Annals of Epidemiology xxx (2017) 1–9

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Review article

Summary measures of agreement and association between many raters’ ordinal classifications

Aya A. Mitani MPH [a,*], Phoebe E. Freer MD [b], Kerrie P. Nelson PhD [a]

[a] *Department of Biostatistics, Boston University School of Public Health, Boston, MA*
[b] *Department of Radiology and Imaging Sciences, University of Utah Hospital and Huntsman Cancer Institute, Salt Lake City*

ARTICLE INFO

Article history:
Received 21 April 2017
Accepted 7 September 2017
Available online xxx

Keywords:
Agreement
Association
Cohen’s kappa
Ordinal classification
Weighted kappa

ABSTRACT

Purpose: Interpretation of screening tests such as mammograms usually require a radiologist’s subjective visual assessment of images, often resulting in substantial discrepancies between radiologists’ classifications of subjects’ test results. In clinical screening studies to assess the strength of agreement between experts, multiple raters are often recruited to assess subjects’ test results using an ordinal classification scale. However, using traditional measures of agreement in some studies is challenging because of the presence of many raters, the use of an ordinal classification scale, and unbalanced data.
Methods: We assess and compare the performances of existing measures of agreement and association as well as a newly developed model-based measure of agreement to three large-scale clinical screening studies involving many raters’ ordinal classifications. We also conduct a simulation study to demonstrate the key properties of the summary measures.
Results: The assessment of agreement and association varied according to the choice of summary measure. Some measures were influenced by the underlying prevalence of disease and raters’ marginal distributions and/or were limited in use to balanced data sets where every rater classifies every subject. Our simulation study indicated that popular measures of agreement and association are prone to underlying disease prevalence.
Conclusions: Model-based measures provide a flexible approach for calculating agreement and association and are robust to missing and unbalanced data as well as the underlying disease prevalence.

Introduction

Studies of agreement between expert raters are often conducted to assess the reliability of diagnostic and screening tests. Many screening and diagnostic test results are classified using an ordered categorical scale. For example, radiologists use the Breast Imaging Reporting and Data System (BI-RADS) scale to classify breast density from mammography screenings. BI-RADS is an ordinal classification scale with four categories ranging from A (almost entirely fatty) to D (extremely dense) to reflect increasing breast density [1]. Measures of agreement and association provide useful summaries for ordinal classifications. Measures of agreement focus on assessing the levels of exact concordance (i.e., where raters assign the exact same category to a subject’s test result), whereas measures of association also take into account the degrees of disagreement among raters’ classifications. For example, the level of disagreement is higher between two raters who each independently classify the same mammogram into categories A and D respectively, compared with the level of disagreement between two raters who each independently classify the same mammogram into categories A and B. Measures of association are sometimes considered as weighted measures of agreement in which higher weight (“credit”) is assigned to pairs of raters’ classifications that are more similar.

Cohen’s kappa statistic is a popular summary measure of agreement, but is limited to assessing agreement between two raters’ ordinal classifications [2,3]. However, various extensions of Cohen’s kappa that provide summary measures of agreement (and association) among multiple raters have been developed. These include Fleiss’ kappa for multiple raters [4], the intrarater correlation coefficient, also known as the ICC [5], and weighted (and unweighted) kappas by Meilke et al. [6]. Despite the availability of these extended measures, many agreement studies report the average or the range of pairwise Cohen’s kappas and weighted kappas when assessing the agreement and association respectively among more than two raters

Conflict of interest: The authors have no conflict of interest to report.

[*] Corresponding author. Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118. Tel.: (617) 638-5172; fax: (617) 638-6484.

E-mail address: amitani@bu.edu (A.A. Mitani).

https://doi.org/10.1016/j.annepidem.2017.09.001
1047-2797/© 2017 Elsevier Inc. All rights reserved.

[7–12]. This can lead to complexities in interpretation and is infeasible in studies with a large number of raters.

A model-based approach can flexibly accommodate ordinal classifications of many expert raters and can provide a comprehensive summary agreement measure. Results can be extended to the general populations of raters and subjects. Nelson and Edwards [13,14] recently proposed a population-based measures of agreement and association for ordinal classification. Their model-based approach is based on the observed agreement between raters (where raters assign a subject's test result to the same category) from a generalized linear mixed model (GLMM), while minimizing the impact of chance agreement (where raters assign the same category to a subject's test result because of pure coincidence). Their approach produces easily interpretable single summary measures of agreement and association over all raters' classifications, and unbalanced and missing data can be flexibly accommodated [14]. Furthermore, Cohen's kappa and its variants have several vulnerabilities including susceptibility to extreme prevalence of the underlying disease rate (where prevalence is defined as the probability of being classified into each disease category) while the model-based approach is robust to these effects. Although restricted to small number of raters, other model-based approaches based on the generalized estimating equations also provide summary measures of agreement and association [15]. In contrast, Nelson's model-based approach is applicable to small (at least three) and large numbers of raters.

In this article, we demonstrate how various agreement and association measures can be applied in three real large-scale screening test studies, each based on incorporating many raters' ordinal classifications. Specifically, we apply average pairwise weighted and unweighted Cohen's kappas, Fleiss' kappa, ICC, Mielke's weighted and unweighted kappa for multiple raters, and Nelson and Edwards' model-based measures of agreement and association to three clinical screening test studies of breast cancer, uterine cancer, and skin disease. The rest of the article is constructed as follows: In Methods, we provide a brief description of some of the existing summary measures of agreement and association. In Results, we demonstrate how these summary measures can be implemented in three screening test studies and present results from a simulation study. Finally, in Discussion, we provide some concluding remarks and recommendations.

Methods

Measures of agreement

The conventional interpretation of an agreement or association measure according to Landis and Koch [16] is as follows; <0.00 indicate poor agreement, 0.00–0.20 indicate slight agreement, 0.21–0.40 indicate fair agreement, 0.41–0.60 indicate moderate agreement, 0.61–0.80 indicate substantial agreement, and 0.81–1.00 indicate almost perfect agreement.

Cohen's kappa

Cohen's kappa [2] is a popular measure of agreement between a pair of raters classifying I subjects into C categories adjusting for chance agreement, that is, the chance probability that two raters independently classify subjects into the same category by pure coincidence [2]. The general formula of Cohen's kappa is

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

where p0 is the proportion of observed agreement between the two raters and p_c is the proportion of chance agreement. The statistic ranges from −1 to 1 where 1 indicates complete agreement, −1 indicates complete disagreement, and 0 indicates agreement that is no better than chance. Cohen's kappa is easy to compute and is widely used in agreement studies. However, it has been noted to be vulnerable to extreme prevalence of the underlying disease rate and the marginal distribution of raters (raters' tendencies to classify the test results in a certain way) [17]. Because Cohen's kappa is designed for measuring the agreement between two raters, many authors report the average or the range of the $\binom{J}{2}$ kappa statistics computed from each possible pair of raters when a study involves multiple raters ($J > 2$) [7–12]. However, such a range or average of many kappa statistics can be complicated and difficult to interpret and is impractical in studies with a large number of raters, say $J \geq 5$. In R [18], the *psych* package can be used to compute Cohen's kappa and in SAS (SAS Institute Inc., Cary, NC), the FREQ procedure has options to compute Cohen's kappa.

Fleiss' kappa

Fleiss extended Scott's pi statistic [19] to account for the case of more than two raters classifying multiple subjects using a scale with more than two categories [4]. Fleiss' kappa is also a function of observed agreement corrected for chance agreement. Let I be the total number of subjects, K be the number of ratings for each subject, and C be the number of categories in the ordinal classification scale. Also, let n_{ic} be the number of raters who assign the ith subject to the cth category. Then, Fleiss' kappa is defined as

$$\kappa_F = \frac{\sum\limits_{i=1}^{I} \sum\limits_{c=1}^{C} n_{ic}^2 - IK\left[1 + (n-1)\sum\limits_{c=1}^{C} p_c^2\right]}{IK(K-1)\left(1 - \sum\limits_{c=1}^{C} p_c^2\right)}$$

where $p_c = \frac{1}{1K}\sum_{i=1}^{I} n_{ic}$.

A formula to calculate the corresponding variance of Fleiss' kappa is also available [20]. Because of a multiplicative factor of the sample sizes of raters and subjects on the denominator of this formula, the variance yields disproportionately small values, consequently producing an extremely narrow 95% confidence interval (CI) for Fleiss' kappa, and increasingly so for large sample sizes of raters and subjects. Fleiss' kappa ranges from 0 to 1 where 0 indicates no agreement and 1 indicates perfect agreement. Fleiss' kappa is straightforward to compute but is limited to balanced data where each subject's test result is classified by the same number of raters [3], which can be problematic in many real life studies where the ratings of some subjects' test results are missing. Similar to Cohen's kappa, Fleiss' kappa is also vulnerable to extreme prevalence of the underlying disease rate. To compute Fleiss' kappa in R, the *irr* package can be used, and in SAS, there is a user-written macro, MKAPPA [21].

Model-based kappa statistic

The model-based kappa statistic, which is based on a GLMM was recently introduced by Nelson and Edwards [13]. Suppose, we have a sample of J ($j = 1, \ldots, J$) raters each independently classifying a sample of I ($i = 1, \ldots, I$) subjects using an ordered classification scale with C ($c = 1, \ldots, C$) categories. We denote the rating on the ith subject's test result classified by the jth rater into the cth category as $Y_{ij} = c$. An ordinal GLMM with a probit link and a crossed random effect structure can be used to model the cumulative probability that a subject's test result, Y_{ij}, is classified as category c or lower. Then the probability that a subject's test result is classified as category c can be computed by

$$\Pr(Y_{ij} = c|u_i, v_j) = \Phi(\alpha_c - (u_i + v_j)) - \Phi(\alpha_{c-1} - (u_i + v_j)) \qquad (1)$$

where Φ is the cumulative distribution function of the standard normal distribution, $\alpha_0, \ldots, \alpha_c$ are the thresholds that estimate the