# MRMC analysis of agreement studies

Brandon D. Gallas[a], Amrita Anam[ac], Weijie Chen[a], Adam Wunderlich[a], Zhiwei Zhang[b]

[a]CDRH/OSEL Division of Imaging, Diagnostics, and Software Reliability, 10903 New Hampshire Ave, Silver Spring, MD, 20993;
[b]CDRH/OSB Division of Biostatistics, 10903 New Hampshire Ave, Silver Spring, MD, 20993;
[c]UMBC, Department of Information Systems, 1000 Hilltop Cir, Baltimore, MD 21250

## ABSTRACT

The purpose of this work is to present and evaluate methods based on U-statistics to compare intra- or inter-reader agreement across different imaging modalities. We apply these methods to multi-reader multi-case (MRMC) studies. We measure reader-averaged agreement and estimate its variance accounting for the variability from readers and cases (an MRMC analysis). In our application, pathologists (readers) evaluate patient tissue mounted on glass slides (cases) in two ways. They evaluate the slides on a microscope (reference modality) and they evaluate digital scans of the slides on a computer display (new modality). In the current work, we consider concordance as the agreement measure, but many of the concepts outlined here apply to other agreement measures. Concordance is the probability that two readers rank two cases in the same order. Concordance can be estimated with a U-statistic and thus it has some nice properties: it is unbiased, asymptotically normal, and its variance is given by an explicit formula. Another property of a U-statistic is that it is symmetric in its inputs; it doesn't matter which reader is listed first or which case is listed first, the result is the same. Using this property and a few tricks while building the U-statistic kernel for concordance, we get a mathematically tractable problem and efficient software. Simulations show that our variance and covariance estimates are unbiased.

## 1. PURPOSE AND PROBLEM STATEMENT

The purpose of this work is to present and evaluate methods to compare intra- or inter-reader agreement across different imaging modalities or reading conditions. We apply these methods to multi-reader multi-case (MRMC) studies. We measure reader-averaged agreement with concordance and estimate its variance accounting for the variability from readers and cases (an MRMC analysis). In our future application, pathologists (readers) evaluate patient tissue mounted on glass slides (cases) in two ways. They evaluate the slides on a microscope (reference modality) and they evaluate digital scans of the slides on a computer display (new modality). In the evaluations, the readers count mitotic figures (cells undergoing cell division). These counts are believed to correlate with cancer prognosis. The goal of the MRMC study is to demonstrate the new modality is noninferior to the reference in the mitotic figure counting task.

In this work, we proceed with the understanding that we do not have an independent reference (ground truth) by which we can evaluate the accuracy of mitotic counts. The counts from pathologists using microscopes are the best we can do and we shall use these as the reference. Survival time would be a worthy reference, but this changes the problem from determining mitotic count accuracy to determining the prognostic value of mitotic counts. Additionally the time and cost to obtain survival data can be much to bear. Another option is to use experts or a panel of experts to establish the reference counts. However, it is challenging to define an expert, it can be challenging to recruit an expert, and it is likely that there is reader variability across and within experts and this variability should be understood and counted.

Here we consider a fully-crossed study where every reader ($r = 1, 2, ...N_R$) provides a count for every case ($c = 1, 2, ..., N_C$) using the reference modality ($m = \mathrm{A}$) and the new modality ($m = \mathrm{B}$). A particular count is denoted by $X_{mrc}$. In some situations that we describe below, we may also need replicate counts (a second set of counts) using the reference modality ($m = \mathrm{A}^*$). Given this data, we can pair up every reader and modality with every other reader and modality (Fig. 1). For each of these pairs, we can calculate an agreement between the paired scores. The types of agreement we get fall into four categories:

Further author information: (Send correspondence to BDG.)
E-mail: brandon.gallas@fda.hhs.gov, Telephone: 1 301-796-2531

| Inter-Modality Agreement | | Reference Modality, *A* | | | | |
|---|---|---|---|---|---|---|
| | | Dr. 1 | Dr. 2 | Dr. 3 | ... | Dr. 8 |
| New Modality, *B* | Dr. 1 | $Q_1^{AB}$ | $P_{12}^{AB}$ | $P_{13}^{AB}$ | ... | $P_{18}^{AB}$ |
| | Dr. 2 | $P_{21}^{AB}$ | $Q_2^{AB}$ | $P_{23}^{AB}$ | ... | $P_{28}^{AB}$ |
| | Dr. 3 | $P_{31}^{AB}$ | $P_{32}^{AB}$ | $Q_3^{AB}$ | ... | $P_{38}^{AB}$ |
| | ... | ... | ... | ... | ... | ... |
| | Dr. 8 | $P_{81}^{AB}$ | $P_{82}^{AB}$ | $P_{83}^{AB}$ | ... | $Q_8^{AB}$ |

| Intra-Modality Agreement | | Reference Modality, A | | | | |
|---|---|---|---|---|---|---|
| | | Dr. 1 | Dr. 2 | Dr. 3 | ... | Dr. 8 |
| Reference Modality, *A* or *A** | Dr. 1 | $Q_1^{AA^*}$ | $P_{12}^{AA}$ | $P_{13}^{AA}$ | ... | $P_{18}^{AA}$ |
| | Dr. 2 | $P_{21}^{AA}$ | $Q_2^{AA^*}$ | $P_{23}^{AA}$ | ... | $P_{28}^{AA}$ |
| | Dr. 3 | $P_{31}^{AA}$ | $P_{32}^{AA}$ | $Q_3^{AA^*}$ | ... | $P_{38}^{AA}$ |
| | ... | ... | ... | ... | ... | ... |
| | Dr. 8 | $P_{81}^{AA}$ | $P_{82}^{AA}$ | $P_{83}^{AA}$ | ... | $Q_8^{AA^*}$ |

Figure 1: Inter- and intra-modality agreement for a set of readers. Intra-reader comparisons are on the diagonal. Inter-reader comparisons on on the off-diagonal.

1. **Intra-reader inter-modality agreement**: pair the counts of one reader using the reference modality (A) with the counts of the same reader using the new modality (B). These agreement values $Q_r^{\mathrm{AB}}$ are depicted as the gray boxes on the diagonal in the table on the left of Fig. 1. We average these over all the readers to obtain $\overline{Q}_{\cdot}^{\mathrm{AB}}$.

2. **Intra-reader, intra-modality agreement**: pair the counts of one reader using the reference modality (A) with *replicate* counts of the same reader using the reference modality (A*). These agreement values $Q_r^{\mathrm{AA}^*}$ are depicted in gray on the diagonal in the table on the right of Fig. 1. We emphasize that these agreement values require replicate readings from all the readers. We average these over all the readers to obtain $\overline{Q}_{\cdot}^{\mathrm{AA}^*}$.

3. **Inter-reader, inter-modality agreement**: pair the counts of one reader using the reference modality (A) with the counts of a different reader using the new modality (B). These agreement values $\overline{P}_{rr'}^{\mathrm{AB}}$ are depicted as the white off-diagonal boxes in the table on the left of Fig. 1. We average these over all pairs of readers to obtain $\overline{P}_{\cdot\cdot}^{\mathrm{AB}}$.

4. **Inter-reader, intra-modality agreement**: pair the counts of one reader using the reference modality (A) with the counts of a different reader also using the reference modality (A). These agreement values $\overline{P}_{rr'}^{\mathrm{AA}}$ are depicted as the white off-diagonal boxes in the table on the right of Fig. 1. Note that these agreement values do not require replicate readings. We average these over all pairs of readers to obtain $\overline{P}_{\cdot\cdot}^{\mathrm{AA}}$.

All four types of agreement use the counts from the reference modality. As such, we identify each as a measure of performance. For each inter-modality agreement measure, there is an intra-modality agreement measure. The inter-modality agreement measures characterize the performance of the new modality and the intra-modality agreement measures characterize the performance of the reference modality. Consequently, we view the intra-modality agreement measures as the baseline agreement levels for evaluating the inter-modality agreement measures.

In light of the discussion above, we would like to determine if the new modality is "equivalent" to the reference by comparing the reader-averaged inter-modality agreement to that of the intra-modality agreement measure. We can do this with non-inferiority hypothesis tests comparing the true underlying population means. Specifically, we can conduct one of these two null hypotheses

$$H_0 : \mu_Q^{\mathrm{AB}} < \mu_Q^{\mathrm{AA}^*} - \delta, \tag{1}$$

$$H_0 : \mu_P^{\mathrm{AB}} < \mu_P^{\mathrm{AA}} - \delta, \tag{2}$$

where $\delta$ is a (positive) non-inferiority margin, and the $\mu$'s are the true population means. We want to reject the null hypothesis that agreement with the new modality is worse than the baseline by the amount $\delta$.

From a practical standpoint, the *inter*-reader analysis is less burdensome than the *intra*-reader analysis. This is because the inter-reader analysis does not require replicated readings (multiple reading sessions) while the intra-reader analysis does. From a mathematical standpoint, the inter-reader analysis is more burdensome than the intra-reader analysis. This is because the inter-reader analysis requires averaging over all pairs of readers while the intra-reader analysis requires only a single average over readers. The double-average over the readers adds complexity to the problem in the form of additional correlations to be treated.

In what follows we focus on the more complex inter-reader analysis that will utilize a test statistic given by

$$t = \left(\overline{P}_{\cdot\cdot}^{\mathrm{AB}} - \overline{P}_{\cdot\cdot}^{\mathrm{AA}}\right) \Bigg/ \sqrt{\mathrm{var}\left(\overline{P}_{\cdot\cdot}^{\mathrm{AB}} - \overline{P}_{\cdot\cdot}^{\mathrm{AA}}\right)} \,. \tag{3}$$

To this end we will introduce the concordance agreement measure and discuss the U-statistic estimates of the reader-averaged concordance and corresponding variance. The intra-reader analysis can be developed following the priniciples and methods of the inter-reader agreement method.

## 2. METHODS

### 2.1 Agreement By Concordance

In this work, we measure agreement with concordance. Specifically, we measure the concordance that is defined as the probability that the counts from two readers on two cases are in the same order.[1] We estimate this concordance with the corresponding U-statistic, the unbiased non-parametric estimate.[2] Heuristically, this estimate compares the readers' counts of all unique pairs of cases. The fraction of times that the two readers' counts for a pair cases are in the same order is the U-statistic estimate of concordance (see mathematical expression below). There are four other possible outcomes for a comparison: the pair of counts from the two readers are in the opposite order (discordant), 2) the pair of counts are tied for the first reader but not the second, 3) the pair of counts are tied for the second reader but not the first, 4) the pair of counts are tied for both readers.

We consider the U-statistic estimate of concordance because it has some nice properties: it is unbiased, asymptotically normal, and its variance is given by an explicit formula.[2] To make use of the U-statistic properties, we need to identify the kernel of the U-statistic. To this end, let $x, x'$ be the counts for the "first" reader on two cases and $y, y'$ be the counts for the "second" reader on the same two cases. Then the U-statistic kernel for concordance is

$$S\left(x, x'; y, y'\right) = \begin{cases} 1.0 & \left(x - x'\right)\left(y - y'\right) > 0 \\ 0.0 & \left(x - x'\right)\left(y - y'\right) \leq 0 \end{cases} . \tag{4}$$

Notice that this kernel is symmetric with respect to the cases, as is required for it to be a U-statistic kernel: $S\left(x, x'; y, y'\right) = S\left(x', x; y', y\right)$. It is also zero when a case is compared to itself.

We can add some notation and generalize the concordance to treat the same reader under two reading conditions: different modalities or different observations from a single modality (replications). Specifically, we define concordance values for every reader $r$ or every pair of readers $r, r'$ in Fig. 1, and we label them with $m, m'$ according to the modality or reading condition:

$$Q_r^{mm'} = \frac{2}{N_C\left(N_C - 1\right)} \sum_c \sum_{c' > c} S\left(x_{mrc}, x_{mrc'}; x_{m'rc}, x_{m'rc'}\right), \tag{5}$$

$$P_{rr'}^{mm'} = \frac{2}{N_C\left(N_C - 1\right)} \sum_c \sum_{c' > c} S\left(x_{mrc}, x_{mrc'}; x_{m'r'c}, x_{m'r'c'}\right), \tag{6}$$

where $N_C$ is the number of cases. For this work, $m, m'$ represent the reference modality (A), a replicate of the reference modality (A$^*$), or the new modality B.

Notice for a moment the normalization in the expressions above. The normalization $N_C\left(N_C - 1\right)/2$ equals the binomial coefficient $\binom{N_C}{2}$, the number of ways to pick an unordered pair (without replacement) from $N_C$

cases. This normalization corresponds to the number of elements in the upper triangle of an $N_C \times N_C$ matrix (not counting the diagonal). Since the kernel is symmetric with respect to cases and it is zero when $c = c'$, we can instead sum over all $N_C \times N_C$ combinations of $c$ and $c'$ and normalize the result by the number of ways to pick an ordered pair of cases with replacement from $N_C$ cases: $2! \times \binom{N_C}{2}$. Converting the traditional U-statistic expression, which only sums over unique pairs of observations, into sums over all the elements, is a trick that will simplify some extra complex expressions later on.

### 2.2 Reader-Averaged Concordance

The reader-averaged inter-reader concordance $\overline{P}_{\cdot\cdot}^{mm'}$ is a two-sample U-statistic; it is an average over the readers (sample 1) and the cases (sample 2), assuming the modalities are fixed. Furthermore, $\overline{P}_{\cdot\cdot}^{mm'}$ is degree 2 for the reader sample and degree 2 for the case sample; you need two different readers and two different cases to calculate one observation of inter-reader concordance. ($\overline{Q}_{\cdot}^{mm'}$ is also a two sample U-statistic; however, it is only degree 1 for the reader while still degree 2 for the cases.) Functionally, the MRMC kernel for $\overline{P}_{\cdot\cdot}^{mm'}$ takes any two unique readers ($r' \neq r$) reading any two unique cases ($c' \neq c$). If $m' = m$, the kernel estimates intra-modality concordance. If $m' \neq m$, the kernel estimates inter-modality concordance. We write the inter-reader kernel as

$$\Phi_{rr'cc'}^{mm'} = \frac{1}{2}\left(S\left(X_{mrc}, X_{mrc'}; X_{m'r'c}, X_{m'r'c'}\right) + S\left(X_{mr'c}, X_{mr'c'}; X_{m'rc}, X_{m'rc'}\right)\right) \times \left(1 - \delta_{rr'}\right)\left(1 - \delta_{cc'}\right) \tag{7}$$

where $\Phi$'s and $X$'s are random variables ($\phi$'s and $x$'s are observed data), and $\delta_{jj'} = 1$ if $j = j'$ and 0 otherwise. The expectation of $\Phi_{rr'cc'}^{mm'}$ is the true population mean of the inter-reader concordance $\mu_P^{mm'}$. (There is an analogous kernel and mean for $\overline{Q}_{\cdot}^{mm'}$.) The first quantity on the right functionally provides concordance. It has two terms that symmetrize the kernel with respect to the reader-modality combination $\Phi_{rr'cc'}^{mm'} = \Phi_{r'rcc'}^{mm'}$. The MRMC kernel is already symmetric with respect to the cases thanks to $S$. The next two terms on the right (with the delta functions) functionally evaluate to zero for $r' = r$ and for $c' = c$; they set the "diagonals" of the kernel to zero. We don't technically need these because we only compare pairs of unique readers and pairs of unique cases for inter-reader concordance. We also don't need $\delta_{cc'}$ for the kernel in Eq. 4, because it actually evaluates to zero when the cases are the same. However, here we want to emphasize that our methods below require the diagonals to be zero. Zeros on the diagonals will help us convert sums over unique observations to sums over all observations.

The textbook expression for our U-statistic only sums over $r' > r$ and $c' > c$. However, thanks to the symmetries and zero-diagonals built into our U-statistic, we can write this with complete sums over $r, r', c, c'$ :

$$\overline{P}_{\cdot\cdot}^{mm'} = \frac{1}{2!\binom{N_C}{2}} \frac{1}{2!\binom{N_R}{2}} \sum_r \sum_{r'} \sum_c \sum_{c'} \phi_{rr'cc'}^{mm'}, \tag{8}$$

where $N_R$ is the number of readers.

### 2.3 MRMC (Co)variance of Reader-Averaged Concordance

Thanks to the properties of U-statistics, the covariance of $\overline{P}_{\cdot\cdot}^{mm'}$ and $\overline{P}_{\cdot\cdot}^{m^*m^{**}}$ can be written as

$$\operatorname{cov}\left(\overline{P}_{\cdot\cdot}^{mm'}, \overline{P}_{\cdot\cdot}^{m^*m^{**}}\right) = \sum_{k=0}^{2} \sum_{k'=0}^{2} \frac{\binom{2}{k}\binom{N_R-2}{2-k}\binom{2}{k'}\binom{N_C-2}{2-k'}}{\binom{N_R}{2}^{-1}\binom{N_C}{2}^{-1}} \left(M_{kk'}^{mm',m^*m^{**}} - M_{00}^{mm',m^*m^{**}}\right), \tag{9}$$

which is built from nine constituent moments

$$M_{kk'}^{mm',m^*m^{**}} = E\left(\Phi_{rr'cc'}^{mm'}\Phi_{r^*r^{**}c^*c^{**}}^{m^*m^{**}} \;\middle|\; \begin{array}{c} k \text{ readers in common} \\ \text{and } k' \text{ cases in common} \end{array}\right). \tag{10}$$

The expressions above are a bit complex, but they are clear and explicit. To help parse them, here are some notes to help:

1. The variance of a reader-averaged inter-reader concordance is the covariance of that concordance with itself; namely, $\text{var}\left(\overline{P}_{\cdot\cdot}^{\text{AB}}\right) = \text{cov}\left(\overline{P}_{\cdot\cdot}^{\text{AB}}, \overline{P}_{\cdot\cdot}^{\text{AB}}\right)$.
2. "$k$ readers in common" is an abbreviation for "the sets $\{r, r'\}$ and $\{r^*, r^{**}\}$ have $k$ readers in common, where $k = 0, 1,$ or 2."
3. "$k'$ cases in common" is an abbreviation for "the sets $\{c, c'\}$ and $\{c^*, c^{**}\}$ have $k'$ cases in common, where $k' = 0, 1,$ or 2."
4. $M_{00}^{mm',m^*m^{**}} = \mu_P^{mm'}\mu_P^{m^*m^{**}}$ is analogous to the mean squared term of a variance.
5. Traditional U-statistics books use constituent covariances $\zeta$ (central second-order moments) instead of non-central second-order moments: $\zeta_{kk'}^{mm',m^*m^{**}} = M_{kk'}^{mm',m^*m^{**}} - M_{00}^{mm',m^*m^{**}}$.

### 2.4 (Co)variance Estimation

Simply put, we will estimate the covariance above using U-statistics. Specifically, we will apply U-statistics to each constituent moment and insert the estimates into Eq. 9. The challenge, however, is executing the estimate in a reasonable amount of time. To demonstrate this challenge, let's consider $M_{00}^{\text{AA,AB}}$, which looks simple enough given Note #4 above, but is actually the most complex term to estimate with U-statistics.

To begin, we point out that $\overline{P}_{\cdot\cdot}^{\text{AA}}\overline{P}_{\cdot\cdot}^{\text{AB}}$ is an estimate of $M_{00}^{\text{AA,AB}}$, but it is biased (the mean of the product does not equal the product of the means). Consequently, $\overline{P}_{\cdot\cdot}^{\text{AA}}\overline{P}_{\cdot\cdot}^{\text{AB}}$ is not the U-statistic estimate.

To derive the U-statistic estimate, we first need to specify the kernel. In the current case, we know that the kernel is a symmetrized version of $\Phi_{rr'cc'}^{\text{AA}}\Phi_{r^*r^{**}c^*c^{**}}^{\text{AB}}$. The subscripts of this random variable identify four unique readers and four unique cases (none of the readers or cases are common in the 00 term). The corresponding U-statistic for $M_{00}^{\text{AA,AB}}$ is a two sample U-statistic of degree **four** for both samples. Since there are 4! ways to order the four reader subscripts, it takes 24 terms to (blindly) symmetrize the kernel with respect to readers. Likewise, it takes 24 terms to symmetrize the kernel with respect to cases, and 24×24 terms to symmetrize the kernel for both readers and cases. Consequently, the U-statistic estimate of this moment sums this complicated symmetric kernel over all combinations of four different readers ($N_R = 4$) and four different cases ($N_C = 4$). In short, the estimate has eight sums. Specifically, if we collect a dataset from 10 readers reading 50 cases in 2 modalities, there are on the order of $\binom{10}{4} \times \binom{50}{4} \times 2 \times 24 \times 24 = 55.7$ billion operations.

The calculation above is what you get from a brute force approach. When we apply some thought to the process, we notice that, of the 24 terms needed to symmetrize with respect to the readers, only 6 terms are unique thanks to the symmetries in $\Phi$. Likewise, there are 6 unique terms needed to symmetrize the kernel with respect to cases. This means that there are "only" 36 terms needed to symmetrize the kernel for readers and cases.

While the 36-term kernel still sounds unbearable, we can apply the trick outlined above. We set the diagonals of the kernel to zero; we use delta functions to force the kernel to be zero whenever the same reader appears twice or the same case appears twice. We can do this because the default U-statistic expression is only evaluated for different readers and different cases. Now we can sum over all the readers and all the cases instead of only summing unique combinations. The magic is that now all 36 terms of the symmetric kernel sum to the same result. For the dataset mentioned above, there are more operations than when we started $\left(10^4 \times 50^4 \times 2 = 125 \text{ billion}\right)$, but the kernel is just one term and a bunch of delta functions.

In the last step, we use the delta functions to eliminate summations and create perfect squares. At the end of this step, the U-statistic estimate of $M_{00}^{\text{AA,AB}}$ is a linear combination of nine sums of squares. The number of operations required of the sums of squares is the square root of the number of operations if the sums cannot be reduced to sums of squares. For the dataset mentioned above, the total number of operations is now on the order of $10^2 \times 50^2 \times 2 = 500\,000$. We essentially visit each pair of readers and pair of cases one time.

The full derivation of the U-statistic estimate is quite tedious and beyond the scope of this paper. It will be published at a later date.

### 2.5 Simulation Model

The simulation model that we present here does not wholly model our future application: counting mitotic figures. In this early stage of our work, we need a simpler simulation that is not confounded by ties generated by discrete data and the lower limit of zero. As such we present a simulation of continuous random variables that starts with the familiar MRMC ROC (receiver operating characteristic) simulation model of Roe and Metz (R&M).[3] We then discuss how we adapt this simulation model to better represent an agreement study.

The R&M simulation model is given by

$$X_{ijkt}^{\text{R\&M}} = \mu_t + \tau_{it} + R_{jt} + C_{kt} + [RC]_{jkt} + [\tau R]_{ijt} + [\tau C]_{ikt} + [\tau RC]_{ijkt} + E_{ijkt} \tag{11}$$

where $X_{ijkt}$ denotes the value of the ROC rating (score) for modality $i$, reader $j$, case $k$, and truth state $t$. Modality and truth are fixed factors, and reader and case are random factors. Consequently, the Greek terms are fixed effects: a baseline mean per truth $\mu_t$, and modality-specific means per truth $\tau_{it}$. The remaining terms are random, modeled as independent zero-mean Gaussian random variables, each with its own variance. The seven terms include a reader effect $R$, a case effect $C$, a reader-case interaction $[RC]$, and interactions with modality $\tau$. The last term $E_{ijkt}$ is an independent error term attributable to reader jitter, internal noise. Since we consider replicated readings for the intra-reader intra-modality agreement analysis, we assume the ROC rating $X$ and the independent error term $E$ include an subscript $m$ for replicated readings.

The first step we take to adapt the R&M model for an MRMC agreement study is the treatment of truth. In an ROC study, there are two truth states. In an agreement study, there is generally a spectrum of truth, and the truth does not depend on modality. In this new setting, we let $\mu_k$ be the true state of case $k$ and we eliminate the $t$ subscript. We also eliminate the fixed effect $\tau_{it}$ as it implies that truth changes with modality. For the mitotic counting application, $\mu_k$ is the true number of mitotic figures in the sample field of view of case $k$. In the current model, we will let $\mu_k$ be sampled from a standard normal distribution.

In the next step of our adaptation, we reconfigure how the reader and case effects are treated. For the reader terms $R_j, [\tau R]_{ij}$, we recognize that they actually have no impact on concordance as they appear above. The difference between two cases in the calculation of concordance (Eq. 4) erases the reader terms. Since the reader terms above have no effect, we must find a different way to include reader variability in the new model. For the case terms, the pure case term in the R&M model $C_k$ is redundant with the true state of the case $\mu_k$, while the modality-case interaction term $[\tau C]_{ik}$ does allow for there to be a modality effect on agreement. In the new model, we treat the case terms in a similar fashion to how we treat the reader terms.

The new model has reader-case interaction effects defined from separate reader and case effects:

$$[RC]_{jk} \sim N\left(0, (R_j + C_k)^2\right), \tag{12}$$

$$[\tau RC]_{ijk} \sim N\left(0, \left([\tau R]_{ij} + [\tau C]_{ik}\right)^2\right), \tag{13}$$

where

$$R_j \sim \exp(\text{ mean} = \mu_R), \tag{14}$$

$$[\tau R]_{ij} \sim \exp(\text{ mean} = \mu_{\tau R}), \tag{15}$$

$$C_k \sim \exp(\text{ mean} = \mu_C), \tag{16}$$

$$[\tau C]_{ik} \sim \exp(\text{ mean} = \mu_{\tau C}). \tag{17}$$

The reader and case effects are now specified by means $\mu_R, \mu_{\tau R}, \mu_C, \mu_{\tau C}$ instead of variances. We consider the reader terms to represent the variability of a reader's measurement of the truth. Higher values for $[R]_j, [\tau R]_{ij}$ correspond to readers that give noisier counts. We consider the case terms to reflect the variability a case evokes

from readers. Higher values of $C_k$ or $[\tau C]_{ik}$ correspond to cases that are more challenging to count, cases that elicit more variable counts by the readers.

In the final step of our adaptation, we recognize that the only difference between the replicated counts from a reader is the independent error term $E_{ijkm}$. These independent error terms are all (presumably) drawn from a single distribution in the R&M model; there is no dependence on the reader or case. Consequently the true, or expected, intra-modality concordance of all readers is the same. In other words, the R&M model of the independent error term yields an MRMC agreement study with no reader variability. To rectify this, we model the replication error term in the same way as the reader case interaction effects:

$$[RCE]_{jkm} \sim N\left(0, \left([RE]_j + [CE]_k\right)^2\right), \tag{18}$$

$$[\tau RCE]_{ijkm} \sim N\left(0, \left([\tau RE]_{ij} + [\tau CE]_{ik}\right)^2\right), \tag{19}$$

where

$$[RE]_j \sim \exp\left(1/\mu_{RE}\right), \tag{20}$$

$$[\tau RE]_{ij} \sim \exp\left(1/\mu_{\tau RE}\right), \tag{21}$$

$$[CE]_k \sim \exp\left(1/\mu_{CE}\right), \tag{22}$$

$$[\tau CE]_{ik} \sim \exp\left(1/\mu_{\tau CE}\right). \tag{23}$$

As above, we model the impact on the replication error from the readers and cases with exponential distributions specified by means $\mu_{RE}, \mu_{\tau RE}, \mu_{CE}, \mu_{\tau CE}$, and we interpret the terms as specifiying the replication variability of each reader and the replication variability evoked by each case.

The final model is then

$$X_{ijkm} = \mu_k + [RC]_{jk} + [\tau RC]_{ijk} + [RCE]_{jkm} + [\tau RCE]_{ijkm}, \tag{24}$$

which requires the specification of eight parameters: 4 parameters are interpreted as the mean variability of a reader and 4 parameters are interpreted as the mean variability evoked by a case.

## 3. RESULTS

### 3.1 Simulation Experiments

For every simulation configuration in this work, we run $N_{MC} = 10,000$ trials and we investigate two several study sizes including a small study of 6 readers and 60 cases and a large study of 15 readers and 150 cases. We use Monte Carlo means and variances as surrogates for true means and variances. The simulation parameters we choose to explore carve out a small portion of the eight-dimensional space of possibilities. Specifically, we set

$$\mu_R = \mu_{\tau R} = \mu_{RE} = \mu_{\tau RE}, \text{ and} \tag{25}$$

$$\mu_C = \mu_{\tau C} = \mu_{CE} = \mu_{\tau CE}. \tag{26}$$

These constraints to the simulation parameters provide a manageable and interesting portion of the simulation space. It is interesting because the constraint $\mu_{\tau R} = \mu_{RE}$ combined with $\mu_{\tau C} = \mu_{CE}$ also yields distributions where inter-modality concordance equals the intra-modality concordance: $\mu_Q^{\mathrm{AB}} = \mu_Q^{\mathrm{AA}^*}$ and $\mu_P^{\mathrm{AB}} = \mu_P^{\mathrm{AA}}$. In words, the constraints are equating the variability that comes from counting with the new modality (B) to the variability that comes from replicating the counts on the reference modality. The rest of the constraints force a balance between the variability arising from terms with the modality effect $\left([\tau RC]_{ijk}, [\tau RCE]_{ijkm}\right)$ and terms without the modality effect $\left([RC]_{jk}, [RCE]_{jkm}\right)$.

We vary the parameters in a factorial way, exploring $\mu_R = [0.05, 0.2, .4, .8]$ and $\mu_C = [0.05, 0.2, .4, .8]$, for a total of 16 simulation configurations. The levels of the mean reader variability and the mean case-evoked

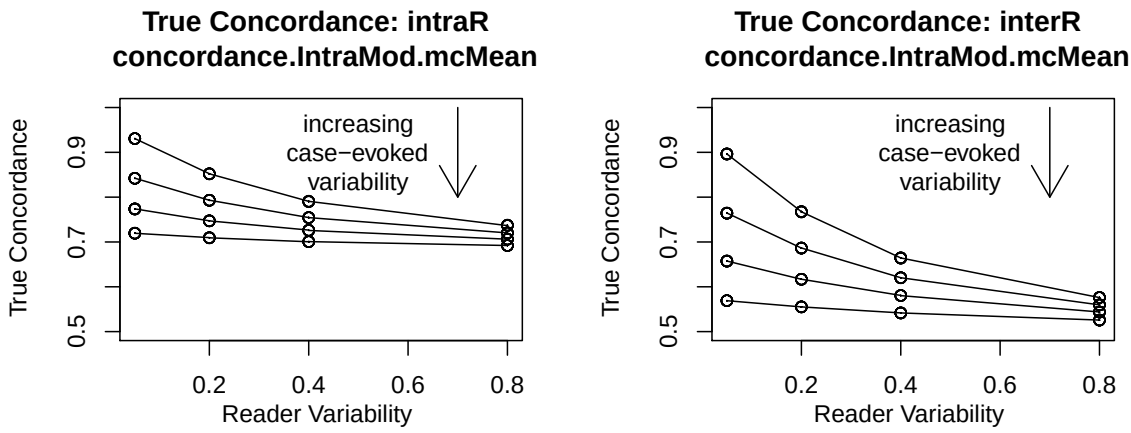


Figure 2: Monte Carlo means of reader-averaged, intra-reader and inter-reader, intra-modality concordances.

variability yield reader-averaged intra-reader intra-modality concordance results ranging from 0.93 to 0.69 and reader-averaged inter-reader intra-modality concordance results ranging from 0.90 to 0.53. These concordance results are shown in Fig. 2. In Fig. 3 we show the corresponding true variance for the small study (top two plots) and the large study (bottom two plots).

In Fig. 4 we show scatter plots of intra- and inter-reader intra-modality concordance of four select readers generated in one Monte Carlo trial ($\mu_R = 0.2,\ \mu_C = 0.05$). The scatter-plots are organized in a manner mimicking the table on the right of Fig. 1, and sorted by intra-reader concordance. The lines in the plots show a band of equivalence to highlight different levels of agreement. The range of intra-reader concordance is higher and wider than inter-reader concordances, though both show a fair amount of reader variability. The plots in the upper-triangle region of the figure are mirror images of those in the lower-triangle region. This would not be the case for inter-modality agreement.

In Figs. 5, 6 we show the performance characteristics of our U-statistic estimator. In Fig. 5 we show the Monte Carlo mean of the variance estimates as a function of the true variance. As expected, our U-statistic estimator of the variance of reader-averaged concordance is unbiased. In Fig. 6 we show the relative standard error of our U-statistic variance estimator for the small study (top two plots) and the large study (bottom two plots). The relative standard error equals the square-root of the Monte Carlo variance of our variance estimates divided by the true variance of concordance and is equivalent to the coefficient of variation for unbiased estimators. These results show the precision of the U-statistic variance estimator. For the small study, the estimator struggles to be precise. For the large study, the relative standard error is between 20-40%.

## 4. DISCUSSION AND CONCLUSION

In this work we present MRMC study designs and analysis methods for intra- and inter-reader agreement. The analysis methods are novel and include point estimates and variance estimates or reader-averaged concordance that are based on U-statistics. We use simulations to evaluate the methods. The simulations show that we can estimate $\text{var}\left(\overline{P}_{\cdot\cdot}^{mm'}\right)$ and $\text{cov}\left(\overline{P}_{\cdot\cdot}^{\text{AB}}, \overline{P}_{\cdot\cdot}^{\text{AA}}\right)$ in an unbiased way. Unfortunately, the precision of the variance estimates is not outstanding and future work needs to improve on the performance shown here.

We focus our attention in this paper on strict concordance. This is appropriate for the continuous data in our simulation. For situations where there are ties in readers' scores, we might want to separately estimate the rate of ties or somehow account for ties in the agreement measure. Knowing the rate of ties is important as it limits the rate of concordance and also says something about reader behavior with the scoring method. To separately estimate the rate of ties (and estimate the corresponding variance), we only need to change the base

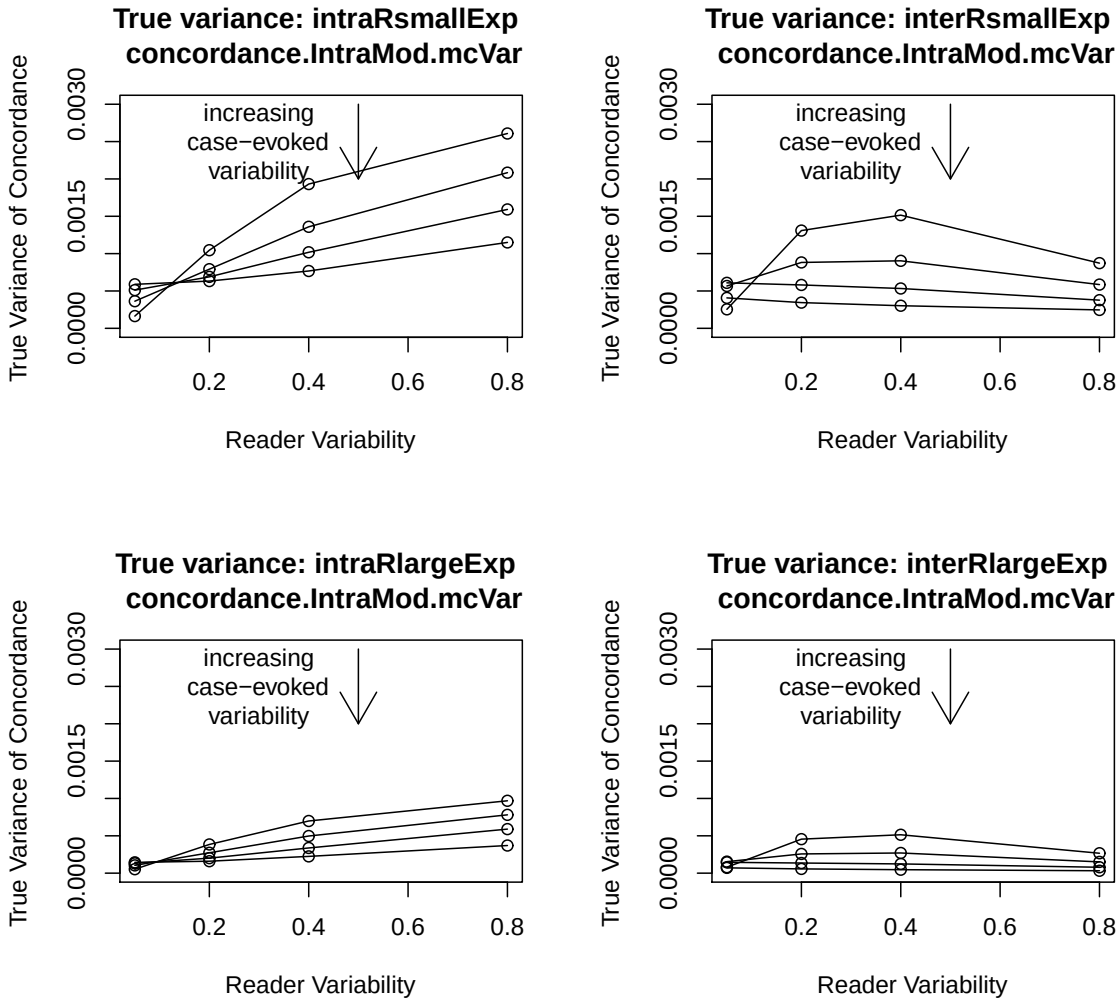


Figure 3: Monte Carlo variances of reader-averaged, intra-reader and inter-reader, intra-modality concordances for a small experiment of 6 readers and 60 cases (top two plots) and a large experiment of 15 readers and 150 cases (bottom two plots).

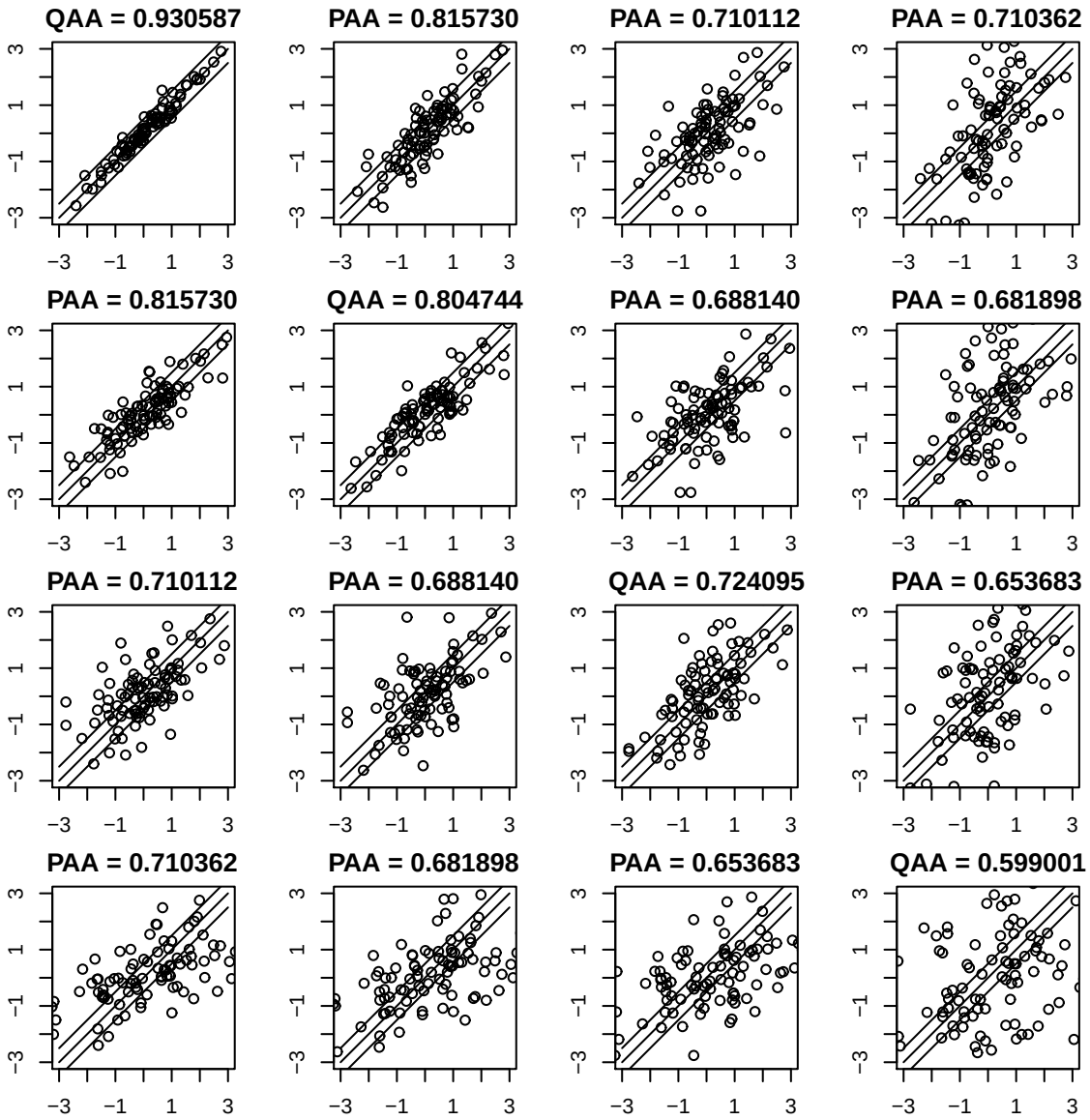


Figure 4: On the diagonal of this array of plots we show scatter-plots of intra-reader intra-modality agreement of four select readers generated in one Monte-Carlo trial. The concordance ranges from 0.93 to 0.60. The lines simply show a band about equivalence to help show different levels of agreement. On the off-diagonal of this array of plots we show scatter plots of inter-reader intra-modality agreement of the same four select readers. The concordance ranges from 0.82 to 0.65. The plots in the upper-triangle region of the figure are mirror images of those in the lower-triangle region. This would not be the case for inter-modality agreement.

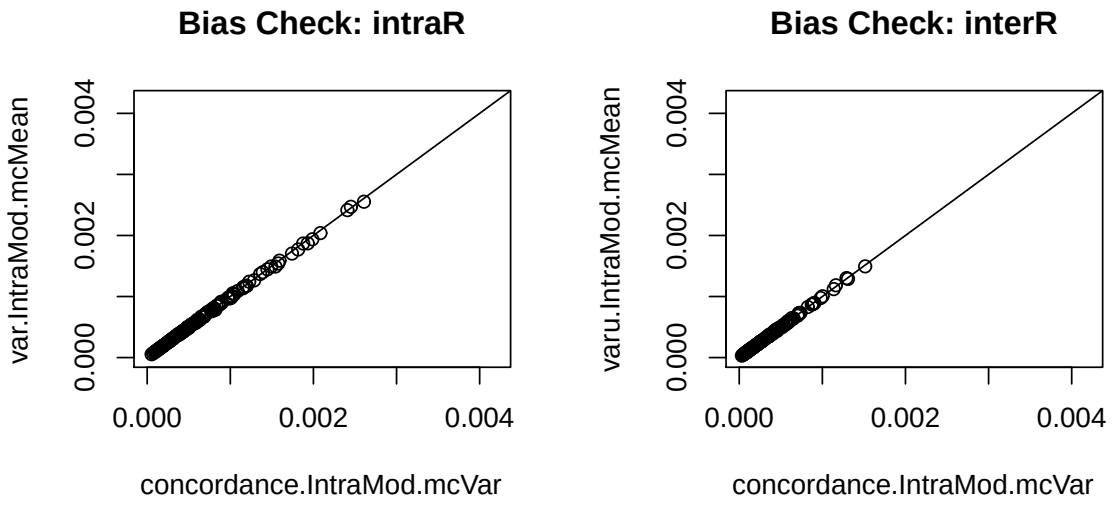


Figure 5: These two plots show estimated variances of intra-reader concordance versus the corresponding "true"variances (intra-modality on left, inter-modality on right).

kernel (Eq. 4) to count ties instead of concordances. An alternative to separately estimating the rate of ties, is to use other estimates of agreement. Some other estimates of agreement can be treated easily by changing the base kernel,[4] and others can be treated by using the Delta method to combine several U-statistics.[1,5–8]

## REFERENCES


1. J. Kim, "Predictive measures of ordinal association," *Am J Sociol* **76**(5), pp. 891–907, 1971.
2. R. H. Randles and D. A. Wolfe, *Introduction to the Theory of Nonparametric Statistics*, John Wiley and Sons, New York, 1979.
3. C. A. Roe and C. E. Metz, "Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic (ROC) data: Validation with computer simulation," *Acad Radiol* **4**(4), pp. 298–303, 1997.
4. N. A. Obuchowski, "An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale," *Stat Med* **25**(3), pp. 481–493, 2006.
5. W. D. Smith, R. C. Dutton, and N. T. Smith, "A measure of association for assessing prediction accuracy that is a generalization of non-parametric ROC area," *Stat Med* **15**(1), pp. 1199–1215, 1996.
6. R. H. Somers, "A new asymmetric measure of association for ordinal variables," *Am Sociol Rev* **27**, pp. 799–811, 1962.
7. M. G. Kendall, *Rank Correlation Methods*, Griffin & Co., London, 1962.
8. L. A. Goodman and W. H. Kruskal, "Measures of association for cross classifications," *J Am Stat Assoc* **49**(268), pp. 732–764, 1954.

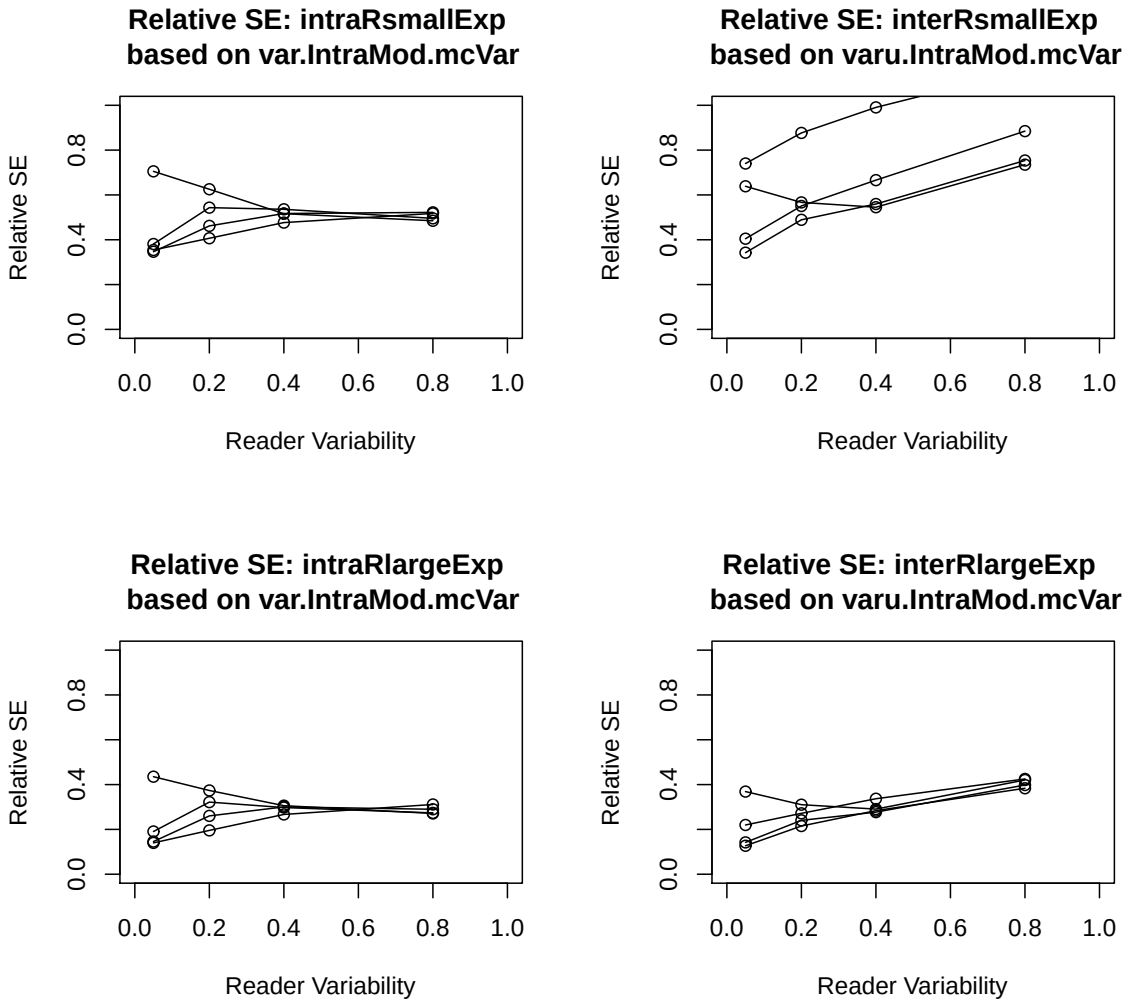


Figure 6: These plots show the relative standard error for a small experiment of 6 readers and 60 cases (top two plots) and a large experiment of 15 readers and 150 cases (bottom two plots). The relative standard error equals the Monte Carlo variance of our variance estimates divided by the true variance of concordance and is equivalent to the coefficient of variation for unbiased estimators.