HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2019 June 21.

Published in final edited form as:

Stat Med. 2018 December 10; 37(28): 4200–4215. doi:10.1002/sim.7912.

Overall indices for assessing agreement among multiple raters

Jeong Hoon Jang[1], **Amita K. Manatunga**[1], **Andrew T. Taylor**[2], and **Qi Long**[3]

[1]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia

[2]Department of Radiology and Imaging Sciences, School of Medicine, Emory University, Atlanta, Georgia

[3]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

Abstract

The need to assess agreement exists in various clinical studies where quantifying inter-rater reliability is of great importance. Use of unscaled agreement indices, such as total deviation index and coverage probability (CP), is recommended for two main reasons: (i) they are intuitive in a sense that interpretations are tied to the original measurement unit; (ii) practitioners can readily determine whether the agreement is satisfactory by directly comparing the value of the index to a prespecified tolerable CP or absolute difference. However, the unscaled indices were only defined in the context of comparing two raters or multiple raters that assume homogeneity of variances across raters. In this paper, we introduce a set of overall indices based on the root mean square of pairwise differences that are unscaled and can be used to evaluate agreement among multiple raters that often exhibit heterogeneous measurement processes in practice. Furthermore, we propose another overall agreement index based on the root mean square of pairwise differences that is scaled and extends the concept of the recently proposed relative area under CP curve in the presence of multiple raters. We present the definitions of overall indices and propose inference procedures in which bootstrap methods are used for the estimation of standard errors. We assess the performance of the proposed approach and demonstrate its superiority over the existing methods when raters exhibit heterogeneous measurement processes using simulation studies. Finally, we demonstrate the application of our methods using a renal study.

Keywords

agreement; coverage probability; multiple raters; root-mean-square difference; unscaled index

Correspondence: Amita K. Manatunga, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322. amanatu@emory.edu.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

1 | INTRODUCTION

In various clinical studies, researchers are often interested in assessing agreement between clinical measurements taken on the same subjects using different raters. For continuous measurements, the use of a graphical method to plot the difference scores of two measurements against the mean for each subject has been advocated.[1] However, this is a purely descriptive method and cannot provide inference regarding agreement. To overcome such limitations, scaled agreement indices, such as intraclass correlation coefficient,[2] concordance correlation coefficient,[3] and their extensions,[4–7] were introduced to assess agreement among two or more raters. Use of these scaled agreement indices has gained popularity in practice for their simplicity and ease of representation.

While being simple, scaled agreement indices have been criticized for several limitations. The main problem with these methods is that they are very sensitive to sample heterogeneity, sometimes resulting in counterintuitive interpretations.[1,8] For example, absurdly high values of intraclass correlation coefficient and concordance correlation coefficient can be obtained even for a highly varied sample, where the relative magnitude of between-subject variability to the total population variability is large. Moreover, scaled indices do not provide the interpretation terms of the original measurement unit, and there is not a set ground for determining how high these indices should be in order to qualify as satisfying agreement.

As a formal alternative, unscaled agreement indices such as total deviation index (TDI)[9] and coverage probability (CP)[10] were introduced. TDI describes an acceptable/tolerable range of absolute difference such that a prespecified proportion of the absolute differences between paired measurements is within the acceptable range. CP is a reciprocal concept in which a proportion of the absolute differences within a prespecified acceptable range is computed. Using unscaled indices has three key advantages: (i) they provide a direct intuitive interpretation tied with the original measurement unit; (ii) satisfactory agreement can be easily determined by directly comparing their values to a prespecified acceptable range of distance or coverage probability; (iii) formal statistical inferences can be made based on their estimates. CP has been recommended as the preferred choice of agreement index for assessing reproducibility in a core lab setting.[11]

All of the aforementioned unscaled agreement indices were only defined in the context of comparing a pair of raters. In the presence of multiple raters and replicated measurements for each subject, several extended unscaled agreement indices such as inter- and total-TDI (inter- and total-CP) have been proposed.[7] These indices were expressed as functions of variance components through a mixed analysis of variance (ANOVA) model. Although it is possible to quantify agreement among multiple raters using inter- and total-TDI (inter- and total-CP), the ANOVA model assumption severely restricts the degree of heterogeneity that actual measurement processes may exhibit. Specifically, this assumption imposes a compound symmetry covariance structure shared by all measurements from different raters. However, it is highly unlikely for the assumption to hold in practice, especially when the goal is to assess inter-rater agreement among newly introduced raters with unknown measurement characteristics.

Consequently, a formal tool that is unscaled in nature and able to assess inter-rater agreement among multiple raters that exhibit highly heterogeneous variabilities in measurement processes would be desirable, but has been lacking in the literature. For example, data from a renal study (details are given in Section 4) demonstrate the need for such a statistical framework. In the absence of a gold standard, it is generally accepted that the best available interpretation of renal scans comes from experienced experts, but interobserver variability still exists as their interpretations do not always agree with each other.[12] Practicing radiologists at US hospitals often have marked variability in their interpretations compared to experienced readers due to the fact that their training in nuclear medicine was limited to 3–4 months.[12,13] An analysis was thus carried out to quantify the interobserver agreement among practicing radiologists and better understand the nature of diagnostic variability present in a real-world clinical practice with renal scans. It is also of interest to determine if a new intervention with educational training called computer-assisted diagnosis (CAD) would reduce the interobserver variability among practicing radiologists. However, every radiologist and expert has different experience and expertise, and there is no unscaled agreement index that can incorporate possible heterogeneous variabilities in respective interpretation processes.

In this paper, we propose a set of overall indices based on the root mean square of pairwise differences (RMSPD) that are unscaled and can be used to assess agreement among multiple raters in the presence of heterogeneity of measurements. Recently, relative area under coverage probability curve (RAUCPC) was introduced as a summary agreement index that is scaled and summarizes agreement based on more than one prespecified absolute differences.[14] Accordingly, we propose another overall agreement index based on RMSPD that is scaled and extends the concept of RAUCPC in the presence of multiple raters. A challenging aspect of using RMSPD to define an agreement index is that its explicit analytical expression for the inverse distribution is unavailable when a general covariance structure is considered. To this end, we propose to adopt an approximate distribution, and the details are described in Section 2. We propose maximum likelihood (ML) and bootstrap approaches for estimation and inference. In Section 3, we conduct simulation studies to evaluate the performance of the proposed approaches. In Section 4, we illustrate the application of our methods via application to a renal study. We present a summary in Section 5.

2 | METHODS

2.1 | Existing unscaled and summary agreement indices for two raters

Let $Y1$ and $Y2$ be measurements from the same subject taken by the first and second raters, respectively. Then, the absolute difference $|D| = |Y1 - Y2|$ represents a distance, the extent to which paired measurements deviate from each other Under this setting, a higher proportion of paired measurements with smaller $|D|$ implies better agreement between the two raters. TDI is defined as the range of absolute difference between paired measurements such that a prespecified proportion (π_0) of observations has absolute differences within that range.[9] In other words, for $0 < \pi_0 < 1$, TDI_{π_0} is defined as the solution to

$\pi_0 = P\left(|D| < \mathrm{TDI}_{\pi_0}\right) = P\left(D^2 < \mathrm{TDI}^2_{\pi_0}\right)$, that is,

$$\mathrm{TDI}_{\pi_0} = \sqrt{G^{-1}(\pi_0)},$$

where $G(\cdot)$ is the cumulative distribution function of D^2, and $G^{-1}(\cdot)$ is the inverse function of $G(\cdot)$.

CP is the reciprocal of TDI.[10] Here, the practitioner first specifies the maximum acceptable/tolerable range of absolute difference between paired measurements and computes the proportion of observations within this predetermined range. Let d denote the prespecified acceptable absolute difference between paired measurements. CP is defined as

$$\mathrm{CP}_d = P(|D| < d) = P\left(D^2 < d^2\right) = G\left(d^2\right).$$

Recently, Banhart[14] proposed the RAUCPC as a summary agreement index between two raters in the presence of multiple acceptable absolute differences. The index is scaled in nature but utilizes information based on a series of estimated coverage probabilities. For example, a practitioner may be interested in quantifying agreement based on certain varying acceptable distance criteria: (i) $100\pi_0^{(1)}\%$ of observations should have absolute difference less than $d^{(1)}$; (ii) $100\pi_0^{(2)}\%$ of observations should have absolute difference less than $d^{(2)}$; (iii) $100\pi_0^{(3)}\%$ of observations should have absolute difference less than $d^{(3)}$. Denote $\delta_{\max}$ as an a priori maximum acceptable range of distance such that $P(D < \delta_{\max}) \approx 1$. Rather than comparing $\mathrm{CP_d}$ to the preset $\pi_0^{(s)}$ at every prespecified absolute difference $d^{(s)}$, $s = 1, 2, 3$, the area under the coverage probability curve $\mathrm{CP}(d) = P(D < d)$, $0 \le d \le \delta_{\max}$, can be used for simultaneous comparison. Specifically, RAUCPC is defined as

$$\mathrm{RAUCPC} = \frac{\int_0^{\delta_{\max}} \mathrm{CP}(x)dx}{\delta_{\max}},$$

where the area under the coverage probability curve is scaled relative to $\delta_{\max}$ so that $0 \le \mathrm{RAUCPC} \le 1$.

2.2 | Overall agreement indices for multiple raters

Let Y_j denote a random variable representing a measurement from rater j, ($j = 1, \ldots, k$). We assume that the $k \times 1$ vector of measurements $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_k]^T$ has finite first and second moments with $k \times 1$ mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_k]$ and $k \times k$ covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ may take an unstructured form so that all k raters can exhibit heterogeneous measurement processes. In this paper, we consider the RMSPD as an extended measure of distance that describes the overall deviance among measurements taken by k (≥ 2) raters, that is,

$$D_k = \sqrt{\frac{2}{k(k-1)} \sum_{1 \le p < q \le k} \left(Y_p - Y_q\right)^2}. \quad (1)$$

D_k is the square root of the average squared difference between all possible pairs of k raters, where the square root is taken to preserve the measurement unit. We note that D_k reduces to $|D|$ when $k = 2$, but in (1), we have expressed the deviation of measurements between any two raters in terms of the squared difference as opposed to the absolute difference used in conventional definitions of TDI and CP. The proposed RMSPD (D_k) basically summarizes the degree of deviation of measurements among multiple raters by taking into account all possible pairwise comparisons of their measurements based on the squared difference.

Based on (1), we propose a novel unscaled agreement index, the overall deviation index (ODI), for measuring agreement among k raters. For $0 < \pi_0 < 1$, $\mathrm{ODI}_{\pi_0, k}$ is defined as the solution to $\pi_0 = P\left(D_k < \mathrm{ODI}_{\pi_0, k}\right) = P\left(D_k^2 < \mathrm{ODI}^2_{\pi_0, k}\right)$, that is,

$$\mathrm{ODI}_{\pi_0, k} = \sqrt{F^{-1}(\pi_0)}, \quad (2)$$

where $F(\cdot)$ is the cumulative distribution of D_k^2, and $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$. Putting into words, this means that $100\pi_0\%$ of observations have RMSPDs among k raters smaller than or equal to $\mathrm{ODI}_{\pi_0, k}$. Thus, the lower the ODI value, the better the agreement among measurements from multiple raters.

As in the case of the original CP, we propose the overall coverage probability (OCP) as the reciprocal of ODI. Initially, the acceptable RMSPD among k raters (d_k) is predetermined. Then, the proportion of observations within this acceptable range is computed to quantify agreement. Specifically, OCP is defined as

$$\mathrm{OCP}_{d_k, k} = P\left(D_k < d_k\right) = P\left(D_k^2 < d_k^2\right) = F\left(d_k^2\right). \quad (3)$$

$\mathrm{OCP}_{d_k, k}$ thus measures the proportion of observations that have RMSPDs among k raters less than or equal to d_k. Thus, a higher OCP value suggests better agreement among measurements from multiple raters.

As for a scaled summary agreement index,[14] we propose to use the relative area under the OCP curve (RAUOCPC) in the presence of multiple raters. For example, consider the three varying acceptable distance criteria as presented in Section 2.1. For the case of multiple raters, each absolute difference $d^{(s)}$ is now replaced by RMSPD $d_k^{(s)}$, $s = 1, 2, 3$. Denote $\delta_{\max,k}$ as an a priori maximum acceptable RMSPD such that $P(D_k < \delta_{\max,k}) \approx 1$. Rather than

comparing $\mathrm{OCP}_k(d_k)$ to the preset $\pi_0^{(S)}$ at every prespecified RMSPD $d_k^{(s)}$, the area under the OCP curve $\mathrm{OCP}_k(d_k) = P(D_k < d_k)$, $0 \le d \le \delta_{\max,k}$, can be used for simultaneous comparison. Specifically, RAUOCPC is defined as

$$\mathrm{RAUOCPC}_k = \frac{\int_0^{\delta_{max,k}} \mathrm{OCP}_k(x)dx}{\delta_{\max,k}}, \quad (4)$$

so that $0 \le \mathrm{RAUOCPC} \le 1$, with higher values indicating better agreement. Therefore, RAUOCPC can be used as a convenient tool to simultaneously compare each OCP to the multiple predetermined acceptable/tolerable RMSPDs.

Note that, when $k = 2$, $\mathrm{ODI}_{\pi_0,2} = \mathrm{TDI}_{\pi_0}$, $\mathrm{OCP}_{d_2,2} = \mathrm{CP}_d$, and $\mathrm{RAUOCPC}_2 = \mathrm{RAUCPC}$. Thus, the ODI, OCP, and RAUOCPC are natural extensions of TDI, CP, and RAUCPC, respectively.

Parameterization—In previous literature,[9,10,14] D was assumed to follow a normal distribution (D^2 to follow a noncentral chi-square distribution) in order to define, estimate, and perform inference on TDI, CP, and RAUCPC. Likewise, formulations of ODI, OCP, and RAUOCPC as in definitions (2), (3), and (4), respectively, require an appropriate parameterization of $F\left(d_k^2\right)$ and its inverse $F^{-1}(\boldsymbol{\pi}_0)$.

Define a $(k-1) \times k$ matrix $\mathbf{A} = \{a_{u,v}\}_{(k-1)\times k}$, where $a_{u,v} = 1$ for $u = v$, $a_{u,v} = -1$ for $u + 1 = v$, and $a_{u,v} = 0$ otherwise. Then, $\mathbf{X} = \mathbf{AY} = [Y_1 - Y_2, Y_2 - Y_3, \ldots, Y_{k-1} - Y_k]^{\mathrm{T}} = [X_1, X_2, \ldots, X_{k-1}]^T$ represents the $(k-1) \times 1$ vector of distinct pairwise differences with $(k-1) \times 1$ mean vector $\boldsymbol{\mu_d} = \mathbf{A}\boldsymbol{\mu} = [\mu_1 - \mu_2, \mu_2 - \mu_3, \ldots, \mu_{k-1} - \mu_k]^T = [\mu_{d1}, \mu_{d2}, \ldots, \mu_{d,k-1}]^T$ and $(k-1) \times (k-1)$ covariance matrix $\boldsymbol{\Sigma_d} = \mathbf{A\Sigma A^T}$. We assume that $\mathbf{X}$ is normally distributed as $\mathrm{MN}_{k-1}(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$, a weaker assumption than imposing normality on $\mathbf{Y}$. Then, D_k^2 can be expressed as a quadratic form in normal variates $\mathbf{X}$ (see Appendix A). Specifically, we have

$$D_k^2 = \frac{2}{k(k-1)} \sum_{1 \le p < q \le k} \left(Y_p - Y_q\right)^2 = \mathbf{X^T}\left\{\frac{2}{k-1}\left(\mathbf{AA^T}\right)^{-1}\right\}\mathbf{X} = \mathbf{X^T B X}, \quad (5)$$

where $\mathbf{B} = \frac{2}{k-1}\left(\mathbf{AA^T}\right)^{-1}$ with rank $(\mathbf{B}) = k - 1$. If $\boldsymbol{\Sigma d}$ is nonsingular, it can be shown that the exact distributional form of D_k^2 can be expressed as a weighted sum of chi-square variables,[15] as follow:

$$D_k^2 \sim \sum_{r=1}^{k-1} \lambda_r \chi^2_{h_r, \delta_r}. \quad (6)$$

The λ_r are the distinct nonzero eigenvalues of $\mathbf{B\Sigma_d}$, the h_r their respective orders of multiplicity, the δ_r are squares of certain linear combinations of $\mu_{d1}, \mu_{d2}, \ldots, \mu_{d,k-1}$, and the $\chi^2_{h_r,\delta_r}$ are independent noncentral chi-square random variables with h_r degrees of freedom and noncentrality parameter δr.

However, computing $F\left(d_k^2\right)$ and its inverse $F^{-1}(\pi_0)$ using exact distributional form (6) is not straightforward except in some special cases. In order to readily define and estimate the proposed overall agreement indices, we propose to adopt the approximate distribution of D_k^2 as opposed to its exact distribution (see Section 5 for a more detailed discussion). Specifically, we propose to approximate $F\left(d_k^2\right)$ and $F^{-1}(\pi_0)$ using a single noncentral chi-square random variable $\chi^2_{l,\delta}$, where the degrees of freedom l and the noncentrality parameter δ are determined by the first four cumulants of D_k^2.[16] Specifically, let κ_t denote the tth cumulant of D_k^2. Then, κ_t can be directly expressed as a function of parameters $(\boldsymbol{\mu_d}, \mathbf{\Sigma_d})$ from the assumed multivariate normal distribution on the distinct pairwise differences,[16,17] that is,

$$\kappa_t\left(\boldsymbol{\mu_d}, \mathbf{\Sigma_d}\right) = 2^{t-1}(t-1)!\left[\text{trace}\left\{\left(\mathbf{B\Sigma_d}\right)^t\right\} + t\boldsymbol{\mu_d}^T\left(\mathbf{B\Sigma_d}\right)^{t-1}\mathbf{B}\boldsymbol{\mu_d}\right].$$

Accordingly, the mean, standard deviation, skewness, and kurtosis of the distribution of D_k^2 can be defined in terms of the cumulants. We omit $(\boldsymbol{\mu_d}, \mathbf{\Sigma_d})$ for ease of representation; thus, we have

$$\mu_Q = \kappa_1, \quad \sigma_Q = \sqrt{\kappa_2}, \quad \beta_1 = \frac{\kappa_3}{\kappa_2^{3/2}} \quad \beta_2 = \frac{\kappa_4}{\kappa_2^2}.$$

We can initially write

$$F\left(d_k^2\right) = P\left(D_k^2 < d_k^2\right) = P\left(\frac{D_k^2 - \kappa_1}{\sqrt{\kappa_2}} < \frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right).$$

Then, the above probability can be approximated using a single noncentral chi-square random variable $\chi^2_{l,\delta}$as

$$P\left(\frac{\chi^2_{l,\delta} - \mu^*}{\sigma^*} < \frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right) = P\left\{\chi^2_{l,\delta} < \left(\frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right)\sigma^* + \mu^*\right\},$$

where $\mu^* = \mathrm{E}\left(\chi^2_{l,\delta}\right) = l + \delta$ and $\sigma^* = \mathrm{SD}\left(\chi^2_{l,\delta}\right) = \sqrt{2(l+2\delta)}$. Here, parameters l and δ are determined so that skewnesses of D_k^2 and $\chi^2_{l,\delta}$ are equal and the difference between their

kurtoses is minimized. Let $s_1 = \kappa_3/\sqrt{8}\kappa_2^{3/2}$, $s_2 = \kappa_4/12\kappa_2^2$ and $a = \sqrt{l + 2\delta}$. It can be shown that if $s_1^2 > s_2$,[16] we have

$$a = \frac{1}{\left(1 - \sqrt{s_1^2 - s_2}\right)}, \quad \delta = s_1 a^3 - a^2 \quad \text{and} \quad l = a^2 - 2\delta,$$

and if $s_1^2 \leq s_2$, we have

$$a = \frac{1}{s_1}, \quad \delta = 0 \quad \text{and} \quad l = a^2.$$

Thus, $F\left(d_k^2\right)$ and its inverse $F^{-1}(\boldsymbol{\pi}_0)$ can be approximated as

$$F\left(d_k^2\right) \approx \chi^2\left\{\left(\frac{d_k^2 - \kappa_1\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)}{\sqrt{\kappa_2\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)}}\right)\sqrt{2}a\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right) + l\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right) + \delta\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right), l\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right), \delta\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)\right\}$$

and

$$F^{-1}\left(\pi_0\right) \approx \frac{\sqrt{\kappa_2\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)}\left[\chi^{2(-1)}\left\{\pi_0, l\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right), \delta\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)\right\} - l\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right) - \delta\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)\right]}{\sqrt{2}a\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right)} + \kappa_1\left(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}\right),$$

where $\chi^2(\cdot, l, \delta)$ is the cumulative distribution function of the noncentral chi-square distribution with l degrees of freedom and noncentrality parameter δ, and $\chi^{2(-1)}(\cdot, l, \delta)$ is the inverse function of $\chi^2(\cdot, l, \delta)$. It is important to note that both quantities are completely determined by the parameters $(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$.

By adopting the proposed parameterization and plugging in approximated values of $F\left(d_k^2\right)$ and $F^{-1}(\boldsymbol{\pi}_0)$, definitions (2),(3), and (4) become

$$\mathrm{ODI}_{\pi_0, k} = \left[\frac{\sqrt{\kappa_2}\left\{\chi^{2(-1)}\left(\pi_0, l, \delta\right) - l - \delta\right\}}{\sqrt{2}a} + \kappa_1\right]^{1/2}, \quad (7)$$

$$\mathrm{OCP}_{d_k, k} = \chi^2\left\{\left(\frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right)\sqrt{2}a + l + \delta, l, \delta\right\}, \quad (8)$$

and

$$\mathrm{RAUOCPC}_k = \frac{\int_0^{\delta_{\max,k}} \chi^2\left\{\left(\frac{x^2-\kappa_1}{\sqrt{\kappa_2}}\right)\sqrt{2}a + l + \delta, l, \delta\right\} dx}{\delta_{\max,k}}. \quad (9)$$

Compound symmetry case—Suppose $\mathbf{Y}$ has a mean vector $\boldsymbol{\mu}$ and a compound symmetry covariance structure $\boldsymbol{\Sigma} = \sigma^2(1-\rho)\mathbf{I_k} + \sigma^2\rho\mathbf{1_k}\mathbf{1_k^T}$. Note that $\mathbf{I_k}$ is a $k \times k$ identity matrix and $\mathbf{1_k}$ is a $k \times 1$ vector with only 1's as its elements. This represents the case in which multiple raters share common variabilities in respective measurement processes. Assume that the vector of distinct pairwise differences $\mathbf{X}$ is normally distributed as $\mathrm{MN}_{k-1}(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ with $\boldsymbol{\Sigma_d} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A^T} = \sigma^2(1-\rho)\mathbf{A}\mathbf{A^T}$. Consequently, by (5) and the relationship between normal and chi-square distributions and noting that $\boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1} = \frac{1}{\sigma^2(1-\rho)}\left(\mathbf{A}\mathbf{A^T}\right)^{-1}$, we have $D_k^2 = \mathbf{X^T}\boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1}\mathbf{X} = \frac{\sum_{1 \le p < q \le k}\left(Y_p - Y_q\right)^2}{k\sigma^2(1-\rho)} \sim \chi^2_{k-1,\gamma}$, where $\chi^2_{k-1,\gamma}$ denotes a noncentral chi-square random variable with $k-1$ degrees of freedom and noncentrality parameter $\gamma = \frac{\sum_{1 \le p < q \le k}\left(\mu_p - \mu_q\right)^2}{k\sigma^2(1-\rho)}$. In other words, when measurements follow a compound symmetry covariance structure, $F\left(d_k^2\right)$ and its inverse $F^{-1}(\pi_0)$ can be computed using exact distributional form as

$$F\left(d_k^2\right) = P\left\{\frac{(k-1)D_k^2}{2\sigma^2(1-\rho)} < \frac{(k-1)d_k^2}{2\sigma^2(1-\rho)}\right\} = \chi^2\left\{\frac{(k-1)d_k^2}{2\sigma^2(1-\rho)}, k-1, \frac{\sum_{1 \le p < q \le k}\left(\mu_p - \mu_q\right)^2}{k\sigma^2(1-\rho)}\right\}$$

and

$$F^{-1}\left(\pi_0\right) = \left\{\frac{2\sigma^2(1-\rho)}{k-1}\right\}\chi^{2(-1)}\left\{\pi_0, k-1, \frac{\sum_{1 \le p < q \le k}\left(\mu_p - \mu_q\right)^2}{k\sigma^2(1-\rho)}\right\}.$$

By adopting the exact parameterization assuming a compound symmetry covariance structure, definitions (2), (3), and (4) become

$$\mathrm{ODI}^{(C)}_{\pi_0,k} = \left[\left\{\frac{2\sigma^2(1-\rho)}{k-1}\right\}\chi^{2(-1)}\left\{\pi_0, k-1, \frac{\sum_{1 \le p < q \le k}\left(\mu_p - \mu_q\right)^2}{k\sigma^2(1-\rho)}\right\}\right]^{1/2}, \quad (10)$$

$$\mathrm{OCP}_{d_k,k}^{(C)} = \chi^2\left\{\frac{(k-1)d_k^2}{2\sigma^2(1-\rho)}, k-1, \frac{\sum_{1\le p<q\le k}\left(\mu_p-\mu_q\right)^2}{k\sigma^2(1-\rho)}\right\}, \quad (11)$$

and

$$\mathrm{RAUOCPC}_k^{(C)} = \frac{\int_0^{\delta_{\max,k}}\chi^2\left\{\frac{(k-1)d_{ik}^2}{2\sigma^2(1-\rho)}, k-1, \frac{\sum_{1\le p<q\le k}\left(\mu_p-\mu_q\right)^2}{k\sigma^2(1-\rho)}\right\}dx}{\delta_{\max,k}}. \quad (12)$$

When there are no replicates from respective raters, definitions (10) and (11) are the same quantities as inter-TDI and inter-CP (or total-TDI and total-CP) proposed by Lin et al,[7] respectively, which are based on the mixed ANOVA model.

2.3 | Estimation

Let $\mathbf{Y}_i\,(i = 1, \ldots, n)$ be the vector of measurements for subject i. Then, $\mathbf{X_i} = \mathbf{AY_i}$ is the vector of distinct pairwise differences for the same subject. Denote $\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ as the (bias-adjusted) ML estimators for the parameters in MN($\boldsymbol{\mu_d}$, $\boldsymbol{\Sigma_d}$) distribution. Specifically, $\hat{\boldsymbol{\mu}}_{\boldsymbol{d}} = \left[\hat{\mu}_{d1}, \hat{\mu}_{d2}, \ldots, \hat{\mu}_{d,k-1}\right]^T = \left[\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{k-1}\right]^T$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{X_i}-\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}\right)\left(\boldsymbol{X_i}-\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}\right)^T$. $= \frac{1}{n}\left[\sum_{i=1}^{n}X_{i1}, \sum_{i=1}^{n}X_{i2}, \ldots, \sum_{i=1}^{n}X_{i,k-1}\right]^T$

We propose to estimate $\mathrm{ODI}_{\pi_0,k}$, $\mathrm{OCP}_{d_k,k}$, and $\mathrm{RAUOCPC}_k$ by replacing parameters $\kappa_1(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$, $\kappa_2(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$, $a(\boldsymbol{\mu_d}, \boldsymbol{\Sigma}_d)$, $l(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$, and $\delta(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ in definitions (7), (8), and (9) by their ML estimates $\hat{\kappa}_1 = \kappa_1\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\right)$, $\hat{\kappa}_2 = \kappa_2\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\right)$, $\hat{a} = a\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\right)$, $\hat{l} = l\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\right)$, and $\hat{\delta} = \delta\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\right)$. Therefore, by the invariance property of ML estimators, we can express the ML estimators of the proposed overall agreement indices as

$$\widehat{\mathrm{ODI}}_{\pi_0,k} = \left[\frac{\sqrt{\hat{\kappa}_2}\left\{\chi^{2(-1)}\left(\pi_0, \hat{l}, \hat{\delta}\right) - \hat{\imath} - \hat{\delta}\right\}}{\sqrt{2\hat{a}}} + \hat{\kappa}_1\right]^{1/2}, \quad (13)$$

$$\widehat{\mathrm{OCP}}_{d_k,k} = \chi^2\left\{\left(\frac{d_k^2-\hat{\kappa}_1}{\sqrt{\hat{\kappa}_2}}\right)\sqrt{2\hat{a}} + \hat{l} + \hat{\delta}, \hat{l}, \hat{\delta}\right\}, \quad (14)$$

and

$$\widehat{\text{RAUOCPC}}_k \frac{\int_0^{\delta_{\max,k}} \chi^2\left\{\left(\frac{x^2-\hat{\kappa}_1}{\sqrt{\hat{\kappa}_2}}\right)\sqrt{2}\hat{a}+\hat{\iota}+\hat{\delta},\hat{l},\hat{\delta}\right\}dx}{\hat{\delta}_{\max,k}}. \quad (15)$$

RAUOCPC can also be estimated using the nonparametric method as suggested by Banhart for the RAUCPC case.[14] We first order the unique observed RMSPDs among k raters as $d_{1,k} < d_{2,k} < \ldots < d_{n,k}$ with $d_{n,k} < \delta_{\max,k}$. Define $ocp_{1,k} < ocp_{2,k} < \ldots < ocp_{n,k}$ as the estimated overall coverage probabilities, where $ocp_{i,k}$ denotes the proportion of all possible RMSPDs less than or equal to $d_{i,k}$, $i = 1, 2, \ldots, n$. Then, the empirical overall coverage probabilities can be drawn as a series of straight lines connecting $(d_{0,k}, ocp_{0,k})$, $(d_{1,k}, ocp_{1,k})$, $\ldots$, $(d_{n,k}, ocp_{n,k})$, $(d_{n+1,k}, ocp_{n+1,k})$, where $d_{0,k} = 0$, $ocp_{0,k} = 0$, $d_{n+1,k} = \delta_{\max,k}$, and $ocp_{n+1,k} = ocp_{n,k}$. The nonparametric estimator for RAUOCPC_k is the area under these straight lines scaled by $\delta_{\max,k}$. Specifically, the estimator is given as

$$\widehat{\text{RAUOCPC}}_k^{\text{non-parm}} = \frac{\sum_{i=1}^{n+1}(d_i - d_{i-1})\left(ocp_{i-1} + \frac{ocp_i - ocp_{i-1}}{2}\right)}{\delta_{\max,k}}. \quad (16)$$

2.4 | Inference

One-sample—Let θ_k be one of the three proposed overall agreement indices and $\boldsymbol{\beta} = (\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ be its associated parameter. Denote $\hat{\theta}_k$ and $\hat{\boldsymbol{\beta}}$ as their ML estimators, respectively. Also, let I be the observed Fisher information matrix for $\boldsymbol{\beta}$ and $G = \frac{\partial g\{\theta_k(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ be the gradient vector of the index evaluated at the ML estimator, where g represents a monotone transformation of the parameter that is adopted to accelerate convergence to asymptotic normality. Since $\text{ODI}_{\pi 0,k} \in [0, \infty)$ and $\text{OCP}_{dk,k}$, $\text{RAUOCPC}_k \in [0, 1]$, we use the natural log transformation for the former index and the logit transformation for the latter indices. Then, from the asymptotic normality of ML estimators and the delta method, we have $g(\hat{\theta}_k) \sim AN(g(\theta_k), G^T I^{-1} G)$, where $(G^T I^{-1} G)^{1/2} = \widehat{SE}\{g(\hat{\theta}_k)\}$ denotes the standard error estimate.

Since the analytical form of $(G^T I^{-1} G)^{1/2}$ is complicated, we propose a bootstrap approach for standard error estimation. Specifically, we can take B bootstrap samples from the observed data at the subject level with replacement, compute $g(\hat{\theta}_k)^{(b)}$ for each bootstrap sample $b = 1, 2, \ldots, B$, and obtain the bootstrap estimate of the standard error, that is,

$$\widehat{\text{SE}}_B\{g(\hat{\theta}_k)\} = \left[\frac{1}{B}\sum_{b=1}^{B}\left\{g(\hat{\theta}_k)^{(b)} - \overline{g(\hat{\theta}_k)_B}\right\}^2\right]^{1/2}, \quad (17)$$

where $\overline{g(\hat{\theta}_k)_B} = \frac{1}{B}\sum_{b=1}^{B} g(\hat{\theta}_k)^{(b)}$. Note that sampling on the subject level is essential as we should account for correlated measurements within a subject.

Suppose we postulate that agreement among k raters based on the ODI is satisfactory if approximately $100\pi_0\%$ of observations have RMSPDs among the raters less than a predetermined constant L_0. Here, L_0 denotes the maximum RMSPD that we are willing to tolerate, and accept that all k raters exhibit homogeneity in the measurement processes. We would accept satisfactory agreement with a Type I error α if the $100(1-\alpha)\%$ upper confidence limit of $\mathrm{ODI}_{\pi_0, k}$, that is,

$$U_{\mathrm{ODI}_{\pi_0, k, 1-\alpha}} = \exp\left\{\log\left(\widehat{\mathrm{ODI}}_{\pi_0, k}\right) + z_{1-\alpha}\left(\widehat{SE}_B\left\{(\log\left(\widehat{\mathrm{ODI}}_{\pi_0, k}\right)\right\}\right\}, \quad (18)$$

is less than L_0, where the bootstrap standard error estimate is calculated from (17) and $z_{1-\alpha}$ denotes the $100(1-\alpha)$th percentile of a standard normal distribution.

For the OCP, the lower confidence limit is preferably calculated because ensuring acceptable agreement with respect to OCP often involves a null proportion π_0, which we would deem too small as to conclude satisfactory agreement. Specifically, using the bootstrap standard error estimate (17), the $100(1-\alpha)\%$ lower confidence limit of $\mathrm{OCP}_{dk,k}$ is computed as

$$L_{\mathrm{OCP}_{d_k, k, 1-\alpha}} = h\left[\mathrm{logit}\left(\widehat{\mathrm{OCP}}_{d_k, k}\right) - z_{1-\alpha}\widehat{SE}_B\left\{\mathrm{logit}\left(\widehat{\mathrm{OCP}}_{d_k, k}\right)\right\}\right], \quad (19)$$

where $h(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$. Then, given a Type I error rate of α and tolerable RMSPD d_k, we accept that k raters produce reasonably homogeneous ratings on a given subject if (19) is greater than π_0.

For the RAUOCPC, consider the three multiple acceptable RMSPDs $\left(d_k^{(1)}, d_k^{(2)}, d_k^{(3)}\right)$ and denote $\delta_{\max,k}$ as an a priori maximum acceptable RMSPD such that $P(D_k < \delta_{\max,k}) \approx 1$. Initially, the area under the $\widehat{\mathrm{OCP}}_k(d_k), 0 \le d_k \le \delta_{\max,k}$, can be visually compared to the area under straight lines that connect points formed by a series of preset RMSPDs with the corresponding overall coverage probabilities, for example, connecting points (0,0), $\left(d_k^{(1)}, \pi_0^{(1)}\right)$, $\left(d_k^{(2)}, \pi_0^{(2)}\right)$, $\left(d_k^{(3)}, \pi_0^{(3)}\right)$, and $(\delta_{\max,k}, 1)$. The larger size of the former area would suggest a satisfying agreement among k raters. Testing whether the difference between the sizes of the two areas is statistically significant is equivalent to testing whether $\mathrm{RAUOCPC}_k$ is greater than T_0, which denotes the size of the latter area scaled by $\delta_{\max,k}$. Thus, we can focus on deriving the $100(1-\alpha)\%$ lower boundary. Specifically, after obtaining the bootstrap standard error estimate from (17), the $100(1-\alpha)\%$ lower confidence limit of $\mathrm{RAUOCPC}_k$ is computed as

$$L_{\mathrm{RAUOCPC}_{k,\,1-\alpha}} = h\left[\mathrm{logit}\left(\widehat{\mathrm{RAUOCPC}_k}\right) - z_{1-\alpha}\widehat{SE}_B\left\{\mathrm{logit}\left(\widehat{\mathrm{RAUOCPC}_k}\right)\right\}\right]. \quad (20)$$

If (20) is greater than T_0, we can conclude a satisfactory agreement among k raters based on the three varying acceptable distance criteria. Note that the $100(1-\alpha)\%$ lower boundary based on the nonparametric RAUOCPC estimate can be computed in a similar manner.

Two-sample—Now, suppose we are interested in comparing degrees of inter-rater agreement among measurements on the same set of subjects between two groups of raters using one of the three proposed overall indices. This scenario often arises when the goal of a study is to evaluate the performance of a group of new raters relative to a group of best standard raters in terms of inter-rater agreement. Suppose $\theta_k^{(1)}$ and $\theta_k^{(2)}$ measure agreement among the first and second groups of k raters, respectively. We form a null hypothesis against the alternative hypothesis

$$H_0: \theta_k^{(1)} = \theta_k^{(2)} \quad \text{or, equivalently,} \quad H_0: g\left(\theta_k^{(1)}\right) = g\left(\theta_k^{(1)}\right),$$

against the alternative hypothesis

$$H_1: \theta_k^{(1)} \neq \theta_k^{(2)} \quad \text{or, equivalently,} \quad H_0: g\left(\theta_k^{(1)}\right) \neq g\left(\theta_k^{(1)}\right).$$

Using the asymptotic property of ML estimators, we can formulate the Wald test statistic as

$$\frac{g\left(\hat{\theta}_k^{(1)}\right) - g\left(\hat{\theta}_k^{(2)}\right)}{\mathrm{SE}\left\{g\left(\hat{\theta}_k^{(1)}\right) - g\left(\hat{\theta}_k^{(2)}\right)\right\}} = \frac{g\left(\hat{\theta}_k^{(D)}\right)}{\mathrm{SE}\left\{g\left(\hat{\theta}_k^{(D)}\right)\right\}} \sim AN(0, 1),$$

under the null hypothesis. Since the analytical form of the standard error is complicated, we estimate the standard error by the bootstrap approach. See Appendix B for detailed steps of the two-sample hypothesis testing procedure.

3 | SIMULATIONS

We conducted simulation studies to assess the performance of the proposed approaches to evaluate agreement via overall agreement indices. We assumed that there are four raters and the data are generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ and covariance matrix $\boldsymbol{\Sigma}$, for both compound symmetry and unstructured scenarios. Under both scenarios, a total of 1000 simulated data sets were generated. We considered sample sizes of 20 and 60. To evaluate the performance of inference based on the bootstrap approach, 1000 bootstrap samples were used to compute standard error estimates and one-sided 95% confidence intervals.

Under the compound symmetry scenario, data were generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4) = (3.5, 3.5, 3.5, 4.2)$ and compound symmetry covariance matrix with common variance $\sigma^2 = 0.25$ and common correlation coefficient $\rho = 0.8$. Specifically, the compound symmetry covariance matrix was defined as $\boldsymbol{\Sigma} = \{\sigma_{u,v}\}$, where $\sigma_{u,v} = 0.25$ if $u = v$ and $\sigma_{u,v} = 0.2$ if $u \neq v$, $u, v = 1, 2, 3, 4$. Thus, the first three raters exhibit homogeneity in respective measurement processes. However, the fourth rater represents a heterogeneous measurement process in which ratings are consistently overestimated as evidenced by its larger mean.

Under the unstructured scenario, data were generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4) = (3.5, 3.5, 3.5, 4.2)$ and unstructured covariance matrix defined as $\boldsymbol{\Sigma} = \{\sigma_{u,v}\}$, where $\sigma_{1,1} = \sigma_{2,2} = \sigma_{3,3} = 0.30$, $\sigma_{1,2} = \sigma_{1,3} = \sigma_{2,3} = 0.24$, $\sigma_{4,4} = 0.15$, and $\sigma_{1,4} = \sigma_{2,4} = \sigma_{3,4} = 0.08$. Again, the first three raters exhibit homogeneity in the measurement process with common covariances among themselves. However, the fourth rater is allowed to exhibit stronger heterogeneity in the measurement process by adding further flexibility in the data generation process. Not only is its mean larger compared to the first three, but its variance is smaller, and its linear relationship with others is relatively weak as evidenced by smaller covariances in relation to other raters.

Table 1 shows the simulation results for ODI and $\text{ODI}^{(C)}$ (see Section 2.2) under both scenarios. We considered two widely used prespecified coverage probabilities: $\pi_0 = 0.80$ and 0.90. Under the compound symmetry scenario, both ODI and $\text{ODI}^{(C)}$ estimates yielded almost identical results, as expected for normal data with a compound symmetry covariance structure. All absolute relative biases are less than 1%, implying that our proposed estimation approach provides reasonable unbiased estimates even for a sample size as small as 20. The 95% coverage is slightly less than the nominal level for a sample size of 20, possibly due to the underestimation of standard errors as compared to empirical counterparts. The 95% coverage is very close to the nominal level for sample sizes of 60 or greater (not shown). Under the unstructured scenario, ODI estimates again have negligible bias. The coverage of true ODI based on a one-sided 95% confidence interval is 92% or less with a sample size of 20, but a coverage of 94% or 95% is generally achieved for sample sizes of 60 or greater. However, $\text{ODI}^{(C)}$ estimates are biased underestimating the true ODI. This results in very poor coverage probabilities, especially when $\pi_0 = 0.90$. Under both scenarios, estimated standard errors rapidly approach their empirical counterparts as the sample size increases, confirming that our bootstrap procedure provides valid and robust standard error estimates.

Table 2 shows the simulation results for OCP and $\text{OCP}^{(C)}$ under both scenarios. We considered two different preset RMSPDs among four raters: $d_4 = 0.8$ and 0.9. Under the compound symmetry scenario, both OCP and $\text{OCP}^{(C)}$ estimates show good performances, as expected under the correct model specification. Each absolute relative bias is less than 1%, indicating the consistency of the point estimates. The 95% coverage is slightly greater than the nominal level, possibly due to the overestimation of standard errors. However, coverage probabilities are all generally around 95% in almost every situation. Under the unstructured scenario, OCP estimates are virtually unbiased even for a small sample size of 20. The 95% coverage of true OCP is close to the nominal level for all sample sizes. However, $\text{OCP}^{(C)}$

estimates are biased overestimating the true OCP, and the bias in fact increases for larger sample sizes. As a result, coverage probabilities are, in general, noticeably below the desired nominal rate, especially when $d_4 = 0.9$. We should also remark that the standard error estimate increases as d_4 increases in any scenario, and this indicates less precision in OCP estimates for a relatively large prespecified RMSPD compared to the range of data. Under both scenarios, estimated standard errors quickly approach their empirical counterparts as the sample size increases.

Table 3 shows the simulation results for RAUOCPC under both scenarios. We first fixed $\delta_{\max} = 1.1$. Under the compound symmetry scenario, both parametric methods provide virtually unbiased estimates for each sample size, as expected for normal data generated under the compound symmetry covariance structure. The empirical coverage rate of the one-sided 95% confidence interval is reasonably close to the nominal value even for a small sample size of 20, indicating fast convergence of the estimates to the normal distribution. On the other hand, the nonparametric method overestimates the true RAUOCPC, but the bias decreases as sample size increases. The 95% coverage rate is 90% or less, though reasonable coverage of 92% is generally achieved for a sample size of 100 (not shown). Under the unstructured scenario, it should be highlighted that the RAUOCPC$^{(C)}$ estimates significantly underestimate the true RAUOCPC. Moreover, their empirical coverage rates indicate that the nominal confidence tends to be overestimated, with values close to 100%. In contrast, the proposed parametric approach results in much smaller values of bias and coverage probabilities that approach 95% as sample size increases. We should also point out that the performance of the nonparametric method is relatively poor for a sample size of 20, but it significantly improves in terms of bias and coverage rate as the sample size increases. Standard error estimates are close to their empirical counterparts in any situation.

Our simulation studies show that estimates of the compound symmetry case overall agreement indices are largely biased and have very poor coverage of the true values when the underlying covariance structure is not strictly compound symmetry. Because it is usually rare to assess agreement among multiple raters that assume homogeneity of all variabilities in measurement processes, we recommend using the proposed approach, defined absent any restriction on the covariance structure, for estimation and inference in most practical situations.

4 | RENAL STUDY

Renal scans in nuclear medicine (diuresis renography) play an important role in the determination of kidney obstruction, which is a condition that may lead to loss of kidney function.[18] Diagnosis of kidney obstruction using renal scans is generally a difficult problem for the following reasons: (i) there is currently no good gold standard for detecting kidney obstruction; (ii) correct interpretation of renal scans requires a deep understanding of renal physiology and technetium-99m-mercaptoacetyltriglycine (99mTc-MAG3) pharmacokinetics.

In the absence of a gold standard, it is generally accepted that the best available interpretation comes from an expert with broad expertise and extensive experience in

academic nuclear medicine. Although interobserver variability still exists between different nuclear medicine experts, it is generally considered to be minimal.[13] The vast majority of the 590 000 renal scans performed annually in the United States are interpreted by general radiologists, who have less than 4 months training and experience in interpretations of renal scans and, thus, have marked variability in their interpretations compared to experienced readers.[12,13] For instance, a survey shows that different practicing radiologist may disagree on the interpretation of the same scan between 9% and 72% of the time.[19]

Given the background, a pilot study was conducted at Emory University with the goal of gaining better insight into the nature of diagnostic variability present in a real-world clinical practice using renal scans. In this respect, the goal of the study is to quantify the interobserver variability in the population of practicing radiologists. The nuclear medicine residents with a minimum of one year of formal training in nuclear medicine were recruited in the pilot study as a surrogate for practicing radiologists. Three nuclear medicine experts who have more than 20 years of experience in nuclear medicine were also recruited. It has been recognized that residents may not perform as well as experts due to their lack of training and limited experience.[12,13,20]

An intervention called CAD was recently introduced by Emory researchers to minimize errors and reduce the interobserver variability among practicing radiologists. The CAD analyzes renal image data and provides a second opinion about the diagnosis with reasoning. Having a second opinion with reasoning is thought to minimize the interobserver variability among radiologists. Thus, it is also of interest to determine if the CAD intervention would reduce the interobserver variability among radiologists.

Thirty-five patients with suspected obstruction (20 females, mean age ± SD, 58.7 ± 15.8y) in either right or left kidney were randomly selected. Their scans from 70 kidneys (35 left kidneys and 35 right kidneys) were independently interpreted by the three groups of raters: (i) three nuclear medicine experts each with over 20 years of experience (“Experts”); (ii) three nuclear medicine residents each having completed a minimum one of their three-year nuclear medicine residency (“Residents”); (iii) same three residents with subsequent access to CAD (“Residents + CAD”). Raters scored each kidney of a scale of −1 to +1.

We first performed a series of preliminary data analyses to investigate the structure of data and check relevant statistical assumptions. The multivariate normality assumption of pairwise differences within each group of raters was assessed based on chi-square Q-Q plots (not shown) and Doornik-Hansen’s test of multivariate normality,[21] and no significant departure from the assumption was found for all five sets of pairwise differences (p value range: 0.07–0.51) except within experts in left kidneys (p value = 0.01). Heterogeneous variabilities appeared to exist, especially among the three residents, whose sample variances ranged from 0.2 to 0.5 in both kidneys. The presence of such heterogeneity was further confirmed by the Morgan-Pitman test for comparing variances between correlated samples (all p values < 0.05).

Table 4 presents the estimates of $ODI_{0.9,3}$ and $OCP_{0.5,3}$ for the three groups of raters and the results of their pairwise across-group comparisons based on two-sample hypothesis tests.

The prespecified RMSPD of $d_3 = 0.5$ was determined based on a clinically important squared pairwise difference between any paired ratings that may disagree on the obstruction status, with the consultation of a nuclear medicine expert, Dr Taylor, from Emory University. Statistical inference is based on 1000 bootstrap samples. Note that all ODI estimates are rounded to one decimal place to reflect the unit of measurement, whereas all OCP estimates are rounded to two decimal places.

The three experts show high agreement for both kidneys as expected. Specifically, based on the ODI estimates, 90% of ratings from the three experts have RMSPDs less than 0.5 (95% upper limit = 0.6) for the left kidney and less than 0.4 (95% upper limit = 0.5) for the right kidney. Equivalently, based on the OCP estimates, about 91% (95% lower limit = 72%) of ratings for the left kidney and almost 98% (95% lower limit = 92%) of ratings for the right kidney have RMSPDs less than 0.5. Agreement among the three residents is significantly worse than that of the three experts for both left and right kidneys as evidenced by significantly higher ODI estimates and significantly lower OCP estimates (all p values $<$ 0.001). However, agreement among residents dramatically improves after given access to the CAD. For this group ("Residents + CAD"), approximately 90% of ratings have RMSPDs less than 0.7 and 0.6 (95% upper limits = 1.0 and 0.8) in left and right kidneys, respectively. Equivalently, at least 72% (95% lower limit = 53%) of ratings for the left kidney and 77% (95% lower limit = 61%) of ratings for the right kidney have RMSPDs less than 0.5. Results of pairwise across-group comparisons based on two-sample hypotheses tests show that agreement among residents with CAD is significantly better than that among the same residents before access to CAD (all p values <0.001). Moreover, the interobserver agreement among the residents with CAD resembles that among the experts, as we find that the differences between two groups are not significant for the left kidney (p values = 0.107 and 0.191).

Figure 1 shows three estimated OCP curves with parametric RAUOCPC estimates (95% lower limits) and the results of their pairwise across-group comparisons for left and right kidneys, respectively. Note that an a priori maximum acceptable difference was set to 1. We can first visually verify that the experts have the highest agreement as evidenced by the largest area under the curve, closely followed by the residents with CAD. The area under the curve is smallest for the residents without access to CAD, indicating that their agreement is worst among the three groups. The three experts have the highest estimated RAUOCPC as 0.71 (95% lower limit = 0.64) for the left kidney and as 0.79 (95% lower limit = 0.74) for the right kidney. The residents have significantly lower estimates as compared to those of experts for both left and right kidneys (all p values <0.001), though the agreement significantly improves with CAD (all p values <0.001). Moreover, agreement among residents with CAD closely matches that of the experts for the left kidney (p value = 0.106). Our preliminary results suggest that although there exists high interobserver variability in the interpretations of renal scans among practicing radiologists, use of CAD significantly reduces their interobserver variability and results in the degree of variability being close to that among the experts.

As suggested by a referee, it is interesting to consider pairwise agreement. The three pairwise TDI values (given a prespecified probability of 0.9) within each group of experts,

residents, and residents with CAD for the right kidney are (0.3, 0.4, 0.5), (0.9, 1.7, 1.4), and (0.8, 0.6, 0.6), respectively. We also computed TDI values between each of the three experts and each of the three residents with/without CAD for the right kidney. Between the experts and the residents, the mean of the nine pairwise TDI values is 0.9 (range: 0.7–1.2; median: 0.9). On the other hand, between the experts and the residents with CAD, the mean of the nine pairwise TDI values is 0.6 (range: 0.5–0.7; median: 0.6). A similar trend is observed for the left kidney. These results further provide supporting evidence for the above conclusion.

5 | DISCUSSION

We have proposed several overall agreement indices (ODI, OCP, and RAUOCPC), which are practically useful for assessing agreement among multiple raters. The key is to quantify disagreement among measurements using a new comprehensive measure of distance that represents the RMSPD among measurements provided multiple raters. The proposed overall indices defined without any assumption regarding the covariance structure among measurements provide great flexibility in a sense that practitioners can quantify agreement even among multiple raters with highly heterogeneous variabilities in respective measurement processes. Especially, unscaled overall indices are tied to the original measurement scale and can be compared against prespecified acceptable/tolerable RMSPDs to determine satisfying agreement. Thus, they are easily explained to nonstatistical practitioners and can serve as a useful alternative or complement to existing scaled indices in various clinical study settings.

Definitions and interpretations of overall agreement indices depend on the choice of an extended measure of distance that quantifies the degree of disagreement among multiple raters. In this paper, we have proposed the use of RMSPD (D_k) as defined in (1). A possible alternative to this measure is the average of pairwise absolute differences among multiple raters, that is, $D_k^* = \frac{2}{k(k-1)} \sum_{1 \le p < q \le k} \left| Y_p - Y_q \right|$. D_k and D_k^* quantify inherently different aspects of the spread of data, and each has its own merits. One advantage of D_k^* is its robustness (less sensitive to the outliers) compared to D_k, which depends on moments of the distribution. The proposed measure D_k, on the other hand, has better mathematical and asymptotic properties in connection with various widely used statistical models and probability distributions.

For practitioners, we note that the square root of acceptable/tolerable squared difference that one is willing to impose between a typical pair of measurements can serve as an interpretable reference level based on RMSPD. In practice, reference standard data can serve as a guidance to set up an acceptable RMSPD value. For instance, expert rating data on selected patients with suspected obstruction have been available over the years at Emory University Hospital, and an RMSPD value based on such data may guide the choice of the acceptable value.

We used the approach in the work of Liu et al[16] to approximate the distribution of D_k^2 (squared RMSPD) under two rationales: (i) we can rely on nearly all statistical packages to easily obtain its quantile function (inverse of a cumulative noncentral chi-square distribution

function), which is used to define a series of overall agreement indices; (ii) the reasonably high accuracy of approximation was demonstrated in simulation studies. We may consider using exact distributional form (6) or other approximation approaches. However, these alternatives either have a complicated form from which the quantile function is not readily obtainable or yield poor approximation results as compared to the approach by Liu et al.[16] For instance, since exact distribution form (6) has as an infinite series representation,[22] we can compute F by truncating the series after N terms, but obtaining F^{-1} is a burdensome task because it involves inverting an infinite series. Another approach uses numerical inversion of the characteristic function of D_k^2 to approximate the cumulative distribution function.[15] However, obtaining F^{-1} is extremely difficult as it involves inverting an analytically intractable integral. There exist other moment-based approximation approaches, such as the extension of Pearson's three-moment central χ^2 method.[15,23] This approach essentially provides a central χ^2 approximation to nonnegative D_k^2 by matching the third-order moments. Its approximate distribution form is simple, as it only requires an inversion of a central χ^2 cumulative distribution function. However, the current approach produces better approximation results for the tail probabilities since it further requires a best match of the fourth-order moments.[16]

Our proposed approach assumes distinct pairwise differences to follow a multivariate normal distribution in defining overall agreement indices, which is a weaker assumption than imposing normality on measurements themselves. Such assumption is a natural conceptual extension to the traditional TDI case in which the difference between paired measurements is assumed to follow a univariate normal distribution. In the event that there are potential violations to the normality assumption, appropriate transformation can be applied to the measurements to ensure that the distributional assumption likely holds. It is important to note that interpretations should then be in terms of the transformed scale. Recently, several novel nonparametric methods for estimation and inference of the TDI were introduced.[24–26] It is of future interest to develop a similar nonparametric approach for our proposed indices to deal with nonnormal data.

ACKNOWLEDGEMENT

This research was supported by the National Institute of Diabetes and Digestive and Kidney Diseases under Grant 1R01DK108070-01A1.

Funding information

National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: 1R01DK108070-01A1

APPENDIX A

DERIVATION OF QUADRATIC FORM (5)

In this section, we derive the quadratic form of the extended measure of distance D_k in terms of distinct pairwise differences $\mathbf{X}$. Firstly, consider

$$\sum_{1 \le p \le q \le k} \left(Y_p - Y_q\right)^2 = \sum_{q = 2, \dots, k} \left(Y_1 - Y_q\right)^2 + \sum_{q = 3, \dots, k} \left(Y_3 - Y_q\right)^2 + \cdots + \sum_{q = k} \left(Y_{k-1} - Y_q\right)^2. \quad \text{(A1)}$$

Also, define a $(k-1) \times (k-1)$ matrix $\mathbf{M_s}$ with the (m, n)th element given as $\{M_s\}_{mn} = 0$ if $\min(m, n) < s - 1$ and $\{M_s\}_{mn} = k - \max(m, n)$ otherwise. Then, we can express the first and second terms on the right-hand side of Equation (A1) as $2\sum_{1 \le p \le q \le k}(Y_p - Y_q)^2 = \mathbf{X^T M_1 X}$ and $\sum_{q=3,\dots,k}(Y_2 - Y_q)^2 = \mathbf{X^T M_2 X}$, respectively. Repeat such computation until the last term on the right-hand side of Equation (A1). Then, add the results to re-express the left-hand side of Equation (A1) as $\sum_{1 \le p < q \le k}(Y_p - Y_q)^2 = \mathbf{X^T}(\sum_{s=1,\dots,k-1}\mathbf{M_s})\mathbf{X}$, where the (m, n)th element of the matrix $\sum_{s=1,\dots,k-1}\mathbf{M_s}$ can be derived as $\{\sum_{s=1,\dots,k-1}M_s\}_{m,n} = m(k-n)$ if $1 \le m \le n \le k-1$ and $\{\sum_{s=1,\dots,k-1}M_s\}_{m,n} = n(k-m)$ if $1 \le n < m \le k-1$. By applying some basic linear algebra, we find that $\sum_{s=1,\dots,k-1}\mathbf{M_s} = \text{adj}(\mathbf{AA^T})$, where $\text{adj}(\mathbf{AA^T})$ denotes the adjugate matrix of $\mathbf{AA^T}$. Consequently, we can derive quadratic form (5) as

$$\sum_{1 \le p < q \le k} \left(Y_p - Y_q\right)^2 = \mathbf{X^T}\text{adj}\left(\mathbf{AA^T}\right)\mathbf{X} \Rightarrow \mathbf{X^T}\left(\mathbf{AA^T}\right)^{-1}\mathbf{X} = \frac{\sum_{1 \le p < q \le k}\left(Y_p - Y_q\right)^2}{k}$$

$$\Rightarrow \mathbf{X^T}\left\{\frac{2}{k-1}\left(\mathbf{AA^T}\right)^{-1}\right\}\mathbf{X} = D_k^2,$$

by noting that $\det(\mathbf{AA^T}) = k$ and $(\mathbf{AA^T})^{-1} = \text{adj}(\mathbf{AA^T})/\det(\mathbf{AA^T})$.

APPENDIX B

STEPS FOR TWO-SAMPLE HYPOTHESIS TESTING IN SECTION 2.4

Step 1: Denote $\mathbf{X}_{\text{obs}}^{(1)}$ as observations from a first group of raters and $\mathbf{X}_{\text{obs}}^{(2)}$ as observations from a second group of raters. Take B bootstrap samples from the observed data matrix $\mathbf{X}_{\text{obs}}^{(1)}$ and $\mathbf{X}_{\text{obs}}^{(2)}$ at the subject level with replacement, respectively.

Step 2: Compute $g\left(\hat{\theta}_k^{(1)(b)}\right)$ and $g\left(\hat{\theta}_k^{(2)(b)}\right)$ for each bootstrap sample $\mathbf{X}_{\text{obs}}^{(\mathbf{1})(\mathbf{b})}$ and $\mathbf{X}_{\text{obs}}^{(\mathbf{2})(\mathbf{b})}$, respectively, $b = 1, 2, \dots, B$.

Step 3: Compute B differences $g\left(\hat{\theta}_k^{(D)(b)}\right) = g\left(\hat{\theta}_k^{(1)(b)}\right) - g\left(\hat{\theta}_k^{(2)(b)}\right)$, $b = 1, 2, \dots, B$.

Step 4: Compute the standard deviation of B differences between the two ODI estimates, which gives a bootstrap estimate of the standard error, that is,

$$\widehat{SE}_B\left\{g\left(\hat{\theta}_k^{(D)}\right)\right\} = \left[\frac{1}{B}\sum_{b=1}^{B}\left\{g\left(\hat{\theta}_k^{(D)(b)}\right) - \overline{g\left(\hat{\theta}_k^{(D)}\right)}_B\right\}^2\right]^{1/2}, \quad \text{(B1)}$$

where $\overline{g\left(\hat{\theta}_k^{(D)}\right)}_B = \frac{1}{B}\sum_{b=1}^{B} g\left(\hat{\theta}_k^{(D)(b)}\right)$.

Step 5: Given a Type I error rate of α and bootstrap standard error estimate (B1), reject the null hypothesis if

$$\left|\frac{g\left(\hat{\theta}_k^{(D)}\right)}{\widehat{\mathrm{SE}}_B\left\{g\left(\hat{\theta}_k^{(D)}\right)\right\}}\right| \geq z_{1-\alpha/2}$$

and conclude that the degrees of agreement are not equal between two groups of raters.

REFERENCES

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307–310. [PubMed: 2868172]
2. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep. 1966;19(1): 3–11. [PubMed: 5942109]
3. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1): 255–268. [PubMed: 2720055]
4. King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Statist Med. 2001;20(14):2131–2147.
5. Banhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics. 2002;58(4):1020–1027. [PubMed: 12495158]
6. Banhart HX, Song J, Haber MJ. Assessing intra, inter and total agreement with replicated readings. Statist Med. 2005;24(9):1371–1384.
7. Lin LI, Hedayat AS, Wu W. A unified approach for assessing agreement for continuous and categorical data. J Biopharm Stat. 2007;17(4):629–652. [PubMed: 17613645]
8. Atkinson J, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. Biometrics. 1997;53(2):775–777.
9. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Statist Med. 2000;19(2):255–270.
10. Lin LI, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. J Am Stat Assoc. 2002;97(457):257–270.
11. Banhart HX, Yow E, Crowley AL. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res. 2016;25(6): 2939–2958. [PubMed: 24831133]
12. Taylor AT, Manatunga AK, Garcia EV. Decision support systems in diuresis renography. Semin Nucl Med. 2008;38(1):67–81. [PubMed: 18096465]
13. Taylor AT, Garcia EV. Computer-assisted diagnosis in renal nuclear medicine: rationale, methodology, and interpretative criteria for diuretic renography. Semin Nucl Med. 2014;44(2): 146–158. [PubMed: 24484751]
14. Banhart HX. Assessing agreement with relative area under the coverage probability curve. Statist Med. 2016;35(18):3153–3165.
15. Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961;48(3–4):419–426.

16. Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data Anal. 2009;4:853–856.
17. Mathai A, Provost S. Quadratic Forms in Random Variables: Theory and Applications. New York, NY: Marcel Dekker; 1992.
18. Taylor AT. Radionuclides in nephrourology, Part 2: pitfalls and diagnostic applications. J Nucl Med. 2014;55(5):786–798. [PubMed: 24591488]
19. Jaksić E, Beatović S, Paunković N, Stefanović A, Han R. Variability in interpretation of static renal scintigraphy findings. Vojnosanit Pregl. 2005;62(3):189–193. [PubMed: 15790046]
20. Erdogan Z, Abdülrezzak U, Silov G, Özdal A, Turhal O. Evaluation of interobserver variability of parenchymal phase of Tc-99m mercaptoacetyltriglycine and Tc-99m dimercaptosuccinic acid renal scintigraphy. Indian J Nucl Med. 2014;29(2):87–91. [PubMed: 24761059]
21. Doornik JA, Hansen H. An omnibus test for univariate and multivariate normality. Oxf Bull Econ Stat. 2008;70:927–939.
22. Kotz S, Johnson NL, Boyd DW. Series representations of distributions of quadratic forms in normal variables II. Non-central case. Ann Math Statist. 1967;38(3):838–848.
23. Pearson ES. Note on an approximation to the distribution of non-central χ^2. Biometrika. 1959;46:364.
24. Choudhary PK. A unified approach for nonparametric evaluation of agreement in method comparison studies. Int J Biostat. 2010;6(1): Article 19.
25. Perez-Jaume S, Carrasco JL. A non-parametric approach to estimate the total deviation index for non-normal data. Statist Med. 2015;34(25):3318–3335.
26. Lin L, Pan Y, Hedayat AS, Banhart HX, Haber M. A simulation study of nonparametric total deviation index as a measure of agreement based on quantile regression. J Biopharm Stat. 2016;26(5):937–950. [PubMed: 26391352]

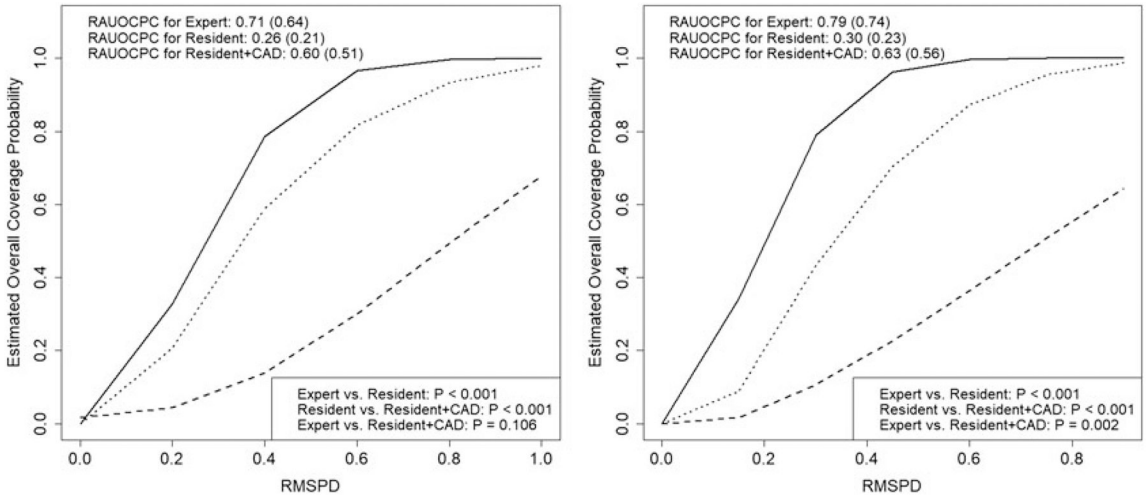

FIGURE 1.

Overall coverage probability curves based on left (Left) and right (Right) kidneys from renal study data. The solid lines indicate the estimated overall coverage probability curves for experts; the dotted lines indicate the estimated overall coverage probability curves for residents + CAD; and the dashed lines are indicate estimated overall coverage probability curves for residents

TABLE 1

Simulation results for the overall deviation index (ODI) based on 1000 simulated data sets under compound symmetry (CS) and unstructured (UN) scenarios

Scenario	*n*	Statistics	True Value	Relative Bias, %[a]	Std of Estimate[b]	Mean of SE Estimate[c]	CP[d]
CS	20	$ODI_{0.80,4}$	0.7056	−0.2474	0.0642	0.0611	0.916
		$ODI^{(C)}_{0.80,4}$		−0.5809	0.0608	0.0597	0.923
		$ODI_{0.90,4}$	0.7827	0.4463	0.0659	0.0631	0.932
		$ODI^{(C)}_{0.90,4}$		−0.7953	0.0608	0.0587	0.922
	60	$ODI_{0.80,4}$	0.7056	−0.0707	0.0369	0.0364	0.942
		$ODI^{(C)}_{0.80,4}$		−0.0731	0.0368	0.0357	0.940
		$ODI_{0.90,4}$	0.7827	0.0493	0.0382	0.0376	0.941
		$ODI^{(C)}_{0.90,4}$		−0.0229	0.0346	0.0348	0.946
UN	20	$ODI_{0.80,4}$	0.8416	0.2438	0.0966	0.0952	0.923
		$ODI^{(C)}_{0.80,4}$		−3.5240	0.0940	0.0896	0.854
		$ODI_{0.90,4}$	0.9855	0.0933	0.1025	0.0948	0.910
		$ODI^{(C)}_{0.90,4}$		−6.7998	0.0918	0.0867	0.764
	60	$ODI_{0.80,4}$	0.8416	−0.0169	0.0565	0.0565	0.944
		$ODI^{(C)}_{0.80,4}$		−3.0690	0.0548	0.0534	0.841
		$ODI_{0.90,4}$	0.9855	0.1131	0.0566	0.0563	0.948
		$ODI^{(C)}_{0.90,4}$		−6.4116	0.0519	0.0515	0.621

[a] Sample mean of 1000 relative biases calculated as $100 * \left\{(\widehat{\text{ODI}} - \text{ODI})/\text{ODI}\right\}$, where $\widehat{\text{ODI}}$s are obtained through antitransformations.

[b] Standard deviation of 1000 ODI estimates with log transformations.

[c] Sample mean of 1000 bootstrap standard error estimates.

[d] Proportion of 1000 estimated 95% upper confidence limits computed using formula (18) that are greater than the true value.

TABLE 2

Simulation results for the overall coverage probability (OCP) based on 1000 simulated data sets under compound symmetry (CS) and unstructured (UN) scenarios

Scenario	*n*	Statistics	True Value	Relative Bias, %[a]	Std of Estimate[b]	Mean of SE Estimate[c]	CP[d]
CS	20	$OCP_{0.80,4}$	0.9162	−0.6576	0.6210	0.6339	0.959
		$OCP^{(C)}_{0.80,4}$		0.2728	0.6090	0.6359	0.953
		$OCP_{0.90,4}$	0.9742	−0.6919	0.8839	0.9061	0.965
		$OCP^{(C)}_{0.90,4}$		−0.1471	0.8606	0.8760	0.951
	60	$OCP_{0.80,4}$	0.9162	−0.3581	0.3485	0.3516	0.958
		$OCP^{(C)}_{0.80,4}$		−0.0220	0.3350	0.3373	0.950
		$OCP_{0.90,4}$	0.9742	−0.2610	0.4860	0.5002	0.963
		$OCP^{(C)}_{0.90,4}$		−0.0546	0.4282	0.4465	0.959
UN	20	$OCP_{0.80,4}$	0.7620	−0.1722	0.4467	0.4683	0.964
		$OCP^{(C)}_{0.80,4}$		1.6989	0.5597	0.6017	0.939
		$OCP_{0.90,4}$	0.8465	−0.4474	0.5333	0.5643	0.963
		$OCP^{(C)}_{0.90,4}$		3.8343	0.6981	0.7353	0.898
	60	$OCP_{0.80,4}$	0.7620	−0.1676	0.2177	0.2214	0.951
		$OCP^{(C)}_{0.80,4}$		2.3916	0.3147	0.3261	0.914
		$OCP_{0.90,4}$	0.8465	−0.2431	0.2996	0.2999	0.950
		$OCP^{(C)}_{0.90,4}$		4.0765	0.3728	0.3840	0.807

[a] Sample mean of 1000 relative biases calculated as $100 * \left\{(\widehat{OCP} - OCP)/OCP\right\}$, where $\widehat{OCP}$s are obtained through antitransformations.

[b] Standard deviation of 1000 OCP estimates with logit transformations.

[c] Sample mean of 1000 bootstrap standard error estimates.

[d] Proportion of 1000 estimated 95% lower confidence limits computed using formula (19) that are smaller than the true value.

TABLE 3

Simulation results for the relative area under overall coverage probability curve (RAUOCPC) based on 1000 simulated data sets under compound symmetry (CS) and unstructured (UN) scenarios

Scenario	*n*	Statistics	True Value	Relative Bias, %[a]	Std of Estimate[b]	Mean of SE Estimate[c]	CP[d]
CS	20	RAUOCPC_4	0.4889	−0.5872	0.1388	0.1318	0.940
		$\text{RAUOCPC}_4^{(C)}$		0.1358	0.1356	0.1326	0.924
		$\text{RAUOCPC}_4^{\text{non-parm}}$		4.0914	0.1332	0.1322	0.848
	60	RAUOCPC_4		−0.1392	0.0789	0.0781	0.948
		$\text{RAUOCPC}_4^{(C)}$		−0.0930	0.0799	0.0777	0.939
		$\text{RAUOCPC}_4^{\text{non-parm}}$		1.4879	0.0792	0.0785	0.892
UN	20	RAUOCPC_4	0.4544	−0.7209	0.2080	0.1978	0.928
		$\text{RAUOCPC}_4^{(C)}$		−5.0134	0.2368	0.2223	0.955
		$\text{RAUOCPC}_4^{\text{non-parm}}$		4.5911	0.2220	0.2199	0.861
	60	RAUOCPC_4		−0.1624	0.1195	0.1167	0.941
		$\text{RAUOCPC}_4^{(C)}$		−5.0183	0.1317	0.1339	0.988
		$\text{RAUOCPC}_4^{\text{non-parm}}$		0.7797	0.1271	0.1301	0.928

[a] Sample mean of 1000 relative biases calculated as $100 * \left\{\left(\widehat{\text{RAUOCPC}} - \text{RAUOCPC}\right)/\text{RAUOCPC}\right\}$, where $\widehat{\text{RAUOCPC}}$s are obtained through antitransformations.

[b] Standard deviation of 1000 RAUOCPC estimates with log transformations.

[c] Sample mean of 1000 bootstrap standard error estimates.

[d] Proportion of 1000 estimated 95% upper confidence limits computed using formula (20) that are greater than the true value.

TABLE 4

Estimated overall deviation indices (ODIs) and overall coverage probabilities (OCPs) from renal study data

Statistics	Kidney	Experts (95% Limit)[a]	Residents (95% Limit)[a]	Resid.+CAD (95% Limit)[a]	Experts vs Residents[b]	Residents vs Resid.+CAD[b]	Experts vs Resid.+CAD[b]
$ODI_{0.9,3}$	L	0.5 (0.6)	1.4 (1.6)	0.7 (1.0)	< 0.001	< 0.001	0.107
	R	0.4 (0.5)	1.3 (1.5)	0.6 (0.8)	< 0.001	< 0.001	0.002
$OCP_{0.53}$	L	0.91 (0.72)	0.21 (0.15)	0.72 (0.53)	< 0.001	< 0.001	0.191
	R	0.98 (0.92)	0.27 (0.18)	0.77 (0.61)	< 0.001	< 0.001	0.014

[a]The 95% upper confidence limits for ODI estimates and 95% lower confidence limits for OCP estimates.

[b]The *p* values from two-sample hypothesis tests.