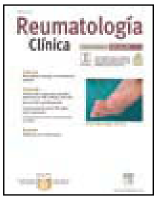
ELSEVIER
DOYMA

Reumatología Clínica

www.reumatologiaclinica.org



Review article

Validation of questionnaires

M. Jesús García de Yébenes Prous, [a,*] Francisco Rodríguez Salvanés, [b] and Loreto Carmona Ortells [a]

[a] *Unidad de Investigación, Fundación Española de Reumatología, Madrid, Spain*
[b] *Agencia Laín Entralgo, Madrid, Spain*

ARTICLE INFO

Article history:
Received July 17, 2008
Accepted September 16, 2008
Online May 20, 2009

Keywords:
Questionnaire
Health indices
Validity
Reliability

Palabras clave:
Cuestionario
Escalas de evaluación
Validez
Fiabilidad

ABSTRACT

The development of a questionnaire or a measuring instrument is a laborious and complex process and requires verification of its usefulness before implementation. We present a methodological work on the psychometric characteristics of assessment instruments and analysis of their main features, reliability, and validity.

Validación de cuestionarios

RESUMEN

El desarrollo de un cuestionario o instrumento de medición es un proceso laborioso y complejo y requiere la comprobación de su utilidad antes de su aplicación. Se presenta un trabajo metodológico sobre las características psicométricas de los instrumentos de evaluación y el análisis de sus principales características: la fiabilidad y la validez.

Introduction

In 1948, the World Health Organization defined “health” as a state of complete physical, mental, and social well-being. Since then, numerous projects have been undertaken to translate this conceptual definition into objective methods that, through questionnaires or other instruments, generate scales and indexes that facilitate the measurements of the dimension of the health status. Along with the interview, the questionnaire is one of the most commonly employed techniques in research. In this article we will consider that questionnaires, scales and instruments are synonymous with the same concept: data collection techniques.

The interview is a data collection technique that requires knowledge of verbal communication techniques, a structured script and a specific objective. It is an excellent tool for qualitative research, designed to quantify and universalize information and standardize the interview procedure. Its objective is to obtain comparative information.[1]

In general, when talking about questionnaires, evaluation scales are often referred to; for example, the quality of life questionnaire, the SF-36 is an evaluation scale. Therefore, evaluation scales are those instruments that allow for a cumulative scaling of its items, and which provide global scores at the end of the evaluation. Its cumulative character makes it different from data collection questionnaires, symptom inventories, standardized interviews, or formularies.

Both interviews and questionnaires base their information on the validity of verbal information perception, feelings, attitudes or conducts transmitted by the person interviewed; information which can be difficult to contrast and translate to a measurement system, in other words, a score; this difficulty leads to the complexity of establishing quality of this type of instrument.

The use of these evaluation scales is based on psychophysics and psychometrics. Psychophysics approximates the quantification process of perception (translating into numbers intangible phenomena such as symptoms, limitation, through the use of analogies). Psychometrics permits us to study the adaptation of the scale to the phenomenon that is the object of measurement and the quality of the measurement.[1]

[*] Corresponding author.
E-mail address: mjesus.garciadeyebenes@ser.es (M.J. García de Yébenes Prous).

1699-258X/$ - see front matter © 2008 Elsevier España, S.L. All rights reserved.

The development of a questionnaire is a laborious process that can take months before reaching a definite version that fills all of the expectations. Therefore, questionnaires that have shown to be useful in other studies must be employed, which allows for the comparison of results. However, there are occasions in which it is inevitable to design new instruments, for example, when they have been shown to be less than satisfactory or have been proven effective in different application media, or when there are no questionnaires useful for the measurement of that which we desire to measure. Under these circumstances the design of a new questionnaire and the evaluation of its usefulness before its application is justified. Questionnaires are instruments that are designed to measure a series of parameters that, in many occasions, are theoretical or abstract concepts. These measurement objects, which are not directly observable, are defined as "constructs."[2]

A valid questionnaire, as with any measurement instrument, must have the following characteristics[1]:

1. Simplicity, viability, and patient, user and researcher acceptance (feasibility).
2. Reliability and precision, in other words, mistake free measurements.
3. Adequate for the problem intended for measure (content validity).
4. Reflect underlying theory in the phenomenon or concept to be measured (construct validity).
5. Capable of measuring change, both in different individuals as in the response of the same individual through time (sensitivity to change).

While reliability and feasibility are necessary requirements of all of the instruments, the importance of other psychometric characteristics depends on the context. For example, sensitivity to change is important if the instrument is applied as part of a response measure in clinical trials, but no if used for a study on opinions or attitudes towards disease.[2]

The analysis of the metric characteristics of the instrument is a complex process that implies the evaluation of feasibility, reliability, validity and sensitivity to change (Table 1).

The objective of this work is to describe the methodology of the feasibility, reliability, and validity of questionnaires as scales or measurement instruments that allow us to obtain and quantify data with the end comparing information. The analysis of sensitivity to change is not a part of this article.

Feasibility

The best instruments are useless if their application is difficult, complex, or costly. Characteristics such as the time used for their application, how simple or approachable the format is, interest, brevity and clarity of the questions as well as the ease with which it can be scored, registered, coded and interpreted are all aspects related to feasibility. This characteristic is studied through the performance of a pilot study in a group of approximately 30 subjects and its results can be employed to carry out timely modifications to the measurement index.

Reliability

Reliability is the degree with which an instrument accurately measures something free of error. Reliability measures the proportion in the variation of measurements that is due to the diversity of values that the variable adopts and that are not a product of error; in other words, reliability measures the proportion of the total variance owed to true differences between the subjects.[2,3] A reliable instrument is precise, providing measurements free of error. Variation due to error can be due to 2 types of mistakes:

1. Systematic or bias: error produced in a systematic manner. For example, an evaluator can always score lower than the rest.
2. Random: error produced by random factors. For example, due to different circumstances, an evaluator can give sometimes give higher scores and others, lower scores to those considered correct. Random error is the one most affecting an instrument.

Reliability of an instrument us evaluated through internal consistency, test-retes or intraobserver reliability or interobserver reliability.

I. Internal consistency

This property refers to the coherency of the measurement instrument components, in other words, that items that measure the same attribute are consistent among themselves. A consistent scale guarantees that every one of its components or items measure a single construct which is homogeneous. If the scale has an elevated internal consistency, the sum of the scores can represent the measurement of a single construct with which in general, it maintains a linear relationship.

Table 1
Characteristics of the measurement instruments

Term	Sinonym	Aspects to consider	Analysis technique
Viability	Feasibility	Time employed Clarity of questions Registry, coding Result interpretation	Pilot study
	Reliability	Internal consistency	Cronbach Alpha
		Intraobserver	ICC, kappa index, Bland and Altman graphic method
		Interobserver	ICC, kappa index, Bland and Altman graphic method
Validity		Logic (*face validity*)	Postulating questions
		Content	Expert opinion
		Construct	Factorial analysis construct
		Criterion	Diagnostic tests
Sensitivity to change	Responsiveness	Intrinsic Extrinsic	In relation to the design and type of change expected

ICC indicates intraclass correlation coefficient.

Questionnaires are developed to separately assess different components or dimensions of a problem. For example, a health questionnaire can be divided into questions on physical and mental health; a good agreement on questions regarding the same component can be expected. Therefore, if a questionnaire is composed by different subscales, each one of which intends to measure a different dimension of the same phenomenon, the internal consistency of each one of them should be evaluated.[2,3] Internal consistency of one evaluation scale depends on the number of items composing the instrument and the mean correlation between them, and is evaluated in a single application of the instrument through the statistical method of Cronbach's alpha,[4] with values between 0 and 1, which are interpreted similarly to a correlation coefficient.

For example, the AUSCAN (Australian/Canadian Osteoarthritis Hand Index) contains 3 subscales that evaluate pain (5 items), stiffness (1 item), and functional capacity (9 items) of patients with osteoarthritis of the hands during the past 48 h. Subscales can be used individually or added to obtain a single global score. An evaluation of the internal consistency of the global score and subscales, Cronbach's alpha of the global score was .96 while pain and functional capacity scores were .93 and .94, respectively.[5]

II. Test-retest or intraobsever reliability

Reproducibility or test-retest reliability is used, in those instances when the same questionnaire is applied to the same population on two separate moments in time, when identical or similar results are obtained; therefore, it measures the stability of the scores given by the same evaluator in the same subjects and with the same methods in separate moments in time. This technique has some practical difficulties. For example, if the time between the 2 applications is very long, the measured phenomenon can suffer variations, while when it is too short it can present a "learning effect," the patient remembering the questions. In both cases a distorted measurement or reproducibility is obtained; in addition, some individuals may not accept a second application of the same questionnaire. However, this methd is useful for biochemical or laboratory measurements. Its analysis is performed using the interclass correlation coefficient (ICC) for quantitative measurement scales and through Cohen's kappa index for qualitative measurements.[6,7]

An important limitation of ICC is its dependency on the variability of the observed values. If the subjects studied have little variation in their score (homogeneous sample), the ICC tends to be low, while in the case of very heterogeneous samples, it tends to be higher. Bland and Altman proposed an alternative graphic method to evaluate agreement in a way that the result did not depend on the nature of the study group. However, the estimation of agreement is subjective and does not provide an objective index such as the ICC.[8]

III. Interobserver reliability

This refers to the degree of agreement between 2 or more evaluators who test the same subjects with the same instruments. The more important problems in the analysis of this dimension of reliability are systematic error and the proportion of agreement owed to randomness. The most commonly employed statistical methods for its evaluation are those commented in the prior paragraph.

In the past few years, echographic examination has awakened a great interest as a method of evaluating activity or therapeutic response of different rheumatic diseases. In this sense, Szkudlarek et al published an interobserver reliability study of the ultrasonographic findings in the joints of the fingers of patients with rheumatoid arthritis that were evaluated by two researchers with different backgrounds. Different parameters were analyzed (bone erosion, synovial membrane thickening, synovial effusion, and Power Doppler signal) that were scored in a semiquantitative scale of 0 to 4, and also as the presence or absence of each alteration. Interobserver reliability for each parameter was calculated through the three proposed statistical methods: ICC, kappa index, and the Bland and Altman graphical method. ICC and the kappa index of the examined parameters showed moderate to good reliability (0,61 to 0,81 and 0,48 to 0,68) with an elevated global agreement (79% to 91%).[9]

Validity

Validity of an instrument refers to its capacity to measure that for which it was designed. As in the case of reliability, there are different dimensions to the validity of an instrument: a logical or apparent dimension, one regarding content, one of construct, and one of criteria.

I. Logical or apparent validity

Logical or apparent validity refers to the degree in which a questionnaire "appears" to measure that for which it was intended in the opinion of experts and the subjects themselves. The decision on whether the questions should be logical or not must be taken before they are redacted. If the questions lack logical validity it is very likely that the subjects under study will refuse the questionnaire. However, in some cases it can be interesting to ask questions that lack logical validity. For example, when very sensitive or conflicting topics are breached, the use of direct questions (with high logical validity) may lead the subject to not answer or lie, making indirect questions, with less logical validity, more useful.[2]

II. Content validity

Validity of content is the degree with which the measurement envelops most of the dimensions of the concept under study; therefore, an instrument is considered valid if it contemplates all of the related aspects of the concept under study. This dimension of validity is related with the composition of the instrument and evaluates if it contains a representative sample (item) of the components of the construct it intends to measure. It consists in a systematic exam of the content of the measurement tool to determine if its items are relevant (if all are related to the concept that is supposed to be measured) as well as representative of the domain being measured (to determine whether they represent the essential characteristics of the construct and whether or not they are within the adequate proportions).

Evaluation of content validity is based on judgments with different origins (medical literature review, expert opinion, pilot studies). This should guarantee, empirically, that the content of the instrument is adequate.

There are some other forms to evaluate the validity of content, such as factorial analysis which explores the answers the to the questions of the questionnaire and attempts to group them in relation to the underlying factors that identify the possible dimensions.

The difference between apparent validity and content validity resides in the fact that the evaluation of the latter is a more exhaustive process, and perhaps more formal, in which both researchers and clinicians should participate, as well as members of the target population.

III. Construct validity

It evaluates the degree to which the instrument reflects the theory of the phenomenon or the concept to be measured. The validity of the construct guarantees that the measurements that the measurements that result from the questionnaire responses can be considered and used as a measurement of the phenomenon under study. It is therefore defined as the capacity of an instrument to adequately measure a theoretical construct. The measurement of theoretical concepts requires the prior identification of the content of the instruments to be used and the elaboration of a conceptual model that helps to interpret the results obtained with these instruments.

The validation of the construct represents the degree in which a measurement is related to other measurements in a manner consistent with the theoretical hypothesis that define the phenomenon or construct that is under measurement, and is one of the most frequent alternatives in case of a reference criteria or external criteria being absent.[10]

A commonly used method to evaluate the validity of a construct is factorial analysis which groups responses in relation to the underlying factors; in these cases it is called factorial validity. Through this technique, the correlations that exist between a set of variables are analyzed to attempt their explanation through the extraction of the factors mentioned.

Another simpler procedure is to examine if the concept in question is related to other measurements consistent with what would be expected through lineal regression or correlation coefficient (convergent validity) analysis.[2,10] For example, echographic validation of synovial inflammation has shown the validity of the construct because in transversal studies it has shown good correlation with clinical indexes of inflammatory activity and in longitudinal studies there has been correlation between synovial echographic changes after treatment and clinical and analytical changes.[11]

IV. Criteria validity

In general, when a new measurement instrument is designed, one has an alternative way to measure the phenomenon under study with a proven validity, which is taken as a reference to determine the validity of the new instrument. When one has an adequate reference method, the validity criteria of the new questionnaire must be evaluated. When attempting to validate a questionnaire, researchers refer to validity criteria. The external criteria or the reference criteria should be independent measurements, in other words, should be obtained by a different method in which the results of the questionnaire do not interfere.

This is the type of validity to which one generally refers when talking about validating an instrument, and must fill the following steps: *a)* identify a relevant and reliable external criteria; *b)* reunite a representative sample of the population in which the instrument will be used; *c)* administer the instrument and obtain a score for each subject; and *d)* evaluate each one of the individuals with the external reference criteria. The prototype of the validity of the criteria is the diagnostic test analysis.

Diagnostic test analysis

A questionnaire or a scale is designed to detect the presence or absence of a determined process. The scale in question is considered valid if the subjects are classified according to whether or not they present or not the process, with as little error as possible. For this reason it is important to determine the degree of similarity between the results obtained in the questionnaire and those obtained from a reliable and widely accepted external reference criteria as a valid measure (always positive in the presence of the process, always negative in the absence of it) of the diagnosis of this process.

The external criterion is a dycotomous one (presence or absence of disease), while the scale of a questionnaire is a continuous measure. In these cases one must choose a value or a cutpoint from which to consider the obtained score as a positive result. By establishing this cutpoint, subjects can be classified as healthy or sick according to whether the value obtained is over or under the cutpoint or threshold chosen. The classification generated when choosing a determined cutpoint contemplates 2 types of error: false positive or healthy subjects diagnosed as sick and false negatives, or sick subjects diagnosed as healthy.[11,12]

Analysis of the validity of a diagnostic test starts by the construction of a 2×2 table (Table 2).

The validity of a diagnostic test is evaluated through the sensitivity and specificity indexes.

Sensitivity

Sensitivity (S) of a diagnostic test is defined as the proportion of subjects with the disease that have a positive test. Tests are very sensitive when they detect most of the sick individuals (few false negatives).

$$S = \frac{\text{True positives}}{\text{Total sick}} = \frac{TP}{TP + FN}$$

Specificity

Specificity (E) of a diagnostic rest is the proportion of subjects without the disease that have a negative result on the test. The most specific tests are those that rule out the disease in most of the healthy subjects (few false positives).

$$E = \frac{\text{True negatives}}{\text{Total not sick}} = \frac{TN}{TN + FP}$$

In general, a diagnostic test has a reasonable validity if its sensitivity and specificity are equal or over 0.80.[2,12,13]

When the diagnostic test provides a quantitative result, the sensitivity and specificity depend on the chosen cutpoint, in other words, the value of the test from which it is considered that the subject has a positive or negative test result. The decision on the cutpoint must be carefully considered as there is an interdependence between sensitivity and specificity, making an increase in one of them lead to a reduction in the other. When choosing a cutpoint, one must take into account the fundamental objective of the test.

Diagnostic efficiency curves

When the values of a diagnostic test follow a quantitative scale, sensibility and specificity vary according to the chosen cutpoint to classify the population as healthy or sick; in other words, they represent indexes of the validity of the diagnostic test for a determined cutpoint. In this situation, a global measure of validity of the test in the universe of all possible cutpoints is obtained through the use of ROC curves (*receiver operating characteristics*) (Figure).[14] To build a ROC curve it is necessary to calculate sensitivity and specificity for all

Table 2
Basic analysis of a diagnostic test

Test result	External reference criteria	
	Not sick	Sick
Positive	FP	TP
Negative	TN	FN
TOTAL	FP+VN	VP+FN

FN indicates false negative; FP, false positive; TN, True negative; TP, true positive.

possible cutpoints. Sensitivity (S) or the proportion of true positives is situated between the Y axis and the complement of specificity is placed on the X axis (1-specificity) or the proportion of false positives; ROC curves are then drawn by joining the pairs of resulting values (1-E; S) which correspond to each cutpoint. The area under the curve (AUC) is defined as the probability of correctly classifying a pain of individuals (one sick, one healthy) randomly selected after applying the test. This type of graphic permits the evaluation of 2 extreme situations:

A test with perfect discrimination (S=1; E=1) will be represented by an ROC curve situated on the left, superior side of the graph.

A test with no diagnostic discrimination (the probability of correctly diagnosing both a healthy subject as well as a sick one will be 0.5; S=0.5; E=0.5) will be represented by a diagonal line on the graph.

The ROC curve facilitates the choice of the cutpoint. In general, if the cost of producing a false positive is similar to that of producing a false negative, the best cutpoint is the one closest to the left superior angle of the graph.[14]

Behavior of a diagnostic test

In addition to the study of the validity of a diagnostic test, it is important to evaluate its behavior when applied to different clinical contexts. To do this it is important to calculate the predictive values and the efficiency of the test:

Positive predictive value

It is the proportion of subjects with the disease within the group of individuals with a positive test result. In other words, it's the probability that an individual with a positive result will have the disease.

$$\text{PPV} = \frac{\text{True positives}}{\text{Total positives}} = \frac{\text{TP}}{\text{TP + FP}}$$

Negative predictive value

It is the proportion of subjects without the disease in the group of individuals with a negative test result. In other words, it is the probability that an individual with a negative result will not have the disease.

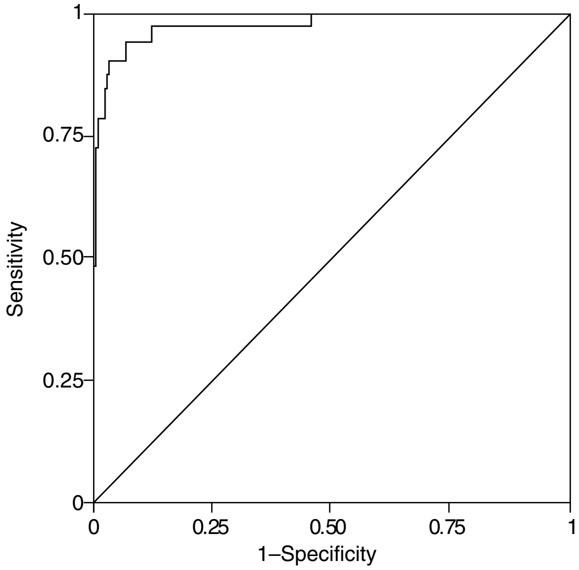


Figure. Diagnostic efficiency curve.

$$\text{NPV} = \frac{\text{True negatives}}{\text{Total negatives}} = \frac{\text{TN}}{\text{TN + FP}}$$

Efficiency or global value

It is the proportion of correctly classified subjects.

$$\text{GV} = \frac{\text{True positives + True negatives}}{\text{Total subjects}} = \frac{\text{TP + TN}}{\text{TP + FP + TN + FN}}$$

It must be taken into account that the predictive values, both positive and negative, are indexes that depend on the prevalence or the previous probability of the disease, evaluating the behavior of the diagnostic test in a population with a determined proportion of healthy subjects. Prevalence is the most determining factor of the predictive values. By being intrinsic characteristics of a measurement, sensitivity and specificity do not experiment great variations according to where they are applied if they are always performed in similar conditions. For this reason, the capacity of a test to predict cannot be evaluated without considering the prevalence of the disease; if it is high, a positive result will tend to confirm its presence, while if the result is negative it will not help to exclude it. On the contrary, when the prevalence is low, a negative result will allow ruling out the disease with an elevated confidence margin, but will not allow affirming its existence. In general, the positive predictive value is reduced when the diagnostic test is applied to populations with a lower prevalence of disease. This is because a test that produces false positives is applied to a population of mostly healthy individuals, making this situation a relatively easy one to obtain many false positives and, therefore, the predictive value for positives is reduced.[15]

Probability reasons

One way to avoid the influence of prevalence on the validity of a diagnostic test is the use of the so called *likelihood ratios* that relate sensitivity and specificity in a single index, without variations due to the prevalence of the process.

Likelihood ratio for a positive result

It is calculated by dividing the proportion of sick subjects with a positive test result (sensitivity) by the proportion of healthy subjects, but with a healthy result also (1-specificity).

$$\text{LR+} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

Likelihood ratio for a negative result

It is the quotient between the sensitivity complement and the specificity.

$$\text{LR}- = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

Of these 2 indexes, the most commonly employed in practice is the likelihood ratio for a positive result, simply known as "likelihood ratio." If, for example, a likelihood ratio of 8 is obtained, this value indicates that in the group of sick persons the probability of finding a positive test result is 8 times higher than in the group of the healthy subjects.

Table 3
Questions used to evaluate a questionnaires validity

Concept	Question	Expected on an adequate fatigue scale
Validity		
Apparent validity	Is the method sensitive?	The language reflects the ideas of patients regarding fatigue
Content validity	Are the questions based on patients?	Patients are the source or reviewers of questions
	Have all of the necessary items been included?	Ex., physical, emotional, cognitive, severity, consequence aspects
	Have all of the confusing items been avoided?	Items that could be mistaken with functional limitation
Criteria validity	Has it been compared to an external criterion of fatigue?	Evaluated versus other fatigue scales
Construct validity	Does it converge with other adequate variables?	Ex., moderate correlation with pain, inflammation, mood, anemia
Reliability		
Internal consistency	Is it internally consistent?	Moderate or elevated interitem consistency
Stability	Is it stable?	Not modified in stable patients
Feasibility	How long does it take to apply	10 to 15 min maximum
	Is it self-administered or is an interviewer employed?	It is better for scales to be self-administered
	Is it easy to score and interpret?	Clear instructions

It is considered that the likelihood ratio is one more measure to evaluate the validity of a diagnostic test. This index has the advantage of relating sensitivity and specificity in a single measurement and, therefore, is independent of the prevalence of the process. Another use of the likelihood ratio is that it also allows calculating the predictive values.

With the objective of calculating the sensitivity and specificity of Doppler echo in the diagnosis of rheumatoid arthritis of the hands and wrists and defining a cutpoint for 2 inflammation indexes (resistance index and color fraction), Terslev et al performed a study on a sample of 88 patients with active rheumatoid arthritis and 27 healthy controls. All of the individuals in the sample were studied using Doppler to calculate the resistance index and the color fraction of the wrist, metacarpophalangeal and proximal interphalangeal joints. ROC curves were constructed for both inflammation parameters and the cutpoints with the greatest sensitivity and specificity were selected. The area under the curve was 0.84 for both indexes. The cutpoint for the color fraction was 0.01 with sensitivity and specificity values of 0.92 and 0.73, respectively. In the case of the resistance index, a cutpoint value of 0.83 was selected, with a sensitivity and specificity of 0.72 and 0.70, respectively. The authors concluded that Doppler echo could detect vascularization of the inflamed synovial membrane with an elevated sensitivity and a moderate specificity.[16]

Conclusion

The use of inadequate or invalid measurement indexes can lead to confusing or unreliable results. In a systematic review of different fatigue measuring indexes in rheumatoid arthritis, the authors found that only 6 of 23 scales had reasonable validity evidence. The authors developed a series of questions that can be very useful and which are presented on Table 3.[17]

References

1. Martín Arribas MC. Diseño y validación de cuestionarios. Matronas profesión [serial online] 2004 [consulted May 19, 2008]; 5:23–9. Available from: http://www.enferpro.com/documentos/validacion_cuestionarioswww.enferpro.com/documentos/validacion_cuestionarios.
2. Argimón Pallás JM, Jiménez Vila J. Métodos de investigación clínica y epidemiológica. 3rd edition. Madrid: Edition Harcourt; 2006.
3. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chron Dis. 1985;38:27-36.
4. Altman DG, Bland JM. Cronbach&apos;s alpha. BMJ. 1997;314:572.
5. Allen KD, Jordan JM, Renner JB, Kraus VB. Validity, factor structure and clinical relevance of the AUSCAN Osteoarthritis hand index. Arthritis Rheum. 2006;54: 551-6.
6. Prieto L, Lamarca R, Casado A. La evaluación de la fiabilidad en las observaciones clínicas: el coeficiente de correlación intraclase. Med Clin. 1998;110:142-5.
7. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med. 1994;13:2465-76.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.
9. Szkudlarek M, Court-Payen M, Jacobsen S, Klarlund M, Thomsen HS, Ostergaard M. Interobserver agreement in ultrasonography of the finger and toe joints in rheumatoid arthritis. Arthritis Rheum. 2003;48:955-62.
10. van der Hofstadt CJ, Rodríguez-Marin J. Adaptación de un cuestionario para la medida de la representación de la enfermedad. Psicothema [serial online] 1997 [cited May 22, 2008]; 9:237–45. Available from: URL: www.psicothema.com.
11. Naredo E. Evaluación de la artritis reumatoide por técnicas de imagen: ecografía. Reumatol Clin. 2006;2:S13-7.
12. Pozo Rodríguez F. La eficacia de las pruebas diagnósticas (I). Med Clin (Barc). 1988; 90:779-85.
13. Pozo Rodríguez F. La eficacia de las pruebas diagnósticas (II). Med Clín (Barc). 1988; 91:177-83.
14. Bargueño MJ, García-Bastos JL, González-Buitrago JM. Las curvas ROC en la evaluación de las pruebas diagnósticas. Med Clín (Barc). 1995;104:661-70.
15. Cabello López JB, Pozo Rodríguez F. Métodos de investigación en Cardiología Clínica (X). Estudios de evaluación de las pruebas diagnósticas en Cardiología. Rev Esp Cardiol. 1997;50:507-19.
16. Terslev L, von der Recke P, Torp-Pedersen S, Koenig MJ, Bliddal H. Diagnostic sensitivity and specificity of Doppler ultrasound in rheumatoid arthritis. J Rheumatol. 2008;35:8-10.
17. Hewlett S, Mehir M, Kirwan JR. Measuring fatigue in rheumatoid arthritis: A systematic review of scales in use. Arthritis Care Res. 2007;57:429-39.