Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(10–11) 3012–3026

DOI: 10.1177/0962280218794733
journals.sagepub.com/home/smm
SAGE

# Asymptotic variability of (multilevel) multirater kappa coefficients

**Sophie Vanbelle**

**Abstract**

Agreement studies are of paramount importance in various scientific domains. When several observers classify objects on categorical scales, agreement can be quantified through multirater kappa coefficients. In most statistical packages, the standard error of these coefficients is only available under the null hypothesis that the coefficient is equal to zero, preventing the construction of confidence intervals in the general case. The aim of this paper is triple. First, simple analytic formulae for the standard error of multirater kappa coefficients will be given in the general case. Second, these formulae will be extended to the case of multilevel data structures. The formulae are based on simple matrix algebra and are implemented in the R package "multiagree". Third, guidelines on the choice between the different mulitrater kappa coefficients will be provided.

**Keywords**

Fleiss's kappa, Conger kappa, pairwise agreement, hierarchical, nested, rater

## 1 Introduction

Reliability and agreement studies are of paramount importance in medical and behavioral sciences. They provide information about the amount of error inherent to any diagnosis, score or measurement. Using unreliable measurement instruments and procedures can lead to incorrect conclusions from scientific studies and unreproducible research while disagreement between physicians can lead, in clinical decision making, to different treatments for the patient. Reliability is classically defined as the ratio between the true score variance and the total variance and is quantified through different versions of the intraclass correlation coefficient (ICC), depending on the study design.[1] When several observers rate subjects, ICCs for consistency are obtained if the systematic shifts between the observers are ignored while ICCs for agreement are obtained if they are taken into account. In parallel to the ICCs, scaled agreement coefficients[2–4] were developed outside the classical test theory and were found to be closely related to ICCs for agreement.

While it is easy to define the agreement between two observers on a categorical scale for a given object (they agree or they don't agree), this is not the case when agreement is searched between several observers ($R > 2$). In this latter case, the agreement can be defined by an arbitrary choice along a continuum ranging from agreement between a pair of observers to agreement among all the $R$ observers, i.e. a concordant classification between $g$ observers ($g = 2, \ldots, R$). Conger[5] formalised this framework by defining the g-wise agreement coefficients, including the less restrictive (pairwise) and the most restrictive (R-wise) definition of agreement. In practice, $g$ is often equal to 2 or to the majority of the observers ($g > R/2$). Mielke and Berry[6] prefer the R-wise definition to take all interactions between the R observers into account. Despite this appealing property, attention is restricted to pairwise agreement coefficients ($g = 2$) in this paper because of their practical interpretation.

The two pairwise agreement coefficients considered in this paper pertain to the kappa coefficient family and were shown to be asymptotically equivalent to ICCs for agreement when the scale is binary. The first agreement coefficient is commonly named Fleiss kappa. It was developed by Fleiss[7] and was shown to be asymptotically equivalent to the ICC for agreement based on a one-way ANOVA design.[8] In a one-way setting, each object is

Department of Methodology and Statistics, CAPHRI, Maastricht University, The Netherlands

**Corresponding author:**
Sophie Vanbelle, Department of Methodology and Statistics, CAPHRI, Maastricht University, P. Debyeplein 1, 6229 HA Maastricht, The Netherlands.
Email: sophie.vanbelle@maastrichtuniversity.nl

rated by a different set of observers, randomly selected in a population. Therefore, the variation due to the observers cannot be separated from the error variation and only ICC for agreement can be determined.[1] The second coefficient is the pairwise kappa coefficient developed by Conger[5] and equivalently by Davies and Fleiss,[9] Schouten[10] and O'Connell and Dobson.[11] This second coefficient will be referred to as 'Conger kappa' to differentiate it from 'Fleiss kappa'. When all objects are classified on a binary scale by the same set of observers randomly selected in a population, Conger kappa is asymptotically equivalent to the ICC for agreement under a two-way ANOVA setting including the observers as systematic source of disagreement.[9,12]

Fleiss kappa coefficient is popular, as assessed by more than 4000 citations of his original paper in Google scholar as compared to the 350 citations of Conger's paper and 300 citations of Davies and Fleiss' paper. The three following issues were identified with the use of multirater kappa coefficients in the literature. First, Fleiss kappa is used independently of the design of the study. The misuse of Fleiss kappa in the two-way ANOVA setting is likely to result in an underestimation of the agreement level,[13] as Fleiss kappa coefficient gives on average smaller values than Conger kappa. In the same way, the misuse of Conger kappa in one-way ANOVA settings is likely to overestimate the agreement level. It is therefore important to use the appropriate multirater kappa coefficient, based on the study design and the corresponding ANOVA model.

Second, main statistical packages (e.g. R package 'irr', STATA, SAS macro MAGREE, SPSS extension STATS_FLEISS_KAPPA) only provide the standard error of Fleiss kappa under the hypothesis that it equals zero, despite the existence of a formula for the general case derived by Schouten.[10] Worse, with the exception of the R package 'magree', Conger kappa coefficient, when available (e.g. R package 'irr', STATA), is reported without standard error, although an asymptotic formula based on the delta method was also provided by Schouten[14] and O'Connell and Dobson.[11]

Finally, there is a need to define multirater kappa coefficients and provide statistical inference in the presence of multilevel data. Multilevel data are commonly encountered in medical and behavioural sciences, where measures are often obtained on persons nested in organisations (e.g. patients in health care centers), on different parts of the body or by repeated measurements over time. For example, in the study motivating this paper, seven groups of four medical observers with different experience levels were asked to assess the presence of crackles and wheezes (yes/no) on the lung sounds of 20 subjects. The lung sounds were recorded with a stethoscope at three locations on each side of the thorax, leading to six observations per subject. The aim of the study was to evaluate the level of agreement within each group of observers. Specific statistical techniques need to be used to account for the dependency between the objects of the same cluster. It was shown in various contexts that ignoring the hierarchical structure of the data can lead to incorrect conclusions (e.g. Hox[15]). Therefore, Barlow et al.[16] and Oden,[17] among others, proposed stratified agreement coefficients. They use a weighted average of the agreement coefficients obtained on each cluster. These coefficients, however, are not asymptotically equivalent to ICCs and possess a less straightforward interpretation than the coefficients considered here.

The aim of this paper is therefore threefold. First, the formula of the standard error derived with the delta method by Schouten[10,14] and O'Connell and Dobson[11] for Fleiss and Conger kappa will be presented in a unified framework using simple notations. Second, these formulas will be extended to the case of multilevel data structures, based on recent work.[18–20] Third, the paper will emphasise the appropriate use and interpretation of Fleiss and Conger kappa depending on the study design. The standard error formulae derived by the delta method are based on simple algebra, easy to program and implemented in the R statistical package 'multiagree' available on Github. As an alternative, the clustered bootstrap method will also be considered and the statistical performances of the two methods will be compared using simulations.

In Section 2, the two multirater kappa coefficients, Fleiss and Conger kappa coefficients are reviewed and the general formula derived by the delta method for their standard error is given. These definitions are generalised to multilevel data in Section 3. The standard error of the multilevel multirater kappa coefficients are derived using the delta method and the clustered bootstrap method in Section 4. Then, the statistical properties of the delta and the bootstrap methods are studied using simulation in Section 5. The methods are illustrated on psychological and medical data in Section 6. Finally, the results are discussed in Section 7.

## 2 Definition of the classical pairwise agreement coefficients

Suppose that a sample of $N$ objects is classified by several observers on a K-categorical scale. Two situations can be distinguished and will lead to different agreement coefficients: (1) each object $i$ ($i = 1, \ldots, N$) is rated by a different random sample of observers of size $R_i$ and (2) the same $R$ observers rate all objects. Fleiss kappa coefficient is an appropriate agreement measure in the first case and Conger kappa coefficient in the second case.

Let the random variable $Y_{ij(r)}$ be equal to 1 when observer $r$ classifies object $i$ in category $j$ ($\sum_{j=1}^{K} Y_{ij(r)} = 1$) and $y_{ij(r)}$ denote the realisation of the random variable $Y_{ij(r)}$ ($i = 1, \ldots, N; j = 1, \ldots, K; r = 1, \ldots, R_i$). Finally, let $n_{ij} = \sum_{r=1}^{R_i} y_{ij(r)}$ be the number of observers classifying object $i$ in category $j$. When each object $i$ ($i = 1, \ldots, N$) is rated by a different random sample of observers, only the $n_{ij}$ are available. The two pairwise kappa coefficients, Fleiss and Conger kappas, denoted, respectively, by $\kappa_1$ and $\kappa_2$, are estimated by

$$\hat{\kappa}_l = \frac{P_o - P_{el}}{1 - P_{el}} \tag{1}$$

The proportion $P_o$ is the observed agreement. It is defined as the mean proportion of agreement between all possible pairs of observers. In the case of Fleiss kappa coefficient, we have

$$P_o = \frac{1}{N}\left\{\sum_{i=1}^{N} \frac{1}{R_i(R_i-1)} \sum_{j=1}^{K} n_{ij}(n_{ij}-1)\right\} = \frac{1}{N}\sum_{i=1}^{N} P_{o,i} \tag{2}$$

For Conger kappa, the same expression is obtained, namely

$$\begin{aligned} P_o &= \frac{1}{NR(R-1)} \sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{r=1}^{R}\sum_{r' \neq r} y_{ij,r} y_{ij,r'} \\ &= \frac{1}{NR(R-1)}\left\{\sum_{i=1}^{N}\sum_{j=1}^{K} n_{ij}(n_{ij}-1)\right\} = \frac{1}{N}\sum_{i=1}^{N} P_{o,i} \end{aligned} \tag{3}$$

The proportion $P_{el}$ ($l = 1, 2$) is the agreement expected under the assumption of statistical independence between any two observers. Its expression differs for the two multirater kappa coefficients $\hat{\kappa}_1$ and $\hat{\kappa}_2$, as explained in the following section.

## 2.1 Fleiss kappa coefficient

The expected agreement was defined by Fleiss[7] under a one-way ANOVA setting, i.e. when the $R_i$ observers are not the same for all objects, as

$$P_{e1} = \sum_{j=1}^{K} p_j^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R_i}\sum_{j=1}^{K} n_{ij} p_j = \frac{1}{N}\sum_{i=1}^{N} P_{e1,i} \tag{4}$$

where $p_j = \sum_{i=1}^{N} n_{ij}/(NR_i)$ is the overall proportion of objects classified in category $j$ ($j = 1, \ldots, K$). When the scale is binary, $\hat{\kappa}_1$ is asymptotically ($N \geq 20$) equivalent to the ICC for agreement corresponding to a one-way random effect ANOVA model including the observers as source of variation in the denominator.[8] The difference with the ICC lies in the definition of the between objects mean sum of squares (i.e. BMS) which is divided by the number of objects $N$ instead of $N-1$. The agreement coefficient $\hat{\kappa}_1$ can be expressed in terms of variance components[21] and reduces to the intraclass kappa coefficient[22] when $R_i = R = 2$.

The asymptotic sampling variance of $\hat{\kappa}_1$ was derived by Schouten[10] and can be written as

$$\text{var}(\hat{\kappa}_1) = \frac{\sum_{i=1}^{N}\left[(1-P_{e1})P_{o,i} - 2(1-P_o)P_{e1,i} - (P_o P_{e1} - 2P_{e1} + P_o)\right]^2}{N^2(1-P_{e1})^4} \tag{5}$$

where $P_{o,i}$ is the observed agreement corresponding to object $i$ defined in equation (2). The quantity $P_{e1,i}$ is the expected agreement for object $i$ defined in equation (4).

Under the null hypothesis that $\kappa_1 = 0$ and an equal number of observers per object ($R_i = R$), the formula reduces to the formula derived by Fleiss[23] and available in statistical software,

$$\text{var}(\hat{\kappa}_1) = \frac{2\left[(1-P_{e1})^2 + P_{e1} - 2\sum_{j=1}^{K} p_j^3\right]}{(1-P_{e1})^2 NR(R-1)} \tag{6}$$

## 2.2 Conger kappa coefficient

The expected agreement is defined as the mean proportion of expected agreement between all $R(R-1)$ pairs of observers[9] and can be expressed as

$$P_{e2} = \frac{1}{R(R-1)} \sum_{j=1}^{K} \sum_{r=1}^{R} \sum_{r' \neq r} p_{j(r)} p_{j(r')} = \sum_{j=1}^{K} p_j^2 - \frac{1}{R(R-1)} \sum_{j=1}^{K} \sum_{r=1}^{R} (p_{j(r)} - p_j)^2 \tag{7}$$

where $p_{j(r)}$ is the proportion of objects classified in category $j$ by observer $r$ ($j = 1, \ldots, K$; $r = 1, \ldots, R$).

For binary scales, Davies and Fleiss[9] have shown that $\hat{\kappa}_2$ is asymptotically ($N > 15$) equivalent to the ICC for agreement corresponding to a two-way random effect ANOVA model[8] including the observers as source of variation. Conger kappa can also be expressed in terms of variance components, the difference with the ICC lies in the denominator. The term $R(JMS - EMS)/N$ in the ICC is replaced by $R(JMS)/(N-1)$, where $JMS$ denotes the between observers mean sum of squares and $EMS$ the mean residual sum of squares. The agreement coefficient $\hat{\kappa}_2$ reduces to Cohen's kappa coefficient when $R = 2$. Davies and Fleiss[9] gave the formula of the large sampling variance in the binary case under the null hypothesis that the agreement coefficient is equal to zero and proposed a FORTRAN program for scales with more than two categories. However, Schouten[14] and O'Connell and Dobson[11] derived a formula in the general case for nominal scales using the delta method, available in the R package 'magree'

$$\text{var}(\hat{\kappa}_2) = \frac{\sum_{i=1}^{N} [(1 - P_{e2}) P_{o,i} - 2(1 - P_o) P_{e2,i}]^2 / N - (P_o P_{e2} - 2P_{e2} + P_o)^2}{N(1 - P_{e2})^4} \tag{8}$$

where

$$P_{e2,i} = \frac{1}{R(R-1)} \sum_{r=1}^{R} \sum_{r' \neq r} \sum_{j=1}^{K} p_{j(r)} y_{ij(r')}$$

## 3 Definition of multilevel multirater pairwise kappa coefficients

Multilevel multirater pairwise kappa coefficients will be defined similarly to the case of two observers.[18–20] Suppose that the population $\mathcal{I}$ of objects possesses a 2-level hierarchical structure in the sense that there are $C$ clusters with $n_c$ objects ($\sum_{c=1}^{C} n_c = N$). If there are $R_i$ observers rating object $i$, $P_i = R_i(R_i - 1)$ pairs of observers can be formed. These pairs will be denoted by the superscript $p = (r_1, r_2)$ where $r_1$ and $r_2$ correspond to the two observers of pair $p$. Let $Y_{ij(r),c}$ equal 1 if object $i$ from cluster $c$ is classified in category $j$ by observer $r$ and $y_{ij(r),c}$ be its realisation. Note that under a one-way design, only $n_{ij,c} = \sum_{r=1}^{R_i} y_{ij(r),c}$ is in general available.

In order to be able to define an overall kappa coefficient, two assumptions are made. First, it is assumed that the objects are homogeneous in each cluster, in the sense that the probability $E(Y_{ij(r_1),c} Y_{ik(r_2),c}) = \pi_{jk,c}^{(p)}$ of being classified in category $j$ by observer $r_1$ and $k$ by observer $r_2$ of pair $p$ is the same for all objects in cluster $c$. This implies that the probability to be classified in category $j$ by observer $r$ is the same for all objects in cluster $c$, namely $\pi_{j(r),c}$. Second, it is assumed that there is no sub-population of objects, i.e. $E(\pi_{jk,c}^{(p)}) = \pi_{jk}^{(p)}$ and therefore also $E(\pi_{j(r),c}) = \pi_{j(r)}$.

Let $v_c = n_c/N$ denote the relative sample size of the $c$th cluster and $p_{jk,c}^{(p)}$ the realisation of $\pi_{jk,c}^{(p)}$. The multilevel observed agreement is defined as the average observed proportion of agreement over all possible pairs of observers. In case of a one-way analysis of variance, this means

$$P_o = \sum_{c=1}^{C} v_c P_{o,c} = \sum_{c=1}^{C} v_c \sum_{i=1}^{n_c} \frac{1}{n_c R_i (R_i - 1)} \sum_{j=1}^{K} n_{ij,c}(n_{ij,c} - 1)$$

The expected agreement for multilevel data is then defined by

$$P_{e1} = \sum_{j=1}^{K} p_j^2$$

with $p_j = \sum_{c=1}^{C} v_c \sum_{i=1}^{n_c} \frac{n_{ij,c}}{n_c R_i} = \sum_{c=1}^{C} v_c p_{j,c}$.

In the case of a two-way ANOVA setting, if $p_{jk}^{(p)} = \sum_{c=1}^{C} \nu_c p_{jk,c}^{(p)}$, the observed agreement is

$$P_o = \frac{1}{P}\sum_{p=1}^{P}\sum_{j=1}^{K} p_{jj}^{(p)} = \frac{1}{P}\sum_{p=1}^{P}\sum_{j=1}^{K}\sum_{c=1}^{C} \nu_c p_{jj,c}^{(p)} = \frac{1}{P}\sum_{p=1}^{P}\sum_{c=1}^{C} \nu_c P_{o,c}^{(p)}$$

In the same way, if the proportion of objects classified in category $j$ is denoted by $p_{j(r_1)} = \sum_{c=1}^{C} \nu_c p_{j(r_1),c}^{(p)} = \sum_{c=1}^{C} \nu_c \sum_{k=1}^{K} p_{jk,c}^{(p)}$ for observer $r_1$ of pair $p$ and $p_{j(r_2)} = \sum_{c=1}^{C} \nu_c \sum_{k=1}^{K} p_{kj,c}^{(p)}$ for observer $r_2$, the expected agreement for multilevel data is defined by

$$P_{e2} = \frac{1}{R(R-1)}\sum_{r_1=1}^{R}\sum_{r_1\neq r_2}\sum_{j=1}^{K} p_{j(r1)}p_{j(r2)}$$

The multilevel counterpart of Fleiss kappa coefficient ($\hat{\kappa}_1$) and Conger kappa coefficient ($\hat{\kappa}_2$) are obtained by using the multilevel expression of $P_o$ and $P_{el}$ in equation (1). They reduce to Fleiss and Conger kappa coefficients when the hierarchical level of the data is ignored.

## 4 Sampling variability

### 4.1 Delta method

We will consider the vector $\widehat{\boldsymbol{\xi}_1}$

$$\widehat{\boldsymbol{\xi}_1} = \begin{pmatrix} P_o \\ \boldsymbol{p} \end{pmatrix} = \sum_{c=1}^{C} \nu_c \begin{pmatrix} P_{o,c} \\ \boldsymbol{p}_c \end{pmatrix}$$

for the one-way ANOVA setting, where $\boldsymbol{p} = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_K)^T$ *and* $\boldsymbol{p}_c = (p_{1,c}, \cdots, p_{K,c})^T$. For the two-way ANOVA setting, let $\boldsymbol{p}_{(r),c}$ be the vector with the marginal classification proportions relative to cluster $c$ and observer $r$, that is $\boldsymbol{p}_{(r),c} = (p_{1(r),c}, \ldots, p_{K(r),c})^T$. The observed agreement between observers $r_1$ and $r_2$ of pair $p$ for cluster $c$ is given by $P_{o,c}^{(p)} = \sum_{j=1}^{K} p_{jj,c}^{(p)}$.

We will consider the vector

$$\widehat{\boldsymbol{\xi}_2} = \begin{pmatrix} P_o^{(1)} \\ \cdots \\ P_o^{(P)} \\ \boldsymbol{p}_{(1)} \\ \cdots \\ \boldsymbol{p}_{(R)} \end{pmatrix} = \sum_{c=1}^{C} \nu_c \begin{pmatrix} P_{o,c}^{(1)} \\ \cdots \\ P_{o,c}^{(P)} \\ \boldsymbol{p}_{(1),c} \\ \cdots \\ \boldsymbol{p}_{(R),c} \end{pmatrix}$$

Similarly to Yang and Zhou[18,19] and to Vanbelle,[20] it can be shown that asymptotically, under mild regularity conditions, $\widehat{\boldsymbol{\xi}_1}$ and $\widehat{\boldsymbol{\xi}_2}$ are asymptotically normally distributed with variance–covariance matrix $\text{var}(\widehat{\boldsymbol{\xi}_j})$, $j = 1, 2$. The elements of $\text{var}(\widehat{\boldsymbol{\xi}_j})$ are estimated in Appendix 1, following the technique of Obuchowski.[24]

The delta method will be applied on successive functions of the vector $\widehat{\boldsymbol{\xi}_j}$ to lead to the standard error of the multilevel Fleiss and Conger kappa coefficients. The aim is to derive the asymptotic variance–covariance matrix of the vector $(P_o, P_{ej})^T$. Then, a last application of the delta method will lead to the asymptotic variance–covariance of $\hat{\kappa}_1$ and $\hat{\kappa}_2$.

#### 4.1.1 Multilevel Fleiss kappa

When the objects are not all classified by the same set of observers, the vector $(P_o, P_{e1})^T$ is a function of the vector $\widehat{\boldsymbol{\xi}_1}$ (i.e. $\widehat{\boldsymbol{\theta}} = f(\hat{\boldsymbol{\xi}}_1)$) fulfilling the conditions of the multivariate delta method. The asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\theta}}$ is, by application of the delta method, given by

$$\text{var}(\widehat{\boldsymbol{\theta}}) = C\boldsymbol{F}\text{var}(\widehat{\boldsymbol{\xi}_1})\boldsymbol{F}^T$$

where $\boldsymbol{F}$ is the Jacobian matrix corresponding to $f(.)$ with respect to $\widehat{\boldsymbol{\xi}_1}$, that is, $F$ is a $2 \times (K+1)$ matrix with null elements except elements (1, 1) equal to 1 and element $(2, 2:(K+1))$ equal to $2\boldsymbol{p}^T$.

#### 4.1.2 Multilevel Conger kappa

When the objects are all classified by the same set of observers, the expected agreement is the average of the expected agreement over all pairs of observers. In matrix notation, the agreement expected under the independence assumption of the two observers of pair $p$ is given by $P_{e1}^{(p)} = \boldsymbol{p}_{(r_1)}^T \boldsymbol{p}_{(r_2)}$. The vector $\widehat{\boldsymbol{\Psi}} = (P_o^{(1)}, \ldots, P_o^{(P)}, P_{e2}^{(1)}, \ldots, P_{e2}^{(P)})^T$ is a function of the vector $\widehat{\boldsymbol{\xi}_2}$ (i.e., $\widehat{\boldsymbol{\Psi}} = m(\widehat{\boldsymbol{\xi}_2})$) fulfilling the conditions of the multivariate delta method. The asymptotic variance–covariance matrix of $\left(P_o^{(1)}, \ldots, P_o^{(P)}, P_{e2}^{(1)}, \ldots, P_{e2}^{(P)}\right)^T$ is, by application of the delta method, given by

$$\text{var}(\widehat{\boldsymbol{\Psi}}) = C\boldsymbol{M}\text{var}(\widehat{\boldsymbol{\xi}_2})\boldsymbol{M}^T$$

where $\boldsymbol{M}$ is the Jacobian matrix corresponding to $m(.)$ with respect to $\widehat{\boldsymbol{\xi}_2}$, that is

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I}_{P\times P} & \boldsymbol{0}_{P\times RK} \\ \boldsymbol{0}_{P\times P} & \boldsymbol{M}_1 \end{pmatrix}$$

where $\boldsymbol{M}_1$ is a $P \times RK$ matrix with null elements except elements $(p, r_1 : (r_1 + K))$ equal to $\boldsymbol{p}_{(r_2)}$ and elements $(p, r_2 : (r_2 + K))$ equal to $\boldsymbol{p}_{(r_1)}^T$, $(r_1, r_2 = 1, \ldots, R)$.

In the same way, the overall observed agreement and expected agreement $P_o$ and $P_{e,2}$ are the average of the observed and expected agreement for all pairs of observers given in $\widehat{\boldsymbol{\Psi}}$, fulfilling the conditions of the multivariate delta method, $(P_o, P_{e,2})^T = q(\widehat{\boldsymbol{\Psi}})$. The asymptotic variance–covariance matrix of $(P_o, P_{e2})^T$ is, by application of the delta method, given by

$$\text{var}((P_o, P_{e2})^T) = \boldsymbol{Q}\text{var}(\widehat{\boldsymbol{\Psi}})\boldsymbol{Q}^T$$

where the matrix $Q$ is the Jacobian corresponding to the function $q(.)$, i.e. a $2 \times 2P$ matrix with null elements except elements $(1, i)$ and $(2, P + i)$ equal to $1/P$ $(i = 1, \ldots, P)$.

#### 4.1.3 Multilevel Fleiss and Conger kappa

Finally, the multilevel multirater kappa coefficient $\hat{\kappa}_1$ and $\hat{\kappa}_2$ are function of the vectors $(P_o, P_{e1})^T$ and $(P_o, P_{e2})^T$, respectively, fulfilling the conditions of the multivariate delta method. $\hat{\kappa}_l = h((P_o, P_{el})^T)$, $l = 1, 2$. The variance–covariance matrix $\text{var}(\hat{\kappa}_l)$ is, by application of the delta method, given by

$$\text{var}(\hat{\kappa}_l) = \frac{1}{C}\boldsymbol{H}\text{var}((P_o, P_{el})^T)\boldsymbol{H}^T \tag{9}$$

with

$$\boldsymbol{H} = \begin{pmatrix} \frac{1}{1-P_{el}} & \frac{P_o-1}{(1-P_{el})^2} \end{pmatrix}$$

When there is only one unit per cluster ($n_c = 1\ \forall c$), the variance given by equation (9) for the multilevel multirater Fleiss and Conger kappa coefficients multiplied by a correction factor, namely $C/(C-1)$, reduces to equation (5) for multilevel Fleiss kappa coefficient and equation (8) for multilevel Conger kappa coefficient, respectively. When there are only two observers, the formula reduces to the formula derived by Yang and Zhou.[18]

### 4.2 The clustered bootstrap method

The clustered bootstrap method was applied by Kang et al.[25] to derive the standard error of the Cohen's kappa coefficient in the presence of multilevel data and by Vanbelle[20] to derive the variance–covariance matrix when comparing several kappa coefficients. The clustered bootstrap consists of three steps:

(1) Draw a random sample with replacement of size $C$ from the cluster indexes.
(2) For each cluster, take all observations belonging to the cluster. If the cluster sizes are different, the sample size of the bootstrap sample could be different from the original sample size $N$.
(3) Repeat steps 1 and 2 to generate a total of $B$ independent bootstrap samples.

Depending on the study design, the multilevel Fleiss or Conger kappa coefficient is determined for each bootstrap sample $b = 1, \ldots, B$: $\hat{\kappa}_l^b$ ($l = 1, 2$). The bootstrap estimate of the agreement coefficient $\kappa_l$ is then defined by[25]

$$\hat{\kappa}_{l,B} = \frac{1}{B}\left(\sum_{b=1}^{B} \hat{\kappa}_l^b\right) \tag{10}$$

with variance

$$\text{var}(\hat{\kappa}_{l,B}) = \frac{\sum_{b=1}^{B} (\hat{\kappa}_l^b - \hat{\kappa}_{l,B})^2}{B-1}$$

Alternatively, percentiles can be considered to construct confidence intervals.

## 5 Simulations

To study the behavior of the type I error rate ($\alpha$), multilevel-dependent binary variables with fixed marginal distribution and dependency between pairs of variables were simulated following the algorithm of Emrich and Piedmonte.[26] Data were simulated under a two-way ANOVA setting, leading Conger kappa coefficient as the appropriate agreement measure. That is, we supposed that $R$ observers each classified $C$ clusters with each $n_c = n$ subjects. For each cluster, a $1 \times n_c R$ vector of binary correlated random variables was generated using the R package 'mvtbinaryEP' version 1.0.1. Note that the behavior of Fleiss and Conger kappa coefficients is very similar since they only differ in the definition of the expected agreement. The two measures coincide if the marginal probability distributions of the observers are exactly the same.

The assessment on a binary scale of $C = 25$, 50 and 100 clusters with each $n_c = n = 1, 2, 5$ or 10 objects by $R = 2$, 5 or 10 observers was simulated. For each cluster, the association structure between the assessments made by the observers can be characterised by two $n \times n$ matrices. The first matrix represents the intra-cluster association structure. The diagonal elements are equal to 1 (same observer, same object) and the off-diagonal elements (same observer, different objects), representing the association strength between members of a same cluster, were fixed to $\kappa_{intra} = 0$, 0.1, 0.3, 0.5 and 0.7. The second matrix gives the inter-observer agreement structure. The diagonal elements, representing the inter-observer agreement levels, were fixed to $\kappa_2 = 0$, 0.2, 0.4, 0.6 and 0.8. The off-diagonal elements, representing the association between the classification of two different objects by two different observers, were randomly chosen in the possible values allowed by the algorithm, given the Fréchet bounds. This represents a total of 180 schemes for each number of clusters $C$.

To allow a wide range of possible agreement values, all observers were assumed to have a uniform marginal probability distribution. This implies that $\kappa_1$ and $\kappa_2$ reduce to the correlation coefficient for the binary case, namely the $\phi$ coefficient.[2] For each simulation scheme, the mean squared error, the mean standard error, and the coverage probability, defined as the number of times the 95% confidence interval covers the theoretical agreement value, were recorded. For the clustered bootstrap method, the coverage was determined for the 95% confidence interval based on mean and standard error and based on percentiles. The clustered bootstrap method was based on $B = 5000$ bootstrap samples. A total of 1000 simulations were performed for each parameter configuration. Therefore, the 95% confidence interval for the nominal coverage level is [0.936; 0.963].

### 5.1 Simulation results for $n_c = 1$ (no multilevel data)

The coverage levels obtained for Conger kappa coefficients when there is no multilevel structure are presented in Figure 1 for observers with uniform marginal probability distribution using the delta and the percentile-based clustered bootstrap method. The complete results are given as supplemental material.

The coverage levels obtained with the delta and the percentile-bootstrap methods are very similar, except for high agreement values where the percentile-bootstrap method performs better. The percentile-bootstrap confidence intervals (CIs) are left-skewed in that case and provide better coverage levels. An important finding is that the coverage is too low when the sample size is small ($C = 25$) and the kappa coefficient is small ($\kappa < 0.2$). This situation worsens when the number of observers increases.

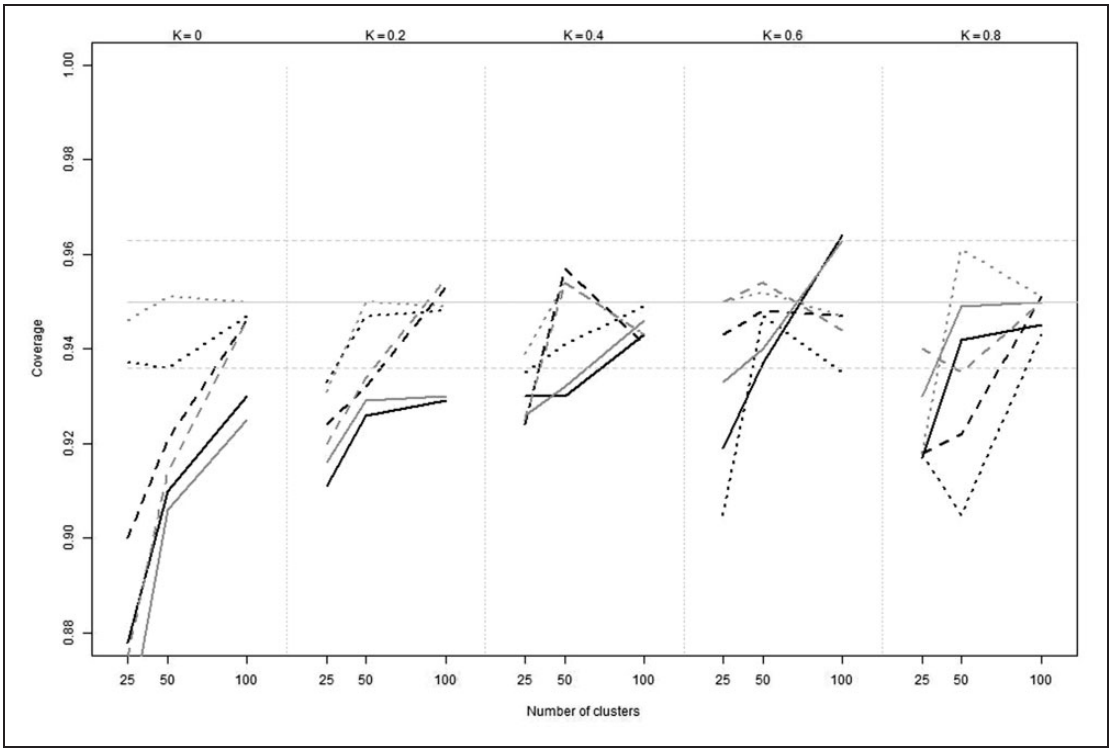


**Figure 1.** Simulations. Coverage for Conger's kappa coefficient against the number of clusters obtained with the delta method (black) and the percentile-based bootstrap method (gray) in the presence of 2 (dotted), 5 (dashed) and 10 (plain) observers with uniform marginal probability distribution. The number of objects per cluster is equal to 1.

### 5.2 Simulations results for $n_c = 2, 5, 10$ (multilevel data)

The results obtained with the delta and the clustered bootstrap methods were very similar and stable across the different number of objects per cluster. Therefore, only the results obtained with the delta method for five objects per cluster ($n_c = 5$) are presented in Figure 2. The complete results can be found in the supplemental material.

As seen in Figure 2, the coverage level becomes closer to the nominal level when the value of Conger coefficient increases, when the intra-cluster association level decreases and when the number of observers decreases. The coverage level was generally within the 95% confidence interval for kappa values above 0.4 and a number of clusters larger than 50. Here too, the percentile cluster bootstrap method provides better coverage levels for high agreement values when the number of clusters is small (see Supplemental material, $C = 25$).

## 6 Examples

### 6.1 Psychiatric diagnosis

This section focuses on the data analysed in the original paper of Fleiss.[7] These data does not present a multilevel structure. A total of six psychiatrists were unsystematically selected from a pool of 43 psychiatrists to give a psychiatric diagnosis to a subject. The set of observers can therefore differ from subject to subject, leading to Fleiss kappa coefficient as agreement measure. A total of 30 subjects were classified as suffering mainly of (1) depression, (2) personality disorder, (3) schizophrenia, (4) neurosis or (5) other psychiatric disorder. The probability to be classified in these categories was respectively $p_1 = 0.144$, $p_2 = 0.144$, $p_3 = 0.167$, $p_4 = 0.306$ and $p_5 = 0.239$ (see Table 1).

Fleiss' conclusion was that agreement was better than chance for all categories. While 0 is indeed not included in the confidence interval for each of the five categories, the lower confidence bound is close to 0 for categories 1 and 2 (see Table 1). The observed agreement $P_o$ varies between 0.78 and 0.87, meaning that when isolating one category, pairs of observers agree, on average, on 78–87% of the patients. However, when considering the five diagnostic categories together, this percentage drops to 56%. This suggests that agreement, when isolating

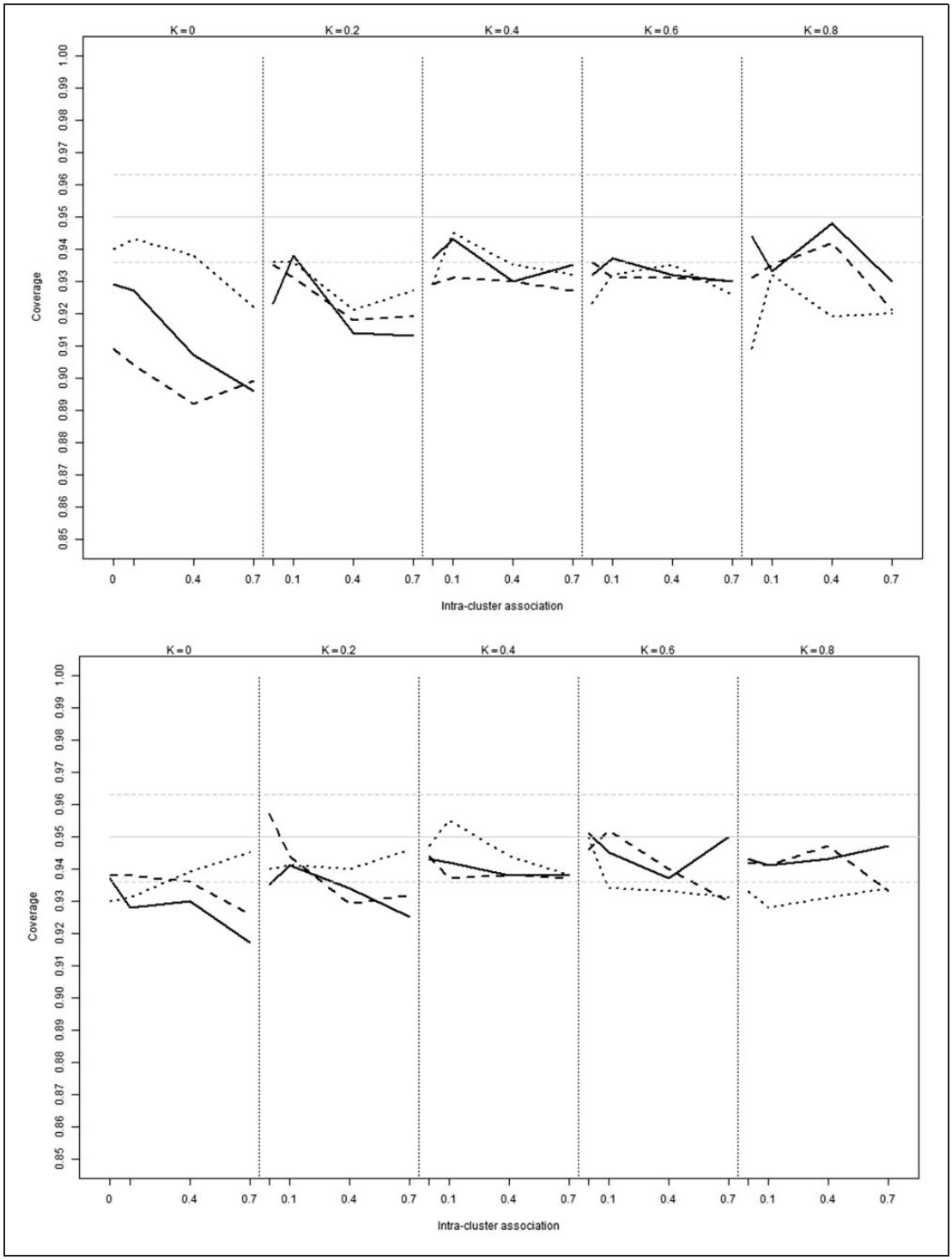


**Figure 2.** Simulations. Coverage for Conger's kappa coefficient according to the delta method in the presence of 2 (dotted), 5 (dashed) and 10 (plain) observers with uniform marginal probability distribution, 25 (up), 50 (middle) and 100 (bottom) clusters and five objects per cluster.

one category, does not mainly occur on the isolated category but rather in the category mixing the other four diagnostic categories. If we focus on the interpretation of the confidence interval for Fleiss kappa coefficient (0.33–0.53), it can be concluded that we are 95% confident that the actual proportion of disagreement is between (1–0.53)100 = 47% and (1–0.33)100 = 67% lower than the proportion of disagreement

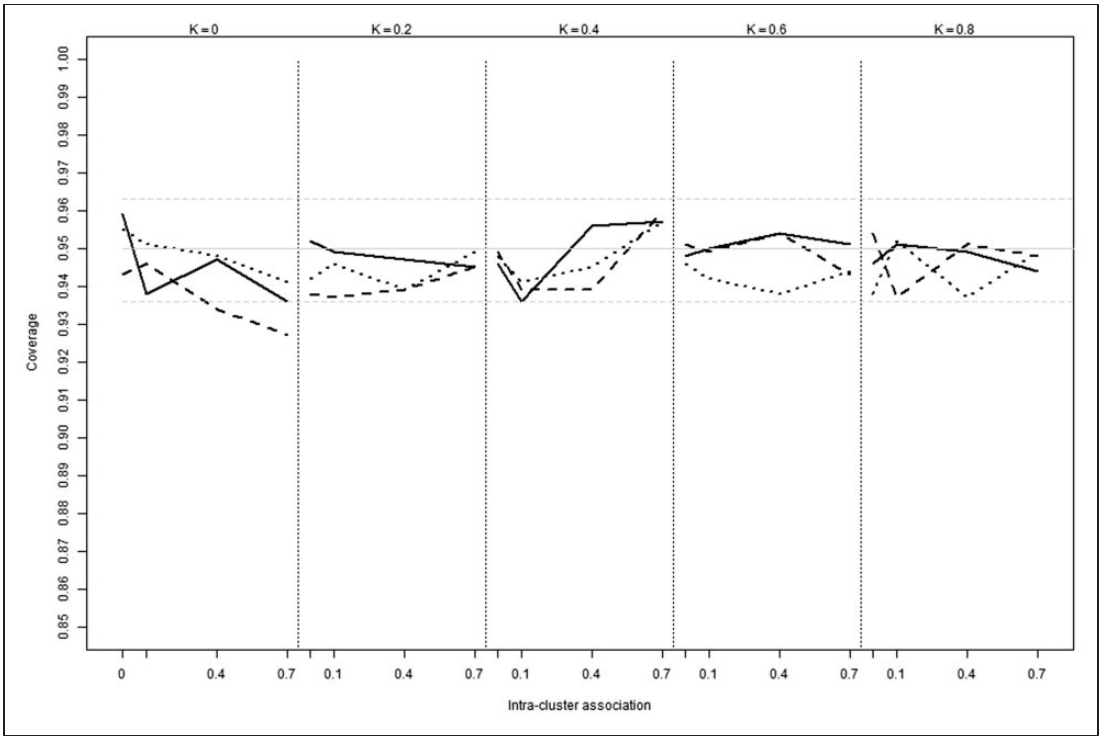


**Figure 2.** Continued.

**Table 1.** Fleiss example.

| | | | | Delta method | | | Bootstrap method | | |
|---|---|---|---|---|---|---|---|---|---|
| Category | $p_j$ | $P_o$ | $P_e$ | $\kappa_1$ (SE) | 95% CI | | $\kappa_1$ (SE) | 95% CI | |
| 1 | 0.144 | 0.813 | 0.753 | 0.245 (0.109) | 0.031 | 0.459 | 0.232 (0.108) | 0.020 | 0.443 |
| 2 | 0.144 | 0.813 | 0.753 | 0.245 (0.115) | 0.020 | 0.470 | 0.231 (0.098) | 0.040 | 0.422 |
| 3 | 0.167 | 0.867 | 0.722 | 0.520 (0.100) | 0.324 | 0.716 | 0.511 (0.078) | 0.358 | 0.664 |
| 4 | 0.306 | 0.776 | 0.576 | 0.471 (0.084) | 0.307 | 0.635 | 0.459 (0.076) | 0.310 | 0.608 |
| 5 | 0.239 | 0.842 | 0.636 | 0.566 (0.115) | 0.341 | 0.791 | 0.550 (0.128) | 0.298 | 0.801 |
| Overall | | 0.556 | 0.220 | 0.430 (0.054) | 0.324 | 0.536 | 0.418 (0.055) | 0.309 | 0.526 |

Note: Summary of the statistics to compute Fleiss kappa for each category separately and overall.

expected under the independence assumption of the observers. Both observed agreement and Fleiss kappa coefficients therefore indicate a non-negligible variability in the psychiatric diagnostic within groups of observers.

## 6.2 Tromsø study (multilevel)

Lung auscultation is routinely used in daily clinical practice by health professionals. While new methodology of chest imaging such as MRI, CT scans and portable ultrasound are now available, the stethoscope remains advantageous when it comes to costs, availability, patient care and training of health professionals to use it. Lung auscultation has proven to be helpful in the diagnosis of several lung and heart related conditions as a part of routine physical examination. However, there is a lack of information about how the presence of wheezes or crackles relates to common heart and lung diseases and the prognostic value these findings might have.

The Tromsø study is a population-based study designed to evaluate abnormal auscultation findings against a wide range of clinical and epidemiological endpoints. Due to the subjective nature of evaluating sounds,

**Table 2.** Tromsø example.

| Body location | EXP | NOR | RUS | WAL | NLD | PLN | STU |
|---|---|---|---|---|---|---|---|
| U | 0.13 | 0.11 | 0.23 | 0.083 | 0.12 | 0.22 | 0.22 |
| L | 0.29 | 0.38 | 0.37 | 0.19 | 0.16 | 0.34 | 0.36 |
| A | 0.016 | 0.048 | 0.3 | 0.029 | 0.046 | 0.12 | 0.22 |
| *P*-value | <0.0001 | <0.0001 | 0.031 | <0.0001 | 0.0015 | <0.0001 | 0.0093 |

Note: Probability to detect crackles according to the location (anterior thorax (A), upper posterior thorax (U) and lower posterior thorax (L). The probabilities are compared among locations using a multilevel probit regression.

**Table 3.** Tromsø example.

| | U | | L | | A | | All | |
|---|---|---|---|---|---|---|---|---|
| Group | $P_o$ | $\kappa_2$ (SE) | $P_o$ | $\kappa_2$ (SE) | $P_o$ | $\kappa_2$ (SE) | $P_o$ | $\kappa_2$ (SE) |
| EXP | 0.88 | 0.65 (0.13) | 0.78 | 0.52 (0.08) | 0.91 | 0.04 (0.06) | 0.86 | 0.56 (0.08) |
| NOR | 0.92 | 0.75 (0.12) | 0.78 | 0.55 (0.10) | 0.85 | 0.10 (0.06) | 0.85 | 0.58 (0.08) |
| RUS | 0.72 | 0.25 (0.08) | 0.64 | 0.26 (0.07) | 0.59 | 0.06 (0.07) | 0.65 | 0.20 (0.05) |
| WAL | 0.86 | 0.48 (0.17) | 0.88 | 0.71 (0.10) | 0.86 | 0.01 (0.05) | 0.87 | 0.53 (0.09) |
| NLD | 0.85 | 0.54 (0.13) | 0.86 | 0.61 (0.12) | 0.85 | 0.07 (0.06) | 0.86 | 0.49 (0.10) |
| PLN | 0.80 | 0.50 (0.14) | 0.76 | 0.49 (0.12) | 0.73 | 0.05 (0.07) | 0.76 | 0.40 (0.09) |
| STU | 0.78 | 0.43 (0.15) | 0.79 | 0.56 (0.11) | 0.63 | 0.02 (0.05) | 0.74 | 0.37 (0.08) |

Note: Proportion of agreement ($P_o$) and Conger's kappa coefficient (standard error) for each group of observers reported overall (All) and at each thorax location (anterior thorax (A), upper posterior thorax (U) and lower posterior thorax (L)).

the inter-observer agreement among medical professionals in classifying lung sounds was studied before the implementation of the Tromsø study.[27]

Seven groups of four observers were asked to assess the presence of crackles and wheezes on the lung sounds of 20 subjects: general practitioners (GPs) from The Netherlands (NLD), Wales (WAL), Russia (RUS), and Norway (NOR), pulmonologists working at the University Hospital of North Norway (PLN), sixth year medical students (STU) at the Faculty of Health Sciences in Tromsø and an international group of experts (researchers) in the field of lung sounds (EXP). Lung sounds were recorded at six different locations, three locations on each side of the thorax (Anterior thorax (A), upper posterior thorax (U) and lower posterior thorax (L)), leading to a multilevel data structure. A more detailed description of the study can be found in Aviles et al.[27]

In this section, we will focus on the detection of crackles. Since the same observers classified all the sounds obtained at the six body places, the multilevel Conger kappa coefficient is adopted. There are two prerequisites to the definition of agreement at the patient level: (1) the absence of patient sub-population in terms of crackles detection and (2) the homogeneity of crackles detection within patients, that is the probability of detecting crackles should be the same for the three thorax locations. Among the 20 subjects, 13 were recruited in a rehabilitation center and 7 in the office environment of the researchers. Although differences in the probability to detect crackles are expected between these two groups of subjects, Conger's kappa coefficient will be computed overall due to the limited sample size. The probability to detect crackles is given for the seven groups of observers and the three thorax locations in Table 2.

Since the probability to detect crackles differs between the three locations, the average proportion of agreement between pairs of observers is reported for each group of observers and each thorax location separately, on top of the overall agreement (see Table 3).

When looking at the individual thorax locations, it can be seen in Table 3 that on average, pairs of general practitioners (NOR, WAL, NLD) agree on the classification of more than 78% of the sounds, independently of the thorax location except for Russian GPs (RUS) where pairs agree, on 59% to 72% of the sounds. This lower agreement level might partially be explained by a confusion with the English nomenclature around the term crackles (see Aviles et al.[27] for more details). The experts agree on average on 78–91% of the sounds, the pulmonologists on 73–80% and the students on 63–79%, depending on the location of the auscultation.

These agreement proportions translate to relatively low Conger kappa coefficients, especially in the anterior thorax location. This can be explained by the low probabilities of detecting crackles (see Table 2) combined with the small sample size. The misclassification of one sound in fact represents a disagreement on 5% of the sounds. The overall and per location multirater agreement levels within groups of GPs were considered satisfactory by the researchers for using lung auscultation in the Tromsø study.

## 7 Discussion

In this paper, the asymptotic formula of the standard error of Fleiss and Conger kappa coefficients using the delta method was presented in a unified framework. The formula was extended to account for multilevel data structures. The formula only involves simple matrix calculations and can be easily implemented in practice. A `R package 'multiagree'` was developed by the author and is available on Github. Code to reproduce the results is available as Supporting Information on the journal's web page.

The scope of this paper was limited to Fleiss and Conger kappa coefficients for two reasons. First, they cover two study designs frequently encountered in practice. Second, both are asymptotically equivalent to ICCs for agreement. Fleiss kappa coefficient was developed as an agreement measure under a one-way ANOVA model, i.e. when the objects are rated by different sets of observers. On the other hand, Conger kappa was developed as an agreement measure under a two-way ANOVA model, i.e. when all objects are rated by the same set of observers. The choice between these two agreement coefficients should therefore be primarily based on the study design.

Two assumptions were made to ensure the existence of an overall multirater multilevel kappa coefficient, i.e. the homogeneity of the members of a cluster and the existence of a common kappa coefficient across the clusters. When there is evidence that the assumptions do not hold, as discussed by Yang and Zhou,[18] a separate multirater multilevel kappa coefficient should be computed for each sub-population identified. In the same way, if sub-groups of observers are identified, it is better to compute agreement separately within the different groups.[14]

The multilevel delta method, although asymptotic, showed similar coverage levels than the clustered bootstrap method. In the presence of more than two observers, good statistical performances of the delta method were observed for moderate number of clusters (e.g. $C = 50$) and multilevel kappa coefficients higher than 0.4, disregarding the cluster size. For two observers, good statistical properties were observed already for small sample sizes ($C = 25$). When the sample size is small, confidence intervals based on the percentile clustered bootstrap method provide better coverage levels for high kappa coefficients ($\kappa = 0.8$).

One extension of the methods presented in this paper is also implemented in the R package 'multiagree'. The results in this paper were combined with the results in Vanbelle[20] to allow the comparison of several (multilevel) multirater agreement coefficients. A further extension could be the inclusion of agreement weights when computing multilevel multirater agreement coefficients.

To summarise, this paper provides two simple methods to compute the standard error of the multirater kappa coefficients that perform well when the number of clusters is moderate ($C = 50$). Only the percentile clustered bootstrap method provided satisfactory coverage levels when the number of clusters was small ($C = 25$) and the agreement was high ($\kappa = 0.8$). The delta and the clustered bootstrap methods should therefore be used with caution when the number of clusters is small.

### Acknowledgments

The author is grateful to Dr. Juan Carlos Aviles (Department of Community Medicine, Tromsø university) and Pr. Hasse Melbye (General practice research unit, Tromsø university) for providing the data of the Tromsø example. The author thanks the two anonymous reviewers whose constructive remarks improved the manuscript.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is part of project 451-13-002 funded by the Netherlands Organisation for Scientific Research.

## Supplemental material

Supplemental material for this article is available online.

## ORCID iD

Sophie Vanbelle http://orcid.org/0000-0001-6584-2522

## References


1. McGraw KO and Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Meth* 1996; **1**: 30–46.
2. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960; **20**: 37–46.
3. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; **70**: 213–220.
4. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**: 255–268.
5. Conger AJ. Integration and generalization of kappas for multiple raters. *Psychol Bull* 1980; **88**: 322–328.
6. Mielke PW, Berry KJ and Johnston JE. Resampling probability values for weighted kappa with multiple raters. *Psychol Rep* 2008; **102**: 606–613.
7. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; **76**: 378–382.
8. Landis JR and Koch GG. A one-way components of variance model for categorical data. *Biometrics* 1977; **33**: 671–679.
9. Davies M and Fleiss JL. Measuring agreement for multinomial data. *Biometrics* 1982; **38**: 1047–1051.
10. Schouten HJA. Measuring pairwise agreement among many observers. ii. Some improvements and additions. *Biometric J* 1982; **24**: 431–435. http://dx.doi.org/10.1002/bimj.4710240502.
11. O'Connell DL and Dobson AJ. General observer-agreement measures on individual subjects and groups of subjects. *Biometrics* 1984; **40**: 973–983.
12. McKenzie DP, McKinnon AJ, Peladeau N, et al. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *J Psychiatr Res* 1996; **30**: 483–492.
13. Shrout PE and Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; **86**: 420–428.
14. Schouten HJA. Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica* 1982; **36**: 45–61.
15. Hox J. *Multilevel analysis: techniques and applications* Mahwah, USA; Quantitative methodology series. Lawrence Erlbaum Associates, 2002).
16. Barlow W, Lai MY and Azen SP. A comparison of methods for calculating a stratified kappa. *Stat Med* 1991; **10**: 1465–1472.
17. Oden NL. Estimating kappa from binocular data. *Stat Med* 1991; **10**: 1303–1311.
18. Yang Z and Zhou M. Kappa statistic for clustered matched-pair data. *Stat Med* 2014; **33**: 2612–2633. http://dx.doi.org/10.1002/sim.6113.
19. Yang Z and Zhou M. Weighted kappa statistic for clustered matched-pair ordinal data. *Computat Stat Data Analysis* 2015; **82**: 1–18.
20. Vanbelle S. Comparing dependent kappa coefficients obtained on multilevel data. *Biometric J* 2017; **59**: 1016–1034.
21. Fleiss JL. *Statistical methods for rates and proportions*, 2nd ed. New York, NY: John Wiley, 1981.
22. Kraemer HC. Ramifications of a population model for $\kappa$ as a coefficient of reliability. *Psychometrika* 1979; **44**: 461–472.
23. Fleiss JL, Nee JCM and Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979; **86**: 974–977.
24. Obuchowski NA. On the comparison of correlated proportions for clustered data. *Stat Med* 1998; **17**: 1495–1507. (http://dx.doi.org/10.1002/(SICI)1097-0258(19980715)17:13<1495::AID-SIM863>3.0.CO;2-I (accessed 19 June 2018)).
25. Kang C, Qaqish B, Monaco J, et al. Kappa statistic for clustered dichotomous responses from physicians and patients. *Stat Med* 2013; **32**: 3700–3719. http://dx.doi.org/10.1002/sim.5796.
26. Emrich LJ and Piedmonte MR. A method for generating high-dimensional multivariate binary variates. *Am Stat* 1991; **45**: 302–304.
27. Aviles-Solis J, Vanbelle S, Halvorsen P, et al. International perception of lung sounds: a comparison of classification across some European borders. *BMJ Open Respir Res* 2017; **4**: e000250.
28. Rao JNK and Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992; **48**: 577–585.

## Appendix 1. Multilevel Fleiss kappa coefficient

Define the vector $\widehat{\xi_1}$ by

$$\widehat{\xi_1} = \begin{pmatrix} P_o \\ \boldsymbol{p} \end{pmatrix} = \sum_{c=1}^{C} \nu_c \begin{pmatrix} P_{o,c} \\ \boldsymbol{p}_c \end{pmatrix}$$

Denote the vector containing the agreement observed in each cluster by $\boldsymbol{P}_o = (P_{o,1}, \ldots, P_{o,C})^T$ and the vector $\boldsymbol{p}_\bullet = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_C)^T$. Further let the vector with the weight relative to each cluster be denoted by $\boldsymbol{v} = (\nu_1, \ldots, \nu_C)^T$ and $\boldsymbol{\Omega} = \left(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{1}_{C\times 1}^T\right)\mathrm{diag}(\nu_1^2, \ldots, \nu_C^2)\left(\boldsymbol{I} - \boldsymbol{1}_{C\times 1}\boldsymbol{v}^T\right)$.

Similarly to Yang and Zhou,[18] it can be shown that under mild regularity conditions, the vector $\widehat{\xi_1}$ is asymptotically normally distributed with variance–covariance matrix given by a four block matrix $W$

$$\mathrm{var}(\widehat{\xi_1}) = \frac{1}{C}\begin{pmatrix} W_A & \boldsymbol{W}_B \\ \boldsymbol{W}_B^T & \boldsymbol{W}_D \end{pmatrix}$$

The elements of $\mathrm{var}(\widehat{\xi_1})$ can be estimated following the techniques of Rao and Scott[28] and Obuchowski.[24] The variance of the observed agreement is given by

$$\hat{W}_A = \tfrac{C^2}{C-1}\boldsymbol{P}_o\boldsymbol{\Omega}\boldsymbol{P}_o^T$$

The variance–covariance relative to the observed agreement and the marginal probability distribution of the observers is given in the $1 \times K$ matrix $\boldsymbol{W}_B$ by

$$\widehat{\boldsymbol{W}}_B = \frac{C^2}{C-1}\boldsymbol{P}_o\boldsymbol{\Omega}\boldsymbol{p}^T$$

Finally, the variance–covariance relative to the marginal probability distribution of the observers is given in the $K \times K$ matrix $\boldsymbol{W}_D$ by

$$\widehat{\boldsymbol{W}}_D = \tfrac{C^2}{C-1}\boldsymbol{p}\boldsymbol{\Omega}\boldsymbol{p}^T$$

## Multilevel Conger kappa coefficient

Denote the vector containing the agreement observed in each cluster by $\boldsymbol{P}_o^{(p)} = (P_{o,1}^{(p)}, \ldots, P_{o,C}^{(p)})$ $(p = 1, \ldots, P)$, the vector with the weight relative to each cluster by $\boldsymbol{v} = (\nu_1, \ldots, \nu_C)^T$, the matrix with the $C$ cluster-specific marginal classification distributions by $\boldsymbol{p}_{r,\bullet} = (\boldsymbol{p}_{(r),1}, \ldots, \boldsymbol{p}_{(r),C})_{K\times C}$ for observer $r$ $(r = 1, \ldots, R)$. Define the vector $\widehat{\xi_2}$ as

$$\widehat{\xi_2} = \begin{pmatrix} P_o^{(1)} \\ \cdots \\ P_o^{(P)} \\ \boldsymbol{p}_{(1)} \\ \cdots \\ \boldsymbol{p}_{(R)} \end{pmatrix} = \sum_{c=1}^{C} \nu_c \begin{pmatrix} P_{o,c}^{(1)} \\ \cdots \\ P_{o,c}^{(P)} \\ \boldsymbol{p}_{(1),c} \\ \cdots \\ \boldsymbol{p}_{(R),c} \end{pmatrix}$$

Using these notations, the vector with the overall marginal classification distribution for the $R$ observers is given by $\boldsymbol{p}_r = \boldsymbol{p}_{r,\bullet}\boldsymbol{v}$. Similarly to Yang and Zhou,[18] it can be shown that under mild regularity conditions, the vector $\widehat{\xi_2}$ is asymptotically normally distributed with variance–covariance matrix given by a four block matrix

$$\mathrm{var}(\widehat{\xi_2}) = \frac{1}{C}\begin{pmatrix} \boldsymbol{V}_A & \boldsymbol{V}_B \\ \boldsymbol{V}_B^T & \boldsymbol{V}_D \end{pmatrix}$$

The elements of $\text{var}(\widehat{\xi}_2)$ can be estimated following the techniques of Rao and Scott[28] and Obuchowski.[24] The variances and covariance between the observed agreement in pair $s$ and $t$ are given in the $P \times P$ $\boldsymbol{V}_A$ matrix.

$$\hat{V}_{A,ss} = \frac{C^2}{C-1}\boldsymbol{P}_o^{(s)}\boldsymbol{\Omega}\boldsymbol{P}_o^{(s)T}, \quad (s = 1, \ldots, P) \quad \text{and}$$

$$\hat{V}_{A,st} = \hat{V}_{A,st} = \frac{C^2}{C-1}\boldsymbol{P}_o^{(s)}\boldsymbol{\Omega}\boldsymbol{P}_o^{(t)T}, \text{ respectively}$$

The variance–covariance part relative to the observed agreement and the marginal probability distribution of the $R$ observers is given in the $P \times R$ matrix $\boldsymbol{V}_B$ by

$$\widehat{\boldsymbol{V}}_{B,sr} = \tfrac{C^2}{C-1}\boldsymbol{P}_o^{(s)}\boldsymbol{\Omega}\boldsymbol{p}_{r,\bullet}^T$$

Finally, the variance–covariance part between the marginal probability distribution of the $R$ observers is given in the $RK \times RK$ matrix $\boldsymbol{V}_D$ by

$$\widehat{\boldsymbol{V}}_{D,ru} = \tfrac{C^2}{C-1}\boldsymbol{p}_{r,\bullet}\boldsymbol{\Omega}\boldsymbol{p}_{u,\bullet}^T$$