Guide méthodologique – Validation d'échelles cliniques et de questionnaires

Introduction

Dans le cadre de la rééducation il est essentiel que les évaluations cliniques soient précises et fiables afin d'orienter et d'optimiser la prise en charge thérapeutique des patients. Les échelles cliniques et questionnaires représentent des outils précieux pour mesurer de manière standardisée des variables cliniques telles que la qualité de vie, la fonction motrice globale, l'équilibre, les capacités de marche ou encore la douleur. La plupart de ces outils sont repris depuis la littérature anglaise, puis appliqués dans de nombreuses autres langues. Cependant, leur utilisation dans des populations variées et pour le suivi des patients nécessite une traduction ainsi qu'une validation rigoureuse afin de s'assurer de leur fiabilité et de leur validité dans différents contextes culturels. Ce processus de validation n'est pas seulement essentiel pour la pertinence clinique, mais aussi pour la recherche scientifique. Des outils validés sont nécessaires au recueil de données fiables et reproductibles, critères indispensables pour la publication dans des revues scientifiques de qualité.

Ce guide méthodologique est conçu pour accompagner les professionnels de santé et les chercheurs désireux de traduire et valider des échelles cliniques et des questionnaires pour leur utilisation transnationale. Par la construction d'outils de mesure fiables et reproductibles, ce guide se veut facilitant pour le développement de pratiques basées sur des preuves scientifiques robustes et pour la contribution à la littérature scientifique internationale.

Les propriétés psychométriques

Les propriétés psychométriques d'une échelle ou d'un questionnaire sont essentielles pour déterminer son utilité et sa fiabilité dans la mesure de variables spécifiques. Ces propriétés peuvent être subdivisées en plusieurs catégories : la fiabilité, la validité et la sensibilité. Ces catégories permettent de s'assurer que l'outil est non seulement capable de mesurer ce qu'il prétend mesurer, mais aussi qu'il le fait de manière précise et constante.

Fiabilité

La fiabilité d'une échelle ou d'un questionnaire est un indicateur de la constance et de la stabilité des mesures qu'il permet. Les trois domaines les plus communément retrouvés dans la littérature scientifique sont la cohérence interne, la fiabilité test-retest et la fiabilité inter-évaluateurs.

Cohérence interne

C'est le degré de corrélation entre les différentes parties d'une échelle, qui indique dans quelle mesure les items qui constituent l'échelle sont cohérents entre eux lorsqu'ils mesurent un même concept, une même dimension, un même aspect.

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

Comment la mesurer ?

Coefficient Alpha de Cronbach : c'est un test statistique qui analyse la variance de chaque item de l'échelle ou du questionnaire et la variance de tous les items combinés. On applique ce test statistique à l'ensemble des items du questionnaire. Le coefficient obtenu va de 0 à 1 et s'interprète de la manière suivante :

- **Alpha < 0.5** : **Inacceptable**. Indique une cohérence interne très faible et suggère que les items ne mesurent pas de manière cohérente le même domaine.
- **Alpha entre 0.5 et 0.6** : **Faible**. Considérée comme une fiabilité marginale. Elle pourrait être acceptable dans les études exploratoires, mais généralement, des améliorations sont nécessaires.
- **Alpha entre 0.6 et 0.7** : **Modérée**. Indique une fiabilité acceptable pour des études exploratoires ou des recherches préliminaires, mais pour des applications cliniques ou des décisions importantes, elle est considérée comme insuffisante.
- **Alpha entre 0.7 et 0.8** : **Bonne**. Indique une bonne fiabilité.
- **Alpha entre 0.8 et 0.9** : **Très bonne**. Très bonne fiabilité pour les évaluations cliniques et les recherches.
- **Alpha > 0.9** : **Excellente**, mais peut indiquer une redondance. Bien que cette valeur élevée montre une excellente fiabilité, elle peut aussi suggérer que certains items sont redondants ou que l'échelle est trop longue.

Fiabilité test-retest

Evalue la stabilité du score à un test administré à deux moments différents, indiquant que le test permet des mesures constantes dans le temps.

Comment la mesurer ?

Le test/l'échelle/le questionnaire est administré par exemple à 50 patients présentant la même condition, deux fois à une à deux semaine d'intervalle, sans qu'aucune intervention susceptible de modifier les résultats n'ai eu lieu entre temps.

Après ces deux passations, plusieurs tests statistiques peuvent être conduit

Coefficient de corrélation de Pearson
Il est particulièrement adapté aux données continues (taille, poids, âge, distance, temps, etc…) où les scores sont supposés suivre une distribution normale. Il mesure la corrélation linéaire entre deux scores à deux moments différents.

Interprétation :

- **0.00-0.19** : Très faible
- **0.20-0.39** : Faible
- **0.40-0.59** : Modérée
- **0.60-0.79** : Forte

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

- **0.80-1.00** : Très forte

Coefficient de corrélation de Spearman

S'applique aux données ordinales (exemple de réponses : très insatisfait, insatisfait, neutre, satisfait, très satisfait) ou lorsque les scores ne suivent pas une distribution normale. Il mesure la corrélation des rangs entre les deux scores mesurés à deux moments différents.

Interprétation : même seuils que pour le coefficient de Pearson, car il est également une mesure de corrélation linéaire (mais pour les rangs)

Kappa de Cohen

Il s'applique au données catégorielles (par exemple lorsque les réponses sont oui/non, ou que vous classez votre population dans plusieurs catégories). Il évalue l'accord entre deux évaluations tout en prenant en compte la part dû au hasard.

Interprétation :

- < 0 : Pas d'accord
- 0.01-0.20 : Léger
- 0.21-0.40 : Limite
- 0.41-0.60 : Modéré
- 0.61-0.80 : Substantiel
- 0.81-1.00 : Presque parfait

Coefficient de corrélation intra-class

Il convient pour les mesures continues lorsque plusieurs évaluateurs sont impliqués. Il peut être utilisé pour évaluer la cohérence (même scores à chaque mesure) ou l'accord absolu (mêmes scores et dans le même ordre).

Interprétation :

- < 0.5 : Faible
- 0.5-0.75 : Modéré
- 0.75-0.9 : Bon
- > 0.90 : Excellent

Fiabilité inter-évaluateurs

Elle évalue la concordance entre deux ou plusieurs évaluateurs qui mesurent la même chose, par exemple le même patient, en même temps, avec la même échelle de mesure et les mêmes instructions pour coter. Il est nécessaire de valider la fiabilité inter-évaluateurs pour toutes mesures où les jugements subjectifs des évaluateurs peuvent influencer les résultats.

Comment la mesurer ?

Coefficient de corrélation intra-class

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

Très souvent utilisé pour la fiabilité inter-évaluateurs, un coefficient élevé permet d'affirmer que la mesure est fiable, quel que soit l'évaluateur qui réalise la mesure. On estime qu'un coefficient supérieur à 0.75 indique une fiabilité élevée.

Kappa de Cohen

Permet de mesurer la concordance entre les deux évaluateurs en excluant la part de similitude qui pourrait être due au hasard. On estime qu'un coefficient supérieur à 0.60 est suffisant.

Exemple pratique

Vous souhaitez vérifier la fiabilité inter-évaluateur d'une échelle clinique dont la passation consiste à donner des consignes au patient, puis d'utiliser une échelle pour noter ses performances au regard des consignes qui lui sont données.

Deux évaluateurs font passer l'échelle au patient en même temps, et l'un des deux évaluateurs est tiré au sort pour donner les consignes. Seul cet évaluateur interagit avec le patient, conformément aux instructions de l'échelle clinique. L'autre évaluateur ne fait que coter les performances du patient conformément à l'échelle de cotation du test.

Une fois que l'effectif est atteint (par exemple 40 patients, avec 20 patients pour lesquels l'évaluateur A était tiré au sort pour donner les consignes, et 20 pour lesquels l'évaluateur B était tiré au sort pour donner les consignes), on peut effectuer les tests statistiques.

Attention, lorsque vous débutez votre étude, le binôme ou les binômes d'évaluateurs doivent toujours être constitués des deux mêmes évaluateurs.

Idéalement, on essaye d'avoir 2 ou 3 binômes.

Validité

La validité permet de déterminer si un outil mesure bien ce qu'il est censé mesurer. Par exemple, est-ce que la Berg Balance Scale mesure bien l'équilibre ?

On peut la diviser en plusieurs type : validité de contenu, de construit ou encore de critère.

Validité de contenu

Elle évalue l'adéquation des items d'un outil de mesure avec le domaine spécifique qu'ils sont censés couvrir. Cela permet de s'assurer que le résultat donné reflète tous les aspects pertinents du domaine que l'outil est censé mesurer. Elle suppose que les questions ou les items de l'échelle/du questionnaire, sont représentatifs du domaine à mesurer, mais également que tous les aspects significatifs de ce domaine sont abordés, sans oublis ou surreprésentation (Haynes, Richard, & Kubany, 1995). Prenons l'exemple de la Berg Balance Scale (BBS). C'est une échelle constituée de 14 items pour mesurer l'équilibre chez les adultes et les personnes âgés, via une série de tâches statiques et dynamiques. Chaque item est noté sur une échelle allant de 0 à 4. Plus le score est élevé, meilleur est l'équilibre du patient.

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

La validité de contenu de la BBS vérifie que les tâches qui font parties de l'échelle couvrent tous les aspects importants de l'équilibre et de la stabilité qu'elle est censée mesurer.

Comment la mesurer ?

Pour déterminer la validité de contenu, un ensemble d'experts va examiner chaque item et déterminer si l'ensemble des tâches de la BBS permet de couvrir tous les aspects de l'équilibre que l'on souhaite évaluer. Ils peuvent par exemple utiliser une échelle de Lickert allant de 0 (pas pertinent) à 4 (très pertinent).

Coefficient de Validité de Contenu (CVI)

Une fois les notes des experts obtenues, on peut utiliser le CVI, un test statistique très répandu pour évaluer la validité de contenu d'un outil. Il évalue le degré d'accord entre les experts par rapport à la pertinence des items.

- I-CVI (Item-level CVI) : pour chaque item, il est calculé en divisant le nombre d'experts ayant attribué une note élevée à l'item par le nombre total d'expert
- S-CVI (Scale-level CVI), avec deux modes de calcul
    - S-CVI/Ave : moyenne des I-CVIs de tous les items de l'échelle.
    - S-CVI/UA (Universal Agreement) : pourcentage d'items pour lesquels tous les experts attribue la meilleure note

Interprétation :

- 0.78 ou plus : acceptable. A noter qu'un minimum de 0.8 est requis pour déterminer la validité de contenu d'une nouvelle échelle
- 0.90 ou plus : excellente validité de contenu

Coefficient Kappa de Fleiss

Il est également possible d'utiliser le coefficient Kappa de Fleiss, qui évalue l'accord entre plusieurs évaluateurs qui utilisent une base nominale pour coter des échelles. C'est une généralisation du Kappa de Cohen pour plus de deux évaluateurs.

Interprétation :

- -1 : désaccord total
- 0 : pas d'accord autre que celui du au hasard
- 0.01 – 0.20 : Accord léger
- 0.21 – 0.40 : Accord juste
- 0.41 – 0.60 : Accord modéré
- 0.61 – 0.80 : Accord substantiel : à partir de ces seuils, les items sont considérés comme pertinents pour mesurer le domaine envisagé.
- 0.81 – 1.00 : Accord presque parfait ou parfait

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

Validité de construit

La validité de construit vérifie quant à elle la capacité d'un outil à mesurer ce qu'il est réellement censé mesurer. Cela implique de vérifier si les mesures réalisées avec l'outil se comportent de manière cohérente avec les théories, hypothèses et contextes d'application du domaine mesuré.

Cette validité va inclure les notions de validité convergente et discriminante.

Validité convergente

Elle évalue si un outil de mesure est en corrélation avec d'autres instruments qui sont censés mesurer la même chose.

Comment la mesurer ?

On compare les scores obtenus sur une cohorte de patients avec d'autres outils reconnus et validés mesurant la même chose. Par exemple, dans le cas de la BBS, on peut la comparer au Time Up And Go.

On utilise ensuite des tests statistiques de corrélations comme les coefficients de Pearson ou Spearman, en fonction de la nature des données recueillies et de la distribution des données.

Validité discriminante (ou divergente)

Elle vérifie que l'outil de mesure n'évalue pas quelque chose qu'il ne devrait pas mesurer. Gardons l'exemple de la BBS. On va cette fois-ci la comparer à, par exemple, un test qui évalue la dépression, car l'état dépressif et le contrôle de l'équilibre sont théoriquement des constructions distinctes.

On réalise entre les données obtenues à la BBS et au test pour la dépression, des coefficients de corrélation. On s'attend à ce que les coefficients soient faibles, indiquant une bonne validité discriminante de la BBS.

Validité de critère

Elle évalue la performance d'un outil avec un outil ou une variable externe considérée comme une évaluation plus directe et valide de la mesure en question. C'est la notion de comparaison à un gold-standard. Tout comme la validité de construit, elle inclut deux notions : la validité concurrente et la validité prédictive.

Différence entre la validité convergente et la validité de critère

La distinction entre ces deux validités peut paraître subtile, mais elles évaluent chacune des aspects différents de la façon dont un outil de mesure se rapporte à d'autres mesures. La différence majeure entre les deux est que la validité de critère implique nécessairement un « gold standard », ce qui n'est pas le cas pour la validité convergente. De plus, la validité de critère implique également la capacité à prédire des résultats futurs sur le domaine évalué par l'outil.

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

Validité concurrente

Elle évalue si les résultats de l'outil correspondent à ceux d'un autre outil ou d'une autre mesure, réalisée simultanément ou quasiment au même moment. L'objectif étant de montrer que l'outil est capable de mesurer une variable de manière aussi efficace que le standard existant.

Comment la mesurer

Par exemple, on fait passer la BBS a une cohorte de patients. Peu avant ou après la BBS, ils réalisent également une posturographie sur une plateforme de force 3D permettant d'évaluer l'équilibre de manière objective et quantifiable grâce aux déplacements du centre de pression.

On réalise ensuite des coefficients de corrélation de Pearson ou de Spearman entre le score à la BBS, et des variables reconnues dérivées des trajectoires du centre de pression. Un coefficient de corrélation supérieur ou égal à 0.70 est considéré comme indiquant une forte validité concurrente. Dans ce cas, on considère que l'outil évalué est un substitut valable au gold standard.

Validité prédictive

La validité prédictive évalue la capacité d'un instrument à prédire des résultats à venir, l'évolution d'un patient en terme d'amélioration ou de détérioration des performances dans le domaine mesuré, ainsi que sa potentielle réponse à un traitement.

Par exemple, dans le cas de la BBS qui est un test d'équilibre, on pourrait évaluer la capacité de l'outil à prédire le risque de chute.

Comment la mesurer ?

Avec l'exemple de la BBS : dans un premier temps, faire passer la BBS à un groupe de patient. Puis, faire un suivi à un an des patients, en recueillant le critère que vous souhaitez prédire grâce à la BBS, en l'occurrence, le nombre de chute.

Mise en place d'un modèle de régression logistique

Cette méthode statistique permet la prédiction d'une variable dépendante binaire : une variable avec deux catégories possible, par exemple « oui/non » ou « chute/pas de chute », à partir d'une ou de plusieurs variables indépendantes.

Dans notre exemple avec la BBS, on commence à définir le modèle :

- Variable dépendante : Chute « oui/non ». Oui : le patient a chuté au moins une fois sur la période considérée.
- Variable indépendante : score total de la BBS

On peut également ajouter des covariables, telles que : l'âge, la condition médicale, le sexe… qui pourraient influencer le risque de chute. On examine ensuite **les coefficients bêta du**

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

modèle et **le rapport de cotes** qui permettent d'évaluer l'influence du score BBS sur le risque de chute.

- Le **coefficient bêta** décrit l'influence des variables indépendantes (ici, le score BBS) sur la variable dépendante (ici, le risque de chute).
- Le **rapport de cotes (odds ratio)** estime de combien la cote (probabilité) de l'issue (ici le risque de chute) augmente (rapport > 1) ou diminue (rapport < 1) lorsque la variable indépendante augmente d'un point.

Par exemple, si le rapport de cote pour la BBS est de 0.95, cela signifie qu'une augmentation d'un point à la BBS est associé à une diminution de 5% du risque de chute (1-0.95).

Création d'une courbe de ROC

Les courbes de ROC (pour Receiver Operating Characteristics) sont des outils graphiques qui permettent d'évaluer la performance d'un test ou d'un modèle prédictif. Elles illustrent la capacité d'un test à distinguer deux conditions (par exemple chute/non-chute) en fonction de différentes variables.

Elle nécessite de calculer la sensibilité (taux de vrais positifs = pourcentage de personnes qui chutent et qui sont identifiées par la BBS comme à risque de chute) et la spécificité (taux de vrais négatifs = pourcentage de personnes qui ne chutent pas et sont identifiées par la BBS comme à faible risque de chute).

$$Sensibilité = \frac{Vrais\ positifs}{Vrais\ positifs + faux\ négatifs}$$

Vrais positifs : nombre de personnes qui ont effectivement chuté et que le test a correctement identifié comme à risque.
Faux négatifs : nombre de personnes qui ont chuté mais que le test n'a pas identifié comme à risque.

$$Spécificité = \frac{Vrais\ négatifs}{Vrais\ négatifs + Faux\ positifs}$$

Vrais négatifs : Nombre de personnes qui n'ont pas chuté et que le test a correctement identifié comme n'étant pas à risque.
Faux positifs : Nombre de personnes qui n'ont pas chuté mais que le test a incorrectement identifié comme à risque.

On trace ensuite la courbe ROC : la spécificité en fonction de la sensibilité (abscisse : spécificité ; ordonnée : sensibilité) puis on calcul l'aire sous la courbe.

Interprétation :

- 0.5 : pas de capacité de discrimination
- 0.6-0.7 : capacité de discrimination faible

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

- 0.7-0.8 : capacité de discrimination acceptable
- 0.8-0.9 : capacité de discrimination forte
- > 0.9 : capacité de discrimination excellente

De plus, le point de la courbe le plus en haut à gauche (qui indique 100% de spécificité et de sensibilité) détermine le seuil optimal de score à l'échelle/à l'outil testé.

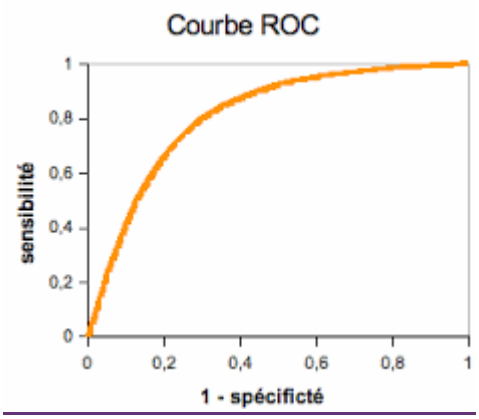

Figure 1 : forme typique d'une courbe de ROC

De ce fait, les courbes ROC permettent non seulement d'évaluer un outil, mais également d'en ajuster la sensibilité et la spécificité en tant qu'outil prédictif, et donc de choisir un score seuil sur l'outil.

Sensibilité

La **sensibilité** d'un instrument mesure sa capacité à détecter des changements dans le temps, ainsi que sa capacité à identifier correctement un état ou une condition d'intérêt. Cela est crucial pour caractériser une population, ainsi que pour évaluer l'efficacité des interventions ou des changements de condition chez le patient. Une bonne sensibilité signifie que l'instrument peut capter même les petits changements qui sont cliniquement significatifs. Elle permet donc de démontrer l'efficacité des interventions thérapeutiques, et suivre la progression des individus.

<u>Comment mesurer la capacité à discriminer différentes populations ?</u>

Comparer les scores de l'outil entre deux populations différentes (par exemple population pathologique vs population asymptomatique) via un test de comparaison de moyenne.

Vous pouvez également réaliser une analyse discriminante pour voir si votre outil permet de classer une population en fonction de ses performances à l'outil. Pour cela, utilisez une analyse de variance (ANOVA) pour évaluer les différences entre les groupes.

<u>Comment mesurer la sensibilité au changement ?</u>

En réalisant une étude longitudinale au cours de laquelle des mesures répétées sont réalisées sur une même population à différents moments, où l'on sait que la performance de la population est censée évoluer (par exemple avant et après une intervention). On utilise en parallèle, si possible, une mesure avec un gold standard ou un autre test dont la sensibilité a déjà été validée.

On utilise ensuite des tests statistiques pour mesures répétées (tests T, ANOVA, ou modèle de régression mixte) pour évaluer si les changements observés sont statistiquement significatifs. On réalise également des analyses de corrélation entre l'évolution des scores à l'outil testé, et l'évolution des scores au gold standard utilisé. On détermine ensuite la taille de l'effet, par exemple le d de Cohen, pour quantifier l'ampleur du changement détecté par l'instrument.

- d = 0.2 : petite taille d'effet
- d = 0.5 : taille d'effet moyenne
- d = 0.8 ou plus : grande d'aille d'effet

Jonathan Pierret, Christelle Requena – Cellule de Recherche Clinique et Innovation – cellule.recherche.ug-ne@ugecam.assurance-maladie.fr

Bibliographie

ANASTASI, A., & URBINA, S. (1997). PSYCHOLOGICAL TESTING (7TH ED.). UPPER SADDLE RIVER, NJ: PRENTICE HALL.

CAMPBELL, D. T., & FISKE, D. W. (1959). CONVERGENT AND DISCRIMINANT VALIDATION BY THE MULTITRAIT-MULTIMETHOD MATRIX. PSYCHOLOGICAL BULLETIN, 56(2), 81-105.

CRONBACH, L. J. (1951). COEFFICIENT ALPHA AND THE INTERNAL STRUCTURE OF TESTS. PSYCHOMETRIKA, 16(3), 297-334.

FLEISS, J. L., LEVIN, B., & PAIK, M. C. (2003). STATISTICAL METHODS FOR RATES AND PROPORTIONS. JOHN WILEY & SONS.

HAYNES, S. N., RICHARD, D., & KUBANY, E. S. (1995). CONTENT VALIDITY IN PSYCHOLOGICAL ASSESSMENT: A FUNCTIONAL APPROACH TO CONCEPTS AND METHODS. PSYCHOLOGICAL ASSESSMENT, 7(3), 238-247.

HOSMER, D. W., LEMESHOW, S., & STURDIVANT, R. X. (2013). APPLIED LOGISTIC REGRESSION (3RD EDITION). WILEY.

KOO, T. K., & LI, M. Y. (2016). A GUIDELINE OF SELECTING AND REPORTING INTRACLASS CORRELATION COEFFICIENTS FOR RELIABILITY RESEARCH. JOURNAL OF CHIROPRACTIC MEDICINE, 15(2), 155-163

MCHUGH, M. L. (2012). INTERRATER RELIABILITY: THE KAPPA STATISTIC. BIOCHEMIA MEDICA, 22(3), 276-282.

POLIT, D. F., & BECK, C. T. (2006). THE CONTENT VALIDITY INDEX: ARE YOU SURE YOU KNOW WHAT'S BEING REPORTED? CRITIQUE AND RECOMMENDATIONS. RESEARCH IN NURSING & HEALTH, 29(5), 489-497.

PORTNEY, L. G., & WATKINS, M. P. (2009). FOUNDATIONS OF CLINICAL RESEARCH: APPLICATIONS TO PRACTICE (3RD EDITION). PEARSON/PRENTICE HALL.

STREINER, D. L., NORMAN, G. R., & CAIRNEY, J. (2015). HEALTH MEASUREMENT SCALES: A PRACTICAL GUIDE TO THEIR DEVELOPMENT AND USE (5TH ED.). OXFORD: OXFORD UNIVERSITY PRESS.

TAYLOR, R. (1990). INTERPRETATION OF THE CORRELATION COEFFICIENT: A BASIC REVIEW. JOURNAL OF DIAGNOSTIC MEDICAL SONOGRAPHY, 6(1), 35-39.

ZAR, J. H. (1972). SIGNIFICANCE TESTING OF THE SPEARMAN RANK CORRELATION COEFFICIENT. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 67(339), 578-580.

ZWEIG, M. H., & CAMPBELL, G. (1993). RECEIVER-OPERATING CHARACTERISTIC (ROC) PLOTS: A FUNDAMENTAL EVALUATION TOOL IN CLINICAL MEDICINE. CLINICAL CHEMISTRY, 39(4), 561-577.