Evaluation en Physiothérapie (N1)

Fiabilité des tests cliniques et précision du diagnostic.

L'examen clinique de l'appareil musculo-squelettique, qu'il soit analytique ou fonctionnel, repose sur de très nombreux tests. Ces évaluations cliniques sont plus ou moins bien validées, et donc plus ou moins fiables, et pourtant nécessaires à l'établissement du Bilan Diagnostic Kinésithérapique.

Partant d'une volonté de "pratique basée sur des preuves" (1), la SPB souhaite proposer une synthèse des tests les mieux validés par la littérature internationale.

Pour ce faire, nous établirons plusieurs fiches de synthèse, par région, par fonction ou par système, afin de rendre lisible les tests à utiliser de manière préférentielle lors du bilan diagnostic.

Afin de mieux comprendre les arguments permettant de donner plus ou moins de poids à la validité d'un test, nous proposons un préambule dont l'objectif est d'acquérir les bases statistiques nécessaires à la lecture des fiches de synthèse.

NB : Il n'est pas nécessaire de comprendre les formules mathématiques des indices proposés ci-dessous pour les utiliser. Cependant, leur lecture permet de constater qu'il est toujours question de dispersion (écart-type ou variance) autour de la moyenne.

Fiabilité d'un test clinique

Les mesures cliniques, guidées par la main du thérapeute et les mouvements (passifs ou actifs) du patient, sont affectés par des erreurs. La quantification de la fiabilité permet de mesurer la *proportion* de ce qui représente la *réalité* et de ce qui représente les *erreurs* dues au hasard, à l'imprécision de l'outil utilisé (la main le plus souvent) ou du thérapeute évaluateur (2).

Deux types de fiabilité doivent être envisagées, la *fiabilité inter-examinateur* (capacité d'un unique évaluateur à obtenir le même résultat plusieurs fois sur un même patient) et la *fiabilité intra-examinateur* (capacité de deux ou plusieurs évaluateurs à obtenir le même résultat sur un même patient).

Pour mesurer cette fiabilité, le choix de l'outil statistique doit être fait en fonction du type de variable considéré.

Variables qualitatives : Coefficient Kappa

Très utilisé, ce coefficient permet de mesurer la *proportion d'accord ou de rejet* des résultats par deux juges (inter-examinateur), ou par le même juge à deux reprises (intra-examinateur) (3;4). La variable étudiée est qualitative, le plus souvent binaire (= dichotomique : Oui / Non) mais parfois ordinale à plusieurs classes (souple, très souple, raide, très raide, pour un exemple à 4 classes).

Voici le mode de calcul du coefficient de Kappa à l'aide d'un exemple (chaque chiffre du tableau de contingence représente un effectif de patients) :

Exemple : Calcul du coefficient de Kappa pour mesurer la fiabilité inter-examinateur d'un test clinique

		Résultat du juge A		
	Réponses	OUI	NON	Total
Résultat du juge B	OUI	**72**	**16**	88
	NON	**25**	**87**	112
	Total	97	103	200

Pour 16 patients, le juge A a obtenu un test négatif alors que le juge B a obtenu un test positif

K = (Po – Pthéo) / (1 – Pthéo)
Po = Concordance **o**bservée = (72 + 87) / 200 = 0.795
Pthéo = (Produits des réponses positives des 2 juges + Produits des réponses négatives des 2 juges) / Carré de l'effectif total = ((88 x 97) + (112 x 103)) / 200^2 = 0.501

K = (0.795 – 0.501) / (1 – 0.501) = **0.589**

Interprétation (3): le Kappa est:

- *Faible* en dessous de 0,5
- *Modéré* entre 0,5 et 0,75
- *Fort* au dessus de 0,75

Dans notre exemple, la fiabilité inter-examinateur est donc Modérée.

NB : Le Kappa peut également être utilisé pour mesurer la force d'un lien entre deux variables qualitatives: comme l'association entre deux symptômes.

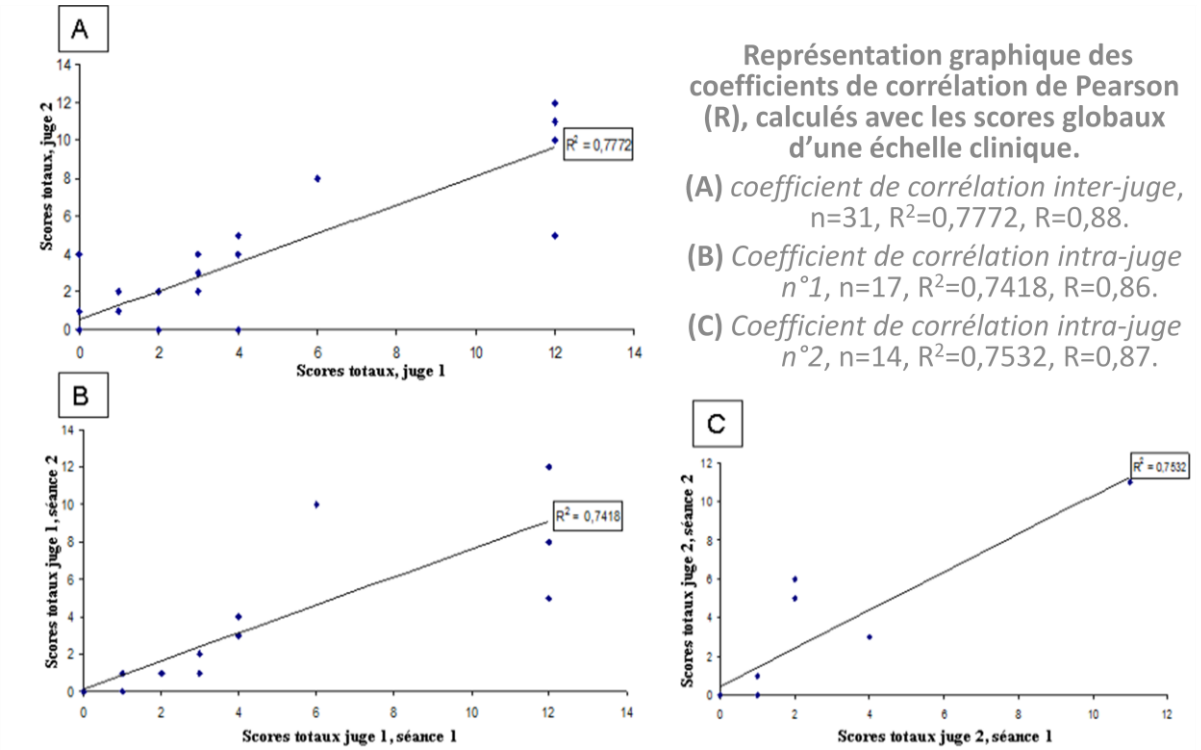
Variables quantitatives : Coefficient de corrélation intra-classe ou coefficient de corrélation de Pearson (5)

Ces deux coefficients peuvent être utilisés dans des cas similaires, mais d'un point de vue statistique, il est préférable d'utiliser le CCI (coefficient de Corrélation Intra-Classe) pour mesurer la variation entre deux mesures (intra-examinateur et inter-examinateur) et le Pearson pour mesurer la force d'un lien entre deux variables quantitatives.

$$r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}}$$

Le calcul du Coefficient de corrélation de Pearson, ou r_P (proche du CCI) est basé sur l'analyse de la variance (dispersion autour de la moyenne) calculée dans les deux variables à corréler.

Exemples de Corrélation : fiabilité inter-examinateur (A) et intra-examinateur (B)(C)



Représentation graphique des coefficients de corrélation de Pearson (R), calculés avec les scores globaux d'une échelle clinique.

(A) *coefficient de corrélation inter-juge*, n=31, R^2=0,7772, R=0,88.

(B) *Coefficient de corrélation intra-juge n°1*, n=17, R^2=0,7418, R=0,86.

(C) *Coefficient de corrélation intra-juge n°2*, n=14, R^2=0,7532, R=0,87.

Iris Gaudot. « Etude du processus de validation d'une échelle gériatrique mesurant la rétropulsion : La Backward Desequilibrium Scale (BDS) » Mémoire de fin d'étude Kinésithérapie. 2012.

Son interprétation est *la même* que pour le coefficient de Kappa :

- *Faible* en dessous de 0,5
- *Modéré* entre 0,5 et 0,75
- *Fort* au dessus de 0,75

Dans notre exemple, tous le "r" sont supérieur à 0,75. L'échelle testée semble donc (fortement) fiable.

NB : Attention, une corrélation avec un seul point à une extrémité de la droite n'a que peu de valeur statistique (exemple C ci-dessus, pas assez de patients avec des scores moyens ou forts à la BDS)

Précision du diagnostic

La précision du diagnostic est la capacité d'un test clinique à plus ou moins bien confirmer ou infirmer la présence d'un trouble spécifique (6). Pour cela, il est nécessaire de comparer les résultats obtenus au test clinique avec une référence standard (ou gold standard). Cette précision est exprimée en terme de Valeurs Prédictives Positives ou Négatives (VPP et VPN), de Sensibilité et de Spécificité, ou encore de Ratio de Vraisemblance (RV).

Pour comprendre ces éléments il suffit de schématiser un tableau d'éventualité :

	Référence standard Résultats +	**Référence standard Résultats -**	
Test clinique Résultat +	**VRAIS POSITIFS a**	FAUX POSITIFS b	VPP = a/(a+b)
Test clinique Résultats -	FAUX NEGATIFS c	**VRAIS NEGATIFS d**	VPN = d/(c+d)
	Sensibilité = a/(a+c)	Spécificité = d/(b+d)	

Les Valeurs Prédictive Positive (VPP) ou Négative (VPN) expriment donc la vraisemblance d'avoir réellement, ou non, la déficience mis en évidence par le test.

La Sensibilité est révélatrice de la capacité du test à bien diagnostiquer les patients porteurs de la déficience. Une bonne sensibilité permet de ne pas passer à côté de cette déficience.

La Spécificité est révélatrice de la capacité du test à exclure les patients non porteurs de la déficience.

En pratique clinique, nous avons besoin de ces deux derniers aspects. Il est donc préférable d'avoir une bonne sensibilité ET une bonne spécificité.

Ces deux grandeurs sont interprétées comme les coefficients précédents :

- *Faible* en dessous de 0,5
- *Modéré* entre 0,5 et 0,75
- *Fort* au dessus de 0,75

Le ratio de vraisemblance (RV) est un calcul "probabilistique" exprimant à la fois la capacité sensible et spécifique d'un test. Les RV positifs (RV+) et les RV négatifs (RV-) sont interprétés selon 4 catégories :

RV +	Interprétation	RV –
➢10	Bonne	< 0,1
5 à 10	Modérée	0,1 à 0,2
2 à 5	Faible	0,2 à 0,5
1 à 2	Très faible	0,5 à 1

Exemple de Précision du diagnostic : Test de Thessaly pour une lésion méniscale

Sensibilité = 0.90 / Spécificité = 0.98
RV+ = 39,3 / RV- = 0,09

= *Bonne précision du diagnostic pour ce test*
(le meilleur pour cette lésion)

Intervalle de confiance

L'Intervalle de Confiance (IC) permet d'estimer la précision d'une estimation, en donnant une borne inférieure et une borne supérieure.

Les IC sont le plus souvent calculés à partir d'un échantillon à un niveau de 95%. **Cela signifie que si l'on prend 100 nouveaux échantillons dans la population, 95 seront contenus dans les bornes de l'intervalle de confiance calculée.**

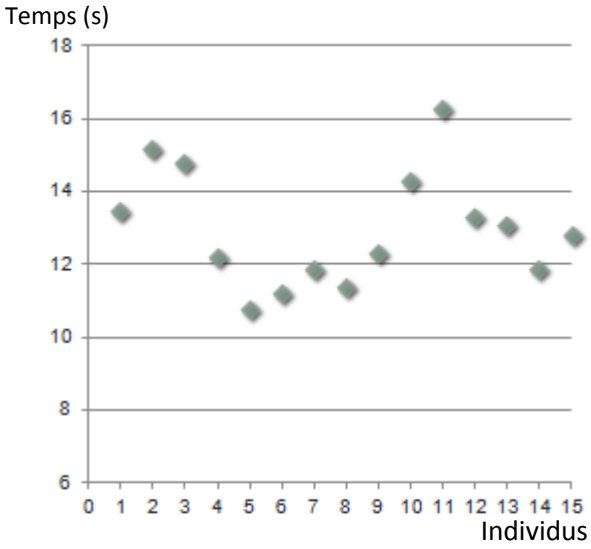
L'IC se calcule avec la moyenne et l'écart-type (dispersion) :

$$I_c = \left[\bar{x} - t_\alpha \frac{s}{\sqrt{n}} \; ; \; \bar{x} + t_\alpha \frac{s}{\sqrt{n}} \right]$$

X barré : moyenne
t_α : 1,96
s : écart-type (racine de la variance s^2)
n : effectif

Exemple : Temps en secondes pour courir un 100m pour 15 individus

individu	Temps sur 100m (s)
1	13,5
2	15,2
3	14,8
4	12,2
5	10,8
6	11,2
7	11,9
8	11,4
9	12,3
10	14,3
11	16,3
12	13,3
13	13,1
14	11,9
15	12,8
moyenne	13
écart-type	1,586550616



Racine (15) = 3,87

IC = [13 – 1,96 (1,59 / 3,87) ; 13 + 1,96 (1,59 / 3,87)]

IC = [12,19 ; 13,80]

Les IC permettent donc de mesurer la dispersion autour d'une moyenne. Cela est important pour tout résultat, en particulier lorsque l'on mesure la fiabilité d'un test clinique.

Exemple d'IC autour d'une mesure de fiabilité:

Pour un test clinique A, le coefficient de Kappa est de 0,55. Pour le Test B le Kappa est à 0,54. A première vue, nous pouvons faire une **confiance modérée** à ces deux tests, les deux semblent aussi fiables l'un que l'autre.

Mais si l'auteur nous donne l'IC:

Test A : K = 0,55 [0,21 ; 0,67]
Test B : K = 0,54 [0,46 ; 0,59]

Pour le test A, 95% des valeurs seront comprises entre 0,21 et 0,67.
Pour le test B, 95% des valeurs seront comprises entre 0,46 et 0,59.
On constate que la dispersion est moins importante pour le test B : le résultat est donc moins variable d'un échantillon à l'autre, on peut accorder plus de confiance à ce test qu'au test A.

Bibliographie

- (1) Regnaux JP, Guay V, Marsal C. Evidence based practice ou la pratique basée sur les preuves en rééducation. Kinesither Rev 2009;(94):55-61
- (2) Clealand J. Orthopaedic Clinical examination : an Evidence-based approach for physical therapists. Copyright 2005 Elsevier Saunders.
- (3) Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to practice*. 2nd Ed. Upper Saddle River, NJ : Prentice Hall Health ; 2000.
- (4) Domholdt E. *Physical Therapy Research*. Ed. Philadelphia, Pa: WB Saunders; 2000.
- (5) Shrout PE, Fleiss JL. *Intraclass correlations : uses in assessing rater reliability*. Psych Bull. 1979 ; 86 : 420-428.
- (6) Bossuyt PMM, Reitsma JB, Bruns DE, et al. *Towards complete and accurate reporting of studies od diagnosic accuracy ; explanation and elaboration*. Clin Chem. 2003 ; 49 : 7-18.

BREF

- La kinésithérapie à compris depuis plusieurs années la nécessité d'orienter le traitement à partir d'un bilan précis (BDK)

- *Problème : De très nombreux tests utilisés ne sont pas fiables, ou manquent de précision (Le test d'Apley, pour rester sur l'exemple de la lésion méniscale, à une sensibilité de 0,22 = ce test ne diagnostique pas une lésion méniscale sur 78% des patients porteurs d'une lésion méniscale)*

- Solution : Quelques bases statistiques permettent de sélectionner les tests les plus fiables et précis, permettant ainsi de *mieux communiquer* entre professionnels et *d'éviter les erreurs de diagnostic* conduisant à des traitements inadaptés