ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - 2026, Volume 21, Issue 1

Tutorial on Sample Size Calculations for the Precision and the Testing of Cohen's Kappa: A Methodological Review and New Proposals

Bruno Mario Cesana[(1*, 2, 3#)], *Paolo Antonelli*[(4*, 5*)]

(1) University of Brescia, Brescia, Italy (ROR: 02q2d2610)
(2) University Vita-Salute, San Raffaele, Milan, Italy (ROR: 01gmqr298)
(3) Department of Clinical Sciences and Community Health, Unit of Medical Statistics, Biometry and Bioinformatics "Giulio A. Maccacaro", Faculty of Medicine and Surgery, University of Milan, Milan, Italy (ROR: 00wjc7c48)
(4) State Industrial Technical Institute (ITIS) Benedetto Castelli, Brescia, Italy
(5) Section of Medical Statistics and Biometry, Department of Molecular and Translational Medicine, Faculty of Medicine and Surgery, University of Brescia, Brescia, Italy (ROR: 02q2d2610)

(*) retired
(#) guest

CORRESPONDING AUTHOR: Bruno Mario Cesana, Department of Clinical Sciences and Community Health, Unit of Medical Statistics, Biometry and Bioinformatics "Giulio A. Maccacaro", Faculty of Medicine and Surgery, University of Milan, Italy. Via Giovanni Celoria 22, 20133 Milan, Italy, Phone: #39-02503 20854 / 02.503.20855.
Email: brnmrcesana@gmail.com / bruno.cesana@guest.unimi.it

SUMMARY

Introduction: Cohen's kappa is the most widely used concordance index for qualitative variables in the biomedical literature, despite its pitfalls and paradoxes associated with low kappa values, even when a relevant proportion of agreement is observed.
However, the approaches proposed for calculating sample sizes for the precision of the Cohen's kappa confidence interval and for comparing two kappa values are not entirely satisfactory.
We have reviewed both approaches proposed in the statistical literature and extended the relevant methodology with the aim of providing more suitable methods.
Methods: We have highlighted the limitations of calculating sample sizes for agreement studies based on the precision of Cohen's kappa statistics and shown how to avoid them while achieving a satisfactory probability of obtaining sample confidence interval widths below the required precision level. Furthermore, we have outlined the main relevant proposals in the literature, such as those of Flack et al. [3], Donner et al. [7-10] Altaye and Donner [11], Altaye et al. [12], and Donner and Rotondi [9], for calculating sample sizes for comparing two kappa values. Next, we have proposed a partial extension of the common correlation model (PCCM) for the 2x2 contingency table to cxc square contingency tables with a common correlation coefficient of the cells on the main diagonal, considered pairwise. Finally, we have devised a full generalization with a common correlation coefficient model (FCCM) for all cells of the square contingency table. As a result, we obtained sample size calculation formulas written in the open source R language to compare two kappa values in the case of two raters (programs are available upon request).
Results: Our PCCM approach produces very similar maximum and minimum values of kappa variance, resulting in only slightly different sample sizes. In contrast, FCCM produces a unique contingency table and, consequently, a unique value of kappa variance under the null and alternative hypotheses, resulting in a unique sample size. More importantly, this latter sample size is within or equal to the sample sizes calculated using the maximum and minimum values of kappa variance according to PCCM. In the case of 2x2 contingency tables, all of our proposed methods for calculating sample sizes (SS-A&C-max, SS-A&C-

DOI: 10.54103/2282-0930/29771
Accepted: 2nd April 2026

min, and SS-A&C-full) produced equal sample sizes, which are also equal to those calculated according to Flack et al. [3]; however, the sample sizes according to Donner et al. [7-10,11-12] are generally larger. For 3x3 contingency tables, the sample sizes calculated according to our proposed models are smaller or equal to those calculated according to Flack et al. [3] and, almost always, smaller than those of Donner et al. [7-10,11-12]. Superimposable results were obtained for 4x4 tables.
Discussion / conclusions: The sample sizes of our proposed models can be generally recommended since they are smaller than or at most equal to those obtained with the approach of Flack et al. [3]. Furthermore, our PCCM procedure provides very similar maximum and minimum sample size values and, in addition, our FCCM procedure gives only one sample size value. Furthermore, the sample sizes proposed by Donner et al. [7-10,11-12] as an extension of their first model with the primary objective of "consensus or not" can be considered if they are lower and, finally, if there are more than two raters. Furthermore, the fact that there is a unique kappa variance value with FCCM, which leads to sample sizes always between those calculated from the maximum and minimum kappa variance values with PCCM, allows us to argue that this model does indeed refer to the population probability table and, therefore, should be highly recommended.
We also provided guidance on how to calculate sample sizes for comparing kappa values for two or more raters that are useful for practitioners.

Keywords: Raters' agreement, Cohen's kappa statistic, Sample sizes calculation

The paper is organized as follows.
The paragraphs: **"***AS. General considerations on sample size calculation", "BS. Effect size", "CS. Power calculation 'a priori' or 'a posteriori'", "DS. Models for Exploring Agreement", "ES. Historical digression on the agreement indices", "FS. Cohen's kappa values: their meaning.", and "GS. Further theoretical details about Cohen's kappa"* useful for having a general overview of the statistical aspects of the sample size calculation and of the agreement indices proposed in the qualitative variable settings have been confined to the supplementary material.
The **Introduction** covers the following topics: *"I.1.Cohen's kappa model", "I.1.2.Cohen's kappa pitfalls" (with details in the following paragraphs of the supplementary material: "IS.1.2.1 Cohen's kappa: Raters bias effect", "IS.1.2.2 Cohen's kappa: Prevalence effect", "IS.1.2.3 Cohen's kappa: Number of categories and marginal uniformity"), "I.1.3.Weighted Kappa"* (with further insights in the supplementary material paragraph: *"IS.1.3 Examples of agreement and disagreement weights for some contingency tables."), "I.1.4.Variance of Kappa and of Weighted Kappa" (with much more details in the paragraphs "IS.1.4.Variance of Kappa and of Weighted Kappa", "IS.1.5 Filling the cells of contingency tables more than 2x2."* in the supplementary material*), "I.2. Sample Size Calculation:"* with its two sections: *"I.2.1 Sample size calculation for the precision of the Cohen's kappa Confidence Interval."* (with details and examples in the supplementary material paragraphs: *"IS.2 Sample Sizes. "IS.2.1 Sample size calculation for the precision of Confidence Interval (CI)"), "I.2.2 Sample size calculation for the power of the statistical test on two kappa".*
Details and insights are in the paragraphs *"IS.2.2 Sample size calculation for the power of the statistical test.", "IS.2.2.1 Sample Size Calculation according to Flack et al. [3]", "IS.2.2.2 .Sample Size Calculation according to Donner et al. [7-10] Altaye et al. [11,12]"* of the supplementary material.
Then, in the section **"Aims of the paper"** there are: (i) the presentation of the mathematical approach of Linear Programming (LP) in order to obtain a well-defined value of the kappa variance, (ii) the extensions of the "common correlation model" and its sample sizes calculation, (iii) the comparison of the sample sizes calculation approaches currently present in the pertinent literature with those obtained from our proposals, (iv) the assessment of the conservativeness of the sample sizes obtained from Flack et al. [3], and, finally, (v) the proposal of giving some suggestions for the sample sizes calculation. It has to be immediately said that the term "conservative" (as opposed to "liberal") is not used in its standard meaning for a statistical test that reject the null hypothesis below the nominal significance value; in fact, in this case, it has been used to define a sample size calculation that is "conservative" since it has been obtained a value that allow to have a power value more than the nominal, leading to a rejection of the null hypothesis probability greater than required.
The introduction ends with the paragraph "Aims of the paper".
The section **"Methods"** considers the sample sizes calculation. Details of the kappa variance calculation are shown in the paragraphs *"MS.1.Variance of Kappa and of Weighted Kappa in contingency tables." with the subsections: "MS.1.1.Variance of kappa in 3x3 contingency tables under the uniform marginal probabilities", "MS.1.2.Variance of kappa in cxc uniform contingency tables.", and "MS.1.3.Variance of kappa in non-uniform cxc contingency tables, under the Partial Common Correlation Model (PCCM) and the Full Common Correlation Model (FCCM)"*. Then, are introduced the Partial Common Correlation Model (PCCM), and the Full Common Correlation Model (FCCM) which form the theoretical basis for our sample sizes calculation. Their theoretical derivations are in the following sections

of the supplementary material: *"MS.2 Correlation Model", "MS.2.1.Common Correlation Model: 2x2 contingency tables.", "MS.2.2 Correlation Model (CM): cxc contingency tables.", "MS.2.2.1 Example of collapsing a 4x4 contingency table." "MS.2.3.Common Correlation Model (CCM) and its extensions.", "MS.2.3.1.Partial Common Correlation Model (PCCM). Partial determination of the correlation structure by assuming a common correlation of the diagonal cells of the contingency table.", "MS.2.3.2.Full Common Correlation Model (FCCM). Full determination of the correlation structure assuming appropriate correlation of all cells in the contingency table.", and "MS.2.3.3. Identifiability of PCCM and FCCM Models".*
The section **"Results"** shows *"R.1.Preliminary considerations: Sample Size Calculation" , "R.2.1 Samples sizes for uniform contingency tables". "R.2.1.Sample sizes: symmetric 2x2 contingency tables.", "R.2.2.Sample sizes: 2x2 contingency table", "R.2.3.Sample sizes: 3x3 contingency table",* and *"R.2.4. Sample sizes: contingency table 4x4".*
The values of the Sample Sizes calculation are shown in their pertinent sample sizes tables.
In the supplementary material we have confined the details of the comparisons between the sample sizes calculated for the above reported contingency tables according to Flack et al. [3] (SS-Flack) and SS-Donner (Altaye-Donner [28,29,30]), between SS-Flack and our proposals (SS-A&C-PCCM-max, SS-A&C-PCCM-min and SS-A&C-FCCM), between SS-Donner and our suggestions (SS-A&C-PCCM-max, SS-A&C-PCCM-min, SS-A&C-FCCM) and, finally, among our sample size calculations suggestions (SS-A&C-PCCM-max, SS-A&C-PCCM-min and SS-A&C-FCCM-min). The pertinent paragraphs are in the section "*RS. Comparisons between the sample sizes*" considered for 2x2,3x3, and 4x4 contingency tables. Furthermore, there is the paragraph:*"RS.2.5. Influence of the non-uniformity of the row and column marginals and of shifting from the unweighted to the weighted kappa on the required sample size under the same null and alternative hypothesis".*
In addition, in the supplementary material section "*RS.2.6 Further sample sizes tables"* there are shown more sample sizes tables calculated also for the unweighted and weighted kappa with linear and quadratic weights.
The section of the paper *"R.2.5 Conclusive considerations"* gives some general results considering particularly the sample size obtained by our models PCCM (max/min) and FCCM". Then there are the paragraphs: *"S.3 Simulation Study: results for 3x3 and 4x4 contingency tables (Flack model)."* and *"S.3.1 Simulation study under PCCM and FCCM models."* with their details shown in the supplementary material paragraphs: *"SS.3. Simulation Study: results from 3x3 and 4x4 contingency tables (Flack model)" and "SS.3.1 Simulation study under PCCM and FCCM models".*
Furthermore, in the supplementary material there are the following Appendices: " **AS.** Linear Programming (a very elegant and interesting approach from the statistical methodological point of view)", "**Appendix BS** Dirichlet Multinomial Distribution", "**Appendix CS.** Sample Sizes shown in the von Eye and Mun's book [31]", "**Appendix DS.** Unique R function for the sample size calculation for testing two kappa with any number of categories and rater with GOF approach", "**Appendix ES.** R function for the sample size calculation…", and "**Appendix FS.** Examples of sample sizes calculation".
Finally, the paper ends with the **Discussion** with some operative conclusions.

INTRODUCTION

A statistical method very frequently used for assessing the agreement of raters on qualitative variables is the Cohen's kappa that is a landmark in the developments of rater agreement theory. However, agreement studies on the Cohen's kappa statistics are unsatisfactorily powered according to different methods leading to different sample sizes.

As it is possible to see on the web site [1] of PASS® (Power Analysis & Sample Size) [2] one of the most famous and valid commercial software available, the sample sizes are needed for the "Kappa Test for Agreement Between Two Raters" and for the "Confidence Intervals for Kappa". The first case is based on the statistical power while the second on the precision of the estimate given usually by the width of the 95% Confidence Interval (CI). However, both approaches require calculating the kappa variance with the immediate complication that it is different for a kappa equal to zero and a kappa non zero as might be the case when calculation the sample size for of the 95%CI precision and, very frequently, also when testing two kappa values under the null and the alternative hypotheses.

Furthermore, even if we can consider a kappa equal to zero for a null hypothesis (H_0) of a statistical test, we must immediately say that the value of kappa=0 is absolutely not interesting since the rationale is to have a base value of kappa that expresses at least the minimum acceptable agreement and, consequently, let's say at least equal to 0.4.

Naturally, from the standard deviation (SD equals to square root of the variance), it is straightforward to calculate the standard error needed to calculate the sample size by dividing the standard deviation by the square root of the sample size.

In fact, PASS® 13/16 [2] calculates the kappa variance for the statistical test according to the standard error maximization method of Flack et al. [3]; differently, for the sample size related to the precision of the confidence interval, the standard error of the estimated κ is provided by Fleiss et al. [4] or according to the relatively simple formula involving only the proportions of observed (Po) and expected-by-chance (Pe) agreement proposed by Cohen [5].

It is important to emphasize that the formula of Fleiss et al. [4] is considered more accurate and based on a better theoretical basis than the original formula of Cohen [5]. However, the formula of Fleiss

et al. [4] requires knowledge of the agreement matrix (contingency table), which is usually not available during the planning phase of a study; one is therefore often led to prefer Cohen's formula [5] also for its simplicity.

Furthermore, the variance of kappa depends on the number of categories leading to square contingency tables from 2x2 onwards, the number of raters, and the actual values of the inner cells. This topic will be covered in a dedicated section *"I.1.4S. Variance of kappa and weighted kappa"* in the supplementary material.

We would like to consider the article by Bujang and Baharum [6] published in this journal a few years ago as a useful starting point, which requires some mandatory clarifications. First, it should be noted that the "minimum sample size for testing Cohen's kappa" indicated by Bujang and Baharum [6] is obtained with the standard error maximization method of Flack et al. [3] in the case of uniform contingency tables with equal row and column marginals as an obvious consequence of the uniformity of the contingency tables.

Indeed, as shown in Figure 1 of the paper by Bujang and Baharum [6], the sample sizes increase in the case of symmetric contingency tables with only the row and column marginals being equal and further increase in the case of asymmetric contingency tables with different row and column marginals. However, the last part of this Figure 1 is not reproducible in PASS®13/16 [2] (the software used by Bujang and Baharum [6] to calculate the sample sizes) since the row and column marginals are not equal as it is requested by PASS®16/13 [2]. In particular, the sample sizes shown in Figure 1 of 35 and 48 [6] can only be obtained with row and column marginals of 0.10, 0.10, 0.10.

The sample size tables shown in the Bujang and Baharum [6] paper are also not reproduced in the case of the 3x3 contingency table shown in Table 1 (page e12267-4), and are meaningless for all cases of k=0 under the null hypothesis and in the cases of very large number of categories (Tables 2, 3 and 4) with, in addition, prevalence values that seem very unusual in biomedical settings. Finally, the Bujang and Baharum [6] paper does not introduce any new methodology on sample size calculation for testing two kappa values.

I.1.Cohen's kappa model

Consider the following 3x3 square contingency table (cxc or rxr) with the judgments of rater A (in the columns) and rater B (in the rows) as to whether or not they belong to three categories simply coded as "1", "2", and "3".

	Rater A			Total
Rater B	"j=1"	"j=2"	"j=3"	
"i=1"	π_{11} (p_{ij})	π_{12}	π_{13}	$\pi_{1\bullet}$
"i=2"	π_{21}	π_{22}	π_{23}	$\pi_{2\bullet}$
"i=3"	π_{31}	π_{32}	π_{33}	$\pi_{3\bullet}$
Total	$\pi_{\bullet 1}$	$\pi_{\bullet 2}$	$\pi_{\bullet 3}$	$\pi_{\bullet\bullet}=1$

The cells report the actual probabilities of the two raters' joint judgment with the subscript "i" for the rows and the subscript "j" for the columns; in the marginal row and column, the "dot" replaces the corresponding subscript ("j" for the rows and "i" for the columns) on which the summation was performed. Naturally, the Latin letter "p" replaces the Greek letter (π) in the case of observed proportions, as shown, for example, in parentheses in cell (1,1). Row and column numbering starts from the top left. Furthermore, contingency tables with equal row and column marginals (marginal/total of the first row equals the marginal/total of the first column, etc.) are defined as "symmetric", and a symmetric table with the same marginal/total of the rows (columns) is defined as a "uniform table"; consequently, a uniform table is by definition also symmetric.

The true proportions (π_{ii}) on the diagonal of this square contingency table represent the proportion of subjects in each category for which the two raters agree; thus, the overall proportion of agreement is given by:

$\pi_o = \sum_i^c \pi_{ii}$, being the observed agreement given by: $p_o = \sum_i^c p_{ii}$.

This is the very first simple index of agreement, but Cohen[5] criticized its use because it does not take into account the agreement by chance (π_e) "resulting from a random choice of the categories by the raters" that, under the assumption of rater independence, is given by the sum of the products of the corresponding rows ($\pi_{i\bullet}$) and columns ($\pi_{\bullet i}$) marginal probabilities:

$\pi_e = \sum_i^c \pi_{i\bullet}\pi_{\bullet i}$ and $p_e = \sum_i^c p_{i\bullet}p_{\bullet i}$ for the theoretical and observed agreement by chance, respectively.

Thus, Cohen [5] proposed an agreement measure "corrected for chance" or, otherwise defined, "the degree of raters' agreement in excess of chance", given by:

$$k = \frac{\pi_0 - \pi_e}{1 - \pi_e} \qquad (1)$$

The kappa estimate from the observed data is:

$$k = \frac{p_0 - p_e}{1 - p_e}$$

A kappa value of 0 indicates that the observed agreement is the same as that expected by chance, and the minimum value of kappa falls between -1.0 and 0.0.

The maximum value of kappa is 1, which results in

perfect agreement; obviously, in a 2x2 contingency table, this case occurs only when $\pi_{12}=\pi_{21}=0$ ($p_{12}=p_{21}=0$).

However, for an asymmetric squared contingency table, the maximum value of kappa may not be equal to 1. Indeed, the maximum proportion of the observable agreement is equal to the sum of the minimum values between the corresponding column and row marginals; namely:

$$\pi_{MAX} = \sum_{i=1}^{c} min(\pi_{\bullet i}, \pi_{i\bullet}) \qquad (2)$$

Feinstein and Cicchetti [13] proposed calculating Cohen's kappa by using the maximum possible value of observed agreement (p_{max}) instead than 1, at the denominator of the k formula. Thus, it is possible to determine what proportion of the maximum observable agreement is actually achieved by the calculated k value in the context of an actual agreement study. Further insights are in the paragraph *"FS. Cohen's kappa values: their meaning."* in the supplementary material.

The maximum obtainable kappa value depends on the marginal probabilities and it can be calculated as:

$$k_{MAX} = \frac{\pi_{max} - \pi_e}{1 - \pi_e}$$

Of course, the "p" Latin letter replaces the "π" Greek letter in the case of the observed values.

A different formula [10] $k_{MAX} = \frac{\pi_0 - \pi_e}{\pi_{max} - \pi_e}$ gives a value corresponding to the percentage of the observed kappa with respect to the maximum possible kappa value given the row and column marginals. Thus, we think that it should be better indicated as $k\%_{MAX}$.

Furthermore, a kappa value of 0 indicates that the observed agreement is the same as that expected by chance, and the minimum value of kappa is between -1.0 and 0.0. Further insights are available in the paragraph *"I.1.2.3S. Cohen's kappa: Number of categories and marginal uniformity"* in the supplementary material.

Another relevant point is that two (or more) raters are expected to be unbiased or that the two (or more) raters classify the same proportion of items (or subjects) into the different categories of the qualitative (nominal or ordinal) variable under consideration, thus obtaining equal row and column marginals.

However, it should be noted that agreement between two (or more) raters does not have to be "perfect", but it is admitted that the agreement may be less than "perfect" as long as the difference from the perfect agreement (k=1 or $k=k_{MAX}$) can be considered not (clinically) relevant. From this perspective it is possible to release the assumption of equal marginals and accept reasonable small differences. However, this fact makes the sample sizes calculation more complicated and likely only feasible by simulating various scenarios (first row marginal equal to the first column marginal minus or plus an irrelevant amount, for example, with the remaining marginals adjusted accordingly).

Furthermore, in the scientific context, it is not possible to demonstrate a null difference from H_0: k=1, as the expression of perfect agreement. Therefore, similarly to non-inferiority clinical trials, the null and alternative hypotheses reverse their roles becoming respectively a hypothesis of difference (the maximum threshold beyond which there is no agreement) and a hypothesis of no difference. Therefore, the maximum difference allowed to consider two drugs "clinically equivalent" becomes the maximum difference allowed to consider two (or more) raters pragmatically "in agreement" or "in almost perfect agreement."

Further details are available in the section *"GS. Further theoretical details on Cohen's kappa"* in the supplementary material.

I.1.2.Cohen's kappa pitfalls

Many factors influence the kappa value, such as the rater bias, category prevalence, and the number of the categories. It should must be stated immediately that an important assumption underlying the use of kappa is that errors associated with raters are independent, as emphasized by Thompson and Walter [14], Shoukri [15], and Brennan and Prediger [16], among others.

The influence of the Raters Bias *(IS.1.2.1 Cohen's kappa: Raters Bias Effect),* the prevalence of the categories *(IS.1.2.2 Cohen's kappa: Prevalence effect),* the number of categories and the uniformity of the row and column marginals *(IS.1.2.3 Cohen's kappa: Number of categories and marginal uniformity)* are considered in the previously cited paragraphs of the supplementary material.

Several authors have suggested alternative indexes and adjustments to overcome these limitations; see, for example, Gwet [17,18] and Klein [19]. In addition, Falcaro and Newson [20] focussed on some nonparametric measures proposed by Svensson [21,22] for paired ordinal variables.

I.1.3.Weighted Kappa

When the categories are naturally ordered, it is immediate to diversify the level of the agreement/ disagreement, since a disagreement between two adjacent categories in the contingency table is less important than a disagreement between categories that are more distant from each other.

Cohen [5,23] introduced the weighted kappa for "nominal scale agreement" with an exemplification based on disagreement weights, assigning a weight equal to 0 to the cells on the main diagonal and values of 1 and of 3 to progressively more distant categories. Furthermore, Cohen [5,23] introduced also the weighted kappa with agreement weights equal to 1 for

cells on the main diagonal and lower weights as cells move symmetrically away from the main diagonal.

It should be noted that the main criticism of weighted kappa is that the weights can be chosen arbitrarily. In conclusion, the weights can be "agreement weights" given by $0 \leq w_{ij} \leq 1$ with $w_{ii} = w_{jj} = 1$ or also "disagreement weights" given by $0 \leq v_{ij} \leq 1$ with $v_{ii} = v_{jj} = 0$.

More details are provided in section *"I.1.3S Examples of agreement and disagreement weights for some contingency tables."* in the supplementary material.

Note that sample size calculations are typically performed for an unweighted kappa. Otherwise, the weighted kappa with its pertinent variance should be considered.

I.1.4.Variance of Kappa and of Weighted Kappa

The problematic aspects of this topic have been very well synthetized in the first sentence of the Fleiss et al.'s paper [24]: "Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third." The above text has been also reported by Kraemer et al. [25] by adding: "This wry comment by Fleiss et al. in 1979 continues to characterize the situation with regard to the kappas coefficients up to the year 2001, including not only derivation of correct standard errors, but also the formulation, interpretation and application of kappas."

Cohen [5,23, equation 7], derived the variance of the kappa (the weighted kappa will be introduced later) by considering p_e "as a constant" and, consequently, a fixed quantity and p_o as "if it were the population value"; furthermore, the above assumption was considered adequate "ordinarily p_e will not vary greatly relative to k, particularly with large n (i.e., ≥300)". Consequently, the variance of p_o is $p_o(1-p_o)/n$ and the kappa variance becomes the variance of p_o divided by the square of the constant value $(1-p_e)$ in the denominator. The square root of the variance is the standard deviation which in turn divided by the sample size (N) is the kappa standard error given by:

$$\sigma_k = \sqrt{\frac{p_o(1-p_o)}{N(1-p_e)^2}}$$

It should to be noted that the use of the Greek letter "σ" is justified since p_o has been defined as "the population value" without using the Greek letter "π" instead of the Latin letter "p". Given this derivation, as N increases, the k sampling distribution can be assumed to be approximated by the Gaussian distribution resulting in the calculation of the 95% and 99% confidence intervals using the 1.96th and 2.58th quantiles of the standard Gaussian distribution, respectively. Therefore, testing two independent kappa coefficients or a kappa against an expected value can be done by an approximate Z-test. However, Fleiss et al. [4] reported that the approximate values of the variance of Cohen's kappa [23] are overestimated and incorrect. Indeed, Cohen's formula [23] can only be used as a preliminary and simple approximation.

It is important to note that the calculation of the kappa variance must take into account the complete structure of the contingency table and therefore all the cells with their variances and covariances which are different in the case of a value of k equal or different from 0. This fact is clearly shown in the paper by Fleiss et al. [4] in which the formulas of the weighted and unweighted kappa variance are shown in the case of $k=0$ and $k \neq 0$.

Further insights are available in sections *"I.1.4S. Variance of Kappa and of Weighted Kappa."* and *"I.1.5S.Filling the cells of the contingency tables more than 2x2."* in the supplementary material.

I.2.Sample Size Calculation

I.2.1 Sample size calculation for the precision of the Cohen's kappa Confidence Interval (CI)

Donner[26], Donner and Rotondi [27] and Rotondi and Donner [28] showed sample size requirements for interobserver agreement studies with a binary outcome in terms of the number of subjects (N) and raters (n_r) that would allow the expected lower bound of a 95% confidence limit for Cohen's kappa to be equal or greater than a required threshold. Of course, the required threshold must be chosen to achieve an adequate level of agreement. Further details are provided in section *"I.2.1S Sample Sizes for the precision of the Confidence Interval (CI)"* in the supplementary material.

Apart from Donner and Rotondi's approach [26-28], usually the precision of the estimate is represented by the width of the 95% confidence interval (CI), calculated by the difference between the upper CI limit $(k+z_{1-\alpha}SD(k)/\sqrt{N})$ and the lower CI limit $(k-z_{1-\alpha}SD(k)/\sqrt{N})$ which is simply equal to $2z_{1-\alpha}SD(k)/\sqrt{N}$. Naturally, SD(k) represents the Standard Deviation of Cohen's kappa and $z_{1-\alpha}$ is the pertinent quantile of the standard Gaussian distribution (equal to 1.96 for a two-sided 95%CI).

This equation equated to the Required Width (RW, say) allows us to obtain after a few simple algebraic steps: $N=[(2z_{1-\alpha}SD(k)/RW]^2$.

Of course, the Required Width must be chosen appropriately by the researchers. Obviously, considering the half-width of the CI (hRW) as the relevant expression of the precision, the formula above becomes: $N=[(z_{1-\alpha}SD(k)/hRW]^2$.

The problem now is to calculate the standard deviation (SD) of kappa. If the kappa and marginal values of the row and column are known (reasonably assumed equal when planning the study, but not

necessarily with equal prevalence of the categories), it is possible to calculate p_e (proportion of agreement due to chance given by the sum of the squared marginal values) and consequently p_O (proportion of expected agreement given by: $p_O = k(1-p_e)+p_e$. Furthermore, p_e can be obtained by the kappa and p_O values by: $(p_o-k) / (1–k)$. Subsequently, it is possible to calculate the Cohen's SD(k), incorrect but very simple approximate, given by: $\sqrt{[p_O(1-p_O) / (1–p_e)^2]}$ and also reported here for ease of reference.

Otherwise, calculating SD(k) according to Fleiss [4] (e.g., among others), requires the full structure of the contingency table with cell proportions in addition to the marginal ones. For this formula, readers are referred to the paragraph *"Sample Size for testing Cohen's kappa"* in the supplementary material.

However, a very important limitation of this approach to calculating sample size for the precision of the confidence interval (CI) is that the required precision (usually the required width) can only be achieved in about 50% of cases resulting in a "Confidence Interval power" of about 0.50. To overcome this drawback, it is necessary to implement a statistical methodology described in the section *"AS. General considerations on sample size calculation"* section in the supplementary material" based on iteratively increasing the calculated sample sizes until the CI power function reaches a satisfactory value (≥0.80).

Furthermore, since a Confidence Interval smaller than the Required Width that does not include the parameter is considered not particularly relevant, the iterative process can be performed until the sample size allows an adequate probability of reaching the joint event of a CI interval smaller than the required width (CI power) and including the parameter (coverage). Examples of sample sizes calculation according to this approach and sample sizes increments are provided in section *"I.2.1S Sample Sizes for the precision of the Confidence Interval (CI)"* in the supplementary material.

I.2.2 Sample size calculation for the power of the statistical test on two kappa

We begin with the very pragmatic presentation of sample size calculation provided by Choudhary and Nagaraja [29], without considering the problems of different variance values, most likely because in the considered case of two raters and two categories the two variances are generally similar.

They used a different formulation of the formula by Fleiss et al. [4] for the kappa variance of a 2x2 contingency table given by:

$$\hat{Var}(k) = \frac{1}{N(1-p_e)^2}\left\{\sum_{i=1}^{2} p_{ii}\left[1-(p_{i\cdot}+p_{\cdot i})(1-k)\right]^2 + (1-k)^2\left[p_{12}(p_{1\cdot}+p_{\cdot 2})^2 + p_{21}(p_{2\cdot}+p_{\cdot 1})^2\right] - \left[k-p_e(1-k)\right]^2\right\}$$

This variance can be used, assuming a Normal distribution in the case of large samples, to calculate approximate confidence intervals and to test the agreement hypotheses (H_0: $k \le k_0$ vs. H_A: $k > k_0$, for example).

Furthermore, the sample size calculation formula (12.28 in the paragraph "12.4.7 Sample Size Calculations" in the book by Choudhary and Nagaraja [29]) is simply given by:

$$N = \hat{\sigma}^2 \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(k_0 - k_A\right)^2}$$

Where is an estimate of the unit variance kappa and $z_{1-\alpha}$ and $z_{1-\beta}$ are the quantiles (1-α) and (1-β) of the standard normal distribution. Interesting is the suggestion by Choudhary and Nagaraja [29] to take into account that the sample size increases as we move away from the situation in which the row (column) marginals are equal to 0.5; in particular, they suggested using preliminary estimates of the prevalence of the classification classes and, then, to insert the increased variance obtained by substituting k_0 or k_A in the sample size formula. Indeed, for row (column) marginals of 0.6 and 0.4, k_0=0.4 and k_A=0.7, α=0.025 and power=0.80, the sample sizes calculated using the variances below k_0 or below k_A are 76 and 46, respectively, instead of 66 and 67 calculated according to Cantor [30] and PASS®13/16 [2], respectively.

It is also worth noting the formula for calculating the sample sizes proposed by Cantor [30] in which the variances under the null and alternative hypothesis are separated along with their pertinent quantiles:

$$N_{Cantor} = \frac{\left(z_{1-\alpha}\sqrt{\hat{\sigma}_0} + z_{1-\beta}\sqrt{\hat{\sigma}_A}\right)^2}{\left(k_0 - k_A\right)^2} \quad (3)$$

Where $\hat{\sigma}_0^2$ is the estimate of the unity kappa variance under H_0 and $\hat{\sigma}_A^2$ is the estimate of the unit kappa variance under H_A. N is, as usual, rounded to the upper integer.

It is important to underline that this formulation, which considers the variances under the null and alternative hypotheses, is more common in sample size calculation and, moreover, allows obtaining only one value of the sample size which is, obviously, intermediate between the sample sizes calculated according to Choudhary and Nagaraja [29], which are too conservative or too liberal for the variance greater and lower, respectively. It is obvious that the suggestion of Choudhary and Nagaraja [29] is not recommended.

Further details are given in the sections *"Sample Size Calculation according to Flack et al. [3]"*, and *"IS.2.2.2 Sample Size Calculation according to Donner et al. [7-10], Altaye et al. [11,12] and Rotondi [31]"* of the supplementary material; further theoretical details are reported in the Appendix BS. Dirichlet Multinomial Distribution in the supplementary material. Finally, details on the Sample Size Calculation according to

Von Eye and Mun [32] are considered in the Appendix CS of the supplementary material.

AIMS OF THE PAPER

First, we will provide a tutorial for calculating sample size based on the precision of the Cohen's kappa confidence interval and the statistical test between two kappa values under the null and alternative hypotheses. Next, we will address some inconsistencies in the literature and, after introducing the Cohen's kappa model, provide a general overview of sample size calculation methods for testing two Cohen's kappa values and propose an original unified approach. Theory and examples for two raters and two or more categories will be presented. Additionally, programs in the open-source R language will be presented to simplify sample size calculation, even for non-professional biostatisticians.

Furthermore, a relevant aim of our research is to provide the mathematical theory supporting Flack et al.'s proposal to obtain the maximum value of the kappa variance [3] using Linear Programming (LP) to directly obtain the contingency table with the maximum value of the kappa variance. However, it should be remembered that using LP it is also possible to obtain different values of the kappa variance, such as the minimum or any predetermined value. The possibility of calculating the maximum and minimum values of the kappa variance allows obtaining intermediate values of sample size that could ensure the effective feasibility of an agreement study. This point is extensively considered in "*Appendix AS. Linear Programming*" in the supplementary material with an example of sample size calculation, to which the reader is referred.

Another important aim is to propose a generalization of the "common correlation model for dichotomous variables" to multinomial variables, which will provide a new theoretical starting point for the sample size calculation procedure. Specifically, with this model, we imposed two constraints: the first, less restrictive, is that only cells on the main diagonal have a common correlation coefficient ρ, giving rise to the "partial common correlation model (PCCM)"; the second, more restrictive, is that all cells in the contingency table not on the main diagonal, have a correlation coefficient $\rho_{ij} = \rho\gamma_{ij}$, giving rise to the "full common correlation model (FCCM)". The important result is that in both models we have: $\rho = k_w$. Naturally, these two models provide two different sample sizes.

Another aim is to compare the sample sizes obtained from the aforementioned PCCM and FCCM models with the sample sizes obtained by Flack et al. [3] or PASS®13/16 [2], by Donner and Rotondi [7-10], by Altaye et al. [11,12] according to their "Goodness-of-fit" (GOF) model and by, including, for completeness, the values calculated with the minimum value of the kappa variance.

An additional aim is to confirm the conservative characteristic of the sample sizes calculated according to Flack et al. [3] or PASS®13/16 [2], by calculating the empirical power by means of a simulation study with sample sizes calculated according to Flack et al. [3]. In addition, we also calculated the 95% Confidence Interval coverage of the sample kappa CI to assess whether the conservative sample sizes guaranteed the nominal coverage.

A final pragmatically relevant aim is to provide some suggestions for obtaining sample sizes suitable for the actual feasibility of an agreement study.

METHODS

The sample size calculation procedure requires knowledge of the variance of k_0 which is different from zero, since the null hypothesis H_0: $k_0 = 0$ is unacceptable, and of k_A under the alternative hypothesis (H_A).

It is important to note that the kappa variance (Var(k)), to be calculated according to Fleiss et al. [4], is known and unique only if all joint probabilities π_{ij} of the contingency table are fixed.

It is known that for 2x2 contingency tables, Cohen's kappa "unity variance" (conveniently defined this way since it is obtained with a sample of only one unit – N=1 - to remove the influence of the sample size on its value) is determined uniquely (see in the supplementary material paragraphs *"MS.1.Variance of Kappa and of Weighted Kappa in contingency tables."* with the subsections: *"MS.1.1.Variance of kappa in 3x3 contingency tables under the uniform marginal probabilities"*, *"MS.1.2.Variance of kappa in cxc uniform contingency tables."*, and *"MS.1.3.Variance of kappa in non-uniform cxc contingency tables, under the Partial Common Correlation Model (PCCM) and the Full Common Correlation Model (FCCM)."*).

Thus, it is possible to calculate the variance values under the null (H_0) and alternative (H_A) hypotheses and insert them into the sample size calculation formula. Otherwise, the kappa variance for square tables cxc(c>2), given the row and column marginals, and the expected agreement proportion or the kappa value are not uniquely determined; thus, it is not possible to calculate a unique Var(k) and, consequently, a well-defined sample size value.

Flack et al. [3] calculated the sample size after determining the cell probabilities so that the kappa variance, proposed by Fleiss et al. [4], is maximized under the null hypothesis (H_0) and the alternative hypothesis (H_A), although the procedure for obtaining such a configuration of the cell probabilities has not been clearly described. Obviously, this approach leads to a conservative sample size calculation.

We improved Flack et al.'s proposal [3] by introducing a mathematical approach based on Linear Programming (LP, also called linear optimization) to obtain the maximum, minimum or any other value of

the variance (see Appendix AS. Linear Programming in the supplementary material). More generally, it is possible to determine the cell probabilities for obtaining a particular value of the kappa variance such as its maximum or its minimum or any other particular value.

Furthermore, to counteract the fact that the sample sizes calculated according to Flack et al. [3] may be too conservative, we extended the "common correlation model" for binary variables to two new models: the Partial Common Correlation Model (PCCM) and the Full Common Correlation Model (FCCM) and calculated their respective sample sizes.

From a theoretical point of view, the PCCM model is a special case of the more general model of Flack et al. [3] since, in addition to the conditions of Flack et al. [3], it adds an additional constraint on the probabilities π_{ij} on the main diagonal. For the PCCM, LP must be used to compute the probabilities of the cells that maximize (or minimize) the variance kappa, both under both H_0 and H_A, and, consequently, the maximum (minimum) sample size has been computed. Finally, the FCCM model is a particular case of the PCCM.

The statistical theory of the Correlation Model and its statistical extensions (PCCM and FCCM) are shown in the supplementary material paragraphs: *"MS.2 Correlation Model", "MS.2.1.Common Correlation Model: 2x2 contingency tables.", "MS.2.2 Correlation Model (CM): cxc contingency tables.", "MS.2.2.1 Example of collapsing a 4x4 contingency table." "MS.2.3.Common Correlation Model (CCM) and its extensions.", "MS.2.3.1.Partial Common Correlation Model (PCCM). Partial determination of the correlation structure by assuming a common correlation of the diagonal cells of the contingency table.", "MS.2.3.2.Full Common Correlation Model (FCCM). Full determination of the correlation structure assuming appropriate correlation of all cells in the contingency table.", and "MS.2.3.3. Identifiability of PCCM and FCCM Models.".*

RESULTS

R.1.Preliminary considerations: Sample Size Calculation"

We calculated the sample sizes according to Flack et al. [3] with the maximum (SS-Flack-max) and minimum value (SS-Flack-min) of the kappa variance, to Donner et al. [7-10] (SS-Donner to Altaye et al. [7,8] and, to our method based on the multinomial "Partial Common Correlation Model" both for a maximum (SS-A&C-PCCM-max), and a minimum (SS-A&C-PCCM-min) value of the kappa variance and, finally, to our method based on the multinomial "Full Common Correlation Model" with the notable property of having only one value of the kappa variance (SS-A&C-FCCM) and, consequently, only one value of the sample size.

We considered four different scenarios of the row (column) marginals: from the uniform case to a noticeable non-uniform pattern. In addition, we considered eighteen null and alternative hypotheses given by three values of k_0=0.4, 0.5, and 0.6 and respective k_A values given by k_0+0.05, k_0+0.10, to 0.90 with steps of 0.10, and, finally, k_A=0.95. Finally, we considered a significance level of α=0.05 (1-sided) and two tailed (not shown), and power=0.80. So, there are seventy-two scenarios of sample size calculations. Thus, we produced sample sizes tables for comparing two kappa values; three tables for unweighted kappa in contingency tables 2x2 (Table SS1), 3x3 (Table SS2), and 4x4 (Table SS3) are shown in the paper. Sample sizes tables for weighted kappa are shown in the section *"RS.2.6 Further sample sizes tables"* of the supplementary material.

Tables SS1, SS2, and SS3 show the sample sizes for testing two unweighted kappas for two raters at a significance level of α=0.05(1-sided) and power 0.80 for 2x2, 3x3, and 4x4 contingency tables. Particularly, the tables show the unweighted k_0 and k_A values, the marginals of rows and columns and the four previously reported sample sizes: SS-Flack-max, SS-Flack-min, SS-Donner, SS-A&C-PCCM-max (SS-A&C-PCCM-min when different from SS-A&C-PCCM-max: Table SS3 for 4x4 contingency tables). After having fixed the appropriate null (k_0) and alternative (k_A) hypotheses on the first two columns on the left, you need to select the appropriate pattern of the row (column) marginals on the next two (three or four) columns on the right; then, it is possible to read for α=0.05 (1-sided), the five sample sizes: SS-Flack-max/min, SS-Donner, SS-A&C-PCCM-max, SS-A&C-PCCM-min (in Table SS3). It is therefore possible to choose the sample size best suited to the actual feasibility of the agreement study. SS-A&C-FCCM values are not reported since they are equal to SS-A&C-PCCM-max/min or more or less than a few units for 4x4 contingency tables (differences in italic and between brackets are reported in Table SS3). Furthermore, a free R source code program for their calculation is provided in Appendix DS of the supplementary material.

Flack et al.'s approach [3], extended with our LP approach [33] to calculate the maximum/minimum variance, leads to have the widest interval of values and, obviously, also the widest interval of sample sizes. Otherwise, our PCCM max/min model gives much narrower intervals that, in addition, are within the Flack et al.'s intervals [3].

The FCCM model allows to uniquely determining the cell probabilities, and consequently the kappa variance and the sample size are equal or within the sample size values calculated according to the PCCM model with the minimum and maximum value of the kappa variance and obviously within those calculated according to Flack et al. [3] with the minimum and maximum value of the kappa variance.

Results of the sample size calculation will be shown by comparing (i) SS-Flack-max/min and SS-Donner, (ii) SS-Flack-max/min and SS-A&C-PCCM-max/min, (iii) SS-A&C-PCCM-max and SS-A&C-PCCM-min and, finally, (iv) SS-A&C-PCCM-max/min and SS-A&C-FCCM. The comparison between SS-Flack-max and SS-Flack-min has been reported just for sake of completeness, being the respective values so different. In addition, SS-Flack-min for not uniform contingency tables are not recommended since it is very likely that the empirical power will be much lower than the theoretical power required.

R.2.1 Samples sizes for uniform contingency tables

For the 2x2 uniform contingency tables considered, Flack et al. [3] and our PCCM and FCCM approaches uniquely determine the kappa variance with unique and equal sample sizes: SS-Flack-max/min, SS-A&C-PCCM-max/min and SS-A&C-FCCM.

The same result applies to the 3x3 and 4x4 uniform contingency tables considered.

Furthermore, also for the symmetric 2x2 contingency tables considered, SS-Flack-max/min, SS–A&C-PCCM-max/min and SS–A&C-FCCM are equal.

For 3x3 tables, we have shown that for PCCM the kappa variance is unique and, consequently, there is only one value of the sample size which is, moreover, equal to the sample size obtained with FCCM.

Here, we will give some results of the pairwise sample sizes differences for asymmetric contingency tables in a narrative way leaving the descriptive statistics for the pairwise sample sizes differences in the supplementary material sections: *"RS.1.Contingency tables 2x2."*, *"RS.2. Contingency Tables 3x3."*, and *"RS.2. Contingency Tables 4x4."*.

R.2.2.Sample sizes: symmetric 2x2 contingency tables

Generally, SS-Flack-max (equal to SS-Flack-min, since for 2x2 contingency tables, the variance is unique) are always lower than SS-Donner in case of 0.4 and 0.6 marginals. Otherwise, SS-Donner are lower in the case of highly non-uniform marginals and of a difference k_A-k_0 equal to 0.05 or ≤0.10. The comparisons between SS-A&C-PCCM-max (equal to SS-A&C-PCCM-min, since for 2x2 contingency tables, the variance is unique) and SS-Donner are the same as those between SS-Flack-max/min and SS-Donner.

R.2.3.Sample sizes: symmetric 3x3 contingency tables

SS-Flack-max and SS-Flack-min are equal for the case of uniform contingency tables. Otherwise, SS-Flack-max are greater and even much greater than SS-Flack-min.

Finally, SS-Flack-min are lower and even much lower than the other sample sizes values.

Generally, SS-Flack-max are smaller in 49 (68.1%) cases than SS-Donner and always lower in case of uniform marginals (0.3(3), 0.3(3), 0.3(3)) and moderately asymmetric marginals (0.5;0.25;0.25). Otherwise, SS-Donner are lower in the case of a greater asymmetry and of a difference k_A-k_0 equal to 0.05 or ≤0.10 and sometimes even to 0.20.

SS-Flack-max are greater than SS-A&C-PCCM-max, but sometimes they are equal in case of maximum values of k_A-k_0 the differences.

R.2.4.Sample sizes: symmetric 4x4 contingency tables

SS-Flack-max and SS-Flack-min are equal for the case of uniform contingency tables. Otherwise, SS-Flack-max are greater and even much greater than SS-Flack-min.

Finally, SS-Flack-min are lower and even much lower than the other sample sizes values.

SS-Flack-max are a little lower than SS-Donner in the case of uniformity and in several cases of symmetry with a difference k_A-k_0 >0.30 (generally).

SS-Flack-max are greater than SS-A&C-PCCM-max and and SS-A&C-PCCM-min.

SS-A&C-PCCM-max/min are lower than SS-Donner except in the case of some difference k_A-k_0 equal to 0.05 or ≤0.10. SS-A&C-FCCM are practically always equal to SS-A&C-PCCM-max/min; however, the differences are shown in the Table SS3 by reporting the lower (numbers within brackets, preceded by the minus sign) and the greater differences (numbers within brackets).

Finally, in the section *"RS.2.6 Further sample sizes tables"* of the supplementary material are reported some other sample sizes tables. Particularly, Table SS2S.A: sample sizes for 3x3 contingency tables: α=0.05 (two-tails), power = 0.80 according to Flack and our PCCM with the maximum value of the variance in the case of unweighted kappa and of linear and quadratic weighted kappa.

Table SS3S.A: sample sizes for 4x4 contingency table: α=0.05 (two-tails), power=0.80 according to Flack and our PCCM with the maximum and minimum value of the variance in the case of unweighted kappa; sample sizes according to our FCCM approach and to Donner are also shown.

Finally, Table SS3S.B and Table SS3S.C: show sample sizes for 4x4 contingency table: α=0.05 (two-tails), power = 0.80 according to Flack and our PCCM with the maximum and minimum value of the variance and according to our FCCM approach. In the case of unweighted kappa, sample sizes according to our FCCM approach and to Donner [7-10] and Rotondi [31] are also shown.

The "Influence of the non-uniformity of the row and columns marginals and of shifting from the unweighted to the weighted kappa on the required sample size under the same null and alternative hypothesis" is

shown in the supplementary material (section RS.3.6 with the same title).

R.2.5 Conclusive considerations

As an important consequence of the "common correlation model for the diagonal cells of the contingency tables of multinomial variables", the difference between the maximum and the minimum variance values is very narrow, being equal to zero for the 2x2 and 3x3 contingency tables. Then, in the case of the 4x4 contingency tables, the mean difference is of 1.12 (min, Q_1, median, Q_3 equal to 0 and max=19.0), leading to a maximum difference between the minimum and maximum sample size values ranging from 0 and 19 occurred in the case of k_0=0.4, k_A=0.45, with marginal probabilities of 0.60, 0.20, 0.20. Generally, difference of more than ten occurred when the difference between k_0 and k_A is only of 0.05. As a conclusion, the maximum and minimum sample sizes calculated under this model are practically the same with differences of only a few units and are lower that the sample sizes calculated according to Flack et al. [3].

Furthermore, the sample size calculated under the "full common correlation model" (SS-A&C-FCCM) are always within the SS-A&C-PCCM-max and SS-A&C-PCCM-min, and allows to have a unique contingency table pattern without recurring to the LP procedure. Finally, in 4x4 contingency table, the sample sizes are in the 77.8% equal to SS-A&C-PCCM-max and in the 86.1% equal to SS-A&C-PCCM-min.

S.3 Simulation Study: results for 3x3 and 4x4 contingency tables (Flack model)

Firstly, it has to be stressed that simulation studies on the kappa properties have to be planned with a well-defined variance-covariance matrix (all π_{ij} have to be known) in addition to the row (column) marginals and the observed agreement proportions, given the Cohen's kappa value. It has to be noted that c-2 cells are "free", being not determined by the constraints given by the fixed row (column) marginals and by the kappa value. Of course, the simulation results will depend on how the "free" π_{ij} are determined. We simulated, starting from the multinomial distribution, under the Flack et al.'s [3] condition of a maximum kappa variance value under which the "free" π_{ij} are fixed. So, the calculated sample size will be the maximum among all sample sizes calculated for contingency tables with the same marginal and observed agreement proportions. Consequently, this is the most conservative condition, leading to the rejection of the null hypothesis more than the researchers had hoped for. Naturally, this situation is not consistent with the principle that, in biomedical research, sample sizes must be adequate to try to demonstrate the primary objective of the research with a satisfactory level of probability, but also as parsimonious as possible according to ethical requirements.

Finally, it should be emphasized that the objectives of the simulation study are to evaluate the actual power of testing kappa against an expected value and the effective coverage of the 95% kappa CI with such conservative sample sizes. Details of this simulation study are shown in the section "*SS.3 Simulation Study: results for contingency tables 3x3 and 4x4*" of the supplementary material.

Table SIM.1AS of the supplementary material for the 3x3 contingency table shows the descriptive statistics (mean, SD, 95%CI limits, Median, Minimum, First Quartile -Q1- Third Quartile -Q3– Maximum) of the raw, percent, absolute, and absolute percent bias of the estimates of k_0 and k_A and of the coverage. In the supplementary material are reported the Tables with the detailed results. Table SIM.1BS shows the bias (raw, raw percent, absolute and absolute percent) of the power for the simulations with sample sizes calculated for a power of 0.80 and of 0.90, respectively.

Table SIM.2AS and Table SIM.2BS show the corresponding results of the simulation for the 4x4 table. The absolute mean bias for k_0 and k_A are very limited.

S.3.1 Simulation study under PCCM and FCCM models

Additionally, the results of a 1,000-sample simulation study, performed according to PCCM and FCCM, are shown in Tables SIM.PCCM.AS and SIM.PCCM.BS of the section SS.3.1 of the supplementary material for uniform or symmetric 3x3 and 4x4 contingency tables. In particular: row (column) marginals, k0 and kA values, the sample sizes calculated according to PCCM, and finally, the empirical significance (α=0.05, 2-sided) values and power obtained under the null (H_0) and the alternative (H_A) hypotheses.

Tables SIM.FCCM.AS and SIM.FCCM.BS in the supplementary material show the same results for FCCM.

DISCUSSION

First, it should be noted that calculating the sample size for Cohen's kappa is not a very simple task. Indeed, while k_0 and k_A under the null and alternative hypotheses can be obtained from a careful search of the relevant literature (k_0) and from the difference that can be considered clinically relevant in the particular context of the agreement study (k_A), the next step of determining the prevalence of the rater's judgment categories will likely require a more laborious search of the relevant literature. Therefore, assuming that the row (or column) margins have also been sensibly set, calculating the sample size is still undefined since

the variance of contingency tables greater than 2x2 cannot be defined unambiguously without imposing further constraints.

We have shown that calculating sample sizes for confidence interval (CI) precision as suggested by Donner et al. [9] gives unsatisfactory results unless the power of the confidence interval is taken into account by increasing the initially calculated sample size until a satisfactory probability value (CI power, say at least 0.80) of achieving the required precision is reached. Furthermore, rather than calculating the sample size so that the lower bound of the kappa 95%CI is greater than a required acceptable lower margin of agreement[9], we have shown that a more appropriate approach is to obtain the combined result of obtaining a sample CI width smaller than the required width and a sample CI that includes the parameter (coverage).

Considering the sample sizes calculation for the power of the statistical test, we clarified some statements contained in Bujang and Baharum[6] paper such as the overly generic recommendation to double the sample sizes in the case of non-uniform contingency table. Indeed, sample sizes must be calculated appropriately starting from the most reasonable population contingency table.

Starting from the sample size calculation approach proposed by Flack et al. [3] based on the maximum value of the kappa variance obtained without a well-defined methodology, we showed a mathematical approach (linear programming Appendix AS in the supplementary material for its theoretical aspects and the R code) to be devised to obtain the maximum (minimum or any other) value of the kappa variance.

Furthermore, since the maximum value of the kappa variance leads to a conservative sample size that may be too demanding for the actual feasibility of an agreement study, we have shown some sensible alternatives based on our new PCCM and FCCM models.

In particular, the FCCM allows to obtain a unique probability table and a single value of the kappa variance without resorting to the LP procedure. Therefore, we can assume that we have obtained less biased estimates of the true population variance.

It is important to note that in the case of 2x2 contingency tables, all our proposed methods (SS-A&C-PCCM-max, SS-A&C-PCCM-min, and SS-A&C-FCCM) provided equal sample sizes which are, in addition, equal to those calculated according to Flack et al. [3]. Conversely, in the case of 3x3 contingency tables, all our proposed methods (SS-A&C-PCCM-max, SS-A&C-PCCM-min, and SS-A&C-FCCM) provided the same sample sizes which are also smaller or, at most, equal to those obtained by the Flack et al.'s approach [3]. Therefore, our methods may be recommended as a suitable alternative. The situation of the 4x4 contingency table is more diverse. However, our sample sizes (SS-A&C-PCCM-max, SS-A&C-PCCM-min) could be recommended as smaller or, at most, equal to SS-Flack. More precisely, SS-A&C-FCCM, always between SS-A&C-PCCM-min and SS-A&C-PCCM-max, might be a more sensible choice.

As a final consideration, it has to be stressed that the very similar maximum and minimum values of the kappa variance under the "partial common correlation model" allows avoiding the big differences between the sample sizes calculated with the maximum or minimum kappa variance according to Flack et al. [3]. Furthermore, the unique kappa variance value under FCCM leads to sample sizes always within those calculated from the maximum and minimum values of the kappa variance under PCCM and allows to argue that this latter model has to be strongly recommended since it avoids the need of calculating a maximum kappa variance value by resorting to LP procedure.

Of course, the sample sizes can be calculated with the minimum value of the kappa variance for the null (k_0) and alternative (k_A) hypothesis, leading to the lowest sample size. Perhaps, in this case there are no problems on the actual feasibility of the study but, nonetheless, there are some concerns on the actual possibility of demonstrating a different agreement clinically relevant; consequently, this approach cannot be recommended.

As a further remark, it has to be considered that the row (column) marginals pattern, the number of the categories of the variable, and also the number of raters influence the sample sizes. Since the required sample size increases as the marginal get more even, being the number of the categories fixed by the kind of the considered variable and the number of raters fixed by the kind of the agreement study, researchers are strongly advised to be very careful on the prevalence of the categories. Perhaps, if the knowledge of the distribution of the prevalence is unknown, a sensitivity sample size calculation starting from the uniform marginal pattern with the minimum sample size to an intermediate pattern has to be warmly recommended to choose the most reasonable sample size.

Then, it has to be concluded that the possibility of having some sample sizes calculation procedures can be considered as a very useful tool in order to choose the sample size value more suitable for the actual feasibility of the study in the particular scenario of row (column) marginals and null and alternative hypothesis.

Finally, owing to the relevant increase of the sample sizes required for the weighted kappa, its use cannot be recommended even if a sensible approach

of weighting the agreement (disagreement).

It is very important to note that we considered the sample size calculation for the case of only two raters, which is also the most frequently considered situation in biomedical research. Indeed, also commercial software such as PASS®13/16 [2] and nQuery® [33] make sample size calculation for only two raters.

In the case of more than two raters it has to resort to the kappaSize package in R from Rotondi [31] for doing the pertinent sample size calculations or to our program, written in the open-source R language, which uses only one function instead of the different formulas, written ad hoc for a fixed number of categories (from 2 to 5) and raters (from 2 to 6) each for a defined number of raters, of the kappaSize package (see Appendix DS in the supplementary material). This package considers multinomial variables instead of the binomial ones proposed by Altaye and Donner [11] that it is based on the "agreement or not"; however, it does not differentiate the levels of disagreement through weighted kappa statistics.

In addition, a very recent paper by Vanbelle et al. [34] shows sample size calculation for a "targeted confidence interval width" and for "testing statistical hypotheses" in the case of binary outcome and 2 or more raters supported by an R package and a shiny app "simpleagree".

More than two raters are considered also by O'Connell and Dobson with the R Package: magree [35] that "Implements the O'Connell-Dobson-Schouten Estimators of Agreement for Multiple Observers", by Fleiss[36-37], Schouten [38], Gwet [17,18], Vanbelle et al. [34] and Moss [38] even if these latter papers are a very difficult statistical level.

Interested readers are referred, among others, to the above papers.

Operative conclusions.

First you have to decide whether the agreement study is with two (A) or more raters (B). This last aspect is treated in this work only with some indications from the literature both because it is decidedly complicated and above all due to lack of space.

A)-Agreement study with only two raters.

In any case, the significance level (α) and the power (1-β) have to be fixed. If the power can be immediately fixed at 0.80, the significance level could be placed in one or two tails. However, since the null hypothesis is of kappa values lower than a minimum acceptable agreement level it could be argued of choosing a unilateral test with the consequent decrease of the necessary sample sizes. As a further relevant point, since the sample sizes required for the weighted kappa are much more than those for the unweighted kappa, it is highly recommended to consider only the unweighted kappa unless in particular situations where the agreement/disagreement must be absolutely weighed. Then, researchers have to fix the proportions of the row and column marginals (usually equal or at most just slightly different) corresponding to the prevalence of the categories in the common population for both raters or in the population of each rater.

This step leads us to decide whether we will be dealing with a 2x2, 3x3, or 4x4 contingency table or more (though not considered in this paper). It will also allow us to understand whether we are dealing with uniform or symmetric tables, which determines the choice of the sample size calculation approach.

Thus, it is possible to calculate the proportion of agreement expected by chance. Following the logical path, after having fixed k_0 (the minimum not interesting agreement value) and k_A (the required interesting agreement value) it is also possible to calculate the observed agreement proportion expected under H_0 (k_0) and under H_A (k_A). However, for calculating the kappa variance under H_0 and H_A, it is necessary to know also the cells probabilities that in 2x2 contingency table can be found quite easily by dividing the observed expected proportion given k_0 and k_A over the cells of the main diagonal and then finding the other given the marginals.

Then we must move on to the following paragraphs A.1.1 or A.1.2 or to those A.2.1 or A.1.2.

A.1.1)-Case of 2x2 uniform contingency tables

SS-A&C-PCCM, SS-A&C-FCCM and SS-Flack-max/min [3] are equal and lower than SS-Donner; so, any one of the four methods can be chosen also for a little non-uniform marginals 0.6; 0.4.

A.1.2)-Case of 2x2 symmetric contingency

SS-Donner can be considered particularly in the case in which they are much lower for highly non-uniform marginals (0.8; 0.2).

A.2.1)-Case of 3x3 and 4x4 uniform contingency tables

Since the probability table 3x3 is unique, SS-Flack-max, SS-A&C-PCCM-max/min and SS-A&C-FCCM are always the same. Then, any one of the methods can be chosen with the advantage for FCCM, if its assumptions are sensible, of not having to calculate the variance of kappa using LP.

A.2.2)-Case of 3x3 and 4x4 symmetric contingency tables

In this case SS-A&C-PCCM-max/min lower than SS-Flack-max can be a sensible choice except for the case of highly non-uniform contingency tabled in which SS-Donner are lower.

It should be emphasized that SS-Flack-min, lower than SS-A&C-PCCM-max/min cannot be recommended due to the risk of obtaining underpowered studies.

However, it is always possible to calculate the sample sizes under the considered models and to select the most suitable for the actual feasibility of the study. Appendix FS in the supplementary material shows some examples of sample sizes calculation.

B)-Agreement study with more than two raters.

We recommend SS-Donner computed by the kappaSize package (see Appendix DS in the supplementary material) or better by the R function we developed for any number of raters. Alternatively, the approach of Vanbelle et al. [33] in the case of binary outcome. Other alternatives are the package of O'Connell and Dobson: magree [34] which "Implements the O'Connell-Dobson-Schouten agreement estimators for multiple observers" and those shown in Fleiss [35], Fleiss [36], Schouten [37], Gwet [14,15] and Moss [38], but this topic is too difficult statistically and computationally.

Table SS1. Sample Sizes for testing unweighted k_0, k_A, marginals of a 2x2 contingency table (shown at each k_0 change); α = 0.05 (1-sided) and power = 0.80

k_0	k_A	π_1	π_2	SS-Flack-max	SS-Flack-min	SS-Donner	SS-A&C-PCCM-max
0.4	0.45	0.5	0.5	2,042	2,042	2,078	2,042
0.4	0.50			501	501	520	501
0.4	0.60			119	119	130	119
0.4	0.70			50	50	58	50
0.4	0.80			26	26	33	26
0.4	0.90			15	15	21	15
0.4	0.95			11	11	18	11
0.5	0.55	0.5	0.5	1,811	1,811	1,855	1,811
0.5	0.60			441	441	464	441
0.5	0.70			103	103	116	103
0.5	0.80			42	42	52	42
0.5	0.90			21	21	29	21
0.5	0.95			15	15	23	15
0.6	0.65	0.5	0.5	1,530	1,530	1,583	1,530
0.6	0.70			368	368	396	368
0.6	0.80			83	83	99	83
0.6	0.90			32	32	44	32
0.6	0.95			21	21	33	21
0.4	0.45	0.6	0.4	2,121	2,121	2,148	2,121
0.4	0.50			520	520	537	520
0.4	0.60			124	124	135	124
0.4	0.70			52	52	60	52
0.4	0.80			27	27	34	27
0.4	0.90			15	15	22	15
0.4	0.95			11	11	18	11
0.5	0.55	0.6	0.4	1,887	1,887	1,924	1,887
0.5	0.60			459	459	481	459
0.5	0.70			107	107	121	107
0.5	0.80			44	44	54	44
0.5	0.90			21	21	31	21
0.5	0.95			15	15	24	15
0.6	0.65	0.6	0.4	1,597	1,597	1,645	1,597
0.6	0.70			384	384	412	384

k_0	k_A	π_1	π_2	SS-Flack-max	SS-Flack-min	SS-Donner	SS-A&C-PCCM-max
0.6	0.80			87	87	103	87
0.6	0.90			33	33	46	33
0.6	0.95			22	22	34	22
0.4	0.45	0.8	0.2	3,110	3,110	3,032	3,110
0.4	0.50			766	766	758	766
0.4	0.60			183	183	190	183
0.4	0.70			77	77	85	77
0.4	0.80			39	39	48	39
0.4	0.90			22	22	31	22
0.4	0.95			16	16	26	16
0.5	0.55	0.8	0.2	2,839	2,839	2,783	2,839
0.5	0.60			692	692	696	692
0.5	0.70			162	162	174	162
0.5	0.80			66	66	78	66
0.5	0.90			32	32	44	32
0.5	0.95			22	22	35	22
0.6	0.65	0.8	0.2	2,437	2,437	2,418	2,437
0.6	0.70			586	586	605	586
0.6	0.80			133	133	152	133
0.6	0.90			51	51	68	51
0.6	0.95			33	33	50	33
0.4	0.45	0.9	0.1	5,418	5,418	5,092	5,418
0.4	0.50			1,338	1,338	1,273	1,338
0.4	0.60			321	321	319	321
0.4	0.70			134	134	142	134
0.4	0.80			69	69	80	69
0.4	0.90			38	38	51	38
0.4	0.95			28	28	43	28
0.5	0.55	0.9	0.1	5,060	5,060	4,786	5,060
0.5	0.60			1,235	1,235	1,197	1,235
0.5	0.70			289	289	300	289
0.5	0.80			117	117	133	117
0.5	0.90			57	57	75	57
0.5	0.95			40	40	60	40
0.6	0.65	0.9	0.1	4,397	4,397	4,221	4,397
0.6	0.70			1,058	1,058	1,056	1,058
0.6	0.80			239	239	264	239
0.6	0.90			91	91	118	91
0.6	0.95			59	59	87	59

SS-A&C-FCCM are the same as SS-A&C-PCCM-max and SS-A&C-PCCM-min and SS-Flack-max and SS-Flack-min, since there is only one variance estimate. Sample sizes for α=0.05 (two-sided) are larger by approximately 10 subjects when k_A-k_0 >0.30, but are also larger by 100 subjects when k_A-k_0 $\leq$0.10

Table SS2. Sample Sizes for testing unweighted k_0, k_A, marginals of a 3x3 contingency table (shown at each k_0 change); α = 0.05 (1-sided) and power= 0.80

k_0	k_A	π_1	π_2	π_3	SS-Flack-max	SS-Flack-min	SS-Donner	SS-A&C-PCCM
0.4	0.45	0.33	0.33	0.33	1,321	1,321	1,336	1,321
0.4	0.50				326	326	334	326
0.4	0.60				79	79	84	79
0.4	0.70				33	33	38	33
0.4	0.80				17	17	21	17
0.4	0.90				10	10	14	10
0.4	0.95				7	7	12	7
0.5	0.55	0.33	0.33	0.33	1,214	1,214	1,237	1,214
0.5	0.60				297	297	310	297
0.5	0.70				70	70	78	70
0.5	0.80				29	29	35	29
0.5	0.90				14	14	20	14
0.5	0.95				10	10	16	10
0.6	0.65	0.33	0.33	0.33	1,057	1,057	1,089	1,057
0.6	0.70				255	255	273	255
0.6	0.80				58	58	69	58
0.6	0.90				22	22	31	22
0.6	0.95				15	15	23	15
0.4	0.45	0.50	0.25	0.25	1,551	926	1,429	1,426
0.4	0.50				381	233	358	352
0.4	0.60				91	58	90	85
0.4	0.70				38	25	40	36
0.4	0.80				20	13	23	19
0.4	0.90				11	7	15	11
0.4	0.95				8	6	12	8
0.5	0.55	0.50	0.25	0.25	1,398	963	1,325	1,311
0.5	0.60				341	239	332	320
0.5	0.70				80	58	83	76
0.5	0.80				33	24	37	31
0.5	0.90				16	12	21	15
0.5	0.95				11	8	17	11
0.6	0.65	0.50	0.25	0.25	1,196	919	1,167	1,140
0.6	0.70				288	224	292	275
0.6	0.80				65	52	73	63
0.6	0.90				25	20	33	24
0.6	0.95				16	13	24	16
0.4	0.45	0.60	0.30	0.10	2,058	1,098	1,699	1,719
0.4	0.50				504	271	425	424
0.4	0.60				120	65	107	102
0.4	0.70				50	28	48	43

k_0	k_A	π_1	π_2	π_3	SS-Flack-max	SS-Flack-min	SS-Donner	SS-A&C-PCCM
0.4	0.80				26	15	27	22
0.4	0.90				15	9	17	12
0.4	0.95				11	7	15	9
0.5	0.55	0.60	0.30	0.10	1,801	1,002	1,572	1,572
0.5	0.60				437	245	393	384
0.5	0.70				102	60	99	90
0.5	0.80				41	25	44	37
0.5	0.90				20	13	25	18
0.5	0.95				14	9	20	13
0.6	0.65	0.60	0.30	0.10	1,501	875	1,378	1,359
0.6	0.70				360	217	345	328
0.6	0.80				81	51	87	75
0.6	0.90				31	20	39	29
0.6	0.95				20	13	29	19
0.4	0.45	0.80	0.10	0.10	2,893	200	2,589	2,731
0.4	0.50				714	66	648	675
0.4	0.60				171	22	162	163
0.4	0.70				72	10	72	68
0.4	0.80				37	6	41	35
0.4	0.90				21	3	26	20
0.4	0.95				15	2	22	14
0.5	0.55	0.80	0.10	0.10	2,671	690	2,449	2,554
0.5	0.60				652	184	613	624
0.5	0.70				153	49	154	147
0.5	0.80				62	21	69	59
0.5	0.90				30	10	39	29
0.5	0.95				21	7	31	20
0.6	0.65	0.80	0.10	0.10	2,307	1,010	2,174	2,231
0.6	0.70				555	255	544	538
0.6	0.80				126	62	136	122
0.6	0.90				48	24	61	46
0.6	0.95				31	15	45	30

SS-A&C-FCCM are the same as SS-A&C-PCCM-max and SS-A&C-PCCM-min since there is only one variance estimate. Sample sizes for α=0.05 (two-sided) are larger by approximately 10 subjects when k_A-k_0 >0.30, but are also larger by 100 subjects when k_A-k_0 ≤0.10

Table SS3. Sample Sizes for testing unweighted k_0, k_A, marginals of a 4x4 contingency table (shown at each k_0 change); α = 0.05 (1-sided) and power = 0.80

k_0	k_A	π_1	π_2	π_3	π_4	SS-Flack-max	SS-Flack-min	SS-Donner	SS-A&C-PCCM-max	SS-A&C-PCCM-min	SS-A&C-FCCM
0.4	0.45	.25	.25	.25	.25	1,081	1,081	1,089	1,081	1,081	1,081
0.4	0.50					268	268	273	268	268	268
0.4	0.60					65	65	69	65	65	65
0.4	0.70					28	28	31	28	28	28
0.4	0.80					14	14	18	14	14	14
0.4	0.90					8	8	11	8	8	8
0.4	0.95					6	6	9	6	6	6
0.5	0.55	.25	.25	.25	.25	1,015	1,015	1,031	1,015	1,015	1,015
0.5	0.60					249	249	258	249	249	249
0.5	0.70					59	59	65	59	59	59
0.5	0.80					24	24	29	24	24	24
0.5	0.90					12	12	17	12	12	12
0.5	0.95					9	9	13	9	9	9
0.6	0.65	.25	.25	.25	.25	899	899	924	899	899	899
0.6	0.70					218	218	231	218	218	218
0.6	0.80					50	50	58	50	50	50
0.6	0.90					19	19	26	19	19	19
0.6	0.95					13	13	19	13	13	13
0.4	0.45	.4	.3	.2	.1	1,520	811	1,197	1,208	1,194	1,1984
0.4	0.50					372	203	300	299	295	296
0.4	0.60					89	50	75	72	71	72
0.4	0.70					37	22	34	31	30	30
0.4	0.80					19	12	19	16	16	16
0.4	0.90					11	7	12	9	9	9
0.4	0.95					8	5	10	7	7	7
0.5	0.55	.4	.3	.2	.1	1,337	812	1,128	1,123	1,115	1,117
0.5	0.60					325	201	282	275	273	274
0.5	0.70					76	48	71	65	65	65
0.5	0.80					31	20	32	27	27	27
0.5	0.90					15	10	18	13	13	13
0.5	0.95					11	7	14	9	9	9
0.6	0.65	.4	.3	.2	.1	1,121	758	1,006	986	982	983
0.6	0.70					269	185	252	238	238	238
0.6	0.80					61	43	63	55	54	54
0.6	0.90					23	17	28	21	21	21
0.6	0.95					15	11	21	14	14	14
0.4	0.45	.6	.2	.1	.1	1,889	513	1,463	1,511	1,496	1,499
0.4	0.50					463	134	366	373	370	371
0.4	0.60					110	35	92	90	89	90

k_0	k_A	π_1	π_2	π_3	π_4	SS-Flack-max	SS-Flack-min	SS-Donner	SS-A&C-PCCM-max	SS-A&C-PCCM-min	SS-A&C-FCCM
0.4	0.70					46	16	41	38	38	38
0.4	0.80					24	9	23	20	20	20
0.4	0.90					13	5	15	11	11	11
0.4	0.95					10	4	13	8	8	8
0.5	0.55	.6	.2	.1	.1	1,665	658	1,384	1,405	1,397	1,399
0.5	0.60					405	167	346	344	342	342
0.5	0.70					94	42	87	81	81	81
0.5	0.80					38	18	39	33	33	33
0.5	0.90					19	9	22	16	16	16
0.5	0.95					13	6	18	12	11	11
0.6	0.65	.6	.2	.1	.1	1,395	706	1,233	1,231	1,227	1,228
0.6	0.70					335	176	309	297	296	297
0.6	0.80					76	42	78	68	68	68
0.6	0.90					29	17	35	26	26	26
0.6	0.95					19	11	26	17	17	17
0.4	0.45	.7	.1	.1	.1	2,056	314	1,760	1,839	1,839	1,839
0.4	0.50					506	89	440	455	455	455
0.4	0.60					121	26	110	110	110	110
0.4	0.70					51	12	49	46	46	46
0.4	0.80					26	6	28	24	24	24
0.4	0.90					15	4	18	13	13	13
0.4	0.95					11	3	15	10	10	10
0.5	0.55	.7	.1	.1	.1	1,881	644	1,676	1,726	1,726	1,726
0.5	0.60					459	168	419	423	423	423
0.5	0.70					108	43	105	100	100	100
0.5	0.80					44	18	47	41	41	41
0.5	0.90					21	9	27	20	20	20
0.5	0.95					15	6	21	14	14	14
0.6	0.65	.7	.1	.1	.1	1,617	813	1,498	1,517	1,517	1,517
0.6	0.70					389	203	375	366	366	366
0.6	0.80					88	49	94	83	83	83
0.6	0.90					34	19	42	32	32	32
0.6	0.95					22	12	31	21	21	21

SS-A&C-FCCM (last column) is lower than SS-A&C-PCCM-max in 32 (22.2%) cases and greater than SS-A&C-PCCM-min in 20 (13.9%) cases; SS-A&C-FCCM is equal in the remaining 112 (77.8%) and 124 (86.1%) cases, respectively. Sample sizes for α=0.05 (two-sided) are larger by approximately 10 subjects when k_A-k_0 >0.30, but are also larger by 100 subjects when k_A-k_0 ≤0.10

REFERENCES

1. https://www.ncss.com/software/pass/pass-documentation/
2. PASS®13 *Power Analysis and Sample Size* Software 2014. NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass. PASS®16 *Power Analysis and Sample Size Software 2018*. NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass.
3. Flack VFA, Afifi A, Lachenbruch PA, Schouten HJ. A. Sample Size Determinations for the two Rater Kappa Statistic. Psychometrika, 1988; 53(3): 321-325.
4. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa Psychol Bull. 1969, 72,5: 323-327.
5. Cohen JA. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20: 213–220.
6. Bujang MA, Baharum N. Guidelines of the minimum sample size requirements for Cohen's Kappa. Epidemiol Biostat Public Health 2017, 14 (2): e12267-1-10.
7. Donner A, Eliasziw M. A Goodness-Of-Fit Approach to Inference Procedures for the Kappa Statistic: Confidence and Sample Size Estimation Interval Construction, Significance-Testing. Statist Med. 1992; 11: 1511-1519.
8. Donner A, Eliasziw M. Statistical Implications of the Choice Between a Dichotomous or Continuous Trait in Studies of Interobserver Agreement. Biometrics. 1994; 50: 550-555.
9. Donner A, Eliasziw M, Klar N., Testing the Homogeneity of Kappa Statistics. Biometrics. 1996; 52: 176-183
10. Donner A. Sample Size Requirements for the Comparison of two or more Coefficients of Inter-Observer Agreement. Statist. Med. 1998;17: 1157-1168.
11. Altaye M, Donner A, Klar N. Inference procedures for assessing interobserver agreement among multiple raters. Biometrics. 2001; 57: 584-588.
12. Altaye M., Donner A., Eliasziw M. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Statist. Med. 2001; 20:2479–2488 doi: 10.1002/sim.911.
13. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. Resolving the paradoxes. J Clin Epidemiol. 1990;43, 6: 543-549.
14. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol. 1988;41:949–958.
15. Shoukri MM. *Measures of Interobserver Agreement*. 2004 Boca Raton, Fla: Chapman & Hall/CRC.
16. Brennan RL, Prediger DJ. Coefficient kappa: Some uses, misuses, and alternatives. Educ Psychol Meas. 1981; 41: 687–699.
17. Gwet K. Computing inter-rater reliability and its variance in presence of high agreement. Br. J. Math. Stat. Psychol. 2008; 61: 29-48.
18. Gwet KL. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. 4th ed. Gaithersburg, MD: Advanced Analytics. 2014.
19. Klein D. Implementing a general framework for assessing interrater agreement in Stata. SJ 2018;18: 871–901. doi: 10.1177/1536867x1801800408. 2018.
20. Falcaro M, Newson RB. tabagree: Nonparametric measures of agreement and disagreement in paired ordinal data. SJ 2025; 25(3):627–645 doi: 10.1177/1536867X251365495.
21. Svensson E. *Analysis of Systematic and Random Differences Between Paired Ordinal Categorical Data*. Stockholm: Almqvist and Wiksell 1993.
22. Svensson E. A coefficient of agreement adjusted for bias in paired ordered categorical data. Biom J. 1997;39:643–657. doi:10.1002/bimj.4710390602.
23. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213–220.
24. Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psychol Bull. 1979;86:974–977.
25. Kraemer HC, Periyakoil VS, Nod A. Chapter 1.3 Agreement Statistics; Tutorial In Biostatistics; Kappa coefficients in medical research. *Tutorials in Biostatistics Volume 1:* Statistical Methods in Clinical Studies Edited by RB. D'Agostino 2004 John Wiley & Sons, Ltd.
26. Donner A. Sample size requirements for interval estimation of the intraclass kappa statistic. Commun Stat-Simul C Journal 1999; 28(2): 415-429.
27. Donner A, Rotondi MA. Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters. Int J Biostat. 2010; 6(1) Article 31.
28. Rotondi MA, Donner A. A Confidence Interval Approach to Sample Size Estimation for Interobserver Agreement Studies with Multiple Raters and Outcomes. J. Clin. Epidemiol. 2012;65:778-784.
29. Choudhary PK, Nagaraja HN. *Measuring Agreement: Models, Methods, and Applications*. 2017 John Wiley & Sons, Inc. Hoboken, USA.
30. Cantor AB. Sample-Size Calculations for Cohen's Kappa. Psychol. Methods. 1996; 1(2):150-153.
31. Rotondi MA. Package 'kappaSize', Title: Sample Size Estimation Functions for Studies of Interobserver Agreement Version 1.1 released on February 20, 2015 and Version 1.2 released on November 26, 2018, Author: Michael A. Rotondi, Maintainer: Michael A. Rotondi
32. Von Eye A, Mun EY. *Analyzing Rater Agreement: Manifest Variable Methods*. Lawrence Erlbaum Associates; 2005.
33. https://cran.r-project.org/web/packages/lpSolve/index.html Package 'lpSolve', January 24, 2020, Version 5.6.15 Title Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs URL https://github.com/gaborcsardi/lpSolve.
34. Elashoff JD. nQuery Advisor ® 2007 Version 7.0 User's Guide. Statistical Solutions Ltd., Republic of Ireland. http://www.statsolusa.com info@statsolusa.com.
35. Vanbelle S, Hernandez Engelhart C, Blix E. Meas-

uring Agreement in Diagnostics: A Practical Guide for Researchers (Tutorial in Biostatistics). Stat Med, 2025;44, e70299
36. https://cran.r-project.org/web/packages/magree/index.html (via r-universe) July 23, 2025.
37. Fleiss JL. Measuring Nominal Scale Agreement among Many Raters. Psychol. Bull. 1971; 76(5): 378-382.
38. Fleiss J. *Statistical Methods for Rates and Proportions*. Wiley, New York. 1981.
39. Schouten HJA. Measuring pairwise agreement among many observers Biom J. 1880;22(6): 497-504].
40. Moss J. Measures of Agreement with Multiple Raters: Fréchet Variances and Inference Psychometrika 2024,89(2): 517–541. https://doi.org/10.1007/s11336-023-09945-2.