ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 983–990

Journal of Clinical Epidemiology

Quality assessment of ordinal scale reproducibility: log-linear models provided useful information on scale structure

Fabien Valet[a,b,*], Christiane Guinot[b,c], Khaled Ezzedine[d,e], Jean-Yves Mary[a]

[a]*Inserm U717, Département de Biostatistique et Informatique Médicale, Saint-Louis Hospital, 1 Avenue Claude Vellefaux, F-75010 Paris, University Paris 7, France*

[b]*Biometrics and Epidemiology Unit, CE.R.I.E.S,[1] Neuilly sur Seine, France*

[c]*Computer Science Laboratory, University François Rabelais of Tours, Tours, France*

[d]*Department of Dermatology, ULB - Erasme Hospital, Brussels, Belgium*

[e]*Unité Mixte de Recherche INSERM/INRA/CNAM, Paris, France*

Accepted 9 November 2007

Abstract

Objective: In health research, ordinal scales are extensively used. Reproducibility of ratings using these scales is important to assess their quality. This study aimed to compare two methods analyzing reproducibility: weighted Kappa statistic and log-linear models.

Study Design and Setting: Contributions of each method to the reproducibility assessment of ratings using ordinal scales were compared using intra- and interobserver data chosen in three different fields: Crow's feet scale in dermatology, dysplasia scale in oncology, updated Sydney scale in gastroenterology.

Results: Both methods provided an agreement level. In addition, log-linear models allowed evaluation of the structure of agreement. For the Crow's feet scale, both methods gave equivalent high agreement levels. For the dysplasia scale, log-linear models highlighted scale defects and Kappa statistic showed a moderate agreement. For the updated Sydney scale, log-linear models underlined a null distinguishability between two adjacent categories, whereas Kappa statistic gave a high global agreement level.

Conclusion: Methods that can investigate level and structure of agreement between ordinal ratings are valuable tools, since they may highlight heterogeneities within the scales structure and suggest modifications to improve their reproducibility.

Keywords: Rating scales; Reproducibility; Agreement; Kappa; Distinguishability; Log-linear models

1. Introduction

In health research, ordinal rating scales (ORS) are measurement instruments that have been extensively used over the past decades. Initially developed in psychometrics to assess the severity of behavioral troubles or disturbances [1,2], ORS became essential in health research tools to measure clinical outcomes such as symptoms [3,4], pathologist findings [5,6], disease severity [7,8], treatment response [9,10], and health-related quality of life [11,12]. The use of ORS allows clinicians to classify patients, in an objective and homogeneous way, into different patients' categories for which standardized treatment and management could be defined. Given that an ORS is valid if it really measures what it is intended to measure, its validity is an essential issue in the quality of such measurement instruments [13]. Sensitivity to changes is also required to detect clinically significant changes due to disease evolution. In addition, reproducibility of the ratings is also necessary.

Reproducibility can be defined by the ability to obtain similar results when several measurements of the same objects are performed. In particular for patients, the reproducibility of ratings made using an ORS is a major issue, because their classification into one of the different categories may have important consequences on their therapeutic follow-up, and possibly on their quality of life. Therefore, it is of prime importance to analyze the variability of ratings resulting from the use of an ORS and to investigate the reproducibility of these ratings as a component of the quality of this ORS.

To analyze the variability of ratings resulting from the use of an ORS, the same objects are usually rated by the same observer at two distant times (intraobserver ratings)

[1] CE.R.I.E.S. is a research center on human skin funded by CHANEL.

* Corresponding author. Tel.: +33-1-42-49-97-98; fax: +33-1-42-49-97-45.

E-mail address: fabien.valet@paris7.jussieu.fr (F. Valet).

0895-4356/08/$ – see front matter
doi: 10.1016/j.jclinepi.2007.11.004

or independently by two or more observers (interobserver ratings). Then, to estimate the degree of agreement between these ratings, weighted Kappa statistic [14] is widely used because it appears as a simple index of agreement, easily computable using statistical software. For an ORS, the weighted Kappa statistic treats disagreements in adjacent categories as less severe than disagreements in more distant categories. Moreover, these weights can be chosen in accordance with the importance given to the disagreements.

More recently, log-linear models have been developed in an attempt to explain the structure of agreement among two or more observers using an ORS [15–18]. Lately, Valet et al. [19] introduced new developments of log-linear models to investigate the quality of an ordinal scale through the analysis of the degrees of distinguishability between its adjacent categories, that is to say the ability for the two observers to distinguish between these adjacent categories. Actually, this new method provides a description of the structure of agreement, which may highlight where the defects of the scale are located.

In this paper, we described Kappa statistic and log-linear models methods that have been introduced by Valet et al., to analyze the reproducibility of two ratings made using an ORS. For that purpose, both methods were used to assess the reproducibility of ratings made with three different ORS coming from dermatology, oncology, and gastroenterology.

2. Materials and methods

2.1. Estimating the degree of agreement between two ordinal ratings: weighted Kappa statistic

Ratings made using an ORS by two independent observers A and B, or by the same observer at two distant times, can be summarized in a contingency table (Table 1). To estimate the degree of agreement between ratings n_{ij}, where n_{ij} is the number of objects rated i by A and j by B, the Cohen's weighted Kappa statistic [14] is generally used. This statistic is more informative than simple proportions of agreement, as it accounts for the proportion of agreement that is expected by chance only. Moreover, the weighted Kappa statistic is an improvement of the usual Kappa statistic, which is used for qualitative nonordinal ratings, since it can attribute "full credit" for complete agreement (maximal weight $w_{ii} = 1$ attributed to the corresponding n_{ii} diagonal cells) and varying amounts of "partial credit" for different disagreements (weights $w_{ij} < 1$, attributed to discordant ratings n_{ij}, decreasing as the distance between i and j increases). The weighted Kappa statistic is defined by:

$$\kappa_w = \frac{p_{ow} - p_{ew}}{1 - p_{ew}},$$

where $p_{ow} = (\sum_{i,j}^{I} w_{ij} n_{ij})/(N)$ and $p_{ew} = (\sum_{i,j}^{I} w_{ij} n_{i\cdot} n_{\cdot j})/(N^2)$.

In 1973, Fleiss and Cohen proposed the "square error weights" [20], defined by:

$$w_{ij} = 1 - \frac{(i-j)^2}{(I-1)^2},$$

where I is the number of categories.

Theoretical Kappa values range from −1 to 1, but in practice, they are usually nonnegative. Values close to 0 indicate agreements that can be explained by chance only and a value equal to 1 accounts for a perfect agreement. To interpret the level of agreement, a five-level nomenclature proposed by Landis and Koch [21] is generally used: "slight agreement" for Kappa values inferior or equal to 0.2, "fair agreement"]0.2–0.4], "moderate agreement"]0.4–0.6], "substantial agreement"]0.6–0.8], and "almost perfect agreement"]0.8–1].

The weighted Kappa statistic can be computed, for example, using SAS© (*FREQ* procedure option AGREE) or R (*Kappa* function in package *vcd* or *wkappa* function in package *psy*) statistical software.

2.2. Estimating the degree of distinguishability between adjacent categories: log-linear nonuniform association models

2.2.1. Theoretical aspects

When the same objects are rated by two different observers A and B (or by the same observer at two distant

Table 1
Contingency table: classification of N objects by two observers A and B (or by one observer at two distant times A and B), using an ordinal rating scale with I categories

		B							
	i/j	*1*	*2*	…	*i*	…	*I-1*	*I*	Total
A	*1*	n_{11}	n_{12}		n_{1i}		n_{1I-1}	n_{1I}	$n_{1.}$
	2	n_{21}	n_{22}		n_{2i}		n_{2I-1}	n_{2I}	$n_{2.}$
	…								
	i	n_{i1}	n_{i2}		n_{ii}		n_{iI-1}	n_{iI}	$n_{i.}$
	…								
	I-1	n_{I-11}	n_{I-12}		n_{I-1i}		n_{I-1I-1}	n_{I-1I}	$n_{I-1.}$
	I	n_{I1}	n_{I2}		n_{Ii}		n_{II-1}	n_{II}	$n_{I.}$
	Total	$n_{.1}$	$n_{.2}$		$n_{.i}$		$n_{.I-1}$	$n_{.I}$	$n_{..} = N$

Cell counts n_{ij} indicate the number of objects rated i for measure A and j for measure B.