Méthodologie

Fiabilité d'un test, d'une mesure ou d'une procédure d'évaluation

Reliability of tests, measures and evaluation procedures

PAUL VAUCHER (PhD, MSc Clinical Trials, Ostéopathe CDS-GDK)

Haute Ecole de Santé de Fribourg, Haute Ecole Supérieure Suisse Occidentale (HES-SO)

La rédaction de cet article a été financée par la HES-SO

L'auteur déclare n'avoir aucun conflit d'intérêts en relation avec cet article

Keywords
Methodology, reliability, consistency, stability, repeatability

Mots clés
Méthodologie, fiabilité, reproductibilité, stabilité, précision, cohérence, répétabilité

Abstract

Reliability plays a central role in the importance we are willing to accord to clinical tests. Empirical observations suggest that tests in manual therapy are more complex than we believed. They apparently do not follow a simplistic dichotomist model defining the presence or absence of somatic dysfunctions. Exploring more complex dimensions of our clinical tests could help understand discrepancies of interpretations between different observers. By defining and standardising our methods of investigation, we could reduce errors and improve reliability. More than ever, it is therefore essential for clinicians to have a clear understanding of what reliability is and how it is measured. This article presents basic concepts about reliability such as absolute and relative reliability, and internal consistency.

Résumé

Les notions de fiabilité sont essentielles pour mieux décider de la place que l'on veut accorder aux tests cliniques. En thérapies manuelles, les observations empiriques semblent suggérer que la plupart des tests cliniques sont bien plus complexes que ce que l'on croyait. Ils ne répondent apparemment pas à un modèle dichotomique définissant la présence ou l'absence d'une dysfonction somatique. L'exploration des dimensions plus complexes de nos tests pourrait nous aider à mieux comprendre le manque de cohérence des interprétations entre observateurs. En définissant plus clairement ce que l'on mesure et en standardisant la méthode de nos tests cliniques, on peut espérer réduire leurs erreurs et améliorer leur fiabilité. Plus que jamais, il semble donc important de bien comprendre ce qu'est la fiabilité et comment nous la mesurons. Cet article présente quelques concepts de base tels que la fiabilité absolue, la fiabilité relative et la cohérence interne.

Introduction

Depuis la fin des années quatre-vingt, les thérapies manuelles semblent manifester des difficultés à résoudre un défi de taille : modéliser la place des tests cliniques dans le raisonnement clinique.[1] Qu'on adopte une approche par arbre décisionnel,[2] une approche Bayésienne,[3] un modèle narratif,[4] ou même une approche par logique floue,[5] le problème de base reste le même : la subjectivité de ces tests et l'importance des erreurs de mesure nuisent à la qualité de l'information disponible. Le fait que plusieurs praticiens n'arrivent pas à s'entendre sur leurs interprétations, sur la qualité de la mobilité d'une articulation,[6] sur la position relative de structures anatomiques,[7] ou sur la présence de mouvements plus fins[8] remet entièrement en question l'utilité de ces tests dans le raisonnement clinique. Afin de faire face à ce défi de taille, cet article propose d'introduire les bases méthodologiques de la fiabilité des tests.

Si l'on admet que pour tout test, il existe un résultat vrai (cf. théorie classique des tests, théorie de la généralisabilité,[9] théorie de réponse d'item[10]), on présuppose alors que ce que nous observons est le résultat de la vraie mesure auquel est ajouté une ou plusieurs erreurs (Figure 1). Ceci suggère également que toute mesure présente un certain degré d'imprécision. Les mesures de fiabilité quantifient alors l'importance de ces erreurs de mesure. Elles permettent de mieux connaître la stabilité des réponses d'un test dans le temps, le degré d'équivalence entre différentes versions d'un test, et la cohérence interne entre les réponses de différents tests lorsque ceux-ci sont combinés ensemble. Cet article va donc développer les notions de fiabilité absolue (précision ou stabilité) et relative (utilité) et celle de cohérence interne (batterie de tests), avant de donner un bref aperçu des méthodes statistiques existantes.

Sources possibles d'erreurs de mesure

- Les conditions environnementales (ex. température, bruits, luminosité, etc...).
- Le contexte du test (ex. présence de douleur, état d'inflammation, étirement préalable, plaie superficielle, etc...).
- L'état psychologique du sujet testé (ex. niveau d'apprentissage, état d'habituation, niveau d'attention, état de fatigue, état de stress, etc...).
- L'état de l'observateur (ex. niveau d'expérience, capacité de concentration, etc...).
- Les propriétés du test qui changent pour certaines sous-populations (ex. enfants versus personnes âgées, personnes raides versus personnes souples, etc...).
- Temporalité des tests (ex. période de la prise de mesure en présence d'une fluctuation naturelle circadienne ou menstruelle, régression vers la moyenne, etc...).
- Erreurs de lecture.

› Figure 1 : sources possibles d'erreurs de mesure

Développement

Précision ou fiabilité absolue

Pour tout test qui fournit un résultat sous forme de score ou d'unité de mesure, on est en droit de s'attendre à recevoir des informations sur sa précision. Celles-ci donnent généralement des indications sur la marge d'erreur que l'on peut s'attendre à avoir avec l'instrument. Plus cette marge est petite, plus l'instrument devient précis. Un tel instrument peut alors identifier de faibles différences ou fluctuations (par exemple amélioration ou péjoration d'une atteinte, présence d'asymétrie, écart par rapport à une norme attendue).

Pour mesurer la marge d'erreur d'un instrument, on répète les mesures afin de vérifier leur stabilité. On donne ensuite une valeur exprimant l'étendue des erreurs observées. Ces valeurs correspondent à la fiabilité absolue (Tableau 1).

Young et al [11] se sont intéressés à la fiabilité des mesures d'amplitude de mouvement des genoux. En répétant trois mesures sur 29 sujets, ils ont pu montrer que la précision d'un goniomètre manuel en plastique (SEM= 0.86 degrés) était tout aussi bonne que celle observé avec un goniomètre électronique de type Lamoreux (SEM=0.99 degrés). Comme clinicien, nous savons maintenant que l'instrument n'est donc pas utile pour détecter une progression, ou une péjoration, qui serait inférieure à 2.4 degrés pour le goniomètre manuel et 2.7 degrés pour le goniomètre électronique.

En connaissant la précision nécessaire, on peut choisir le meilleur instrument en fonction des circonstances. Le but étant que l'instrument puisse au moins détecter la différence minimale que l'on cherche à observer et qui aurait un sens clinique [12]. Le meilleur test étant celui qui donne suffisamment de précision tout en étant simple d'utilisation, bon marché, confortable, exempt d'effets secondaires et rapide.

Mesurer la fiabilité absolue peut s'avérer difficile, notamment lorsque la mesure elle-même détruit ou altère l'objet que l'on mesure.

Interprétation clinique et fiabilité relative

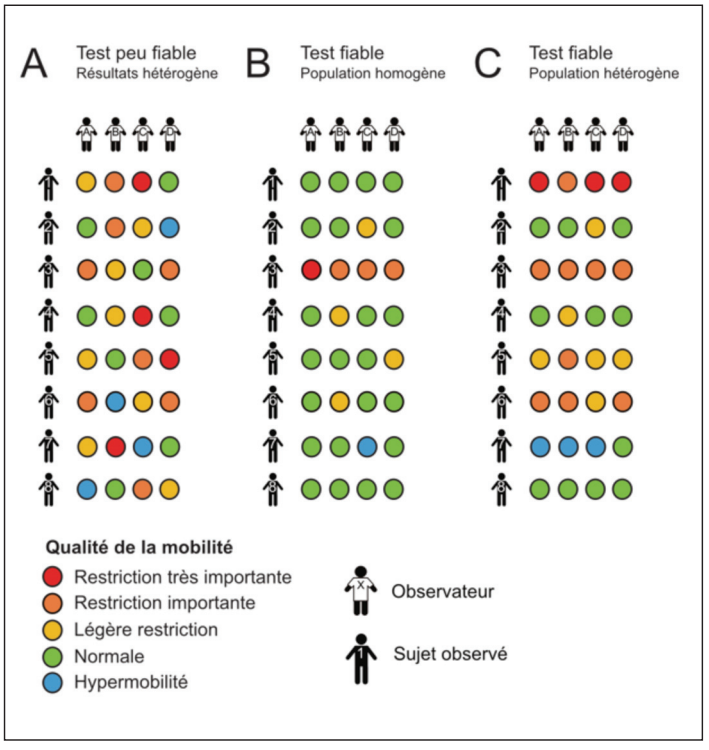
Une autre manière d'approcher la fiabilité est de contextualiser l'utilisation d'un instrument de mesure et d'évaluer sa capacité à détecter des différences entre des patients. On parle alors de fiabilité relative. Pour mesurer la fiabilité relative, on distingue généralement la fiabilité intra-testeurs de la fiabilité inter-testeurs. La première évalue la stabilité des mesures pour un seul observateur, alors que la seconde évalue la stabilité des mesures entre plusieurs observateurs.

Pour les estimer, on peut effectuer une analyse de variance [12]; on observe toutes les variations entre de multiples mesures prises sur de multiples personnes (Figure 2). Un test est considéré comme étant fiable lorsque la variation totale des mesures est davantage expliquée par des différences entre les patients (par ex. colonnes dans la Figure 2) que par les différences entre les mesures prises sur le même patient (par ex. rangées dans la Figure 2). Ceci revient à dire qu'un test commence à devenir utile à partir du moment où la fluctuation des mesures prises sur différents patients dépasse la marge d'erreur de l'instrument de mesure. Si tel n'est pas le cas, on ne peut pas savoir si le résultat observé correspond à l'état du patient ou non (par ex. Situation A dans la Figure 2).

	Abréviation (anglais)	Valeurs (min; max)	Interprétation
Fiabilité absolue			
Erreur type de mesure* (standard error of measurment)	SEM	± en unité mesurée	Correspond à l'écart type de la distribution théorique des erreurs de mesure.
Limites de l'entente* (limits of agreement)	LoA	Unité mesurée (95%inf - 95%sup)	Donne l'intervalle d'erreur qu'on est en droit d'attendre. Ceci correspond aux limites dans lesquelles se trouveraient 95% des valeurs de la distribution des erreurs de mesure.
Coefficient de variation (coefficient of variation)	CV	(0 ;∞)	Correspond au rapport entre l'erreur de mesure et la moyenne des mesures. Un indice bas indique une haute
Fiabilité relative			
Indice de kappa* (kappa coefficient)	ϰ	(0;1)	Correspond à la proportion de sujets pour lesquels deux examinateurs s'accordent sur le résultat d'un test après avoir corrigé pour les ententes qu'on peut s'attendre à trouver par hasard.
Corrélation intra-classe* (intra-class correlation coefficient)	ICC	(0;1)	Proportion de la variante expliquée par des différences entre les sujets observés. Un indice de 1 indique une parfaite entente entre les observateurs.
Cohérence interne			
Alpha de Cronbach* (Cronbach's alpha)	α	(0;1)	Proportion de la variance partagée entre différents items d'un score par rapport à la variance totale. Un indice de 1 indique que tous les items d'un score fluctuent ensemble de manière totalement prédictible. Un score de 0 indique que chaque item varie indépendamment des autres.
Coefficient KR-20 (Kuder-Richard formula 20)	r	(0;1)	L'équivalent de l'alpha de Cronbach pour les variables dichotomiques (ex. positif/négatif)

› Tableau 1 : méthodes statistiques communes utilisées pour mesurer la fiabilité

* Présume que l'erreur est la même indépendamment de la valeur de la mesure (homoscédasticité)



› Figure 2 : fiabilité relative et principe de l'analyse de variance

Moran et Gibbons[13] ont testé la fiabilité de la palpation du rythme crânio-sacré en comparant le rythme perçu (cycle par minute) entre deux praticiens expérimentés qui testaient 11 sujets. Ils ont observés une très mauvaise fiabilité inter-testeur de la palpation du rythme au crâne avec un $ICC_{(2,1)}$ de 0.05. Ceci signifie que la différence de rythme observée entre les sujets (SD=0.5 cyc.min^{-1} pour un observateur et SD=1.0 cyc. min^{-1} pour l'autre) étaient tout aussi importante que la différence observée entre les deux testeurs. La fiabilité intra-testeur était cependant bien meilleure avec un $ICC_{(2,1)}$ de 0.47 pour le premier testeur et de 0.73 pour le second. Ceci suggère que la palpation d'un même clinicien peut être reproductible mais pas celle entre différents praticiens. On trouve une erreur systématique entre les praticiens qui ne remet cependant pas nécessairement entièrement en question l'utilité de la palpation pour identifier une progression dans le temps.

Contrairement à la fiabilité absolue, la fiabilité relative dépend de la population étudiée et de la gravité attendue de l'affection recherchée. Les exigences sont bien plus élevées pour détecter une affection qui serait rare, et qu'on cherche à détecter à un état précoce (par ex. situation B dans la Figure 2), que pour détecter une atteinte plus fréquente avec une grande variété de gravité de manifestations (par ex. situation C dans

la Figure 2). La distinction entre une attitude scoliotique et une scoliose vraie est bien plus difficile lorsque l'on cherche à les distinguer en début d'apparition que lorsque la scoliose est bien plus importante. La fiabilité relative du dépistage est bien meilleure à l'hôpital orthopédique que lors du dépistage scolaire[14].

En thérapie manuelle, une grande partie des affections recherchées sont très nuancées. Il s'en suit que pour être utile, notre méthode de dépistage doit donner des résultats très précis pour être fiable. Pour certaines atteintes, ceci peut être réalisé en combinant ensemble plusieurs tests évaluant la même chose. On parle alors d'une batterie de tests.

Batterie de tests et cohérence interne

Une batterie de tests permet de combiner les résultats de différents tests pour ensuite conclure à la présence ou la gravité d'une affection. La hiérarchisation des tests, partant de tests plus globaux pour aboutir à des tests spécifiques, est un exemple d'une batterie de tests. Ce modèle, utilisé entre autre en ostéopathie,[15] est une forme d'arbre décisionnel qui devrait suivre des règles bien précises pour être fiable.

Une autre façon de construire une batterie de tests est de combiner ensemble différents tests ou différentes mesures qui ont comme objectif de tester la même chose. On cherche ensuite à déduire le résultat selon l'ensemble des observations plutôt que chaque résultat pris individuellement. Ceci a l'avantage de diluer les erreurs de chacun des tests qui aurait peu de sens pris isolément. Ainsi des tests peu fiables peuvent contribuer à une mesure générale qui peut devenir très fiable.

En 1992, *Stiell et al.* ont développé la règle d'Ottawa permettant d'exclure la présence d'une fracture de la cheville en explorant 32 signes cliniques sur 900 patients se présentant aux urgences après un traumatisme de la cheville[16]. Ils ont ainsi pu définir une règle simple permettant d'exclure une fracture basée sur quatre signes seulement; avoir moins de 55 ans, l'absence d'une douleur postérieure à la palpation sur les six derniers centimètres de la malléole latérale, l'absence de douleur dans la même zone de la malléole médiale, ainsi que la possibilité de se mettre en charge immédiatement après le traumatisme et au moment de la consultation. La composante clinique combinée a une bonne fiabilité (k=0.72) et la règle, qui a une sensibilité de 100% et une spécificité de 40%, permet d'éviter les coûts inutiles d'une investigation plus poussée[17]. Le succès de cette règle vient du fait d'avoir pu isoler les quatre signes des 28 autres. Ces signes présentent en effet une très bonne cohérence interne.

Pour aboutir à une mesure, les batteries de tests doivent répondre à plusieurs critères : 1) chaque item de la batterie doit contribuer à la mesure en question, 2) chaque item doit partager un lien commun avec tous les autres items, 3) le score résultant de la batterie de tests doit représenter un seul univers. Lorsque l'on mesure la qualité de vie, on cherche à ce que l'ensemble des indicateurs mesurent le même concept. Dans le cas du SF-12, deux univers différents ont été identifiés : la santé physique et la santé mentale, chacune comprenant six items sous forme de questions[18]. Chacune de ces six questions représente une même dimension de la qualité de vie, et elles sont cohérentes entre elles. Lorsque l'on parle de cohérence interne, on cherche donc à évaluer à quel point les différents items d'un instrument contribuent à faire fluctuer le score final.

Classiquement, on utilise généralement des analyses de covariance pour mesurer la cohérence interne (par ex. alpha de Cronbach ou KR-20. Cette approche présume que l'association entre les items reste toujours identique indépendamment de la sévérité de l'affection; ce qui n'est souvent pas le cas. L'approche Bayésienne avec la théorie de réponse d'item et les analyses Rasch[19] permettent bien plus de souplesse en hiérarchisant chaque item d'un instrument de mesure. La méthode d'analyse Rasch a été développée en éducation. Dans un examen scolaire, on imagine facilement que certaines questions peuvent être plus difficiles que d'autres. Pour réussir, l'étudiant doit non seulement réussir les questions faciles mais également celles qui sont plus difficiles. Plus un étudiant est doué, plus il a de chance de répondre positivement aux questions difficiles. Un examen peut être analysé pour s'assurer qu'il contient un gradient de complexité. De même, un examen clinique investigue l'importance d'une affection. Plus elle est sévère, plus il y a de chances que des signes tardifs apparaissent. La hiérarchisation des symptômes selon la sévérité des cas peut donc être très utile pour évaluer la fiabilité d'un instrument à détecter différents niveaux d'état. Cette approche est cependant encore peu utilisée, mais mérite d'être considérée pour la recherche en thérapies manuelles.

Méthodes statistiques habituelles

La fiabilité absolue est souvent décrite par l'erreur type de mesure, par les limites de l'entente et par le coefficient de variation (Tableau 1).

L'erreur type de mesure (SEM) peut être mesurée en utilisant son indice de fiabilité relative :

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

SEM = erreur type de mesure
S = écart type des mesures du test
r = indice de fiabilité du test

Les limites de l'entente sont calculées le plus souvent lorsque l'on compare deux tests évaluant la même chose[20]. On peut aussi l'utiliser pour voir s'il existe une différence systématique entre deux observateurs. En premier lieu, on mesure pour chaque sujet observé, la différence entre les deux mesures. On calcule ensuite la fourchette dans laquelle se retrouvent 95% des mesures, chaque borne se trouvant :

$$LoA = m +/- 1.96xS$$

avec m = différence moyenne, S = écart type et n = nombre de sujets observés.

Lorsque l'on mesure les limites de l'entente, il est toujours utile de représenter graphiquement les valeurs (graphique de Bland-Altman) afin de vérifier visuellement que la dispersion des erreurs est homogène indépendamment de la valeur de la

mesure (par ex. homoscédasticité). On peut aussi facilement voir si les différences entre les mesures sont systématiques ou pas.

Finalement, on peut mesurer le coefficient de variation, qui donne une idée de l'importance de l'erreur par rapport à la valeur attendue des mesures. Le coefficient de variation est simplement :

$$CV = S/m$$

Avec S = écart type des erreurs et m = moyenne des mesures

La fiabilité relative est la plupart du temps exprimée par un coefficient qui va de 0 à 1 (Tableau 1). Pour le mesurer, on utilise l'indice de kappa (variables dichotomique ou nominale) ou les coefficients de corrélation intra-classe (variable continue ou ordinale).

Il existe deux indices de *kappa*, celui de *Cohen* et celui de *Fleiss*[21]. Le premier est utilisé pour comparer deux observations, alors que le second permet de comparer un nombre plus élevé d'observations. L'indice de *kappa* dépend grandement de la probabilité d'avoir une réponse concordante par hasard. La taille de l'échantillon et la prévalence de l'atteinte sont donc déterminantes.

Les méthodes statistiques pour mesurer les coefficients de corrélation intra-classe les plus fréquemment utilisées sont les modèles à effet fixe et à effet aléatoire (fixed or mixed effect models)[22]. Le premier permet de mesurer la fiabilité intra-testeur ou test-retest ($ICC_{(3,1)}$) et le second la fiabilité inter-testeur ($ICC_{(2,1)}$). La plupart des logiciels statistiques permettent ces analyses.

Lorsque l'on planifie une étude de fiabilité, il est important de s'assurer qu'on dispose de suffisamment de puissance pour avoir des mesures de fiabilité qui ont un sens. La puissance va dépendre du nombre de sujet observés, du nombre d'observateurs, et du nombre d'observations par observateur[23-25].

Interprétation des résultats

Les mesures de fiabilité absolues donnent une idée assez précise de ce que l'on peut attendre des mesures. Hormis le coefficient de variation, elles restent cependant vulnérables à l'hétéroscédasticité (c.à-d. l'importance de l'erreur dépend de la magnitude de la mesure) d'où l'importance de réaliser des études de fiabilité qui comprennent des personnes hétérogènes avec un spectre large de sévérité de l'atteinte investiguée.

L'interprétation des mesures de fiabilité relative est plus délicate. Il est important de ne pas extrapoler les résultats sur d'autres populations que celles étudiées. La raison principale étant que ces mesures dépendent grandement des caractéristiques de la population étudiée notamment en ce qui concerne l'hétérogénéité de la sévérité des atteintes. Plus une population est homogène, plus l'ICC ou le Kappa d'un instrument sera faible. L'interprétation du sens clinique dépend également du contexte et des conséquences de faussement classer une personne comme saine ou malade. *Landis & Koch*, [26] puis *Fleiss*[27] ont déterminé arbitrairement une classification des valeurs pour les indices de *kappa* (Tableau 2). Ces interprétations sont à utiliser avec précaution et nécessitent également de tenir compte de la qualité des études. On trouve également des valeurs similaires pour le coefficient de corrélation intra-classe. Un test est généralement considéré comme étant cliniquement utile s'il a un ICC ≥ 0.75 et comme étant utilisable en recherche pour servir d'étalon-or s'il a un ICC ≥ 0.9.

Discussion

La place de la fiabilité en thérapies manuelles

Dans sa maturation moderne, l'éveil des thérapies manuelles ne se fait pas sans chamboulements. Nous commençons à peine de réaliser que ce que l'on traite est complexe et difficilement identifiable[28,29], que nous devons également constater l'urgence d'actualiser nos modèles de processus décisionnels pour identifier les troubles que nous pensons traiter.[30]

Empiriquement, les tests de mobilité et de position prennent une place importante en thérapie manuelle sans pour autant que nous puissions réellement justifier leur place. Le recours à ces tests pour identifier la présence ou l'absence d'une affection en utilisant une approche dichotomique (c.-à-d. test positif ou négatif) est dépassé[28]. Vu les multiples tentatives vaines de soutenir la fiabilité de nos tests, il reste peu de perspectives d'ouverture, ce qui met à mal nos professions. Contextualiser nos tests à la lumière des connaissances actuelles devient alors essentiel. Notre réflexion peut s'appuyer sur les acquis méthodologiques des autres disciplines pour nous aider à mieux identifier ce que l'on teste, comment, pourquoi et quand. Les études de fiabilité pourraient alors se concentrer sur des modèles plus complexes et plus proches de la pratique avancée. Comprendre l'essence de nos méthodes d'évaluation pourrait aider à justifier les approches manuelles de demain.

Landis & Koch[26]		Fleiss[27]	
Valeur de $\varkappa$	**Sens**	**Valeur de $\varkappa$**	**Sens**
0.81 – 1	Presque parfaite	0.76 – 1	Excellente
0.61 – 0.8	Substantielle	0.4 – 0.75	Acceptable à bon
0.41 – 0.6	Modérée	0 – 0.39	Faible
0.21 – 0.4	Acceptable		
0.01 – 0.2	Faible		
≤ 0	Absente		

› Tableau 2 : interprétation qualitative de l'indice de Kappa

Limitations

La fiabilité en psychométrie n'est pas une science exacte. Elle souffre principalement du fait que, pour l'évaluer, on doit répéter plusieurs fois les mêmes tests. Or la répétition est une source d'erreur et modifie les résultats du test (ex. assouplissement, inflammation, apprentissage, fatigue). On se voit donc forcé d'admettre que les mesures de fiabilité sont des estimations et non pas des valeurs précises.

La fiabilité se limite à évaluer la précision et la stabilité d'une mesure, mais n'informe nullement si cette mesure correspond réellement à ce que l'on cherche. La validité de nos mesures est donc un défi supplémentaire qui reste à surmonter.

L'étude de la fiabilité ne permet pas d'identifier de nouveaux modèles à explorer; elle permet simplement d'identifier ceux auxquels on devrait arrêter de se fier. L'étude de la fiabilité nécessite donc une exploration préalable plus approfondie. Mieux comprendre les différentes dimensions que le praticien en thérapie manuelle perçoit et interprète lors des tests peut servir pour mieux développer les futurs instruments.

Conclusions

La fiabilité est une composante importante des tests pour nous permettre de nuancer leurs résultats dans notre processus décisionnel. Aujourd'hui, l'enseignement des tests en thérapie manuelle peut difficilement se passer d'une explication documentée de leurs limites. Comprendre la méthodologie liée à la fiabilité permet donc non seulement de mieux explorer notre futur mais également de mieux intégrer notre passé. La richesse de l'approche manuelle pourrait venir de la subjectivité actuelle de nos tests. Il est fort probable que ce que l'on perçoit comme étant des erreurs de mesure aujourd'hui devienne la base pour expliquer la diversité, les nuances et la source de précision des interprétations possibles de nos tests. A nous d'éclairer ceci pour mieux préparer les praticiens de demain.

Implications pour la pratique

- La fiabilité se traduit par le degré de confiance que l'on peut accorder au résultat d'un test.
- La fiabilité absolue correspond à la stabilité ou à la précision d'une mesure, alors que la fiabilité relative correspond à l'utilité d'un test pour identifier des différences dans une population donnée.
- Plus une atteinte est rare, plus il est important de recourir à un test ayant une haute fiabilité pour la détecter.

Contact

Paul Vaucher
paul.vaucher@hes-so.ch

Quizz (testez vos connaissances)

Pour chacune des cinq questions suivantes, choisissez la réponse qui vous paraît la plus juste.

1. La fiabilité relative permet de

A. savoir si l'on peut différencier des personnes dans un groupe.
B. quantifier la magnitude des erreurs de mesure.
C. déterminer si l'on détecte bien ce que l'on cherche.
D. connaître la précision et la stabilité d'une mesure.
E. s'assurer qu'un test est reproductible dans toutes les circonstances.

2. L'indice de kappa mesure

A. la proportion de fois où plusieurs examinateurs s'entendent sur leur résultats.
B. la fiabilité relative d'une mesure continue.
C. la probabilité de s'entendre sur un résultat au-delà de la chance.
D. la cohérence interne entre les composantes d'un test.
E. la part d'incertitude liée à un test.

3. La puissance d'une étude mesurant la fiabilité relative d'un test ne dépend pas

A. du nombre d'observateurs.
B. du nombre de sujets observés.
C. du nombre d'observations par sujet.
D. de la validité du test.
E. de l'hétérogénéité des réponses.

4. La vraie mesure

A. est associée à une erreur qui est stable.
B. est dépendante des circonstances dans laquelle on la prend.
C. est influençable par la fatigue ou le manque de concentration.
D. est estimée par l'erreur type de la mesure.
E. est exempte d'imprécision.

5. Pour combiner plusieurs tests en une seule mesure, il faut

A. que chaque item soit lui-même très fiable.
B. vérifier leur cohérence interne.
C. que chaque item mesure exactement la même chose.
D. qu'il y ait autant de dimensions à étudier que de nombre d'item.
E. Préalablement s'assurer que chaque item repose sur la même échelle.

Réponses

1. A; 2. C; 3. D; 4. E; 5. B

Références

1. Vaucher P. Questioning the rationality of clinical osteopathic tests : future perspectives for research. Mains Libres. 2016;33(1):33-37.
2. Bae JM. The clinical decision analysis using decision tree. Epidemiol Health. 2014;36:e2014025.
3. Soltani A, Moayyeri A. Deterministic versus evidence-based attitude towards clinical diagnosis. J Eval Clin Pract. 2007;13(4):533-537.
4. Mahr G. Narrative medicine and decision-making capacity. J Eval Clin Pract. 2015;21(3):503-507.
5. Minutolo A, Esposito M, De Pietro G. A fuzzy framework for encoding uncertainty in clinical decision-making. Knowledge-Based Systems. 2016;98:95-116.
6. Nyberg RE, Russell Smith A, Jr. The science of spinal motion palpation: a review and update with implications for assessment and intervention. J Man Manip Ther. 2013;21(3):160-167.
7. Haneline MT, Young M. A Review of Intraexaminer and Interexaminer Reliability of Static Spinal Palpation: A Literature Synthesis. Journal of manipulative and physiological therapeutics. 2009;32(5):379-386.
8. Sommerfeld P, Kaider A, Klein P. Inter- and intraexaminer reliability in palpation of the "primary respiratory mechanism" within the "cranial concept". Manual Therapy. 2004;9(1):22-29.
9. Brennan RL. Generalizability Theory and Classical Test Theory. Applied Measurement in Education. 2010;24(1):1-21.
10. Raykov T, Marcoulides GA. On the Relationship Between Classical Test Theory and Item Response Theory. Educational and Psychological Measurement. 2016;76(2):325-338.
11. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther. 1994;74(8):777-788.
12. 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):231-240.
13. Moran RW, Gibbons P. Intraexaminer and interexaminer reliability for palpation of the cranial rhythmic impulse at the head and sacrum. JMPT. 2001;24(3):183-190.
14. Plaszewski M, Bettany-Saltikov J. Are current scoliosis school screening recommendations evidence-based and up to date? A best evidence synthesis umbrella review. 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. 2014;23(12):2572-2585.
15. Dinnar U, Beal M, Goodridge J, et al. Classification of diagnostic tests used with osteopathic manipulation. JAOA. 1980;79(7):451-451.
16. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Annals of emergency medicine. 1992;21(4):384-390.
17. Stiell IG, McKnight RD, Greenberg GH, et al. Implementation of the Ottawa ankle rules. JAMA : the journal of the American Medical Association. 1994;271(11):827-832.
18. Ware JE, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity. Medical Care. 1996;34(3):220-233.
19. Wright BD, Mok MM. An overview of the family of Rasch measurement models. In: Smith EV, Smith RM. Introduction to Rasch measurement. Maple Grove: Journal of Applied Measurement press. 2004
20. Sedgwick P. Limits of agreement (Bland-Altman method). BMJ : British Medical Journal. 2013;346.
21. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. Statistics in Medicine. 2002;21(14):2109-2129.
22. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979;86(2):420-428.
23. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy. 2005;85(3):257-268.
24. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med. 2012;31(29):3972-3981.
25. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Statistical Methods in Medical Research. 2004;13(4):251-271.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977:159-174.
27. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. Third ed. New Jersey: John Wiley & Sons; 2003.
28. Fryer G. Somatic dysfunction: An osteopathic conundrum. Int J Ost Med. 2016;22:52-63.
29. Penney JN. The Biopsychosocial model: Redefining osteopathic philosophy? Int J Ost Med. 2013;16(1):33-37.
30. Thomson OP, Petty NJ, Moore AP. Clinical reasoning in osteopathy – More than just principles? Int J Ost Med. 2011;14(2):71-76.