MÉTHODOLOGIE

L'auteur ne déclare aucun conflit d'intérêt en relation avec cet article.

Article reçu le 22 avril, accepté le 16 mai.

Les propriétés des outils de mesure : les questions utiles à se poser

Properties of measurement tools : Useful questions to ask

Claude Pichonnaz[1,2] (PhD, PT)

MOTS-CLÉS
propriétés de mesure / validité / fiabilité / sensibilité au changement d'état

KEYWORDS
measurement properties / validity / reliability / responsiveness

RÉSUMÉ

Contexte : Des outils de mesure de plus en plus nombreux sont mis à disposition des cliniciens. De nombreuses propriétés doivent être testées pour s'assurer que le résultat mesuré par l'outil soit sensé, représente la réalité et que son degré de précision soit acceptable.

Objectif : Cet article vise à présenter les diverses questions qui se posent au clinicien lorsqu'il recourt à un outil de mesure, et à expliquer quelle propriété de mesure est rattachée à chaque question.

Développement : Qu'ils soient des appareils, des questionnaires ou des scores de performance physique les outils de mesure doivent faire preuve de multiples qualités pour garantir que le résultat représente correctement la réalité. Ces qualités sont résumées par les notions de validité (l'outil mesure-t-il ce qu'il prétend mesurer ?), de fiabilité (quelle sont ses marges d'erreur ?) et de sensibilité au changement d'état (détecte-t-il les changements d'état du patient ?). De plus, il est utile de connaître certaines valeurs-seuil utiles à l'interprétation des résultats.

Discussion : Un processus exhaustif de validation requiert de nombreuses investigations. Il ne faut pas perdre de vue que les propriétés d'un outil de mesures varient en fonction de la population, de son degré d'atteinte et du contexte d'application. De plus, les aspects pratiques sont importants à considérer lors du choix d'un outil de mesure clinique.

Conclusion : Un bon outil mesure effectivement ce qu'on souhaite mesurer, produit un résultat correct et stable, et reflète les évolutions du patient. Ces qualités doivent être remplies pour que les mesures contribuent à des prises de décisions cliniques adéquates.

ABSTRACT

Context: More and more measurement tools are being made available to clinicians. Many properties need to be tested to ensure that the result measured by the tool makes sense, represents reality and has an acceptable degree of accuracy.

Objective: This article aims to outline the various issues that arise for the clinician when using a measurement tool and to explain which measurement property is attached to each issue.

Development: Whether they are devices, questionnaires or physical performance scores, measurement tools must demonstrate adequate measurement properties to ensure that the result correctly represents reality. These properties encompasses the notions of validity (does the tool measure what it claims to measure?), reliability (what are its margins of error?) and sensitivity to change of state (does it detect changes in the patient's state?). In addition, it is useful to have established threshold values for interpreting the results.

Discussion: A comprehensive validation process therefore requires many investigations. It should be kept in mind that the properties of a measurement tool vary according to the population, its degree of impairment and the context of application. Furthermore, practical issues are important to consider when choosing a clinical measurement tool.

Conclusion: An effective measurement tool measures what it is intended to measure, produces correct and consistent results, and reflects the patient's progress. These qualities must be fulfilled if measurements are to contribute to appropriate clinical decision-making.

[1] HESAV Haute Ecole de Santé Vaud, HES-SO//Haute Ecole Spécialisée de Suisse Occidentale, Lausanne, Suisse.
[2] Service d'orthopédie et de traumatologie, Département de l'appareil locomoteur, CHUV et Université de Lausanne, Lausanne, Suisse.

Mains Libres 2021 ; 2 : 121-127
En ligne sur : www.mainslibres.ch

CONTEXTE

Les outils de mesure peuvent être des appareils, des questionnaires ou des scores de performance physique. Selon leur objectif, ils permettent de quantifier soit des paramètres objectifs, tels que la force ou le risque de chute, soit des paramètres subjectifs, tels que la douleur ou l'anxiété. En synthétisant un état sous forme quantitative, ils jouent un rôle précieux pour situer le niveau et l'évolution des patients, ainsi que pour communiquer sur des critères partagés entre professionnels.

Un outil de mesure doit faire preuve de multiples qualités pour garantir que le résultat qu'il donne représente correctement la réalité. Ces qualités sont résumées par les notions de validité, de fiabilité et de sensibilité au changement d'état[1].

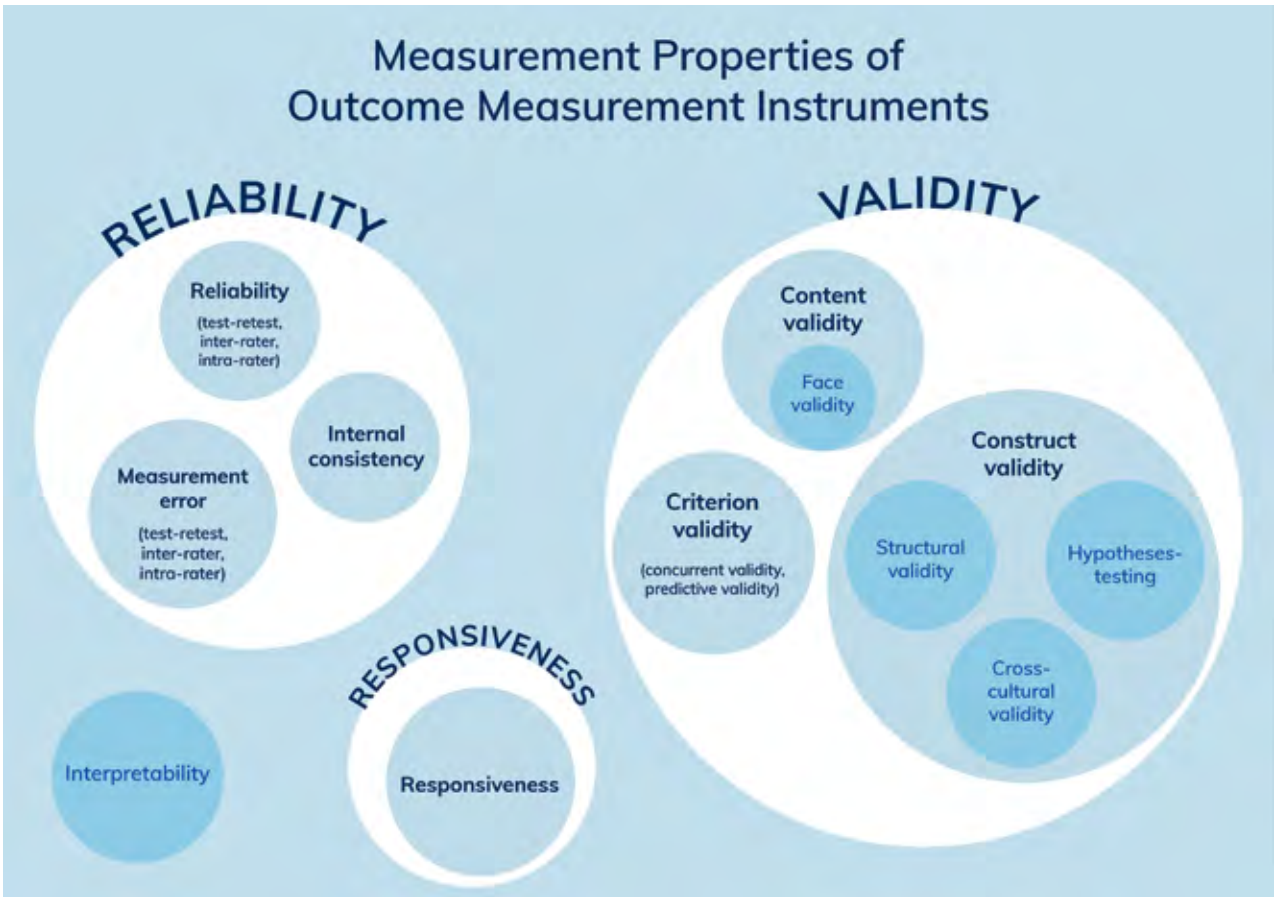
La terminologie concernant les propriétés des outils de mesure est sujette à certaines controverses. Cependant, les définitions proposées par l'initiative COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments), basées sur un consensus interdisciplinaire d'experts, peuvent être considérées comme une référence solide[1]. C'est cette terminologie qui est utilisée comme base pour cet article. La Figure 1 présente un aperçu de la classification COSMIN qui résume les domaines, les propriétés de mesure et leurs aspects utilisés pour définir la qualité d'un outil. La terminologie anglaise étant couramment employée, elle est présentée entre parenthèse pour chaque propriété. Lorsque plusieurs traductions existent en français, elles sont également notées.

Bien que la terminologie et les méthodes statistiques utilisées pour définir les propriétés d'un outil de mesures soient parfois complexes, les questions qui se posent au clinicien qui choisit et utilise un outil de mesure sont essentiellement basées sur le bon sens : mon outil mesure-t-il bien ce qu'il est censé mesuré ? est-il adapté au patient et à sa situation ? la mesure représente-t-elle bien la réalité ? est-elle stable ? permet-elle de détecter les évolutions de mon patient ? quelles sont les marges d'erreur ? quelles valeurs correspondent à un niveau de symptôme acceptable ou à un changement significatif du point de vue du patient ?

Cet article vise à proposer des réponses aux diverses questions qui se posent au clinicien lorsqu'il recourt à un outil de mesure, et à expliquer quelles propriétés sont rattachées à ces questions. Les critères d'appréciations considérés sont également présentés (Tableau 1).

Figure 1

Classification selon le COSMIN



Source : https://www.cosmin.nl/tools/cosmin-taxonomy-measurement-properties/

Tableau 1

Critères d'appréciation des propriétés de mesure

Analyse	Valeurs seuil	Références
Aire sous la courbe ROC*	0,90-1,00: excellent 0,80-0,90: bon 0,70-0,80: acceptable 0,60-0,70: faible 0,50: aucune capacité de discrimination Sensibilité au changement adéquate si aire sous la courbe ≥ 0,70	(4,17,29)
Correlation	0,00 to 0,30 négligeable 0,30 to 0,50 faible 0,50 to 0,70 moyenne 0,70 to 0,90 forte 0,90 to 1,00 très forte Les corrélations entre les mesures de résultats et entre les scores de changement sont adéquates lorsque r ≥ 0,50	(30)
Coefficient de corrélation intraclasse	≥ 0,75 seuil acceptable ≥ 0,90 seuil excellent	(20,31)
Effet plancher/plafond	L'effet est présent lorsque ≥ 15 % des répondants atteignent le plus élevée ou le plus bas possible	(4)
Cohérence interne	Un alpha de Cronbach de 0,70 à 0,90 est généralement considérée comme une mesure de bonne cohérence interne. Une valeur plus faible indique un manque d'homogénéité entre les items; une valeur plus élevée indique une redondance entre les items	(4)

* Receiver operating characteristic (caractéristique de fonctionnement du récepteur).

DÉVELOPPEMENT

Validité : l'outil mesure-t-il bien ce qu'il est censé mesurer ?

La validité (validity) est le degré auquel un outil mesure ce qu'il prétend mesurer[1]. Il s'agit d'une qualité fondamentale, car il n'y aurait aucun sens à mesurer ou étudier un outil que ne rendrait pas le service qu'on lui demande. La validité comprend trois propriétés de mesure, à savoir la validité de contenu (content validity), la validité de construit (construct validity) et la validité de critère (criterion-based validity).

La validité de contenu : l'outil est-il adapté à l'objectif de la mesure ?

La validité de contenu (content validity) fait référence à la relation entre l'outil et ce qu'il vise à mesurer[1]. Elle est donc liée à la question du sens de ce que l'on mesure, et est principalement évaluée dans les situations où la relation entre la mesure et l'outil est complexe. Dans certains cas, cette relation est évidente. Par exemple, la relation entre la mobilité du genou et une mesure goniométrique va de soi. Il est par contre plus difficile de définir des concepts (à comprendre comme étant « l'objet de la mesure » dans cet article) qui laissent place à une interprétation subjective. Par exemple, il n'est pas aisé de définir clairement ce qu'est la qualité de vie puis de démontrer que l'outil est capable de l'évaluer sans confusion avec d'autres concepts.

Une première étape dans l'établissement de la validité de contenu consiste à estimer sa validité apparente, qui est le degré auquel un outil semble effectivement refléter de manière adéquate le concept à mesurer. Son établissement est fondé sur l'avis subjectif des personnes reconnues comme connaissant bien le concept à mesurer[1,2]. Bien qu'assez basique, cette évaluation est une étape initiale importante à considérer avant de lancer des études de validation plus complexes[2].

La validité de contenu à proprement parler fait appel à des méthodes plus élaborées afin de démontrer la pertinence et l'exhaustivité des items de l'outil de manière détaillée[2,3]. La définition claire du concept est donc un préalable incontournable pour déterminer la validité de contenu d'un nouvel outil de mesure. Elle repose sur des études qualitatives auprès des personnes concernées, qui peuvent être des patients ayant acquis des connaissances par leur expérience de la maladie, ou des professionnels de la santé, qui les ont acquises par leur formation et leurs interactions avec les patients[4]. Par exemple, un chercheur qui voudrait développer un nouveau questionnaire de qualité de vie pour les patients atteints de sclérose en plaque pourrait débuter son travail par des études qualitatives auprès des patients et des experts concernés, afin d'identifier les éléments qui impactent la qualité de vie dans ce contexte. Les résultats obtenus lui permettront d'élaborer une première version de son questionnaire dont les items cernent correctement la problématique investiguée. Il devra ensuite tester ce questionnaire auprès de la population concernée pour s'assurer que c'est réellement le cas, et affiner le contenu du questionnaire au besoin.

Il est important de saisir que chaque outil a été développé pour une utilisation dans un contexte déterminé, et ne peut pas être utilisé dans une autre population ou un autre contexte sans nouvelle validation. Notamment, le fait que le contenu soit adéquat pour des patients légèrement atteint ne signifie pas que l'outil est également utilisable chez des patients sévèrement atteints, avant preuve du contraire. Si l'outil n'est pas adapté au niveau du patient, des effets de plancher ou de plafond peuvent en effet apparaître.

En cas d'effet plancher une échelle contenant des items trop difficiles ne sera pas sensible à la détérioration des patients dont les performances sont faibles, car ceux-ci auront déjà obtenu le score minimum avant la détérioration. A l'inverse, un effet plafond est observé lorsque

l'amélioration des patients n'est pas détectée, car ceux-ci ont déjà atteint la valeur maximale du score avant une amélioration. Typiquement, un outil conçu pour des patients présentera un effet plafond chez les athlètes au niveau de performance élevé, et sera donc incapable de distinguer les niveaux de performance au sein de ce dernier groupe[4].

La validité de construit : la construction de l'outil est-elle cohérente ?

Contrairement à la validité de contenu, la validité de construit (construct validity) implique des analyses statistiques pour déterminer objectivement dans quelle mesure l'outil est cohérent avec le concept à mesurer. Elle englobe plusieurs propriétés de mesure, à savoir la validité structurelle (structural validity), la vérification des hypothèses (hypotheses-testing) et la validité interculturelle (cross-cultural validity)[4].

Etablir la validité structurelle implique d'évaluer la relation que les items d'un questionnaire ont entre eux et comment ils se regroupent autour d'une seule ou de plusieurs dimensions, qui peuvent être identifiées au travers d'une méthode statistique appelée analyse factorielle[1,3]. Si le concept n'englobe qu'une seul idée, l'analyse factorielle devrait montrer que les résultats des items convergent tous. C'est le cas par exemple du questionnaire de Rolland-Morris pour l'évaluation de la capacité fonctionnelle chez le patient lombalgique[5]. A l'inverse on peut s'attendre à ce que l'analyse factorielle montre des regroupements autour de plusieurs dimensions par exemple pour l'Hospital anxiety and depression scale, qui vise à évaluer l'anxiété et la dépression dans une seule échelle, sans pour autant les assimiler[6].

Certaines hypothèses, qui sont posées au moment de la création d'un outil de mesure doivent également être confirmées par des analyses statistiques. On s'attend notamment à ce qu'il y ait des différences significatives entre des groupes clairement différents, par exemple entre les patients et les personnes saines (validité des groupes connus – known-group validity). Un nouvel outil d'évaluation devrait également être positivement corrélé à d'autres outils poursuivant le même objectif (validité convergente – convergent validity), et négativement corrélée avec des outils poursuivant un objectif inverse (validité divergente – divergent validity[1,3]. A titre d'exemple, un outil qui vise à évaluer la capacité fonctionnelle de l'épaule devrait logiquement être corrélé positivement avec d'autres outils poursuivant le même but, et négativement corrélés avec des outils qui évaluent le handicap.

La traduction d'un questionnaire dans différentes langues implique un processus de traduction rigoureux, avec des allers-retours de traduction pour garantir l'équivalence linguistique et culturelle entre la version originale et la version traduite[2,7]. Idéalement, toutes les propriétés de la version traduite devraient ensuite être confirmées comme identiques à la version originale.

Validité de critère : les résultats sont-ils comparables à d'autres outils ?

La validité de critère (concurrent validity – validité concourante) consiste à évaluer si les résultats obtenus à l'aide d'un nouvel outil concordent avec ceux d'un outil de référence, considéré comme un «étalon-or» (gold standard) [1]. Typiquement, il s'agit de démontrer qu'un nouvel outil, développé comme une alternative plus simple, moins chère ou plus pratique à utiliser, produit des résultats comparables à l'outil de référence. Lorsqu'aucun « étalon-or » n'existe, le nouvel outil sera comparé à d'autres outils concurrents, comme dans l'évaluation de la validité convergente. La relation entre l'outil de référence et l'outil testé est évaluée en calculant la corrélation entre les outils et en utilisant des graphiques de dispersion. Dans certains cas, le critère de référence est la survenue d'événement dans le futur ; on parle alors de validité prédictive[3]. Par exemple, un bon d'outil d'évaluation du risque de chute sera capable de prédire correctement qui va chuter prochainement.

Fiabilité : quelles sont les marges d'erreur de l'outil de mesure ?

La notion de fiabilité (reliability) fait référence à la proportion d'erreur par rapport à la variation réelle entre les mesures[1]. Un bon outil devrait produire des résultats proches lors de mesures répétées de personnes dont l'état est stable, donc avec une proportion d'erreur minimale. Les différents items qui investiguent un concept devraient varier de manière coordonnée (cohérence interne), et ceci tant au fil du temps (fiabilité test-retest – reproductibilité test-retest), qu'entre différents évaluateurs (fiabilité inter-évaluateur – reproductibilité inter-évaluateur), ou qu'entre plusieurs mesures effectuées par le même évaluateur (fiabilité intra-évaluateur – reproductibilité intra-évaluateur)[1]. De plus, des valeurs telles que l'erreur standard de mesure (ESM), le changement minimal détectable (CMD) et les limites de l'agrément vont donner des indications utiles sur les marges d'erreur des outils.

Cohérence interne : les composantes de l'outil sont-elles cohérentes ?

La cohérence interne fait référence à la force de la relation entre les items d'un score ou d'un questionnaire. Dans un questionnaire bien conçu, chaque item apporte une information supplémentaire sur le concept mesuré, de manière complémentaire et sans redondance. La cohérence interne est fréquemment évaluée à l'aide du coefficient alpha de Cronbach, qui évalue le degré d'inter-corrélation entre les items[8]. Elle peut aussi être appréciée en utilisant une méthode plus complexe, l'analyse Rash, qui tient compte du fait que chaque item n'est pas de difficulté égale. Cette méthode produit des courbes de réponse indicatives de la difficulté de chaque item[9].

Fiabilité test-retest, intra- et inter-évaluateurs : quelle est la part d'erreur dans les mesures ?

La fiabilité test-retest, intra- et inter-évaluateurs (test-retest, intra- et inter-rater reliability) s'intéressent toutes au degré d'erreur lors de mesures répétées, en considérant que le résultat de la mesure est constitué d'une combinaison du score réel et d'un degré d'erreur plus ou moins grand[3]. La fiabilité test-retest concerne les situations où aucun évaluateur externe n'intervient, comme lors du test d'un appareil ou d'un auto-questionnaire. Dans cette situation, les différences entre test et retest sont dues aux variations réelles entre les mesures (p. ex. les variations journalières) et aux erreurs induites par les outils de mesure eux-mêmes. La fiabilité intra-évaluateur est influencée par ces mêmes éléments, auxquels s'ajoute l'erreur induite par l'évaluateur. La fiabilité inter-évaluateurs, additionne encore l'erreur dues aux variations entre les évaluateurs aux sources d'erreur mentionnées précédemment[2].

Le coefficient de corrélations intraclasse (intraclass correlation coefficient – CCI) est la statistique la plus fréquemment utilisée pour évaluer la fiabilité des variables continues (p. ex. un score chronométré), tandis que le coefficient Kappa est utilisé pour les variables dichotomiques (p. ex. réponse oui/non) et le Kappa pondéré pour les variables ordinales (p. ex. cotation faible – moyen – fort)[2,3,10]. Le CCI est une indication de la capacité d'un test à différencier les individus[11]. Il a l'avantage d'être sensible aux différences systématiques, telles que des sur- ou sous-évaluations constantes, contrairement aux corrélations simples de Spearman ou Pearson. Cela peut par exemple être important pour la détection des effets liés à l'entraînement ou à la fatigue lorsqu'on procède à des mesures répétées dans un même groupe.

Erreur standard de mesure : de combien est l'erreur typique ?

L'erreur standard de mesure (standard error of measurement – ESM) est une indication de la précision d'une mesure. Elle consiste à définir un intervalle de confiance autour des valeurs mesurées, c'est-à-dire une fourchette dans laquelle le résultat réel d'un sujet se situe probablement[11]. L'ESM est donc représentative de la marge d'erreur typique d'une mesure. Pour l'ESM95 %, les limites d'erreur spécifiées indiquent l'intervalle dans lequel un clinicien peut être sûr à 95 % que le résultat réel se situe. Ainsi, lors de mesures répétées, un évaluateur peut être raisonnablement sûr que les résultats vont en très grande majorité se situer dans une fourchette de « valeur réelle » +/- ESM95 %[12].

Changement minimal détectable : quelle différence reflète un réel changement ?

Le changement minimal détectable (minimal detectable change – CMD) indique la valeur au-delà de laquelle la différence peut être raisonnablement considérée comme réelle[13]. Le CMD est lié à la valeur de l'ESM car l'expression mathématique pour calculer le CMD est la suivante : $CMD_{95\%} = 1{,}96 \times \sqrt{2} \times ESM$.

On considère que les différences supérieures au $CMD_{95\%}$ ont une probabilité de 95 % d'être dues à une différence réelle[14]. Un clinicien qui mesure une différence plus grande que la $CMD_{95\%}$ aura donc une bonne certitude que son patient a effectivement évolué, car cette différence sera plus grand que l'erreur de mesure.

Analyse de Bland et Altman : quelles sont les valeurs des marges d'erreur et de la différence systématique

Bland et Altman (B&A) ont proposé une procédure pour calculer les limites de l'agrément (limits of agreement) et le biais, qui définissent l'intervalle de confiance à 95 % des différences entre deux mesures et leur différence systématique[15]. L'analyse selon B&A peut être effectuée pour les mesures test-retest, intra- et inter-évaluateur.

Lors de la réalisation d'une analyse B&A, un graphique est produit, ce qui permet de vérifier si les différences entre les mesures ne changent pas en fonction des valeurs mesurées[2,16]. Par exemple, il peut arriver que les erreurs augmentent conjointement à l'augmentation du résultat mesurée.

Sensibilité au changement : les changements d'état du patient sont-ils détectés par l'outil ?

La sensibilité au changement (réactivité – responsiveness) représente la capacité d'un outil à détecter les évolutions qui surviennent au cours du temps[1]. Cette propriété est cruciale pour les interventions de santé visant à améliorer l'état du patient[17]. Plusieurs méthodes sont utilisées pour évaluer la sensibilité au changement : la taille de l'effet, la réponse moyenne standardisée, la corrélation entre les scores de changement et l'analyse des courbes ROC (de l'anglais *receiver operating characteristic*, pour « caractéristique de fonctionnement du récepteur »)[18,19].

Taille de l'effet et réponse moyenne standardisée : quelle est l'amplitude du changement mesuré ?

La taille de l'effet (effect size) de Cohen est un indicateur sans unité de la sensibilité au changement, qui est calculé en divisant le changement moyen par la déviation standard cumulée des mesures[3]. Il est défini qu'une taille d'effet de ≤ 0,20 représente un petit changement, 0,50 représente un changement modéré et ≥ 0,80 représente un grand changement[18,20]. Cependant, l'interprétation de ces valeurs standard doit être pondérée en fonction du contexte.

La comparaison des tailles d'effet de plusieurs outils de mesure dans les mêmes conditions est utile pour déterminer lequel a la meilleure sensibilité au changement. L'outil le plus sensible obtiendra une taille d'effet plus élevée que les autres pour la mesure du même phénomène[17,21].

La réponse moyenne standardisée (RMS, standardised response mean) est basée sur une approche statistique proche de celle de la taille de l'effet. Elle est calculée en divisant le changement moyen par la déviation standard du changement entre les mesures. Les critères de Cohen pour les tailles d'effet petites, moyennes et grandes s'appliquent également à cet indice[20].

Corrélation entre les changements de scores : le changement mesuré correspond-il avec le changement d'un outil de référence ?

Lorsqu'il existe un outil de référence, la corrélation entre le changement de score mesuré sur une référence et celui de l'outil sous investigation peut être utilisée comme indicateur de la sensibilité au changement. Une limitation de cette approche réside dans le fait que, souvent, il n'existe pas d'étalon-or pour une mesure. Il reste alors possible de démontrer que le changement mesuré sur l'outil sous investigation est lié à celui d'un autre outil qui n'est pas un étalon-or, mais dont la sensibilité au changement a été démontrée précédemment[17].

Courbe ROC : l'outil différencie-t-il les personnes améliorées et non améliorées ?

La courbe ROC illustre la relation entre la capacité à classer correctement et incorrectement les patients qui présentent un changement, pour chaque valeur mesurée par l'outil. Une valeur de détection optimale, qui est celle qui représente le rapport le plus élevé entre les deux, peut également être déterminé à partir de la courbe[3,18]. Il est également possible de mesurer l'aire sous la courbe ROC, qui indique globalement la performance de l'outil pour différencier les participants qui se sont améliorés ou pas[17].

Valeurs cliniquement utiles : quels sont les seuils importants pour le patient ?

Toutes les notions précédemment présentées sont utiles pour comprendre les propriétés des outils de mesures, mais ne donnent pas d'indications concernant la perception que le patient peut avoir de son état et de son évolution. La différence minimale cliniquement significative (Minimal Clinically Important Différence – MCID), l'amélioration cliniquement significative (Minimal Clinically Important Improvement – MCII) et le niveau de symptôme acceptable pour le patient (Patient Acceptable Symptom State – PASS) donnent des informations utiles concernant les valeurs importantes pour les patients.

Différence et amélioration minimale cliniquement significative : quel changement est important pour le patient ?

La différence minimale cliniquement importante (MCID) est une valeur qui indique à partir de quelle différence pré-vs. post-traitement le changement est perçu comme important par le patient[12,22,23]. Il peut en effet arriver qu'un traitement fasse une différence significative d'un point de vue statistique, alors que les patients considèrent que son effet n'est pas suffisamment marqué pour être intéressant.

La MCID prend en compte les patients qui se sont améliorés et ceux qui se sont aggravés, alors que la MCII considère uniquement les patients améliorés. La 2e est plus pertinente, les patients étant intéressés par l'amélioration et non pas par le simple changement[23].

La méthode la plus recommandée pour établir la MCII consiste à évaluer sur une échelle de Lickert le degré d'amélioration des patients, puis de prendre la valeur qui correspond au percentile 75 de ceux qui se considèrent comme au moins légèrement améliorés[24].

La valeur de MCII doit être supérieure au changement minimal détectable pour être considérée comme valide, car il serait contradictoire de définir une valeur comme étant importante pour le patient alors qu'elle est inférieure au seuil de détection des changements de l'outil[25].

Niveau de symptôme acceptable pour le patient : les manifestations de la maladie sont-ils acceptables pour le patient ?

Une autre propriété de mesure qui tient compte du point de vue du patient est le niveau de symptôme acceptable pour le patient PASS (Patient Acceptable Symptom State), qui indique à partir de quelle valeur mesurée les patients estiment que le résultat est acceptable, selon leur perception[26]. La méthode la plus souvent utilisée pour le déterminer est basée sur le calcul du 75e percentile des patients qui rapportent un niveau acceptable de symptômes[26-28].

DISCUSSION

Une quantité considérable d'investigations doit être menée avant que l'on puisse affirmer qu'un nouvel outil a subi un processus exhaustif de validation. Malgré les efforts déployés pour standardiser les approches visant à déterminer les propriétés de mesure, certaines controverses existent sur les méthodes à utiliser, ce qui rend parfois complexe l'analyse de la littérature. Cet article a visé à présenter les approches les plus répandues, sans faire état des controverses.

Il ne faut pas perdre de vue que les propriétés d'un outil de mesure ne sont pas déterminées dans l'absolu. Elles varient en fonction de la population mesurée, de son degré d'atteinte et du contexte d'application. Par conséquent, il est important de considérer les conditions dans lesquelles les propriétés de mesure ont été établies avant d'utiliser un outil de mesure dans un contexte clinique.

L'amélioration de la qualité des outils de mesure présente un intérêt pour le patient, car des décisions importantes le concernant sont prises sur la base de mesures de résultats. Par exemple, les décisions de poursuivre ou d'arrêter son traitement ou de lui permettre de retourner à domicile sont influencées par les résultats fonctionnels mesurés. La validité, la fiabilité et la sensibilité de la mesure contribuent donc à la prise de décisions équitables concernant le patient. Une évaluation correcte facilite également l'allocation pertinente des ressources en fonction des besoins des patients.

Si la qualité des propriétés des mesures est fondamentale pour garantir la qualité des mesures, il est également important de prendre en compte les aspects pratiques lors du choix d'un outil de mesure En effet, la plupart des mesures sont réalisées dans des contextes où le temps, le coût, la facilité d'interprétation et la charge de travail sont importants à considérer[32].

CONCLUSION

Cet article visait à effectuer un bref tour d'horizon des nombreuses propriétés qui doivent être investiguées pour connaitre les qualités et limites d'un outil de mesure. Ces connaissances sont indispensables pour que les cliniciens à la recherche d'un outil adapté à leurs besoins puissent faire leur choix en connaissance de cause, et qu'ils puissent ensuite interpréter les résultats avec le recul nécessaire. Au-delà des aspects techniques, il faut garder à l'esprit que trois aspects essentiels sont attendus d'un outil de mesure : qu'il mesure effectivement ce qu'on souhaite mesurer, qu'il produise un résultat correct et stable, et finalement qu'il soit capable de refléter les évolutions du patient.

Lorsque ces conditions sont remplies, l'outil de mesure permet de refléter correctement l'état du patient et son évolution, ce qui est un prérequis incontournable pour prendre des décisions adéquates concernant son état de santé.

IMPLICATIONS POUR LA PRATIQUE

- **Un outil de mesure doit passer par de nombreuses investigations avant que ses propriétés de mesure soient établies**
- **Les propriétés de mesures fondamentales sont la validité, la fiabilité et la sensibilité au changement**
- **Les propriétés d'un outil de mesure sont valables pour des utilisations proches du contexte dans lequel il a été validé**
- **Une mesure de qualité contribue à prendre des décisions cliniques adéquates et équitables**

Pour aller plus loin

Deux sites gratuits qui répertorient des outils de mesure et décrivent leur propriétés de mesure

- **https://www.sralab.org/rehabilitation-measures**
  Rehabilitation Measures Database
  (note : cliquer sur population pour voir apparaître le détail des propriétés de mesure)
- **https://www.cosmin.nl/**
  Site de référence sur les propriétés des outils de mesure
- **http://strokengine.ca/assess/**
  Stroke Engine, a site for individuals who have experienced stroke, their families and health professionals who work in the field of stroke rehabilitation.

Contact

Claude Pichonnaz | E-mail : claude.pichonnaz@hesav.ch

Références

1. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.

2. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide: Cambridge University Press; 2011.

3. McDowell I. Measuring health: a guide to rating scales and questionnaires. Oxford: Oxford University Press; 2006.

4. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.

5. Yamato TP, Maher CG, Saragiotto BT, Catley MJ, McAuley JH. The Roland-Morris Disability Questionnaire: one or more dimensions? European spine journal: official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2017;26(2):301-8.

6. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand. 1983;67(6):361-70.

7. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, *et al.* Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health. 2005;8(2):94-104.

8. Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of applied psychology. 1993;78(1):98.

9. van Alphen A, Halfens R, Hasman A, Imbos T. Likert or Rasch? Nothing is more applicable than good theory. Journal of advanced nursing. 1994;20(1):196-201.

10. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, *et al.* Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106.

11. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. Journal of strength and conditioning research / National Strength & Conditioning Association. 2005;19(1):231-40.

12. Michener LA. Patient- and clinician-rated outcome measures for clinical decision making in rehabilitation. J Sport Rehabil. 2011;20(1):37-45.

13. Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, *et al.* Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. J Rheumatol. 2001;28(2):400-5.

14. van Kampen DA, Willems WJ, van Beers LW, Castelein RM, Scholtes VA, 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.

15. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307 – 10.

16. Giavarina D. Understanding Bland Altman analysis. Biochemia Medica. 2015;25(2):141-51.

17. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Responsiveness. In: Terwee CB, Knol DL, de Vet HCW, Mokkink LB, editors. Measurement in Medicine: A Practical Guide. Practical Guides to Biostatistics and Epidemiology. Cambridge: Cambridge University Press; 2011. p. 202-26.

18. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459-68.

19. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res. 2003;12(4):349-62.

20. Portney LG, Watkins MP. Foundations of Clinical Research: Applications To Practice. 3th ed. Philadelphia: F.A. Davis Company/ Publishers; 2015.

21. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. BMC Medical Research Methodology. 2011;11(1):152.

22. de Vet H, Terwee C, Ostelo R, Beckerman H, Knol D, Bouter L. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health and Quality of Life Outcomes. 2006;4(1):54.

23. Tubach F, Ravaud P, Martin-Mola E, Awada H, Bellamy N, Bombardier C, *et al.* Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: Results from a prospective multinational study. Arthritis Care Res (Hoboken). 2012;64(11):1699-707.

24. Tubach F, Wells GA, Ravaud P, Dougados M. Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues. J Rheumatol. 2005;32(10):2025-9.

25. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Interpretability. In: Terwee CB, Knol DL, de Vet HCW, Mokkink LB, editors. Measurement in Medicine: A Practical Guide. Practical Guides to Biostatistics and Epidemiology. Cambridge: Cambridge University Press; 2011. p. 227-74.

26. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, *et al.* Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis. 2005;64(1):34-7.

27. Tubach F, Ravaud P, Beaton D, Boers M, Bombardier C, Felson DT, *et al.* Minimal clinically important improvement and patient acceptable symptom state for subjective outcome measures in rheumatic disorders. J Rheumatol. 2007;34(5):1188-93.

28. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, *et al.* Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis. 2005;64(1):29-33.

29. Pines JM, Carpenter CR, Raja AS, Schuur JD. Evidence-based emergency care: diagnostic testing and clinical decision rules. 2nd ed. New York: John Wiley & Sons; 2012.

30. Hinkle DE, Wiersma W, Jurs SG. Applied statistics for the behavioral sciences. 5th ed. Boston: Houghton Mifflin, 2003.

31. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine. 2016;15(2):155-63.

32. Valderas JM, Ferrer M, Mendivil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value Health. 2008;11.