Published in final edited form as:
Stat Med. 2013 September 20; 32(21): . doi:10.1002/sim.5796.

Kappa statistic for the clustered dichotomous responses from physicians and patients

Chaeryon Kang[a], **Bahjat Qaqish**[b], **Jane Monaco**[b], **Stacey L. Sheridan**[c,d], and **Jianwen Cai**[b,*]

[a]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.

[b]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

[c]Department of Medicine, Division of General Medicine and Clinical Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, U.S.A.

[d]Center for Health Promotion and Disease Prevention, University of North Carolina, Chapel Hill, NC, U.S.A.

Abstract

The bootstrap method for estimating the standard error of the kappa statistic in the presence of clustered data is evaluated. Such data arise, for example, in assessing agreement between physicians and their patients regarding their understanding of the physician-patient interaction and discussions. We propose a computationally efficient procedure for generating correlated dichotomous responses for physicians and assigned patients for simulation studies. The simulation result demonstrates that the proposed bootstrap method produces better estimate of the standard error and better coverage performance compared to the asymptotic standard error estimate that ignores dependence among patients within physicians with at least a moderately large number of clusters. An example of an application to a coronary heart disease prevention study is presented.

Keywords

Kappa statistic; a dichotomous response for clusters; bootstrap resampling for clusters

1. Introduction

The kappa statistic is a commonly used measure of inter-rater agreement. The kappa statistic was originally proposed by Cohen [1] to quantify the degree of agreement beyond chance when two raters simultaneously score the same subjects on a nominal or ordinal scale. Inter-observer reliability is measured by comparing the observed proportion of agreement, P_o, with the proportion of agreement expected by chance, P_e, and scaling the difference so that a value of one indicates perfect agreement and a value of zero indicates no agreement beyond that expected by chance.

Many extensions to the kappa statistic have been proposed including weighting to account for the seriousness of misclassification errors [2, 3], allowing more than two raters [4], using

*Correspondence to: Department of Biostatistics, CB#7420, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A. cai@bios.unc.edu.

stratified samples [5], comparing a single rater to a consensus group of raters [3], and allowing for multiple observations per subject and multiple categorizations [6].

Recent work has focused on the study of correlated kappa statistics. For example, when two or more raters evaluate the same subjects at different time points or under different conditions and the equality of the kappa statistics is of interest, the correlation between the kappa statistics must be addressed. McKenzie [7] proposed resampling techniques for comparing correlated kappa statistics, considering the case of pairwise comparisons of kappa statistics when the observations were evaluated by three raters. VanBelle and Albert [8] extended these methods using the bootstrap method to allow the comparison of more than two kappa statistics. Model-based procedures for comparing two dependent kappa statistics calculated from two observers from the same data when the outcome is binary were proposed by Donner, Shoukri, Klar and Bartfay [9]. GEE methods for comparing correlated kappa statistics have been proposed for dichotomous [10] or categorical [11] outcomes. Methods for weighted correlated kappas have also been developed [12] using GEE methods which allow for modeling of covariate effects. Barnhart and Williamson [13] proposed a weighted least squares method for comparing correlated kappa coefficients in the case of categorical covariates. Rather than comparing correlated kappa statistics, our research focuses on inference for individual kappa parameters in the presence of correlated binary outcomes.

Clustered binary data can result, as in our motivating example, when investigators are interested in the agreement between ratings by physicians and their patients regarding the same event. Since each physician can see more than one patient, patients seen by the same physician form a cluster. In other words, the responses of a physician for his/her patients tend to be more similar than those from other physicians, which results in clustered responses within a physician. We propose a bootstrap method to address the correlated data structure in such cases.

In our example, 24 physicians and their 157 patients evaluated clinical discussions regarding coronary heart disease (CHD) prevention [14]. Cluster sizes ranged from 1 to 20 patients per physician. Following the physician-patient discussion, each participant was surveyed regarding discussion content and resulting decisions. Each member of a physician-patient pair reported whether or not CHD was discussed. For pairs in which at least one of the physician/patient pair reported that CHD was discussed, agreement was evaluated for several outcome measures including whether medication was recommended or change in the patient's lifestyle for CHD prevention was recommended. The kappa statistic was used as a measure of physician-patient agreement regarding their discussion, and the bootstrap method was used to estimate the standard errors accounting for the clustered data structure.

While methods for analysis have been proposed for clustered data, research has focused primarily on extensions to McNemar's test or estimates of association, such as the odds ratio [15, 16, 17] rather than the kappa statistic. Oden [18] and Schouten [19] proposed a pooled kappa for situations in which two raters evaluate a set of paired units, such as pairs of eyes.

Bootstrap confidence intervals for kappa statistics have been proposed previously to address small sample sizes when observations are independent [20] and for comparing correlated kappa statistics [7, 8]. To our knowledge, the properties of the bootstrap method have not been studied for inference on individual kappa parameters in the context of clustered data.

In this paper, we evaluate the bootstrap method for calculating the kappa statistic and estimating its standard error in the presence of clustered binary outcomes. In Section 2, we provide background information regarding the kappa statistic, its asymptotic standard error, and the bootstrap method. We describe the generation of correlated dichotomous responses

specifically for our simulations in Section 3. Simulation results are presented in Section 4. An applied example of the method follows in Section 5. Finally, Section 6 contains a discussion of the results.

2. Method

2.1. The kappa statistic and its asymptotic standard error (ASE)

In this section, we briefly describe the kappa statistic and its asymptotic standard error (ASE) for the case of independent subjects. Suppose that (Y, X) represents a pair of dichotomous responses from two raters, for example, a physician and a patient. Data on N subjects can be depicted in a 2 × 2 table. Let n_{00}, n_{01}, n_{10} and n_{11} represent the cell counts for (Y, X) = (0, 0), (0, 1), (1, 0), (1, 1), respectively. Let $n_{\cdot 0} = n_{00} + n_{10}$, $n_{0\cdot} = n_{00} + n_{01}$, $n_{\cdot 1} = n_{01} + n_{11}$, and $n_{1\cdot} = n_{10} + n_{11}$, and $n_{1\cdot}=n_{10}+n_{11}.N=\sum_{i,j=0}^{1} n_{ij}$ denotes the number of subjects under study. Define $P_o=\frac{n_{00}+n_{11}}{N}$ and $P_e=\frac{(n_{\cdot\mathbf{0}})\times(n_{0\cdot})+(n_{\cdot\mathbf{1}})\times(n_{1\cdot})}{N^2}$, the kappa statistic $\widehat{\kappa}$ introduced by Cohen [1] is calculated as follows:

$$\widehat{\kappa}=\frac{P_o-P_e}{1-P_e}. \quad (1)$$

We also define

$$\begin{aligned} q= &\ \tfrac{n_{00}}{N}\times\left\{1-\left(\tfrac{n_{0\cdot}}{N}+\tfrac{n_{\cdot\mathbf{0}}}{N}\right)\times(1-\widehat{\kappa})\right\}^2+\tfrac{n_{11}}{N}\times\left\{1-\left(\tfrac{n_{1\cdot}}{N}+\tfrac{n_{\cdot\mathbf{1}}}{N}\right)\times(1-\widehat{\kappa})\right\}^2,\\ r= &\ (1-\widehat{\kappa})^2\times\left\{\tfrac{n_{01}}{N}\left(\tfrac{n_{\cdot\mathbf{0}}}{N}+\tfrac{n_{1\cdot}}{N}\right)^2+\tfrac{n_{10}}{N}\left(\tfrac{n_{\cdot\mathbf{1}}}{N}+\tfrac{n_{0\cdot}}{N}\right)^2\right\},\\ s= &\ (\widehat{\kappa}-P_e\times(1-\widehat{\kappa}))^2. \end{aligned}$$

Following [21] and [22], *ASE* of the kappa statistic can be estimated by

$$ASE\,(\widehat{\kappa})=\sqrt{\frac{q+r-s}{N\times(1-P_e)^2}}. \quad (2)$$

Note that (1) and (2) can be obtained using SAS PROC FREQ.

2.2. Bootstrap sampling algorithm

The ASE calculation introduced in the previous section was developed based on the assumption of independence. Therefore, ASE is not appropriate for the clustered data since responses within clusters tend to be positively correlated which results in underestimating standard error of kappa statistic. To resolve this problem, we propose adopting a bootstrap method [23] by randomly sampling the clusters with replacement and taking all observations belonging to the sampled clusters. This bootstrap method is called the *cluster bootstrap* method since bootstrap sampling is conducted on clusters only [24, 25, 26]. In our study, a cluster is a physician, and observations within the cluster are patients.

2.2.1. Bootstrap sampling of clusters (physicians)—

1. 1. Assume that there are n clusters (physicians), and they are indexed by {1, … , n}. Draw a random sample of n clusters with replacement from the original data. The selected clusters are indexed by {1*, 2*, … , n*}, where the i* (i = 1, … , n) are elements of {1, … , n}.

2. For each sampled cluster, take all observations belonging to cluster i^*. Let n_{i^*} denote the size of cluster i^*, $Y^{i^*} = (y_{i^*,1}, \ldots, y_{i^*,n_{i^*}})^T$ and $X^{i^*} = (x_{i^*,1}, \ldots, x_{i^*,n_{i^*}})^T$. In our example, Y^{i^*} represents a vector of responses from physician i^* for his/her n_{i^*} patients and X^{i^*} represents a vector of responses from the n_{i^*} patients of physician i^*. The bootstrap sample $\mathbf{Z}$ consists of $(\mathbf{Y}^*,\mathbf{X}^*)$, where $\mathbf{Y}^* = (Y^{1^*T}, \ldots, Y^{n^*T})^T$, and $\mathbf{X}^* = (X^{1^*T}, \ldots, X^{n^*T})^T$.

3. Repeat steps 1 and 2 B times to generate B independent bootstrap samples $\mathbf{Z^1}$, …, $\mathbf{Z^B}$.

4. Calculate the kappa statistic $\widehat{\kappa}^b$ corresponding to each bootstrap sample, $\mathbf{Z^b}$, following formula (1).

5. Calculate bootstrap estimate by $\widehat{\kappa}_B = \sum_{b=1}^{B} \frac{\widehat{\kappa}^b}{B}$

6. Estimate bootstrap standard error by $\widehat{SE}(\widehat{\kappa}_B) = \sqrt{\frac{\sum_{b=1}^{B}(\widehat{\kappa}^b - \widehat{\kappa}_B)^2}{B-1}}$.

2.2.2. Confidence interval. The 95% confidence intervals are obtained by—

$$\begin{aligned} 95\%\ \text{bootstrap confidence interval based on normal approximation} &= \left(\widehat{\kappa}_B - 1.96\,\widehat{SE}(\widehat{\kappa}_B), \widehat{\kappa}_B + 1.96\,\widehat{SE}(\widehat{\kappa}_B)\right), \\ 95\%\ \text{bootstrap confidence interval based on percentiles} &= \left(\widehat{\kappa}_B^{(0.025)}, \widehat{\kappa}_B^{(0.975)}\right), \end{aligned} \quad (3)$$

where $\widehat{SE}(\widehat{\kappa}_B)$ denotes bootstrap standard error estimate of $\widehat{\kappa}_B$, and $\widehat{\kappa}_B^{(1-\alpha)}$ is the 100(1 − α)$^{\text{th}}$ empirical percentile using the bootstrap samples.

In addition to the two confidence intervals above, we calculate bias-corrected and accelerated (BC_a) intervals which is an improved percentile method by automatically correcting the bias of the bootstrap estimator and provides second-order accuracy [23, 27, 28]. We computed the BC_a confidence interval following [23] with some modification since our resampling unit is clusters (physicians), not individual subjects. Let $\widehat{G}(c)$ denote the empirical cumulative distribution of c, $\widehat{G}(c) = \sum_{b=1}^{B} \frac{1\left\{\widehat{\kappa}^b < c\right\}}{B}$. Define $\widehat{\kappa}_{BC_a}^{(\alpha)} = \widehat{G}^{-1}\left\{\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - \alpha(z_0 + z^{(\alpha)})}\right)\right\}$, where Φ denotes the standard normal distribution (CDF), $z^{(\alpha)}$ denotes the 100α^{th} percentile point of a standard normal distribution, and for some z_0 and a. Then, 95% bootstrap confidence interval using the BC_a method is defined as follows:

$$95\%\ \text{bootstrap confidence interval based on } BC_a = \left(\widehat{\kappa}_{BC_a}^{(0.025)}, \widehat{\kappa}_{BC_a}^{(0.975)}\right). \quad (4)$$

The constant $\widehat{z}_0$ can be computed by $\widehat{z}_0 = \Phi^{-1}\left(\sum_{b=1}^{B} \frac{1\left\{\widehat{\kappa}^b < \widehat{\kappa}\right\}}{B}\right)$. Next, we calculate a following [23]. Since our resampling unit is a cluster (physician) $i = 1, \ldots, n$, we define $U_i = \widehat{\kappa}_{(\cdot)} - \widehat{\kappa}_{(-i)}$, where $\widehat{\kappa}_{(\cdot)} = \sum_{i=1}^{n} \widehat{\kappa}_{(-i)}/n$ and $\widehat{\kappa}_{(-i)}$ is a kappa statistic computed by the original sample deleting all subjects belonging to i^{th} cluster. Then we compute $\widehat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n} U_i^3}{\left(\sum_{i=1}^{n} U_i^2\right)^{3/2}}$.

Note that both the standard and the percentile methods are not second-order accurate, so relatively larger number of bootstrap replications are required for the BC_a method compared to the standard and the percentile methods. Efron and Tibshirani [23] suggest that at least 1, 000 bootstrap replications are needed for the BC_a method.

3. Simulation set-up

In this section, we provide a detailed description of the data generation procedure for the simulation study based on the clustered data structure in which the cluster is a physician and observations within a cluster are the patients of the physician. The calculation of the kappa statistic, estimation of standard error of the kappa statistic, and construction of the confidence intervals of the kappa statistic follows. Suppose that a pair of dichotomous responses is obtained for each physician-patient encounter. For example, the dichotomous response could denote survey-response of the physician-patient discussion or an assessment of the treatment.

3.1. Generating dichotomous responses for physician-patient pairs

3.1.1. Notation and assumptions—Suppose we have n clusters representing n physicians, and each cluster consists of m pairs of dichotomous responses from the physician-patient pairs. For patient i of a physician, let Y_i and X_i be random variables representing the physician's assessment and the patient's assessment of the same discussion, respectively. Note that $Y_i \in \{0, 1\}$ and $X_i \in \{0, 1\}$ with $Y_i = 1$ or $X_i = 1$ denoting "yes" for a given question. Let $Y = (Y_1, \ldots, Y_m)^T$ and $X = (X_1, \ldots, X_m)^T$ denote the random vectors representing dichotomous responses for a physician and his/her patients, and $\mu_y = (\mu_{y1}, \ldots, \mu_{ym})^T = (P\{Y_1 = 1\}, \ldots, (P\{y_m = 1\})^T = E[Y]$ and $\mu_x = (\mu_{x1}, \ldots, \mu_{xm})^T = (P\{X_1 = 1\}, \ldots, P\{X_m = 1\})^T = E[X]$ denote the corresponding marginal mean vectors. The correlation matrix of the response vectors is define as $R_c = (r_{ij}) = \begin{pmatrix} R_{yy} & R_{xy} \\ R_{xy} & R_{xx} \end{pmatrix}$, where $R_{yy} = (Corr(Y_i, Y_j))$, $R_{xy} = (Corr(Y_i, X_j))$, and $R_{xx} = (Corr(X_i, X_j))$.

In this simulation, a homogeneous cluster size is assumed for simplicity although the same simulation procedure can be applied to the case of the heterogeneous cluster size. We assume that all physicians have the same mean vector, $\mu_{y1} = \ldots = \mu_{ym} = \mu_y$, and all patients have the same mean vector, $\mu_x 1 = \ldots = \mu_{x_m} = \mu_x$. Also, all physicians have the same correlation matrix and same strength of agreement with patients. An exchangeable correlation structure is assumed within-physician and between physician and patient. Hence we define $\rho_w = Corr(Y_i, Y_j)$, $i \neq j$ to be the within-physician correlation and $\rho_b = Corr(Y_i, X_i)$ to be the physician-patient correlation. The parameter ρ_b is related to kappa as explained in subsection 3.1.3. Since all physicians are assumed to have the same mean and correlation matrix, we generate n independent sets of responses for the n physicians by repeating the following data generating procedure n times independently.

3.1.2. Generating correlated dichotomous responses within physicians—Note that each physician could have their own practice pattern, so it is reasonable to assume that the responses from a physician for different patients are correlated. We generate an $m \times 1$ vector of correlated dichotomous responses Y for each of the n physicians following Qaqish [29]. Qaqish [29] introduced the conditional linear family of multivariate Bernoulli distributions which is useful for simulating correlated binary random variables with specified marginal mean vector $\mu = (\mu_1, \ldots, \mu_m)^T$ and correlation matrix, $R = (r_{ij})$. The algorithm has been implemented in the R-package **binarySimCLF** [30], which we used to generate responses for physicians.

Before proceeding with the description of the data generating procedure, we briefly describe some restrictions that any valid (μ, R) should satisfy as [29] noted. First, the correlation matrix $R = (r_{ij})$ should be positive definite, and secondly, restriction on r_{ij} are imposed by μ_i and μ_j. Defining $\psi_i = \sqrt{\frac{\mu_i}{1-\mu_i}}$, the correlations must satisfy

$$\max\left(-\psi_i\psi_j, \frac{-1}{\psi_i\psi_j}\right) \le r_{ij} \le \min\left(\frac{\psi_i}{\psi_j}, \frac{\psi_j}{\psi_i}\right) (i \ne j). \quad (5)$$

For example, if we assume an homogeneous mean $\mu_Y = 0.4$ for all Y_i, (5) implies $-\frac{2}{3} \le \rho_w \le 1$. Since we assume an exchangeable and positive correlation between physician's responses within a physician, both conditions are satisfied.

3.1.3. Generating responses for patients given responses for physician—Once dichotomous responses for physicians, Y, are generated, dichotomous responses for patients given responses for physicians, X, can be generated in such a way that kappa and the marginal means have their stipulated values. We assume that responses for patients are conditionally independent given the physician's responses, and naturally, this implies that, marginally, responses for patients are correlated within physicians.

Let us consider a 2 × 2 table, where Y_i denotes dichotomous response for a physician about patient i, and X_i denotes the corresponding patient's response. Then, $a \equiv P(Y_i = 0, X_i = 0)$ and $c \equiv P(Y_i = 1, X_i = 0)$. Also, d and b can be expressed as follows:

$$\begin{aligned} d &= P(Y_i{=}1, X_i{=}1) = \mu_y\mu_x + \rho_b\sqrt{\mu_y(1-\mu_y)}\sqrt{\mu_x(1-\mu_x)}, \\ b &= P(Y_i{=}0, X_i{=}1) = \mu_x - d, \end{aligned}$$

and we define b_0 and b_1 by

$$\begin{aligned} b_0 &= \frac{b}{1-\mu_y} = \frac{P(Y_i=0,X_i=1)}{P(Y_i=0)} = P(X_i{=}1|Y_i{=}0), \\ b_1 &= \frac{d}{\mu_y} - b_0 = \frac{P(Y_i=1,X_i=1)}{P(Y_i=1)} - b_0 = P(X_i{=}1|Y_i{=}1) - P(X_i{=}1|Y_i{=}0). \end{aligned} \quad (6)$$

This implies that $E[X_i|Y_i] = b_0 + b_1 Y_i$. Therefore, we generate X_i, $i = 1 \ldots, m$ as independent Bernoulli variables with conditional means $b_0 + b_1 Y_i$, $i = 1, \ldots, m$. The correlation coefficient between responses for physician and patient should satisfy the restriction given in (5). In the simulation, we set $\mu_Y = 0.4$ and $\mu_X = 0.5$, so $\psi_y = \sqrt{\frac{\mu_y}{1-\mu_y}} = \sqrt{\frac{2}{3}}$, and $\psi_x = \sqrt{\frac{\mu_x}{1-\mu_x}} = 1$. Therefore,

$$-\sqrt{\frac{2}{3}} \le \rho_b \le \sqrt{\frac{2}{3}} = 0.816497.$$

We calculate κ_0 and ρ_b as follows:

$$\begin{aligned} \kappa_0 &= \frac{a+d-[\mu_y\mu_x+(1-\mu_y)(1-\mu_x)]}{1-[\mu_y\mu_x+(1-\mu_y)(1-\mu_x)]}, \\ \rho_b &= \frac{d-\mu_y\mu_x}{\sqrt{\mu_y(1-\mu_y)}\sqrt{\mu_x(1-\mu_x)}}. \end{aligned}$$

From the above two formulae, κ_0 and ρ_b are related by $\kappa_0 = \rho_b \left\{ \frac{1}{2} \left(\frac{\psi_y}{\psi_x} + \frac{\psi_x}{\psi_y} \right) \right\}^{-1}$.

Setting $\mu_y = 0.4$ and $\mu_x = 0.5$, the maximum value of available ρ_b is 0.816497, and hence the maximum value of κ_0 we can use is 0.816497/1.02062 = 0.8. This is a reasonable boundary in practice. For more details on parameter restrictions, see [29]. Detailed calculation for the marginal correlation between patients within a physician is given in APPENDIX 1. Using this simulation algorithm, for each configuration we generated $M = 1,000$ independent data sets (Monte-Carlo simulations), with n clusters each.

3.2. Calculate the kappa statistic and the bootstrap kappa statistic

After generating (**Y, X**), we calculate the kappa statistic $\hat{\kappa}$ assuming independent observations by formula (1) and $ASE(\hat{\kappa})$ by formula (2). A 95% confidence interval is constructed as $(\hat{\kappa} - 1.96 \times ASE(\hat{\kappa}), \hat{\kappa} + 1.96 \times ASE(\hat{\kappa}))$. We calculate the bootstrap kappa statistic, $\hat{\kappa}_B$, and bootstrap standard error estimate, $\widehat{SE}(\hat{\kappa}_B)$, as described in Section 2.2. 95% bootstrap confidence intervals based on the normal approximation, 95% bootstrap confidence interval based on percentiles, and 95% bootstrap confidence interval by using the BC_a method are constructed by formula (3) and (4). M independent data sets are simulated, and the coverage rate can be calculated as follows:

$$\text{Coverage rate}\,(\%) = \frac{\text{Number of times the true kappa value}\,(\kappa_0)\,\text{lies within the confindence interval}}{\text{Total number of simulations (M)}} \times 100. \quad (7)$$

A method whose coverage rate is closer to the target nominal coverage probability, for example, 95%, has better coverage performance.

3.3. The number of Monte-Carlo simulations

The number of Monte-Carlo simulations, denoted by M, can be determined by $M = \left(\frac{Z_{1-\frac{\alpha}{2}} \sigma}{\delta} \right)^2$ following [31], where σ denotes the variance of κ, δ is the permissible difference between κ_0 and $\hat{\kappa}$, and $Z_{1-\frac{\alpha}{2}}$ denotes the $\left(1 - \frac{\alpha}{2}\right)^{\text{th}}$ quantile of the standard normal distribution. The standard error estimates of $\hat{\kappa}$ and $\hat{\kappa}_B$ from data analysis results presented later in this study are between 0.053 ~ 0.091. With $M = 500$, the permissible difference between κ_0 and $\hat{\kappa}(\delta)$ is between 0.0046 ~ 0.0080. In our simulation study, $M = 1,000$ and the corresponding δ is 0.0033 ~ 0.0056, which is a reasonably small difference.

4. Simulation results

In this section, we present simulation results for examining the performance of the bootstrap estimates with clustered data under various scenarios of the number of physicians (number of clusters), the number of patients per physician (cluster size), and strength of agreement between physician and patient (kappa value). We also present the results under independent assumption. The simulation was performed using R (Version 2.15.1) and the R-package **binarySimCLF**.

4.1. Determine the number of bootstrap replications B

Before we proceed with comparing two methods, a simulation study was conducted to decide the number of bootstrap replications B. We generated 1, 000 independent data sets as

described in Section 3. Each data set consists of 25 clusters (physicians), and each cluster consists of 20 pairs of the dichotomous responses for physician and patient. For each simulated data set, different numbers of replications B ranging from 50 to 2, 000 were explored. For this simulation study, marginal mean of the dichotomous responses for physicians (μ_y) and patients (μ_x) were 0.4 and 0.5, respectively. The correlation coefficient within a physician (ρ_w) was 0.3. We calculated Monte-Carlo Standard Error estimate (MCSE) of $\widehat{\kappa}_B$ and $\widehat{SE}(\widehat{\kappa}_B)$ over M = 1, 000 simulations by $\sqrt{\frac{\sum_{m=1}^{M}\left(\widehat{\theta}^{(m)}-\overline{\widehat{\theta}}\right)^2}{M(M-1)}}$, where $\overline{\widehat{\theta}}=\sum_{m=1}^{M}\frac{\widehat{\theta}^{(m)}}{M}$ and $\widehat{\theta}^{(m)}$ denotes estimate of parameter of interest, θ, obtained from the m^{th} simulation.

Table 1 provides summary statistics using the bootstrap method for various number of bootstrap replications B. Bootstrap estimate of kappa statistic (mean and MCSE of $\widehat{\kappa}_B$), bootstrap error estimate of $\widehat{\kappa}_B$ (mean and MCSE of $\widehat{SE}(\widehat{\kappa}_B)$), and 95% confidence interval coverage rate using the bootstrap confidence interval based on normal approximation $\left(CR_B^S\right)$, the bootstrap confidence interval based on percentiles $\left(CR_B^P\right)$ and the bootstrap confidence interval using the BC_a method $\left(CR_B^{BC_a}\right)$ for $\widehat{\kappa}_B$ are presented. We observed that there was no notable change in $\widehat{\kappa}_B$, $\widehat{SE}(\widehat{\kappa}_B)$, and coverage rates based on the bootstrap confidence interval using normal approximation and percentile methods by increasing the numbers of bootstrap replications. However, coverage rate based on the BC_a method was improved by increasing the number of bootstrap replications until B = 1, 000 which is the minimum number of bootstrap replications suggested by [23]. Therefore, we concluded that 1, 000 is a reasonable number for bootstrap replications, and simulation results with B = 1, 000 are presented for the remaining simulation results.

4.2. Varying strength of agreement between raters, the number of clusters, and cluster size

4.2.1. Simulation set-up and summary statistics—We generated 1, 000 independent data sets, and each data set consists of different numbers of clusters n = (10, 25, 50, 100) and different cluster size m = (5, 20, 50, 100). Poor (κ_0 = 0), fair (0.3), moderate (0.5) and substantial (0.8) strength of agreement associated with kappa statistics between two raters (physician and patient) were investigated. Marginal mean of the dichotomous response for physicians (μ_y) and patients (μ_x) were 0.4 and 0.5, respectively. The correlation coefficient of the responses for physician within a physician (ρ_w) was fixed at 0.3 (exchangeable correlation structure). As we discussed in Section 3.1.3, 0.8 is the maximum possible value for κ_0 for given μ_y = 0.4, μ_x = 0.5 and ρ_w = 0.3.

Table 2 provides the kappa statistics assuming independent observations (mean and MCSE of $\widehat{\kappa}$), asymptotic standard error estimates of the kappa statistics assuming independence (mean and MCSE of $ASE(\widehat{\kappa})$), empirical standard deviations of the kappa statistics using 1, 000 simulations ($std(\widehat{\kappa})$), 95% confidence interval coverage rates using the 95% confidence interval based on normal approximation (CR^S) in addition to summary statistics using bootstrapping on the physicians introduced in Section 4.1.

4.2.2. Simulation result—Figures 1(a) and 1(b) display average of bias of $\widehat{\kappa}$ over 1,000 Monte-carlo simulations. Both methods produced better point estimates of κ with larger number of physicians and number of patients for each physician. No marked difference between the two methods was observed in the point estimates, but kappa statistics assuming

independent observations were slightly closer to κ_0 than those by the bootstrap method over all strength of agreements between raters, especially with small number of physicians (n = 10). This bias is negligible according to [23].

Table 2 and Figures 1(c), 1(d), 1(e), and 1(f) present 95% coverage rates, the average ratio of the $ASE(\hat{\kappa})$ to the empirical standard deviation $std(\hat{\kappa})$ and the average ratio of the bootstrap standard error estimate $\left(\widehat{SE}(\hat{\kappa}_B)\right)$ to the empirical standard deviation for the number of clusters n = (25, 50, 100). Empirical standard deviation was calculated based on 1, 000 estimates of κ, which can be considered as the 'true' standard deviation of kappa statistics. Overall the average ratio of the $ASE(\hat{\kappa})$ to $std(\hat{\kappa})$ decreased rapidly while the average ratio of $\widehat{SE}(\hat{\kappa}_B)$ to $std(\hat{\kappa})$ stayed close to 1 as we increased the strength of the agreement between physician and patient. This indicates that standard error calculation of kappa statistics assuming independent observations tends to underestimate the standard error of kappa statistics on average, particularly when there is a strong agreement between raters. The underestimated standard error of the kappa statistic assuming independent observations resulted in a narrower confidence interval as the strength of agreement between raters increased while the coverage rate based on the bootstrap method was close to the target nominal level (95%) as long as the number of physicians was not very small, even under substantial agreement between raters. The advantage of the cluster bootstrap method was more obvious in the larger number of physicians and larger number of patients for each physician (Figures 1(e) and 1(f)). Overall BC_a method produced confidence interval which is similar to or slightly better than percentile and normal approximation methods except for the case with small number of physicians (n = 10) under no agreement between raters.

4.3. Varying strength of correlation within cluster (physician), the number of clusters, and cluster size

So far, we examined the performance of the two methods to calculate kappa statistics for various strength of agreement between raters while keeping within-cluster correlation fixed. In this section, we varied values of the correlation coefficient within cluster (ρ_w) under fixed moderate strength of agreement between raters (κ_0 = 0.5). Table 3 and Figure 2 provide the simulation results for ρ_w = 0.1, 0.3, 0.5 and 0.8. Calculating ASE of the kappa statistic assuming independent observations tended to underestimate the standard error on average for at least moderately positively correlated subjects within physician while the clustered bootstrap method did not. This pattern is more obvious in Figure 2. Similar to the result by varying strength of agreement between raters, however, the bootstrap method produced poorer coverage rates than those obtained by using ASE assuming independent observations when both the number of physicians and the number of patients for each physician were very small under weak within-physician correlation.

5. Example

5.1. Data Description

Clustered data structure occurred in our motivating example in which investigators were interested in the agreement between physicians and their patients regarding the content of discussions about CHD during a clinic visit. Our study consisted of subset of physician/patient pairs from a larger randomized trial. This larger study [32] compared control patients who received usual care with intervention patients who received a computerized decision aid regarding heart disease prevention followed by several adherence reminders. The patients in the larger study were seen for three visits and evaluated at the third visit for predicted CHD risk and adherence.

Our nested study focused on the discussions between 24 physicians and their 157 patients during the second visit. Complete eligibility criteria, described elsewhere [14], included no history of CHD or diabetes but at least a moderate risk of developing CHD within the next 10 years as predicted by risk factors such as age, gender, smoking status, diabetes status, presence/absence of enlarged heart, blood pressure, total cholesterol and LDL using Framingham risk calculator. Among the physician/patient pairs who participated in the larger study and also eligible for the sub-study, almost all participated in this smaller study (100% of physicians and 97% of patients).

Following the second clinic visit, each physician and his/her patient was surveyed regarding the content of their discussion during that visit. Patients were surveyed immediately following the visit, and the majority of the physicians completed the survey on the same day as the clinic visit. Items of interest were the agreement between the patient and physician about whether CHD was discussed, whether the physician recommended taking medicine, and whether the physician recommended changing lifestyle. For each of these binary measures, the kappa statistic was computed assuming independence. Because each physician could have more than one patient (1-20 patients per physician), the cluster bootstrap method was used to account for the correlated data structure in calculating the standard error for the kappa statistic. Physicians were sampled with replacement, and all patients corresponding to the selected physician were included. Tables 4, 5, and 6 provide cell counts of the 2 × 2 table for three topics: Discussed CHD, MD recommended medication, and MD recommended lifestyle change.

5.2. Analysis Results

Table 7 presents the kappa statistics to describe agreement between physician and patient regarding CHD discussion for both the method assuming independence and the cluster bootstrap method as we presented in Section 4. Of the 157 discussions, 103 physician/patient pairs agreed that CHD had been discussed and 27 pairs agreed that CHD was not discussed, resulting in a moderate agreement between physician and patient. The 130 discussions in which at least one of the physician/patient members reported that CHD was discussed were further analyzed. Moderate agreement was found in whether the physician recommended taking medicine and fair agreement in whether the physician recommended lifestyle modification. We note that the SEs associated with the kappa statistic for “Discussed CHD” and “MD recommended medication” are smaller based on the bootstrap method than those based on the kappa method assuming independent patients while the SE for “MD recommended lifestyle change” shows greater SE using the bootstrap method. The kappa statistic is a function of the first two moments, hence its asymptotic standard error involves third- and fourth-order moments which could make the bootstrap standard error larger or smaller than those assuming independence.

5.3. Additional simulation to understand data analysis result

To further demonstrate the possibility to have smaller bootstrap standard error than ASE in a particular data analysis, we conducted a small investigation through simulation in which data were generated mimicking “Discussed CHD” data in terms of the marginal mean of the dichotomous responses (the marginal mean responses for physicians and patients were 0.752 and 0.732, respectively), kappa statistic (0.55) and heterogeneous number of patients per physician. The number of physicians and average number of patients per physician in the data are 24 and 6.5, respectively. We compared standard error estimates and coverage rates by varying the number of physician (24 or 72), the average number of patients per physician (6.5 and 13) and within-physician correlation (ρ_W = 0.1, 0.3, 0.7).

Table 8 summarizes results of the simulation study mimicking “Discussed CHD” data. With the same number of physician ($n = 24$) and the average number of patients per physician ($m = 6.5$) with “Discussed CHD” data, the reduction in estimating standard error of the kappa statistic by using the cluster bootstrap method was slightly greater on average than ASE assuming independent observations under weak within-physician correlation. The bootstrap method improved upon ASE of kappa statistic assuming independent observations by increasing the number of physicians or within-physician correlation rather than the number of patients per physician. 53.7%, 41.9%, and 19.9% of simulation results showed greater ASE than the bootstrap SE for strong ($\rho_w = 0.7$), fair (0.3), and weak (0.1) within-physician correlation, respectively. This simulation result shows the possibility that bootstrap standard error estimates of kappa statistic can be smaller than ASE of kappa statistics assuming independence for a particular data analysis, especially for the data set with small number of clusters and small cluster size.

6. Discussion

In this article, our study focuses on the evaluation of the bootstrap method to calculate kappa statistic and its standard error for clustered dichotomous responses. For the simulation study, we adopted a computationally efficient procedure to generate correlated dichotomous responses for physicians and their patients. Through simulation studies, we have demonstrated that the asymptotic standard error of the kappa statistic assuming independent observations tends to underestimate the standard error of the kappa statistic on average, particularly when there is a strong agreement between physician and patient or a strong within-physician correlation. This underestimation yields confidence intervals that are too narrow and have poor coverage performance. The proposed bootstrap method produced nearly unbiased standard error estimates and coverage rates that are close to the target nominal level by taking account of correlation within physicians except for the case when both the number of physicians and the number of patients per physician are small.

Bootstrap methods could be computationally intensive. One alternative to the bootstrap method is to use asymptotic approximation. To do that, one needs to establish the asymptotic distribution of the kappa statistic with clustered data. The most difficult part would be to determine the asymptotic variance of the kappa statistic taking into account the correlation within clusters. In the case of independent samples, formulas for the variance of the kappa statistic in situations with different sets of raters [33] or with small number of subjects [34] have been proposed. Feder [35] proposed a Taylor linearization to estimate the variance of kappa statistic when each individual is interviewed on two occasions. The kappa statistic is a function of the first two moments, hence its asymptotic standard error involves third- and fourth-order moments, which are not easily modeled or estimated. The bootstrap offers an automatic way of adjusting the standard errors without explicit modeling of high-order moments. We will explore developing variance formulae for the kappa statistic in clustered data setups in future research.

Acknowledgments

We are grateful to the editors and the reviewers for their insightful comments which have led to important improvements in the paper. This research was partially supported by the National Institutes of Health grants: UL1 RR025747, R01 HL57444, P01 CA142538, and K23 HL074375.

Appendix1: 1. Marginal correlation between responses from two patients of the same physician

The marginal correlation between responses of any two patients of a physician can be expressed as follows:

$$Corr\left(X_i, X_j\right) = \frac{E\left(X_i X_j\right) - \mu_x^2}{\mu_x\left(1-\mu_x\right)},$$

and

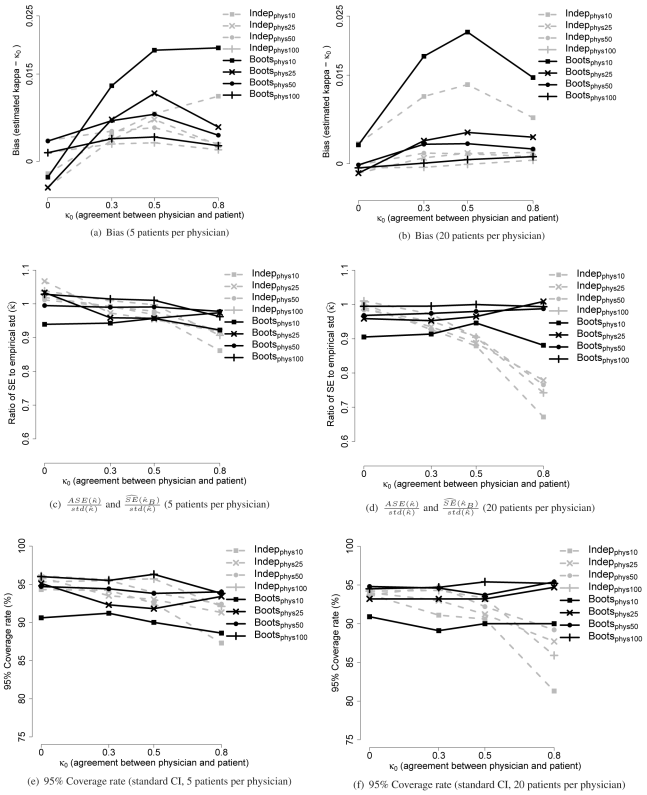
$$\begin{aligned} E\left(X_i X_j\right) &= E\left[E\left[X_i X_j | Y_i, Y_j\right]\right] \\ &= E\left[\left(b_0 + b_1 Y_i\right)\left(b_0 + b_1 Y_j\right)\right] \\ &= b_0^2 + 2 b_0 b_1 E\left(Y_i\right) + b_1^2 E\left(Y_i Y_j\right) \\ &= b_0^2 + 2 b_0 b_1 \mu_y + b_1^2 d. \end{aligned} \tag{8}$$

The second equation holds because of the assumption that $X_i \perp X_j | Y_i, Y_j$. Given marginal means μ_y, and μ_x, correlation within physicians ρ_w and correlation between physician's and patient's responses ρ_b (or equivalently κ), the marginal correlation between any two patients within physicians is automatically determined by the above relationship. These marginal correlations generally increase with ρ_w and ρ_b. The within-physician patient-patient correlation is approximately $\rho_w \rho_b^2$.

References

1. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1):37–46.
2. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70(4):213–220. [PubMed: 19673146]
3. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]
4. Fleiss J. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76(5):378–382.
5. Barlow W, Lai M, Azen S. A comparison of methods for calculating a stratified kappa. Statistics in Medicine. 1991; 10(9):1465–1472. [PubMed: 1925174]
6. Kraemer H. Extension of the kappa coefficient. Biometrics. 1980; 36(2):207–216. [PubMed: 7190852]
7. McKenzie D, Mackinnon A, Péladeau N, Onghena P, Bruce P, Clarke D, Harrigan S, McGorry P. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? Journal of Psychiatric Research. 1996; 30(6):483–492. [PubMed: 9023792]
8. Vanbelle S, Albert A. A bootstrap method for comparing correlated kappa coefficients. Journal of Statistical Computation and Simulation. 2008; 78(11):1009–1015.
9. Donner A, Shoukri M, Klar N, Bartfay E. Testing the equality of two dependent kappa statistics. Statistics in Medicine. 2000; 19(3):373–387. [PubMed: 10649303]
10. Klar N, Lipsitz S, Ibrahim J. An estimating equations approach for modelling kappa. Biometrical Journal. 2000; 42(1):45–58.
11. Williamson J, Manatunga A, Lipsitz S. Modeling kappa for measuring dependent categorical agreement data. Biostatistics. 2000; 1(2):191–202. [PubMed: 12933519]
12. Gonin R, Lipsitz S, Fitzmaurice G, Molenberghs G. Regression modelling of weighted κ by using generalized estimating equations. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2000; 49(1):1–18.

13. Barnhart H, Williamson J. Weighted Least-Squares Approach for Comparing Correlated Kappa. Biometrics. 2002; 58(4):1012–1019. [PubMed: 12495157]
14. Behrend L, Maymani H, Diehl M, Gizlice Z, Cai J, Sheridan S. Patient–physician agreement on the content of CHD prevention discussions. Health Expectations. 2010; 14(1):58–72. [PubMed: 20673244]
15. Eliasziw M, Donner A. Application of the McNemar test to non-independent matched pair data. Statistics in Medicine. 1991; 10(12):1981–1991. [PubMed: 1805322]
16. Donald A, Donner A. Adjustments to the Mantel–Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. Statistics in Medicine. 1987; 6(4):491–499. [PubMed: 3629050]
17. Donald A, Donner A. A simulation study of the analysis of sets of 2 × 2 contingency tables under cluster sampling: Estimation of a common odds ratio. Journal of the American Statistical Association. 1990; 85(410):537–543.
18. Oden N. Estimating kappa from binocular data. Statistics in Medicine. 1991; 10(8):1303–1311. [PubMed: 1925161]
19. Schouten H. Estimating kappa from binocular data and comparing marginal probabilities. Statistics in Medicine. 1993; 12(23):2207–2217. [PubMed: 8310190]
20. Klar N, Lipsitz S, Parzen M, Leong T. An exact bootstrap confidence interval for κ in small samples. Journal of the Royal Statistical Society: Series D (The Statistician). 2002; 51(4):467–478.
21. Cunningham, M. Statistics and Data Analysis. SAS Global Forum; 2009. More than Just the Kappa Coefficient: A Program to Fully Characterize Inter-Rater Reliability between Two Raters.
22. Fleiss, J.; Levin, B.; Paik, M.; Wiley, J. Statistical Methods for Rates and Proportions. 3rd Edition. John Wiley & Sons, Inc.; Hoboken, New Jersey: 2003.
23. Efron, B.; Tibshirani, R.; Tibshirani, R. An Introduction to the Bootstrap. Chapman & Hall/CRC; 1993.
24. Davison, A.; Hinkley, D. Bootstrap Methods and their Application. Cambridge Univ Pr; 1997.
25. Field C, Welsh A. Bootstrapping clustered data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69(3):369–390.
26. Monaco J, Cai J, Grizzle J. Bootstrap analysis of multivariate failure time data. Statistics in medicine. 2005; 24(22):3387–3400. [PubMed: 16237657]
27. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1986; 1(1):54–75.
28. DiCiccio T, Efron B. Bootstrap confidence intervals. Statistical Science. 1996:189–212.
29. Qaqish B. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika. 2003; 90(2):455–463.
30. By, K.; Qaqish, B. binarySimCLF: Simulates Correlated Binary Data. version 1.02009. URL http://CRAN.R-project.org/package=binarySimCLFr package
31. Burton A, Altman D, Royston P, Holder R. The design of simulation studies in medical statistics. Statistics in medicine. 2006; 25(24):4279–4292. [PubMed: 16947139]
32. Sheridan S, Behrend L, Pignone M, Keyserling T, Rimer B, Simpson R, Bangdiwala K, Cai J, Gizlice Z. A randomized trial of an intervention to improve adherence to effective heart disease prevention medications. BMC Health Services Research. 2011; 11(331)
33. Fleiss J, Nee J, Landis J. Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin. 1979; 86(5):974.
34. Gross S. The kappa coefficient of agreement for multiple observers when the number of subjects is small. Biometrics. 1986; 42(4):883–893. [PubMed: 3814729]
35. Feder, M. Variance estimation of the survey-weighted kappa measure of agreement. ASA Section on Survey Research Methods. The Joint Statistical Meetings; Seattle, WA. August, 2006;



(a) Bias (5 patients per physician)

(b) Bias (20 patients per physician)

(c) $\frac{ASE(\hat{\kappa})}{std(\hat{\kappa})}$ and $\frac{\overline{SE}(\hat{\kappa}_B)}{std(\hat{\kappa})}$ (5 patients per physician)

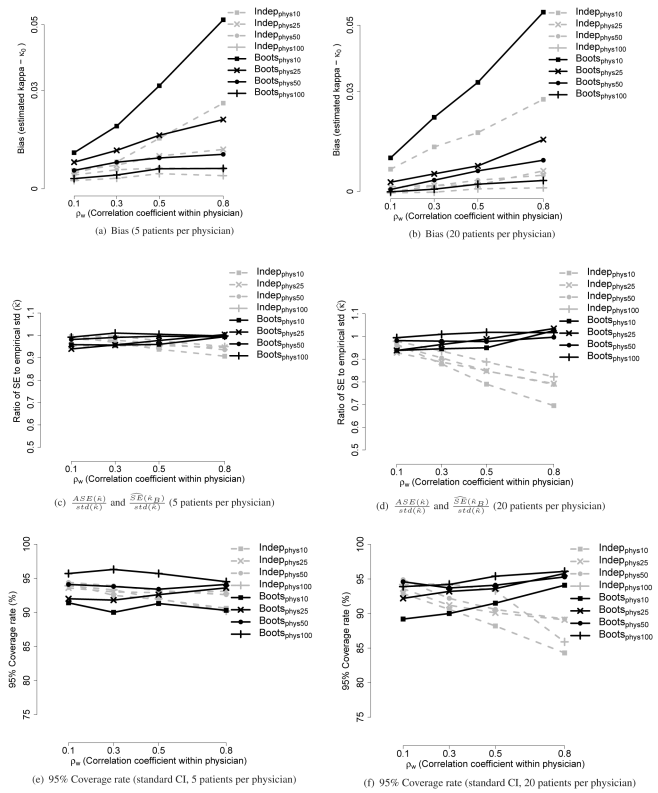
(d) $\frac{ASE(\hat{\kappa})}{std(\hat{\kappa})}$ and $\frac{\overline{SE}(\hat{\kappa}_B)}{std(\hat{\kappa})}$ (20 patients per physician)

(e) 95% Coverage rate (standard CI, 5 patients per physician)

(f) 95% Coverage rate (standard CI, 20 patients per physician)

Figure 1.

Comparisons between $\hat{\kappa}$ assuming independent observations (denoted by $\text{Indep}_{\#\text{ of physicians}}$) and $\hat{\kappa}_B$ using the cluster bootstrap method (denoted by $\text{Boots}_{\#\text{ of physicians}}$) for the numbers of physicians=(10,25,50,100) at different κ_0 values (strength of agreement between physician and patient).



(a) Bias (5 patients per physician) (b) Bias (20 patients per physician)

(c) $\frac{ASE(\hat{\kappa})}{std(\hat{\kappa})}$ and $\frac{\widehat{SE}(\hat{\kappa}_B)}{std(\hat{\kappa})}$ (5 patients per physician) (d) $\frac{ASE(\hat{\kappa})}{std(\hat{\kappa})}$ and $\frac{\widehat{SE}(\hat{\kappa}_B)}{std(\hat{\kappa})}$ (20 patients per physician)

(e) 95% Coverage rate (standard CI, 5 patients per physician) (f) 95% Coverage rate (standard CI, 20 patients per physician)

Figure 2.

Comparisons between $\hat{\kappa}$ assuming independent observations (denoted by $\text{Indep}_{\#\text{ of physicians}}$) and $\hat{\kappa}_B$ using the cluster bootstrap method (denoted by $\text{Boots}_{\#\text{ of physicians}}$) for the numbers of physicians=(10,25,50,100) at different ρ_W values (within-physician correlation coefficient).

Table 1

Simulation results for determining the number of bootstrap replications B (μ_Y = 0.4, μ_X = 0.5, ρ_W = 0.3, κ_0 = 0.5, the number of cluster= 25, and cluster size= 20)

Bootstrap B		$\hat{\kappa}_B$	$\widehat{SE}(\hat{\kappa}_B)$	95% confidence interval coverage rate		
				CR_B^S	CR_B^P	$CR_B^{BC_a}$
50	Mean	0.4947	0.0407	91.9	92.7	90.4
	MCSE	0.0014	0.0002			
100	Mean	0.4946	0.0408	92.6	93.2	91.8
	MCSE	0.0014	0.0002			
200	Mean	0.4947	0.0410	92.9	92.2	92.5
	MCSE	0.0014	0.0002			
300	Mean	0.4947	0.0411	93.1	92.3	92.7
	MCSE	0.0014	0.0002			
500	Mean	0.4947	0.0412	93.2	92.9	92.6
	MCSE	0.0014	0.0002			
1000	Mean	0.4947	0.0412	93.2	93.0	93.1
	MCSE	0.0014	0.0002			
1500	Mean	0.4948	0.0411	93.2	93.0	93.5
	MCSE	0.0013	0.0002			
2000	Mean	0.4947	0.0411	93.2	93.2	93.1
	MCSE	0.0013	0.0002			

Note: Mean of $\hat{\kappa}_B$ denotes the average bootstrap kappa statistic.

MCSE of $\hat{\kappa}_B$ denotes the Monte-Carlo Standard Error estimate of the bootstrap kappa statistic.

Mean of $\widehat{SE}(\hat{\kappa}_B)$ denotes the average bootstrap standard error estimate of bootstrap kappa statistic.

MCSE of $\widehat{SE}(\hat{\kappa}_B)$ denotes the Monte-Carlo Standard Error estimate of $\widehat{SE}(\hat{\kappa}_B)$.

CR_B^S denotes the 95% confidence interval coverage rate (%) using the 95% confidence interval based on normal approximation.

CR_B^P denotes the 95% confidence interval coverage rate (%) using the 95% confidence interval based on percentiles.

$CR_B^{BC_a}$ denotes the 95% confidence interval coverage rate (%) using the 95% confidence interval based on BC_a.

Table 2

Summary of the simulation results for comparing the bootstrap method and the method assuming independent observations with various levels of agreement between physician and patient for $\mu_y = 0.4$, $\mu_x = 0.5$, and $\rho_w = 0.3$.

Setting			kappa assuming independent observations						kappa using the bootstrap method						
# of physicians	# of patients	κ_0	$\hat{\kappa}$		$ASE(\hat{\kappa})$		$std(\hat{\kappa})$	CR^S	$\hat{\kappa}_B$		$\widehat{SE}(\hat{\kappa}_B)$		CR_B^S	CR_B^P	$CR_B^{BC_a}$
			Mean	MCSE	Mean	MCSE			Mean	MCSE	Mean	MCSE			
25	5	0.0	0.004	0.0026	0.086	0.000107	0.081	95.9	0.004	0.0025	0.083	0.000362	95.1	94.8	94.2
		0.3	0.296	0.0027	0.083	0.000105	0.085	93.5	0.293	0.0027	0.081	0.000346	92.3	91.9	92.2
		0.5	0.493	0.0025	0.076	0.000127	0.080	93.0	0.488	0.0025	0.076	0.000331	91.8	92.5	93.2
		0.8	0.797	0.0018	0.053	0.000198	0.057	91.3	0.794	0.0018	0.056	0.000316	93.4	93.7	94.2
	20	0.0	0.002	0.0014	0.043	0.000037	0.044	94.0	0.002	0.0014	0.042	0.000202	93.2	92.4	91.6
		0.3	0.299	0.0014	0.042	0.000028	0.045	93.0	0.296	0.0014	0.042	0.000186	93.2	93.4	93.3
		0.5	0.498	0.0013	0.038	0.000032	0.043	91.2	0.495	0.0014	0.041	0.000189	93.2	93.0	93.1
		0.8	0.798	0.0011	0.026	0.000062	0.034	87.7	0.796	0.0011	0.034	0.000200	94.7	94.6	94.4
50	5	0.0	−0.004	0.0019	0.061	0.000049	0.061	95.3	−0.004	0.0019	0.060	0.000194	94.7	94.1	93.6
		0.3	0.295	0.0019	0.059	0.000050	0.059	95.5	0.293	0.0019	0.059	0.000184	94.4	94.3	94.2
		0.5	0.494	0.0017	0.054	0.000061	0.055	93.9	0.492	0.0017	0.054	0.000172	93.8	93.8	93.7
		0.8	0.797	0.0013	0.037	0.000098	0.041	92.4	0.795	0.0013	0.040	0.000168	94.0	94.2	94.0
	20	0.0	0.000	0.0010	0.031	0.000018	0.031	93.7	0.000	0.0010	0.030	0.000101	94.8	94.4	93.5
		0.3	0.298	0.0010	0.030	0.000012	0.031	94.8	0.297	0.0010	0.030	0.000095	94.6	94.6	95.0
		0.5	0.498	0.0009	0.027	0.000016	0.030	92.2	0.497	0.0009	0.029	0.000098	93.7	93.7	93.7
		0.8	0.799	0.0008	0.019	0.000031	0.024	89.2	0.798	0.0008	0.024	0.000097	95.4	95.0	95.2
100	5	0.0	−0.002	0.0013	0.044	0.000024	0.042	96.1	−0.001	0.0013	0.043	0.000098	96.0	95.9	95.9
		0.3	0.297	0.0013	0.042	0.000023	0.041	95.6	0.296	0.0013	0.042	0.000091	95.5	95.2	95.0
		0.5	0.497	0.0012	0.038	0.000029	0.038	95.7	0.496	0.0012	0.039	0.000084	96.3	96.1	96.4
		0.8	0.798	0.0009	0.026	0.000048	0.029	92.1	0.797	0.0009	0.028	0.000086	93.8	93.6	93.9
	20	0.0	0.001	0.0006	0.020	0.000008	0.019	94.2	0.001	0.0006	0.019	0.000045	94.5	94.6	94.1
		0.3	0.301	0.0007	0.021	0.000006	0.022	94.3	0.300	0.0007	0.021	0.000049	94.7	94.2	93.7
		0.5	0.500	0.0006	0.017	0.000006	0.019	93.3	0.499	0.0006	0.019	0.000046	95.4	95.4	94.9

Setting			kappa assuming independent observations						kappa using the bootstrap method						
# of physicians	# of patients	κ_0	$\hat{\kappa}$		$ASE(\hat{\kappa})$		$std(\hat{\kappa})$	CR^S	$\hat{\kappa}_B$		$\widehat{SE}(\hat{\kappa}_B)$		CR_B^S	CR_B^P	$CR_B^{BC_a}$
			Mean	MCSE	Mean	MCSE			Mean	MCSE	Mean	MCSE			
		0.8	0.799	0.0005	0.012	0.000013	0.016	85.9	0.799	0.0005	0.016	0.000045	95.2	94.8	94.5

Note: κ_0 denotes strength of agreement between physician and patient.

Mean (MCSE) of $\hat{\kappa}$ denotes the average (Monte-Carlo Standard Error estimate of) kappa statistic assuming independence.

Mean (MCSE) of $ASE(\hat{\kappa})$ denotes the average (Monte-Carlo Standard Error estimate of) asymptotic standard error of kappa statistic assuming independence.

$std(\hat{\kappa})$ denotes the empirical standard deviation of kappa statistics.

CR^S denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}$ using the 95% confidence interval based on normal approximation assuming independence.

Mean (MCSE) of $\hat{\kappa}_B$ denotes the average (Monte-Carlo Standard Error estimate of) bootstrap kappa statistic.

Mean (MCSE) of $\widehat{SE}(\hat{\kappa}_B)$ denotes the average (Monte-Carlo Standard Error estimate of) bootstrap standard error estimate of bootstrap kappa statistic.

CR_B^S denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on normal approximation.

CR_B^P denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on percentiles.

$CR_B^{BC_a}$ denotes the 95% confidence interval coverage (%) rate for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on BC_a.

Stat Med. Author manuscript; available in PMC 2014 September 20.

Table 3

Summary of the simulation results for comparing the bootstrap method and the method assuming independent observations with various levels of correlation coefficient between any two responses of a physician ($\mu_y = 0.4$, $\mu_x = 0.5$, $r_b = 0.29$, and $\kappa_0 = 0.5$).

Setting			kappa assuming independent observations						kappa using the bootstrap method						
# of physicians	# of patients	ρ_w	$\hat{\kappa}$		$ASE(\hat{\kappa})$		$std(\hat{\kappa})$	CR^S	$\hat{\kappa}_B$		$\widehat{SE}(\hat{\kappa}_B)$		CR_B^S	CR_B^P	$CR_B^{BC_a}$
			Mean	MCSE	Mean	MCSE			Mean	MCSE	Mean	MCSE			
25	5	0.1	0.495	0.0025	0.076	0.000127	0.079	93.6	0.492	0.0025	0.075	0.000326	92.0	92.6	92.9
		0.3	0.493	0.0025	0.076	0.000127	0.080	93.0	0.488	0.0025	0.076	0.000331	91.8	92.5	93.2
		0.5	0.490	0.0025	0.076	0.000126	0.079	92.9	0.484	0.0025	0.078	0.000336	92.6	93.4	93.6
		0.8	0.488	0.0025	0.076	0.000128	0.080	93.1	0.479	0.0025	0.081	0.000358	93.6	93.1	94.2
	20	0.1	0.499	0.0013	0.038	0.000032	0.041	93.6	0.497	0.0013	0.038	0.000181	92.2	91.8	92.1
		0.3	0.498	0.0013	0.038	0.000032	0.043	91.2	0.495	0.0014	0.041	0.000189	93.2	93.0	93.1
		0.5	0.498	0.0014	0.038	0.000033	0.045	90.1	0.492	0.0014	0.044	0.000231	93.6	93.1	92.5
		0.8	0.494	0.0015	0.038	0.000034	0.048	89.1	0.484	0.0015	0.050	0.000336	95.8	94.4	94.1
50	5	0.1	0.496	0.0017	0.054	0.000061	0.054	94.4	0.494	0.0017	0.054	0.000164	94.1	93.6	93.8
		0.3	0.494	0.0017	0.054	0.000061	0.055	93.9	0.492	0.0017	0.054	0.000172	93.8	93.8	93.7
		0.5	0.494	0.0018	0.054	0.000062	0.056	93.2	0.491	0.0018	0.055	0.000169	93.4	93.5	94.2
		0.8	0.494	0.0018	0.054	0.000062	0.057	92.6	0.489	0.0018	0.057	0.000181	94.1	93.6	93.7
	20	0.1	0.500	0.0009	0.027	0.000015	0.028	94.9	0.499	0.0009	0.027	0.000090	94.6	94.6	95.0
		0.3	0.498	0.0009	0.027	0.000016	0.030	92.2	0.497	0.0009	0.029	0.000098	93.7	93.7	93.7
		0.5	0.497	0.0010	0.027	0.000016	0.032	90.6	0.494	0.0010	0.031	0.000115	94.1	94.0	94.2
		0.8	0.495	0.0011	0.027	0.000017	0.034	89.2	0.491	0.0011	0.034	0.000168	95.3	94.0	94.5
100	5	0.1	0.498	0.0012	0.038	0.000030	0.038	95.4	0.497	0.0012	0.038	0.000085	95.7	95.6	95.3
		0.3	0.497	0.0012	0.038	0.000029	0.038	95.7	0.496	0.0012	0.039	0.000084	96.3	96.1	96.4
		0.5	0.495	0.0012	0.038	0.000029	0.039	94.2	0.494	0.0012	0.039	0.000088	95.7	94.9	95.3
		0.8	0.496	0.0013	0.038	0.000030	0.040	93.6	0.494	0.0013	0.040	0.000094	94.5	94.3	94.8
	20	0.1	0.500	0.0006	0.019	0.000007	0.019	93.9	0.500	0.0006	0.019	0.000045	93.9	94.0	94.0
		0.3	0.500	0.0006	0.019	0.000008	0.020	93.3	0.499	0.0006	0.021	0.000049	94.2	94.2	94.2
		0.5	0.499	0.0007	0.019	0.000008	0.021	92.0	0.498	0.0007	0.022	0.000058	95.4	95.1	94.9

Setting			kappa assuming independent observations						kappa using the bootstrap method						
# of physicians	# of patients	ρ_w	$\hat{\kappa}$		$ASE(\hat{\kappa})$		$std(\hat{\kappa})$	CR^S	$\hat{\kappa}_B$		$\widehat{SE}(\hat{\kappa}_B)$		CR_B^S	CR_B^P	$CR_B^{BC_a}$
			Mean	MCSE	Mean	MCSE			Mean	MCSE	Mean	MCSE			
		0.8	0.499	0.0007	0.019	0.000008	0.023	90.2	0.497	0.0007	0.024	0.000081	96.1	95.2	95.7

Note: ρ_W denotes the correlation coefficient within a physician.

Mean (MCSE) of $\hat{\kappa}$ denotes the average (Monte-Carlo Standard Error estimate of) kappa statistic assuming independence.

Mean (MCSE) of $ASE(\hat{\kappa})$ denotes the average (Monte-Carlo Standard Error estimate of) asymptotic standard error of kappa statistic assuming independence.

$std(\hat{\kappa})$ denotes the empirical standard deviation of kappa statistics.

CR^S denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}$ using the 95% confidence interval based on normal approximation assuming independence.

Mean (MCSE) of $\hat{\kappa}_B$ denotes the average (Monte-Carlo Standard Error estimate of) bootstrap kappa statistic.

Mean (MCSE) of $\widehat{SE}(\hat{\kappa}_B)$ denotes the average (Monte-Carlo Standard Error estimate of) bootstrap standard error estimate of bootstrap kappa statistic.

CR_B^S denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on normal approximation.

CR_B^P denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on percentiles.

$CR_B^{BC_a}$ denotes the 95% confidence interval coverage (%) rate for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on BC_a.

Stat Med. Author manuscript; available in PMC 2014 September 20.

Table 4

Cell counts of the 2 × 2 table for Discussed CHD

		Patient		
		No	**Yes**	
Physician	No	27 (0.172)	12 (0.076)	39 (0.248)
	Yes	15 (0.096)	103 (0.656)	118 (0.752)
		42 (0.268)	115 (0.732)	157

Table 5

Cell counts of the 2 × 2 table for MD recommended medication

		Patient		
		No	**Yes**	
Physician	No	29 (0.223)	19 (0.146)	48 (0.369)
	Yes	17 (0.131)	65 (0.500)	82 (0.631)
		46 (0.354)	84 (0.646)	130

Table 6

Cell counts of the 2 × 2 table for MD recommended lifestyle change

		Patient		
		No	**Yes**	
Physician	No	51 (0.392)	15 (0.115)	66 (0.508)
	Yes	18 (0.139)	46 (0.354)	64 (0.492)
		69 (0.531)	61 (0.469)	130

Table 7

Agreement between physician and patient regarding CHD discussions.

Topic	kappa assuming independent observations			kappa using the bootstrap method				
	$\hat{\kappa}$	$ASE(\hat{\kappa})$	CI^S	$\hat{\kappa}_B$	$\widehat{SE}(\hat{\kappa}_B)$	CR_B^S	CR_B^P	$CR_B^{BC_a}$
Discussed CHD	0.551	0.076	(0.402, 0.700)	0.545	0.053	(0.441, 0.650)	(0.432, 0.648)	(0.440, 0.654)
MD recommended medication	0.400	0.083	(0.237, 0.563)	0.405	0.067	(0.275, 0.536)	(0.287, 0.541)	(0.284, 0.536)
MD recommended lifestyle change	0.492	0.076	(0.342, 0.641)	0.485	0.091	(0.305, 0.664)	(0.303 ,0.662)	(0.322 ,0.680)

Note: $\hat{\kappa}$ denotes the kappa statistic assuming independence.

$ASE(\hat{\kappa})$ denotes the asymptotic standard error of kappa statistics assuming independence.

CI^S denotes the 95% confidence interval based on normal approximation assuming independence.

$\hat{\kappa}_B$ denotes the bootstrap kappa statistic.

$\widehat{SE}(\hat{\kappa}_B)$ denotes the bootstrap standard error estimate of bootstrap kappa statistic.

CI_B^S denotes the 95% bootstrap confidence interval based on normal approximation.

CI_B^P denotes the 95% bootstrap confidence interval based on percentiles.

$CI_B^{BC_a}$ denotes the 95% bootstrap confidence interval based on BC_a.

The number of bootstrap replication B=1,000.

Analysis was performed using R (Version 2.15.1).

Table 8

Summary of the simulation results mimicking "Discussed CHD" data with various the number of clusters, cluster size and within cluster correlation under $\mu_Y = 0.752$, $\mu_X = 0.732$, and $\kappa_0 = 0.55$.

Setting			kappa assuming independent observations						kappa using the bootstrap method						
# of physicians	AVG # of patients	ρ_w	$\hat{\kappa}$		$ASE(\hat{\kappa})$		$std(\hat{\kappa})$	CR^S	$\hat{\kappa}_B$		$\widehat{SE}(\hat{\kappa}_B)$		CR_B^S	CR_B^P	$CR_B^{BC_a}$
			Mean	MCSE	Mean	MCSE			Mean	MCSE	Mean	MCSE			
24	6.5	0.1	0.546	0.0025	0.0760	0.000209	0.0781	94.1	0.541	0.0025	0.0758	0.000504	92.8	91.8	92.1
		0.3	0.537	0.0026	0.0771	0.000271	0.0836	93.7	0.527	0.0026	0.0822	0.000608	92.6	92.0	92.1
		0.7	0.523	0.0032	0.0783	0.000363	0.0990	89.7	0.499	0.0031	0.0984	0.000912	92.6	89.7	91.7
	13	0.1	0.543	0.0018	0.0542	0.000121	0.0567	93.6	0.539	0.0018	0.0545	0.000368	92.5	93.1	92.8
		0.3	0.536	0.0019	0.0547	0.000166	0.0616	91.9	0.526	0.0019	0.0621	0.000522	92.8	91.6	91.5
		0.7	0.521	0.0026	0.0558	0.000249	0.0002	85.7	0.498	0.0026	0.0830	0.000879	92.8	88.3	90.8
72	6.5	0.1	0.548	0.0014	0.0440	0.000070	0.0449	94.5	0.546	0.0014	0.0444	0.000185	94.2	94.3	93.8
		0.3	0.546	0.0015	0.0442	0.000090	0.0478	93.8	0.542	0.0015	0.0466	0.000230	94.2	94.2	94.1
		0.7	0.540	0.0017	0.0446	0.000122	0.0529	91.4	0.532	0.0017	0.0526	0.000356	93.7	92.4	93.0
	13	0.1	0.549	0.0010	0.0312	0.000040	0.0324	93.7	0.547	0.0010	0.0320	0.000131	94.5	94.6	94.0
		0.3	0.545	0.0011	0.0314	0.000056	0.0357	92.0	0.542	0.0011	0.0352	0.000182	94.7	93.4	94.5
		0.7	0.539	0.0014	0.0317	0.000081	0.0436	87.7	0.531	0.0014	0.0436	0.000351	96.6	93.0	95.6

Note: ρ_W denotes the correlation coefficient within a physician.

Mean (MCSE) of $\hat{\kappa}$ denotes the average (Monte-Carlo Standard Error estimate of) kappa statistic assuming independence.

Mean (MCSE) of $ASE(\hat{\kappa})$ denotes the average (Monte-Carlo Standard Error estimate of) asymptotic standard error of kappa statistic assuming independence.

$std(\hat{\kappa})$ denotes the empirical standard deviation of kappa statistics.

CR^S denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}$ using the 95% confidence interval based on normal approximation assuming independence.

Mean (MCSE) of $\hat{\kappa}_B$ denotes the average (Monte-Carlo Standard Error estimate of) bootstrap kappa statistic.

Mean (MCSE) of $\widehat{SE}(\hat{\kappa}_B)$ denotes the average (Monte-Carlo Standard Error estimate of) bootstrap standard error estimate of bootstrap kappa statistic.

CR_B^S denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on normal approximation.

CR_B^P denotes the 95% confidence interval coverage rate (%) for $\hat{\kappa}_B$ using the 95% bootstrap confidence interval based on percentiles.

$CR_B^{BC_a}$ denotes the 95% confidence interval coverage (%) rate for $\widehat{\kappa}_B$ using the 95% bootstrap confidence interval based on BC_a.