closer to *extremely likely* or *extremely unlikely* compared with the assigned pre-measurement chance of improvement.

Another important concept is that the amount of measured improvement should influence the confidence in declaring a patient improved or not. Intuitively, one would expect greater confidence in declaring a patient whose measured value had increased by a large amount to be more likely to have improved than a patient whose

Figure 2 Information gain curves for an RMQ improvement value of 4.

RMQ = Roland-Morris Questionnaire.

measured value had increased by a small amount. To date, threshold value studies have not addressed this issue. However, this could be efficiently communicated by having researchers present a figure that would display the confidence or information gain associated with multiple change values.

To demonstrate the benefit of researchers presenting information for multiple change values, we can use the two examples presented next to compare the interpretations obtained from single and multiple change value reporting for four patients (A, B, C, D). Patient A has a measured improvement of 2 RMQ points, Patient B an improvement of 3 RMQ points, Patient C an improvement of 6 RMQ points, and Patient D an improvement of 7 RMQ points. For Patients A, B, and C, we assume that the clinician assigned a pre-measurement chance of improvement of 50%. For Patient D, we assume that the clinician assigned a pre-measurement chance of improvement of 20%.

Example: reporting a single threshold RMQ improvement value of 4 points

Given that the ROC curve analysis identified an improvement of 4 RMQ points as maximizing both sensitivity and specificity (i.e., the improvement score that generated the data point closest to the upper left-hand corner of [Figure 1](#)), we now examine the interpretations for each of the four patients. Referring to [Figure 2](#) and the method described in the Appendix, we arrive at the following results.

Patient A's measured improvement was 2 RMQ points. Given that this improvement is less than 4 points, we refer

to the curve below the diagonal. For a pre-measurement chance of improvement of 50%, the post-measurement chance of improvement is approximately 19%, suggesting that there is strong evidence that this patient has not improved. Patient B's measured improvement of 3 RMQ points is also less than 4 points. Accordingly, the post-measurement chance of improvement for this patient would also be 19%.

The measured improvement for both Patients C and D was more than 4 RMQ points; however, their pre-measurement chance of improvement differed. For Patient C, whose pre-measurement chance of improvement was 50%, the post-measurement chance of improvement is 84%, a substantial information gain. However, for Patient D, whose pre-measurement chance of improvement was 20%, the post-measurement chance of improvement is 56%, resulting in near-maximum uncertainty as to whether there has been any improvement. Maximum uncertainty exists when there is an equal chance of improvement or not (i.e., 50%). The patients' chances of improvement are shown in [Figure 2](#).

Example: reporting multiple RMQ change values

Referring to the values shown in [Figure 3](#), we see that Patient A has a less than 14% post-measurement chance of having improved; Patient B has a 71% chance of having improved; and Patient C has a 94% chance of having improved; for Patient D, the clinician can identify an improvement (100%).

The benefit of researchers providing information curves for threshold values other than the threshold

Figure 3 Information gain curves for multiple RMQ improvement values.

RMQ = Roland-Morris Questionnaire.

value that maximizes sensitivity and specificity is highlighted when we take a closer look at Patients B and D. Had the researcher provided information curves for an improvement value of only 4 RMQ points, Patient B's measured improvement of 3 RMQ would have been interpreted as no change (i.e., only a 19% chance of improvement). However, we see that an improvement of 3 RMQ points, combined with a pre-measurement chance of improvement of 50%, results in a 71% chance of improvement, an information gain of 21% in favour of declaring the patient improved.

Had the single improvement value of 4 RMQ points been applied to Patient D, who had a pre-measurement chance of improvement of 20%, the post-measurement chance of improvement would have been about 56% (see, in [Figure 2](#), the bold line above the diagonal): considerable uncertainty. However, by including an information curve for an improvement of 7 RMQ points, we see that the clinician can identify an improvement (see, in [Figure 3](#), the top long- and short-dashed line). Thus, presenting information curves for multiple threshold values gives a clinician greater flexibility in interpreting a change score and correctly labelling a patient's change status.

CONCLUSION

This article introduced three questions for both researchers and clinicians to consider. The answers to these questions depend on researchers communicating comprehensive information in a meaningful manner and

on clinicians combining this information with clinical experience and results from prognostic studies. To date, investigators have dutifully provided validity coefficients and threshold values; however, they have not conveyed the potential benefit of increased confidence in decisions when applying these values. In this article, we have demonstrated the benefit of researchers not only presenting the threshold value, which maximizes sensitivity and specificity, but also providing information for multiple change values in a figure that can be interpreted by clinicians without calculations. Armed with this information, clinicians are less likely to make errors in labelling a patient's change status (e.g., Patients B and D) and have greater confidence in their decisions.

KEY MESSAGES

What is already known on this topic

There has been considerable research over the past several decades into identifying threshold change or improvement values for outcome measures of interest to physiotherapists. These studies typically report a validity coefficient associated with change scores, often the area under a receiver operating characteristic curve, and a single threshold change – usually improvement – value, which minimizes misclassification errors. However, they have two important limitations: (1) a failure to report how confidence in clinical decisions can be altered by applying the advocated threshold value and (2) reporting a single threshold value.

What this study adds

In this article, we show how confidence in a patient's change status is informed by the combination of clinical expertise and research findings in the form of the sensitivity and specificity associated with the threshold change value. Second, we demonstrate that confidence is related to the amount of change. Confidence in decisions about a patient's change status would benefit from researchers reporting the consequences of applying multiple threshold values. Finally, we demonstrate how the use of an information gain figure, which combines a clinician's prior impression of change with results from threshold value studies, can be applied to efficiently inform decision making about a patient's change status.

REFERENCES

1. Hoogbeem TJ, de Bie RA, den Broeder AA, et al. The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. *BMC Musculoskelet Disord*. 2012;13(1):117. <https://doi.org/10.1186/1471-2474-13-117>. Medline:22748143
2. Hefford C, Abbott JH, Arnold R, et al. The Patient-Specific Functional Scale: validity, reliability, and responsiveness in patients with upper extremity musculoskeletal problems. *J Orthop Sports Phys Ther*. 2012;42(2):56–65. <https://doi.org/10.2519/jospt.2012.3953>. Medline:22333510
3. Monticone M, Baiardi P, Vanti C, et al. Responsiveness of the Oswestry Disability Index and the Roland Morris Disability Questionnaire in Italian subjects with sub-acute and chronic low back pain. *Eur Spine J*. 2012;21(1):122–9. <https://doi.org/10.1007/s00586-011-1959-3>. Medline:21823035
4. Stevenson TJ. Detecting change in patients with stroke using the Berg Balance Scale. *Aust J Physiother*. 2001;47(1):29–38. [https://doi.org/10.1016/S0004-9514\(14\)60296-8](https://doi.org/10.1016/S0004-9514(14)60296-8). Medline:11552860
5. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. *Eur Spine J*. 2010;19(9):1484–94. <https://doi.org/10.1007/s00586-010-1353-6>. Medline:20397032
6. Hefford C, Abbott JH, Arnold R, et al. The Patient-Specific Functional Scale: validity, reliability, and responsiveness in patients with upper extremity musculoskeletal problems. *J Orthop Sports Phys Ther*. 2012;42(2):56–65. <https://doi.org/10.2519/jospt.2012.3953>. Medline:22333510
7. Stratford PW, Binkley JM, Riddle DL, et al. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Phys Ther*. 1998;78(11):1186–96. <https://doi.org/10.1093/ptj/78.11.1186>. Medline:9806623
8. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine*. 1983;8(2):141–4. <https://doi.org/10.1097/00007632-198303000-00004>. Medline:6222486
9. Messick S. Validity. In: Linn RL, editor. *Educational measurement*. 3rd ed. Phoenix (AZ): Oryx Press; 1993. p. 14.
10. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 4th ed. New York: Oxford University Press; 2008. p. 257–64.
11. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis*. 1986;39(11):897–906. [https://doi.org/10.1016/0021-9681\(86\)90038-X](https://doi.org/10.1016/0021-9681(86)90038-X). Medline:2947907
12. Riddle DL, Stratford PW. *Is this change real? Interpreting patient outcomes in physical therapy*. Philadelphia: F. A. Davis; 2013.
13. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther*. 1997;77(7):745–50. <https://doi.org/10.1093/ptj/77.7.745>. Medline:9225846
14. Stratford PW. Getting more from the literature: estimating the standard error of measurement from reliability studies. *Physiother Can*. 2004;56(1):27–30. <https://doi.org/10.2310/6640.2004.15377>.
15. Binkley JM, Stratford PW, Lott SA, et al. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. *Phys Ther*. 1999;79(4):371–83. <https://doi.org/10.1093/ptj/79.4.371>. Medline:10201543
16. Stratford PW, Riddle DL. When minimal detectable change exceeds a diagnostic test-based threshold change value for an outcome measure: resolving the conflict. *Phys Ther*. 2012;92(10):1338–47. <https://doi.org/10.2522/ptj.20120002>. Medline:22767887
17. Shrout PE, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–8. <https://doi.org/10.1037/0033-2909.86.2.420>. Medline:18839484
18. Kennedy DM, Stratford PW, Riddle DL, et al. Assessing recovery and establishing prognosis following total knee arthroplasty. *Phys Ther*. 2008;88(1):22–32. <https://doi.org/10.2522/ptj.20070051>. Medline:17986495
19. Stratford PW. Diagnosing patient change: impact of reassessment interval. *Physiother Can*. 2000;52(3):225–8.
20. Riddle DL, Stratford PW. Interpreting validity indexes for diagnostic tests: an illustration using the Berg Balance Test. *Phys Ther*. 1999;79(10):939–48. <https://doi.org/10.1093/ptj/79.10.939>. Medline:10499967

APPENDIX: SAMPLE CALCULATION FOR DETERMINING THE CHANCE OF MAKING THE CORRECT DECISION FOR A PATIENT WHOSE REASSESSMENT INTERVAL WAS CONSISTENT WITH A 50% CHANCE OF IMPROVING AND WHO, ON REASSESSMENT, IMPROVED BY 4 ROLAND-MORRIS QUESTIONNAIRE POINTS

The general approach is to multiply the pre-measurement chance of improvement, expressed as an odds, with the likelihood ratio. This product, known as the *posttest odds*, is then converted to the chance of making a correct decision.

Step 1. Convert the 50% pre-measurement chance of improvement to a pretest odds:

$$\begin{aligned} &= \text{chance of improving} \div \text{chance of not improving} \\ &= 50 \div 50 \\ &= 1:1. \end{aligned}$$

Step 2. Calculate the likelihood ratio:

$$= \text{sensitivity} \div (1 - \text{specificity})$$

$$\begin{aligned} &= 0.8015 \div (1 - 0.8438) \\ &= 5.13. \end{aligned}$$

Step 3. Calculate the posttest odds:

$$\begin{aligned} &= \text{pretest odds} \times \text{likelihood ratio} \\ &= 1 \times 5.13 \\ &= 5.13. \end{aligned}$$

Step 4. Convert the posttest odds to the chance of making the correct decision:

$$\begin{aligned} &= 100 \times [\text{posttest odds} \div (\text{posttest odds} + 1)] \\ &= 100 \times (5.13 \div 6.13) \\ &= 83.7\%. \end{aligned}$$

Step 5. Consider the information gain:

$$\begin{aligned} &= \text{post-measurement chance of improvement} - \text{pre-measurement chance of improvement} \\ &= 83.7\% - 50\% \\ &= 33.7 \text{ percentage points.} \end{aligned}$$