Statistical Methods in Medical Research
http://smm.sagepub.com/

Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting

Huiman X Barnhart, Eric Yow, Anna Lisa Crowley, Melissa A Daubert, Dawn Rabineau, Robert Bigelow, Michael Pencina and Pamela S Douglas

Stat Methods Med Res published online 14 May 2014
DOI: 10.1177/0962280214534651

The online version of this article can be found at:
http://smm.sagepub.com/content/early/2014/05/14/0962280214534651

Published by:
SAGE
http://www.sagepublications.com

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20

DOI: 10.1177/0962280214534651
smm.sagepub.com
SAGE

Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting

Huiman X Barnhart, Eric Yow, Anna Lisa Crowley, Melissa A Daubert, Dawn Rabineau, Robert Bigelow, Michael Pencina and Pamela S Douglas

Abstract

Clinical core laboratories, such as Echocardiography core laboratories, are increasingly used in clinical studies with imaging outcomes as primary, secondary, or surrogate endpoints. While many factors contribute to the quality of measurements of imaging variables, an essential step in ensuring the value of imaging data includes formal assessment and control of reproducibility via intra-observer and inter-observer reliability. There are many different agreement/reliability indices in the literature. However, different indices may lead to different conclusions and it is not clear which index is the preferred choice as an overall indication of data quality and a tool for providing guidance on improving quality and reliability in a core lab setting. In this paper, we pre-specify the desirable characteristics of an agreement index for assessing and improving reproducibility in a core lab setting; we compare existing agreement indices in terms of these characteristics to choose a preferred index. We conclude that, among the existing indices reviewed, the coverage probability for assessing agreement is the preferred agreement index on the basis of computational simplicity, its ability for rapid identification of discordant measurements to provide guidance for review and retraining, and its consistent evaluation of data quality across multiple reviewers, populations, and continuous/categorical data.

Keywords

agreement, reliability, reproducibility, method comparison, measurement error

1 Introduction

Clinical core laboratories, such as Echocardiography core laboratories (ECL), are increasingly used in clinical studies with imaging outcomes as primary, secondary, or surrogate endpoints.

Duke Clinical Research Institute, Duke University Medical Center, Durham, USA

Corresponding author:
Huiman X Barnhart, Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke Medical Center, PO Box 17969, Durham NC 27715, USA.
Email: huiman.barnhart@dm.duke.edu

Echocardiography provides non-invasive assessment of cardiac structure, function, and hemodynamics which yield important data on safety and efficacy of drugs and devices as well as insight into disease mechanisms. Measurements from echocardiography can be used for enrollment eligibility, surrogate or primary endpoints in cardiovascular clinical research. The validity and interpretation of the results based on these measurements highly depend on the reliability of the measurements. For example, it is difficult to interpret the results from the Predictors of Response to CRT (PROSPECT) study[1] on investigation of echocardiographic predictors of response to cardiac resynchronization therapy because there was limited agreement between the three ECLs involved and poor reproducibility of some measures.

Because truth or a gold standard is often not available for the measurements obtained in clinical studies, only the reproducibility or precision, rather than the accuracy of the measurements can be assessed and controlled. In order to reduce variability and improve the reproducibility of study results, ECLs have been recommended for use in cardiovascular clinical research and have been increasingly used in clinical trials.[2,3] American Society of Echocardiography (ASE) statement, ''Recommendations for Use of Echocardiography in Clinical Trials,''[2] describes the importance of high-quality imaging for research. A recent ASE consensus statement[3] laid out the echocardiography standards for ECLs. These standards provide guidance on many aspects of the responsibilities and organization of ECLs that include (1) development of the trial or substudy imaging design, (2) training of sonographers and other personnel involved in image acquisition, (3) oversight of acquisition of images, (4) analysis of echocardiographic data, (5) providing quality assurance, (6) information technology services (such as image management (digitization, transfer, storage) and data management), (7) interpretation of data, and (8) preparation of data reports, manuscripts, and support for regulatory submissions. In this paper, we focus on the aspect of (5) providing quality assurance when performing quality assessment (QA) and quality control (QC). The approaches described here are applicable to any clinical core labs where ECLs are used as a motivating example.

While everyone agrees that it is critical to pay special attention to the quality of measurements of the imaging variables, specific standards or a preferred method does not exist for efficient and consistent implementation of QA and QC on intra-reader and inter-reader reproducibility (IIR) in ECLs. There has been vast statistical literature in the last several decades on assessing agreement and reliability.[4] As noted by Barnhart et al.,[4] agreement is a broader concept than reliability in the sense that reliability is a subset of indices for assessing agreement. We use agreement and reliability interchangeably in this paper without getting into the subtle difference between the two because the existing literature often treat the two concepts interchangeably when there is no gold standard or truth available. While some of these indices may lead to different conclusions[5] as shown below, a preferred choice for evaluating and improving IIR in a core lab setting has not emerged. For example, intra-class correlation coefficient (ICC) and kappa statistic have often been used for assessing IIR for continuous and categorical variables, respectively.[6] However, both ICC and kappa are scaled indices that are sometimes counterintuitive and may not be an ideal tool for QC and quality improvement in a core lab setting. Specifically, the measurement error (intuitively understood as the differences between paired measurements on the same subject) may be the same, but the ICC can be high in heterogeneous population when there is large between-subject variability, but low in homogenous population when there is small between-subject variability.[4,5] The following data examples illustrate this issue by using data from two studies conducted at the Duke ECL. Table 1 provides the data of biplane left ventricular ejection fraction (LVEF) measured by the same two sonographers on 10 different echocardiograms from two studies. Table 2 provides the summary statistics and the corresponding ICC values for assessing inter-observer reliability of

Table 1. Biplane LV ejection measurements by the same two sonographers in two different projects.

Project 1				Project 2			
Echo ID	Reader 1	Reader 2	Difference	Echo ID	Reader 1	Reader 2	Difference
101	54.1467	60.8467	−6.70000	201	58.9800	58.5800	0.40000
102	39.5467	36.7600	2.78667	202	63.9100	62.0100	1.90000
103	21.8600	27.2967	−5.43667	203	65.3300	67.1000	−1.77000
104	31.6233	29.7967	1.82667	204	60.4700	66.1600	−5.69000
105	62.7933	62.7800	0.01333	205	60.9400	57.0700	3.87000
106	40.0100	49.8200	−9.81000	206	68.8400	67.6900	1.15000
107	34.7500	36.6533	−1.90333	207	59.3100	53.0400	6.27000
108	30.9667	27.6033	3.36333	208	58.1900	60.2500	−2.06000
109	66.0767	70.6333	−4.55667	209	62.8500	63.6800	−0.83000
110	40.5000	47.2967	−6.79667	210	56.4600	58.5600	−2.10000

Table 2. Summary data of biplane LV ejection by the same two readers in two different projects.

	Project 1 ($N = 10$)	Project 2 ($N = 10$)
LV EF–reader 1		
Mean (SD)	42.2 (14.37)	61.5 (3.73)
Min, Max	21.9, 66.1	56.5, 68.8
LV EF – reader 2		
Mean (SD)	44.9 (15.77)	61.4 (4.79)
Min, Max	27.3, 70.6	53.0, 67.7
Difference (reader 1–reader 2)		
Mean (SD)	−2.7 (4.59)	0.1 (3.40)
Min, Max	−9.8, 3.4	−5.7, 6.3
ICC (95% CI)	0.94 (0.80, 0.99)	0.71 (0.22, 0.92)

ICC: intra-class correlation coefficient; LVEF: left ventricular ejection fraction.

the two sonographers. Based on ICC values of 0.94 and 0.71 in the two projects (although not statistically significant here, but the two ICCs can be statistically significant with large sample size), one would expect to see larger differences in project 2 than project 1. However, the data show the opposite where bigger differences were observed more in project 1 than in project 2 (see Table 1). This counterintuitive observation is confusing in the sense that higher ICC does not always imply that the measurements by two observers are more precise. This is because that ICC is a dimensionless index and its evaluation of measurement error is relative to the magnitude of the between-subject variability, rather than relative to the distance between the measurements. This is why sometimes that ICC is interpreted as the ability to differentiate subjects.[4,7] With higher ICC in project 1 than project 2, this data example indicates that it is easier to differentiate subjects in project 1 than project 2. This may lead to inadvertently accept larger difference for project population with larger between-subject variability. Thus, the commonly used agreement index of ICC may not be the

preferred index to use in terms of data quality indication and quality improvement. Therefore, there is a need to evaluate the available agreement indices in the literature and select a preferred index for QA and QC in terms of IIR in a core lab setting.

In this paper, we specify desirable characteristics of an agreement index in order to choose a preferred agreement index for QA and QC in a research core lab setting with ECL as the motivating example. In section 2, we first discuss the QA/QC setting in a clinical core laboratory that lead to the desirable pre-specified characteristics in an agreement index; then existing agreement/reliability indices are reviewed in terms of these characteristics in order to make a choice. The coverage probability (CP) index would emerge as the preferred choice weighting the pros of cons of these existing indices. Several data examples from the Duke imaging core lab are used in section 3 to illustrate the use of CP and other agreement indices in the IIR assessment process. We conclude with our discussion in section 4.

2 Agreement indices as IIR assessment tool for QA/QC

2.1 Setting in a clinical core laboratory

The ECLs generally need to process multiple large and small studies. Key (primary or secondary) variables can be either continuous or categorical and these variables may or may not be the same across projects. While cardiac magnetic resonance imaging or cardiac catheterization may be treated as an external reference for some echocardiographic variables, these procedures are usually not performed simultaneously with echocardiography and thus a gold standard generally does not exist for echocardiographic variables.

In every project, a limited number of readers may be assigned to read the echocardiograms. Previously assigned readers may leave ECL and new readers may need to be added in the middle of the ongoing projects. A common approach in a core lab may allow the readers to start project reading immediately while the IIR study was carried out as a small substudy by taking a small random sample of the main study. This approach can be problematic because one may need to re-do substantial number of readings resulting in substantial cost and time if the IIR results of small substudy turn out to be unsatisfactory. Ideally, all readers should pass their IIR assessment under the same QC before starting project reading. A cyclical process may be needed before the readers pass their IIR. Specifically, if specific readers do not pass their IIR at their first try, the IIR process should identify specific measurements on specific echo images that resulted in unsatisfactory IIR. This provides an opportunity for the readers to review the measurements in these specific images in order to identify issues for the readers to be trained for improvement. A reader may need to take multiple rounds of IIR assessment in order to meet the standard for acceptable IIR.

A critical component of the IIR process for QA/QC is to determine what level of reproducibility will be acceptable. Without a gold standard, assessment of imaging data quality can only be based on the agreement/reliability of readings between the readers. Degree of sufficient agreement would depend on the statistical approach that is chosen to analyze the IIR data. Barnhart et al.[4] provided an overview of assessing agreement with continuous measurements that combines a large and diverse body of literature on various concepts and associated statistical methodologies that have evolved over several decades. However, a preferred agreement index was not recommended for IIR data because it depends on the subject matter and the goal of the IIR process. We elaborate below the desirable properties of an agreement index in an ECL setting that will be used to choose a preferred agreement index for reliability evaluation.

2.2 Desirable characteristics of an agreement index for the IIR process in a core lab setting

The ultimate goal of conducting an IIR process is to consistently yield reproducible high-quality data. Because we are dealing with the quality of individual measurements, our goal is to ensure that the variability between readers is small for each individual measurement. Different studies have different objectives that entail collection of measurements on different variables from different echocardiography images. Same variables may be measured in different studies, but the patient population may be different due to different objectives of the studies. Thus, it is important for the core lab to maintain the same quality of measurements across different projects with different populations. Both continuous and categorical variables are measured by echocardiography; therefore, it is useful to employ the same agreement index to both types of variables for simplicity. For quality improvement efforts, it is vital to easily and quickly identify the source of measurement disagreement that resulted in unsatisfactory agreement so that retraining can be targeted easily and quickly. All these reasons let us to specify the following desirable characteristics in an agreement index in order to choose a preferred index:

(A) The index is useful in assisting quality control and improvement process with the following properties:
  (1) It assesses the individual differences between multiple observers' measurements on the same subject for easy interpretation.
  (2) The index should have intuitive interpretation in setting satisfactory agreement and can quickly point to all problem observers/readings in case of unsatisfactory agreement.
(B) The index provides an overall consistent statement of data quality and reliability to assist in interpreting results. Aspects of this include the following properties:
  (1) The index should have consistent interpretation for the same variable across different populations in multiple projects.
  (2) The index should have consistent interpretation for both continuous and categorical data.
  (3) The index should have consistent interpretation for various number of observers.

With these pre-specified characteristics in mind, we review the agreement indices for continuous variables first. The corresponding index for categorical variables is indicated if it exists or similar index can be constructed. We elucidate why the CP approach for assessing agreement is the preferred index for setting the quality standard for acceptable IIR in the core lab setting.

2.3 Review of agreement indices for a preferred choice

The agreement indices reviewed in this section are suitable for assessing both intra-observer and inter-observer reliability where intra-observer reliability assesses if the same observer can reproduce his/her own results by taking replicated measurements, while inter-observer reliability assesses if different observers can reproduce from each other. For presentation purpose, we limit our review only on inter-observer agreement and the choice of agreement index for assessing inter-observer agreement would imply the same choice of index for assessing intra-observer agreement as well. We use the following notations throughout. Denote Y_{ij} as the measurement by jth observer for the ith subject, $i = 1, \ldots n, j = 1, \ldots, J$, where mean $E(Y_{ij}) = \mu_j$, variance $Var(Y_{ij}) = \sigma_j^2$ and correlation $Corr(Y_{ij}, Y_{ij'}) = \rho_{jj'}$ are used for continuous measurements. For simplicity, we use $J = 2$ for presentation and discuss only if the formulation does not extend to the case of $J > 2$.

Before getting into specific details of the reviewed agreement indices in terms of their characteristics, Table 3 provides the short summary of each index's pros and cons in a glance.

2.3.1 Pearson correlation coefficient

Pearson correlation coefficient (PCC), $\rho_{jj'}$, evaluates whether one measurement made by observer j, Y_{ij}, is linearly related to another measurement made by observer j', $Y_{ij'}$. The PCC would obtain its largest value of 1.0 if one observer's measurement is linearly related the other observer's measurement even though the measurements, Y_{ij} and $Y_{ij'}$ are far apart. Thus, the PCC does not assess the difference between measurements and it is not appropriate for use to assess IIR.

2.3.2 Mean-squared deviation

The mean-squared deviation (MSD) is defined as the expected squared difference between two readers: $MSD_{jj'} = E\left(Y_{ij} - Y_{ij'}\right)^2 = \left(\mu_j - \mu_{j'}\right)^2 + \sigma_j^2 + \sigma_{j'}^2 - 2\sigma_j\sigma_{j'}\rho_{jj'}$. This index has characteristics of A(1), B(1), and B(3). Specifically, it evaluates the differences of observers' measurements on averaged level of squared difference, rather than on individual level. It has consistent interpretations across different populations and difference number of observers. However, it does not have characteristics A(2) and B(2). First, it is not easy to interpret its magnitude in terms of how large of MSD should be treated as acceptable reliability. One possibility is to use $\leq \delta^2$ as acceptable reliability if δ is the acceptable difference between any two measurements. However, since it evaluates the average of individual squared differences, it is possible to have 50% or more of the individual squared differences to be larger than δ^2 even if the expected MSD is less than or equal to δ^2 ($MSD_{jj'} \leq \delta^2$), and thus it does not have characteristic A(2). Second, it does not have the corresponding quantity for categorical data and thus it does not have characteristic B(2).

2.3.3 ICC and Kappa

The ICC has been the most popular index used to report reliability in medical literature. There are many versions of ICC[4] based on different analysis of variance (ANOVA) model assumptions. We present here the original ICC (based on the original definition of reliability), ICC1, based on one-way ANOVA model. Comparisons of different ICCs can be found in Chen and Barnhart.[8] The ICC1 is a ratio of between-subject variability (σ_B^2) to the total variability (sum of the between-subject variability and the error variability (σ_e^2)), i.e., $ICC1 = \sigma_B^2/\left(\sigma_B^2 + \sigma_e^2\right)$, based on the one-way ANOVA model $Y_{ij} = \mu + \alpha_i + e_{ij}$, where $\mu + \alpha_i$ is the true measurement for subject i and e_{ij} is the error term by reader j on subject i with assumptions of $\alpha_i \sim N\left(\mu, \sigma_B^2\right)$, $e_{ij} \sim N\left(0, \sigma_e^2\right)$. The one-way ANOVA assumptions imply that $E\left(Y_{ij}\right) = \mu_j = \mu$, $Var\left(Y_{ij}\right) = \sigma_j^2 = \sigma_B^2 + \sigma_e^2$, and $corr(Y_{ij}, Y_{ij'}) = ICC1$ for all j and j'. The error variability is often referred as the within-subject variability. For categorical variable, the corresponding index is intraclass kappa which is equivalent to the ICC.[6] The simple kappa (or weighted kappa for ordinal data) corresponds to the other version of ICC where the marginal distribution of the each observer's measurement is not assumed to be the same.

Intuitively if the error variability is small relative to the between-subject variability, then the ICC value will be high (close to 1) and one would imply high reliability due to relatively smallness of the error variability. Landis and Koch[9] provided benchmarks for the range of ICC/Kappa values with 0–0.2 as poor, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.71–0.80 as substantial, and 0.81–1.0 as almost perfect. However, designations of these values for different degree of reliability are very subjective and an almost perfect ICC value may not even correspond to small and acceptable difference between measurements. Because the ICC/Kappa value is dimensionless, it is tempting to use these benchmarks to compare the ICC/Kappa values across different studies with difference

Table 3. Pros and cons of different agreement indices for assessing IIR in ECLs.

Index	Short description	Pros	Cons	Good for both continuous and categorical data?	Good for more than two observers?	Characteristics
Pearson correlation coefficient	Correlation between measurements	Provides the strength of linear relationship	Not appropriate for assessing IIR	No	No	None
Mean-squared deviation	Expected squared difference	Provides the magnitude of squared difference	Hard to set acceptable value	No. For continuous data only	Yes	A(1), B(1), B(3)
ICC or Kappa	Ratio of between-subject variance over total variance	Dimensionless with a value between −1 to 1 for interpretation	Depends on the between-subject variability. Need underlying ANOVA assumption for ICC	Yes where (weighted) Kappa is the ICC for categorical data.	Yes	B(2), B(3)
CCC	1 mimus the ratio of within-subject squared deviation and total deviation	Dimensionless with a value of −1 to 1, a standardized MSD	Depend on between-subject variability although does not require ANOVA assumptions	Yes	Yes	B(2), B(3)
wCV/error rate	Percent of within-subject SD over the overall mean	Dimensionless. Good for comparison of variability assuming the overall means have the same meaning	Dependent on the overall mean and only applicable for ratio variables	No. For continuous data only	Yes	None

(continued)

Table 3. Continued.

Index	Short description	Pros	Cons	Good for both continuous and categorical data?	Good for more than two observers?	Characteristics
CIA	Ratio of squared within-reader difference to the squared between reader difference	Dimensionless; Easy to understand by treating replications as acceptable variability	Need replicates where replication error assumed to be acceptable	Yes	Yes	B(2), B(3)
Limits of agreement ($LOA_{(1-\alpha)\%}$)	Symmetric limits around mean difference within which $(1-\alpha)\%$ of paired differences are expected to fall	Simple, intuitive and easy to understand with Bland and Altman plot	Centered around mean difference and dependent on normality assumption	No. For continuous data only	No	A(1), A(2), B(1)
Coverage probability	Proportion of subjects fall within the pre-set acceptable paired absolute difference	Simple, intuitive and easy to compute	Need to set acceptable difference a priori	Yes	Yes	A(1), A(2), B(1), B(2), B(3)
Total deviation index	Absolute paired difference with the desired CP	Simple, intuitive and easy to understand	Need to set acceptable difference a priori to determine acceptable TDI	No. Only for continuous data	Yes	A(1), A(2), B(1), B(3)

CP: coverage probability; CCC: Concordance correlation coefficient; wCV: within-subject coefficient of variation; TDI: total deviation index; ICC: intra-class correlation coefficient; CIA: coefficient of individual agreement.

populations and even across different variables. However, we need to keep in mind that ICC/Kappa is a relative index. An artificially high value of ICC/Kappa may be obtained if the between-subject variability is very large (e.g., large range of measurements) even though the error variability is not acceptable (see Introduction and example in Table 2). On the other hand, an artificially low value of ICC/Kappa can be obtained if the between-subject variability is small (e.g., homogeneous population) even though the error variability is acceptable. Thus, one should not compare ICC/Kappa values across studies with different populations even if the values are computed for the same variable. Furthermore, one should not compare the ICC/Kappa values across different variables because the between-subject variability has different meanings for different variables. Thus, ICC/Kappa does not have characteristics of A(1), A(2), and B(1).

The implication of high value of ICC as high reliability makes sense only if the corresponding error variability, σ_e^2, is acceptable, assuming one-way ANOVA model holds. To understand whether the error variability is acceptable, one can compute the 95% reproducibility coefficient (RDC) as $RDC_{95\%} = 1.96 \times \sqrt{2}\sigma_e = 2.77\sigma_e$ to represent the expected absolute difference between any two measurements, $Y_{ij} - Y_{ij\prime}$, for 95% of subjects. If $RDC_{95\%} \leq \delta$, where δ is the acceptable difference for an individual measurement, then we can say that the error variability is acceptable. If acceptable difference is not available for the measurement under investigation, then at least we can say that accepting the reported ICC value corresponds to accepting the absolute difference as large as $RDC_{95\%}$. Thus, when reporting ICC value, it is always good to report the estimated error variability so that we can have an intuitive understanding of the expected individual difference for 95% of the data. If the error variability is not reported, but the between-subject variability, σ_B, or the total variability, $Var(Y_{ij}) = \sigma_T^2 = \sigma_B^2 + \sigma_e^2$, is reported, it is still possible to estimate $RDC_{95\%}$. Below, we show how to estimate $RDC_{95\%}$ if at least one of the variability measures is reported along with ICC under the ANOVA model based on normality assumptions. For general $RDC_{(1-\alpha)\%}$ rather than $RDC_{95\%}$, use $\Phi^{-1}\left(\frac{2-\alpha}{2}\right)\sqrt{2}$ to replace 2.77 below:

(a) If the estimated error variability σ_ε^2 is reported along with ICC, we would estimate RDC_95% to be $RDC_{95\%} = 1.96\sqrt{2\sigma_e^2} = 2.77\sigma_e$ because $Var(Y_{ij} - Y_{ij\prime}) = 2\sigma_e^2$.
(b) If the estimated between-subject variability σ_B^2 is reported along with ICC, then by setting $Prob(|Y_{ij} - Y_{ijz'}| \leq RDC) = 0.95$ along with the assumed distribution of $Y_{ij} - Y_{ij'} \sim N\left(0, 2\sigma_B^2 \frac{1-ICC}{ICC}\right)$ based on the one-way ANOVA model, $RDC_{95\%}$ can be estimated as $RDC_{95\%} = 2.77\sigma_B\sqrt{\frac{1-ICC}{ICC}}$.
(c) If the estimated total variability $Var(Y_{ij}) = \sigma_T^2 = \sigma_B^2 + \sigma_e^2$ is reported along with ICC, then by setting $Prob(|Y_{ij} - Y_{ij'}| \leq RDC) = 0.95$ along with the assumed distribution of $Y_{ij} - Y_{ij'} \sim N(0, 2\sigma_T^2(1 - ICC))$, the $RDC_{95\%}$ can be estimated as $RDC_{95\%} = 2.77\sigma_T\sqrt{1 - ICC}$.

We illustrate this with the LVEF example in Table 2 assuming that the data are normally distributed. The SD reported for each of the two sonographers is an estimate of the total variability. Thus, the total variability can be approximated as $\hat{\sigma}_T = \sqrt{(14.37^2 + 15.77^2)/2} = 15.08$ and $\hat{\sigma}_T = \sqrt{(3.73^2 + 4.79^2)/2} = 4.29$ for projects 1 and 2, respectively. The corresponding 95% RDCs are then estimated as $RDC_{95\%} = 2.77 \times 15.08 \times \sqrt{1 - 0.97} = 7.23$ and $RDC_{95\%} = 2.77 \times 4.29 \times \sqrt{1 - 0.71} = 6.40$ for projects 1 and 2, respectively. Thus, the expected individual difference for project 1 is larger than the expected individual difference for project 2 even though the ICC is larger in project 1 than project 2. Similarly, the corresponding 85% RDC can be estimated as $RDC_{80\%} = \Phi^{-1}\left(\frac{2-0.2}{2}\right) \times \sqrt{2} \times 15.08 \times \sqrt{1 - 0.97} = 4.73$ and $RDC_{85\%} = \Phi^{-1}\left(\frac{2-0.2}{2}\right) \times \sqrt{2} \times 4.29 \times \sqrt{1 - 0.71} = 4.15$ for projects 1 and 2, respectively.

Note that the total deviation index, $TDI_{(1-\alpha)\%}$, presented later in this paper, reduces to $RDC_{(1-\alpha)\%}$ if the mean difference between paired measurements on the same subject is zero and this difference follows a normal distribution. Relationship between ICC and TDI or to CP under normality assumption is presented in the later section below.

2.3.4 Concordance correlation coefficient

Concordance correlation coefficient (CCC) is another popular scaled agreement index that was first introduced by Lin[10] for two observers. This is the current exploratory metrics proposed for assessing reader variability in Metrics Champion Consortium for Imaging (http://www.metricschampion.org/), an open, multidisciplinary, non-profit organization comprised biotechnology, pharmaceutical, and service provider organizations. The basic idea of the CCC is as follows. If observer j and observer j' do not agree well, then $E(Y_{ij} - Y_{ij'})^2$ is large. The maximum value of this quantity is the one obtained assuming that observer j and observer j' are independent. The relative scale of these two quantities provides the degree of disagreement. Using one minus this ratio gives the degree of agreement with higher value indicating higher agreement, i.e., the CCC for two observers is defined as

$$CCC = 1 - \frac{E\left(Y_{ij} - Y_{ij'}\right)^2}{E\left[\left(Y_{ij} - Y_{ij'}\right)^2 | Y_{ij}, Y_{ij'} \text{ are independent}\right]} = \frac{2\sigma_j\sigma_{j'}\rho_{jj'}}{(\mu_j - \mu_{j'})^2 + \sigma_j^2 + \sigma_{j'}^2}.$$

The CCC for multiple observers can be defined accordingly.[11] The main advantage of CCC is that it is assumption free while the ICC depends on the ANOVA assumptions. The CCC reduces to a version of ICC if the ANOVA assumptions are valid. In particular, the CCC reduces to the ICC presented above with one-way ANOVA model, if $\mu_j = \mu_{j'} = \mu$, $\sigma_j^2 = \sigma_{j'}^2 = \sigma_T^2 = \sigma_B^2 + \sigma_e^2$. As shown in Chen and Barnhart,[8] the CCC estimate is often smaller or identical to the estimate for some version of ICC, even though the inference is different. Barnhart et al.[4] showed that the CCC depends on the between-subject variability just like the ICC. Therefore, like the ICC, the CCC index does not have characteristics of A(1), A(2), and B(1).

2.3.5 Within-subject coefficient of variation and error rate

The within-subject coefficient of variation (wCV) is defined as the within-subject variability divided by the population mean. Under the ANOVA model (1), the wCV is defined as $wCV = \sigma_e/\mu$. This index is sometimes called error rate as it assesses the measurement error, σ_e, relative to the mean. By definition, the wCV depends on the population mean μ and its value would be small for project with large population mean and large for project with small population mean even if the measurement error, σ_e, remains the same. Thus, this index does not have consistent interpretation across different populations in multiple projects and it would not have characteristic B(1). Furthermore, there is no corresponding wCV for categorical data and thus it also does not have characteristic B(2).

2.3.6 Coefficient of individual agreement

Coefficient of individual agreement (CIA)[12,13] is defined in the situation where there are replicated measurements by the same observer and the repeatability (intra-observer variability) is assumed to be acceptable, or the replicated measurements are not available but the repeatability is known. Assume that there are replicated measurements Y_{ijk} for observer j whose replication error is acceptable. Then, the CIA is defined as $CIA = E\left(Y_{ijk} - Y_{ijk'}\right)^2 / E\left(Y_{ij} - Y_{ij'}\right)^2$, where larger CIA

implies better agreement between observer j and observer $j\prime$. The definition has been extended to the situation with more than two observers[12] and to the situation with categorical data.[13] Barnhart et al.[4] showed that the CIA is less dependent on the between-subject variability. However, the CIA depends on the replication error variability which may be different from different variables and different projects unless the acceptable replication error is pre-specified. Thus, this index likely would not have characteristics of A(1), A(2), and B(1).

2.3.7 Bland and Altman's limits of agreement

The 95% limits of agreement (LOAs) by Altman and Bland[14] are a popular tool for examining agreement between two measurements on the same subject. The 95% LOAs are centered on mean difference and have the interpretation that 95% differences would be within these limits with the assumption that the differences are normally distributed. By assuming the normality on the pairwise difference, $D_{jj\prime} = Y_{ij} - Y_{ij\prime}$, the 95% LOAs can be estimated as

$$\bar{D}_{jj\prime} \pm 1.96\hat{\sigma}_{D_{jj\prime}}$$

where $\bar{D}_{jj\prime}$ is the mean of $(Y_{ij} - Y_{ij\prime})$'s, and $\hat{\sigma}_{D_{jj\prime}}$ is the sample standard deviation of $Y_{ij} - Y_{ij\prime}$. For $(1-\alpha)\%$ LOAs rather than 95% LOAs, one can replace 1.96 above by $\Phi^{-1}\left(\frac{2-\alpha}{2}\right)$.

An intuitive and reasonable criterion to judge whether the agreement is satisfactory would be to require that at least 95% (or $(1-\alpha)\%$) of the differences to be within a pre-set acceptable difference for this variable. Use the LOA approach for implementing this criteria would require that the limits to be within $\pm\delta$, where δ is the absolute acceptable individual difference for 95% (or $(1-\alpha)\%$) of paired differences. There are two drawbacks of using the LOA approach to fulfill this criterion. First, the estimated limits dependence on normality assumption of the difference and the interpretation would not be true if the difference is not normally distributed. Second, because the LOAs are centered around mean difference, rather than zero, thus it is possible that 95% of differences are acceptable, i.e., between $-\delta$ and δ, but one of the 95% LOAs may be outside of $(-\delta, \delta)$ to indicate that the agreement is not acceptable. This point is illustrated here with two simulated data sets. To mimic the LVEF data, Figure 1 displays Bland and Altman plot of 1000 simulated paired differences with mean of 1.2 and standard deviation of 2. The data are uniformly spread with averages of two measurements to range from 30 to 58. The 95% LOAs are estimated to be (−2.62, 5.21). However, the 95% symmetric limits centered at 0 are estimated to be (−4.74, 4.74). For reference, the true LOAs and symmetric limits are (−2.62, 5.21) and (−4.53, 4.53), respectively, based on normal distribution. If 5 is the acceptable absolute difference, then 95% of differences are within (−5, 5). But the LOAs implies that 95% of differences are not within (−5, 5) because the limits are not centered at 0. Thus, LOA approach would imply that the observers' agreement is not acceptable while the symmetric limits would imply that the observers' agreement is acceptable. This phenomenon can be more pronounced if the differences are not normally distributed. Figure 2 displays 1000 simulated differences of two observers' measurements from mixture of normal and log-normal distributed distribution based on the addition of a normal and a log-normal random variables with (0, 1) and (−1, 1.5) as (mean, SD), respectively, i.e., with $D_{ijj\prime} \sim N(0,1) + \log N(-1, 1.5)$ where the averages of the two measurements are simulated from uniform distribution with range from 30 to 60. The 95% LOAs are estimated to be (−4.47, 6.65). However, the 95% symmetric limits centered at 0 are (−4.16, 4.16). Again if 5 is the acceptable absolute difference, then 95% of differences are within (−5, 5). But LOA approach would imply that the observers' agreement is not acceptable with 95% of differences are not in (−5, 5) due to its asymmetric limits.

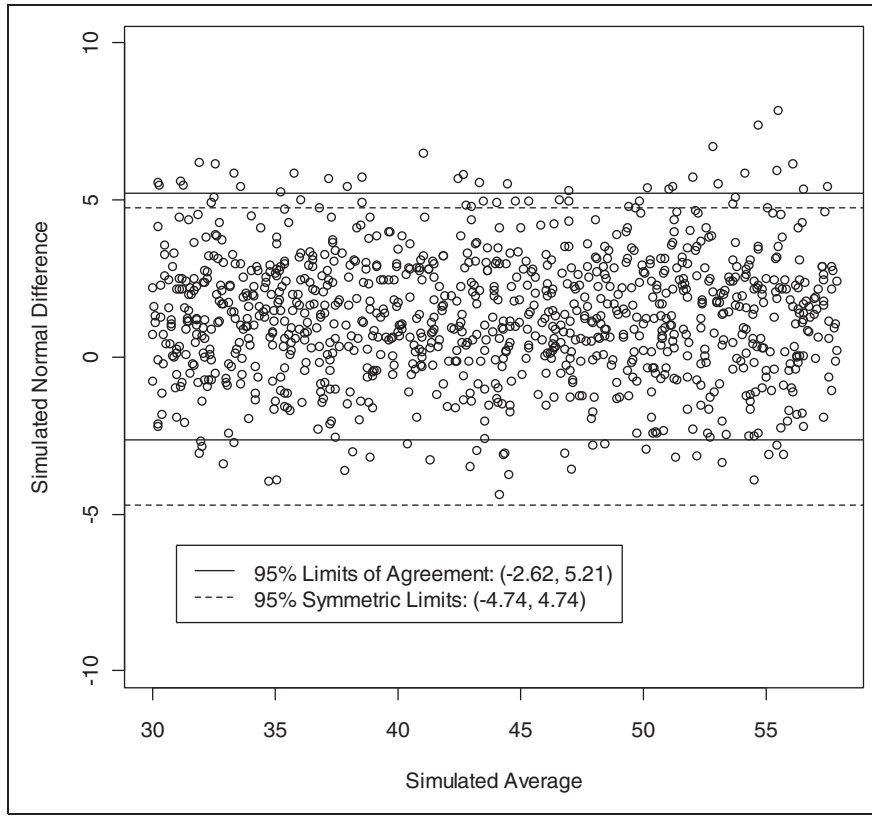

Figure 1. Bland and Altman plot from simulated 1000 normal differences and 1000 uniform averages.

The LOA approach appears to meet the characteristics of A(1), A(2), and B(1) although with the issue identified above. There is no corresponding index for categorical data or for data with more than two observers. Thus, it does not have characteristics of B(2) and B(3).

2.3.8 CP and TDI

CP and TDI are two unscaled agreement indices, with equivalent concepts, to measure the proportion of cases within a boundary for allowed differences.[15–17] The CP for assessing agreement is different from the well-known CP concept used for a confidence interval (CI) which is the proportion of the time that the interval contains the true value of interest. For the CP concept used for assessing agreement, we need to first set the predetermined boundary for the difference, e.g., an acceptable difference $\delta(>0)$. The CP is defined as the probability, π, that the absolute difference between the two measurements, $Y_{ij}, Y_{ij'}$, made on the same subject is less than δ, i.e., $\pi_\delta = \Pr\left(\left|Y_{ij} - Y_{ij'}\right| \leq \delta\right)$. Assuming that all $Y_{ij} - Y_{ij'}$'s are independent random samples from the same distribution, this means that π_δ proportion of the differences are expected to fall within δ, or equivalently, the chance is π_δ where the difference between two measurements make on the same subject would differ by up to δ. Theoretically, π_δ is basically the value of the cumulative distribution function (CDF) evaluated at δ for random variable of absolute difference of measurements between two observers. For example, let Y_j and $Y_{j'}$ be the random variables for the measurements by

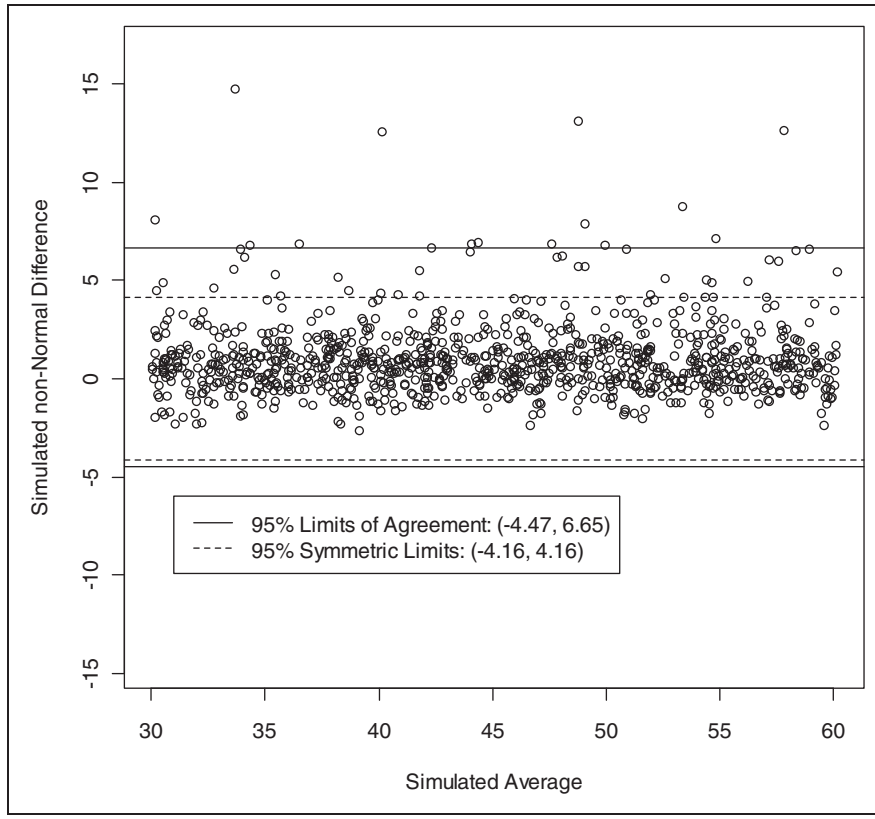

Figure 2. Bland and Altman plot of 1000 simulated non-normal differences and uniform averages.

observers j and $j\prime$. If $(Y_j, Y_{j\prime})\prime$ follows a bivariate normal distribution, then random variable of the absolute difference, $|Y_j - Y_{j\prime}|$, follows a folded normal distribution. A special case of this is when $E(Y_j) = E(Y_{j\prime})$ and $|Y_j - Y_{j\prime}|$ is distributed as a half-normal distribution. Let F(.) be the CDF for $|Y_j - Y_{j\prime}|$, then $\pi_\delta = F(\delta)$ is the population parameter and one can use the folded normal CDF for parametric estimate of CP. In practice, the measurements may not be normally distributed. The nonparametric estimate for CP can be computed by counting the proportion of absolute paired differences that fall within δ or by using SAS procedure GENMOD for binary indicator of difference within δ and empirical inference for correlated outcomes. For TDI, we need to first set the predetermined boundary for the proportion, $\pi = 1 - \alpha$, to represent the majority of the differences, e.g. $1 - \alpha = 0.80$, or 0.95. The TDI is defined as the difference, $TDI_{(1-\alpha)\%}$, that satisfy the equation $1 - \alpha = \Pr\left(\left|Y\_ij - Y\_ij\prime\right| < TDI_{(1-\alpha)\%}\right)$. Intuitively $(1 - \alpha)\%$ of the differences are within the symmetric limits of $(-TDI_{1-\alpha}, TDI_{1-\alpha})$. This is in contrast of $(1 - \alpha)\%$ of the differences are within possibly asymmetric limits from LOAs. If the mean difference is zero and the difference is normally distributed, then the symmetric limits of $\left(-TDI_{(1-\alpha)\%}, TDI_{(1-\alpha)\%}\right)$ corresponds to $(1 - \alpha)\%$ LOAs and $TDI_{(1-\alpha)\%} = RDC_{(1-\alpha)\%}$ as mentioned earlier.

Again an intuitive and reasonable criterion to judge whether the agreement is satisfactory would be to require that an overwhelming majority, e.g., 95% or $(1 - \alpha)\%$ of the differences to be within a pre-set acceptable difference,δ, for this variable, or equivalently it requires that the chance should at least 95% or $(1 - \alpha)\%$ where the difference of any two measurements made on the same subject is

less than or equal to δ. Thus, at least two numbers, the acceptable difference, δ, and the corresponding majority $(1-\alpha)\%$ (or equivalently, a margin of error of α), are needed to pre-specify a quality standard. Either CP or TDI can be used to determine if the pre-specified quality standard is met or not. If CP is used for assessment, then the pre-specified quality standard is met if the CP corresponding to the given acceptable difference is greater than or equal to $(1-\alpha)\%$. Equivalently, if TDI is used for assessment, then the quality standard is met if the TDI corresponding to coverable probability of $(1-\alpha)\%$ is less than or equal to δ. The nonparametric estimate for TDI can be obtained by estimating $(1-\alpha)$th quantile of paired difference and this can be obtained through quantile regression with intercept term in the model (with bootstrap approach for CI if there are correlated differences). Because CP and TDI are equivalent concepts and CP is easier than TDI to estimate, the CP index would be a better index to be used for QA.

The CP index satisfied all characteristics in section 2.2. The CP has characteristic A(1) because obviously by definition it assesses the individual differences between multiple measurements on the same subject. It has characteristic A(2) as well because it has intuitive interpretation in the sense by stating the chance of paired difference falling within pre-specified acceptable difference. In case there is unsatisfactory agreement, the problem observers/readings can be easily identified by picking those paired differences exceeding the pre-set acceptable difference. These readings can then be reviewed to identity reading or training issues. After training, the observers can repeat the process again to see whether the readings meet the pre-specified quality standard. The CP also has characteristic B(1) because the same quality standard can be applied to the same variable across different populations regardless what population a given project is studying and thus the same quality is ensured across different projects. The CP has characteristic B(2) because quality standard is specified the same way for both continuous and categorical with the same intuitive interpretation. Finally, the CP has characteristic B(3) because the index's interpretation does not change due to various numbers of observers.

Because the concepts of CP and TDI have generally not used in the reporting of the clinical research, it would be useful to know the corresponding CP or TDI if other agreement indices were reported. This would provide us understanding how large of difference one is accepting with reported agreement index. Assuming that the difference of any two measurements for the same subject follows a normal distribution, we show below when we can estimate CP and TDI under several scenarios:

(1) If ICC is reported and population standard deviation, σ_T, is available, then we would need to assume that the mean difference between paired measurements is zero and $\sigma_T^2 = \sigma_B^2 + \sigma_e^2$. This implies that $\sigma_e^2 = \sigma_T^2(1-ICC)$, and we have $CP_\delta = \Phi(\frac{\delta}{\sqrt{2\sigma_e^2}}) - \Phi(\frac{-\delta}{\sqrt{2\sigma_e^2}})$, and obtain $TDI_{(1-\alpha)\%}$ by solving equation $1-\alpha = \Phi(\frac{TDI}{\sqrt{2\sigma_e^2}}) - \Phi(\frac{-TDI}{\sqrt{2\sigma_e^2}})$.
(2) If LOAs are reported along with the mean difference, d, and standard deviation of the differences, σ_D, then we have $CP_\delta = \Phi(\frac{\delta-d}{\sigma_D}) - \Phi(\frac{-\delta-d}{\sigma_D})$ and obtain $TDI_{(1-\alpha)\%}$ by solving equation $1-\alpha = \Phi(\frac{TDI-d}{\sigma_D}) - \Phi(\frac{-TDI-d}{\sigma_D})$.
(3) If CV was reported along with the population mean μ, then $\sigma_e^2 = CV^2\mu^2$ and CP and TDI can be estimated by using the formulas in item (1).

3 Examples

We use two echocardiography examples here to illustrate the use of various indices reviewed in section 2 and highlight the benefit of using CP as the preferred agreement index for reproducibility assessment as part of the QC.

3.1 LVEF as continuous measurement

The first example is from the data introduced in section 1 where both the same two sonographers made biplane LVEF measurements on 10 different subjects in each of the two projects. Figure 3 shows the plot of reader 1's LVEF versus reader 2's LVEF in both projects. The plot of average versus difference is shown in Figure 4. Both these figures show that project 1 has wider range of LVEF values where there are larger differences in project 1 than project 2. Table 4 provides the results from various agreement indices discussed in section 2. Because the systematic difference between readers 1 and 2 is small, the PCC, ICC, and CCC all have similar estimates and 95% CIs. These values in project 1 are higher than project 2 that seem to indicate the readers are more reliable in project 1 than readers in project 2. Due to these indices' dependence on between-subject variability, this only means that it is easier to differentiate/discriminate subjects in project 1 than project 2, not because readers are more agreeable in project 1 than project 2 as elaborated in section 3. In contrast, the $LOA_{80\%}$, $TDI_{80\%}$, and CP_5 all indicated that readers differ more in project 1 than project 2 which reflect the intuitive display of the data. The $LOA_{80\%}$ is asymmetric due to reader 2 has a slight tendency of giving higher value than reader 1. Since TDI and CP are equivalent concepts and it is easier to compute CP, thus CP is preferred here for assessment of reproducibility and for development of an action plan. For example, if the acceptable difference is 5 or less and we try to achieve a QC of having 5 or less difference with

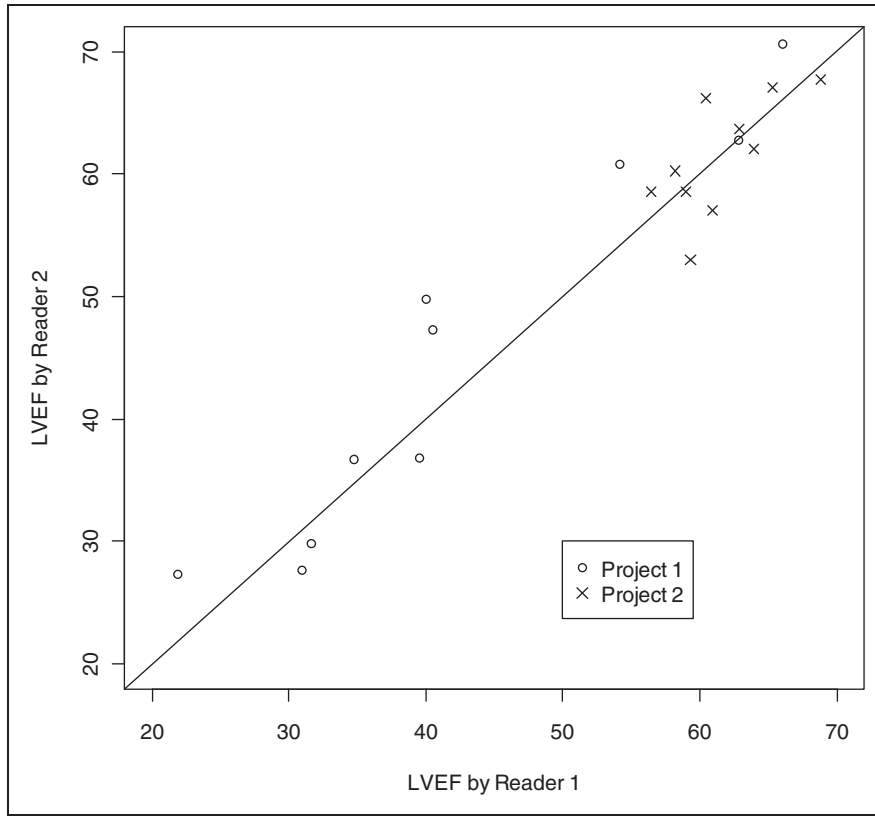

Figure 3. LVEF measurements comparing reader 1 vs. reader 2 in two projects.

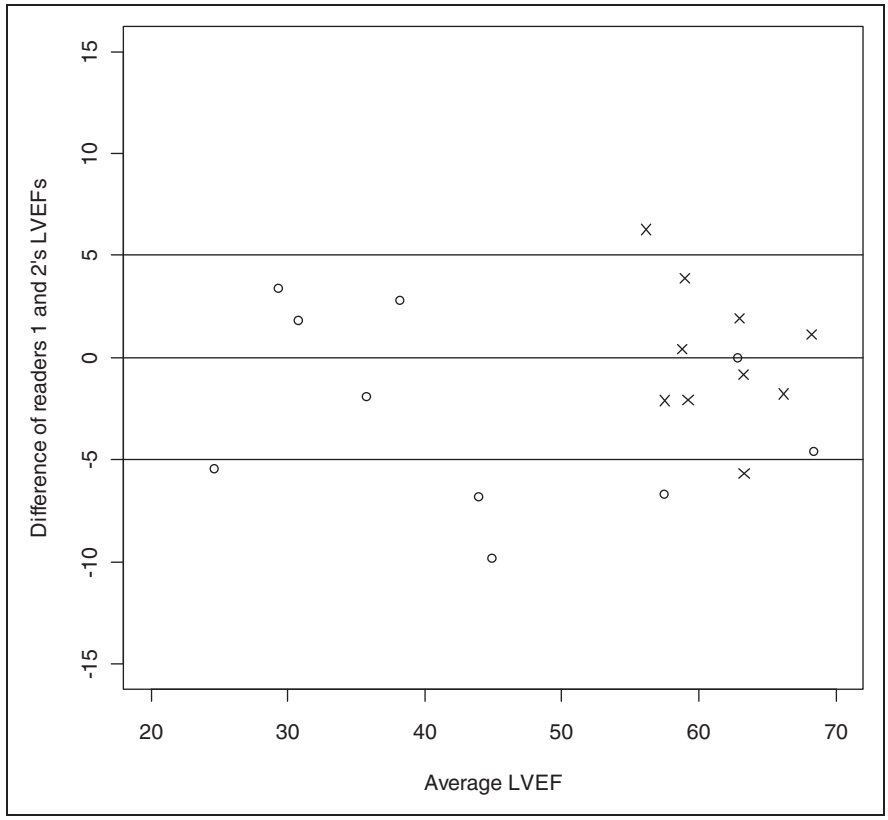

Figure 4. Average LVEF vs. difference of LVEFs in two projects with reference lines of ±5.

Table 4. Agreement indices between readers 1 and 2 in projects 1 and 2.

	Project 1 ($N = 10$)	Project 2 ($N = 10$)
PCC (95% CI)	0.95 (0.81, 0.99)	0.71 (0.10, 0.92)
MSD (95% CI)	29.30 (11.35, 75.62)	11.58 (4.35, 30.86)
ICC (95% CI)	0.94 (0.80, 0.99)	0.71 (0.22, 0.92)
CCC (95% CI)	0.94 (0.79, 0.98)	0.69 (0.19, 0.90)
wCV	8.33%	3.71%
$LOA_{80\%}$(Reader 2–Reader 1)	(−3.16, 8.61)	(−4.47, 4.24)
$TDI_{80\%}$(parametric) (95% CI)	6.9 (4.3, 11.1)	4.4 (2.7, 7.1)
$TDI_{80\%}$(non-parametric) (95% CI)	6.7 (4.1, 9.3)	3.9 (1.8, 5.9)
CP_5 (parametric) (95% CI)	0.60 (0.34, 0.81)	0.80 (0.48, 0.95)
CP_5 (non-parametric) (95% CI)	0.60 (0.30, 0.90)	0.80 (0.46, 0.95)

CP: coverage probability; CCC: Concordance correlation coefficient; wCV: within-subject coefficient of variation; TDI: total deviation index; ICC: intra-class correlation coefficient; LOA: limits of agreement; MSD: mean-squared deviation; PCC: Pearson correlation coefficient.

80% probability (or equivalently 80% of paired differences are within 5 units), then based on the point estimates of CP_5 the readers in project 2 can proceed with readings while readers in project 1 would need to review the four echoes (see Figure 2) where they differed by more than 5 units for evaluation, re-training, and re-do the reproducibility assessment. Both MSD and wCV also indicate that readers in project 2 agree better than readers in project 1. However, the magnitudes of these values are not as intuitive for interpretation.

To illustrate the concepts in the extreme scenario for a homogeneous population, we created an artificial data for project 3 with two readers, where reader 1 has the same LVEF values of X in N images and reader 2 has the same LVEF values of X in N−1 images except value of Y($\neq$X) in one image. It can be shown that ICC = 0 in this situation while CP $\geq$ (N−1)/N. This means that one can have ICC = 0 indicating no reliability while there is almost 100% agreement between the two readers. This extreme situation shows that CP is a much better index for summarizing the agreement between measurements where ICC can fail to show agreement in setting of homogenous population.

3.2 Mitral regurgitation as categorical variable

Table 5 shows two hypothetical data on the measurements of mitral regurgitation by two readers on 10 echoes with four ordinal categories: trace, mild, moderate, and severe. In the first data set, the two readers had perfect agreement on five (50%) of 10 mitral regurgitation measurements and eight (80%) of them are within one category apart. In the second data set, the two readers had perfect agreement on seven (70%) of 10 mitral regurgitation measurements and all 10 (100%) of them are within one category apart. Thus intuitively, the two readers agree better in the second data set than in the first data set. This is also indicated by the CP of 0.8 (95% CI: 0.046, 0.95) and 1.0, respectively, for the first and second data sets based with pre-specified acceptable difference of one category apart. However, the estimated weighted kappa values indicated opposite with higher estimated kappa value in the first data set than in the second data set (the estimated kappa values are 0.43 (95% CI: 0.07, 0.78) and 0.32 (95% CI: −0.16, 0.80), for the first and second data

Table 5. Mitral regurgitation by readers 1 and 2.

	Reader 2			
Reader 1	Trace	Mild	Moderate	Severe
Scenario 1[a]				
Trace	1	1	0	0
Mild	0	2	1	2
Moderate	0	0	1	1
Severe	0	0	0	1
Scenario 2[b]				
Trace	0	0	0	0
Mild	0	0	1	0
Moderate	0	0	6	2
Severe	0	0	0	1

[a]Weighted Kappa (95% CI): 0.43 (0.07, 0.78); CP_1 (95% CI): 0.80 (0.46, 0.95).
[b]Weighted Kappa (95% CI): 0.32 (−0.16, 0.80); CP_1 (95% CI): 1.0 (NA).

sets, respectively). This is because mitral regurgitation measurements in the first data set are more spread out over the four ordinal categories than the ones in the second data set where the latter had very few images in the trace or mild categories. Thus, the weighted kappa exhibits the same issue as in ICC due to its dependency on between-subject variability, while CP approach conveys an intuitive assessment on the extent of agreement. With the CP approach, one can also quickly identity the readings where the two readers differ more than one category apart for review and re-training.

4 Discussion

We have reviewed the existing agreement indices that can be potentially used for assessing and improving measurement reproducibility in a core-lab setting. We pre-specified desirable characteristics in an agreement index in the process of assessing and improving measurement reproducibility and found that the CP approach for assessing agreement is the preferred choice among all indices reviewed. We found that the CP approach is easy to use, intuitive to understand, and consistent in interpretation on multiple fronts. Furthermore, in the event of poor reproducibility, it provides actionable results that guide readers to improve quality in the next round which is an important part of improving reproducibility. See Daubert et al.[18] for an example of implementing this cyclic process for improvement of reproducibility.

Although our review and illustration are on inter-reader agreement, the same argument can also be made about intra-reader agreement as well. It is possible that a large number of readers may be used in a large study where not all of readers are on board during the initiation of the project or existing readers may leave the lab and new readers may be added to the project. In this situation, it will be useful to establish a cohort of readers who met the reproducibility criteria so that the readings of any new readers will be compared to this established cohort of readers. This is a reference cohort approach based on the CP criteria that is discussed in Daubert et al.[18]

With the CP approach, one would need to pre-specify an acceptable difference between any two measurements on the same subject. This pre-specification process is similar to the process of choosing a clinically significant difference when designing a randomized clinical trial. The choice of acceptable difference must be based on clinical knowledge. First, we want to choose an acceptable difference that is small enough so that we can be sufficiently confident that any change that is larger than this difference is not due to measurement error. On the other hand, we do not want to set an acceptable difference that is too small that is humanly impossible to achieve. If some magnitude of difference in the measurement is strongly related to clinical outcome, this would be a good choice for an acceptable difference. Typically, the acceptable difference should be smaller than a clinical significant difference because there is no point on detecting any difference that is due to measurement error.

With the CP approach, a satisfactory agreement is based on sufficiently high satisfactory CP, e.g., 0.95 or 0.8, given a pre-specified acceptable difference. This approach can be extended to situations with multiple pre-specified differences with corresponding coverage probabilities, e.g., 0.8 for difference of 10 and 1.0 for difference of 20 in LVEF. The more differences one specifies, the better control we have on the quality because the satisfactory agreement can be claimed only if the coverage probabilities are met at all pre-specified differences.

In situation, where no information is available on an acceptable difference, and we just want to know what kind of agreement is achievable based on observed differences, one can examine the CP curve, i.e., the cumulative distribution of the absolute difference between any two measurements on the same subject. This kind of curve provides a full spectrum of measurement error.[19]

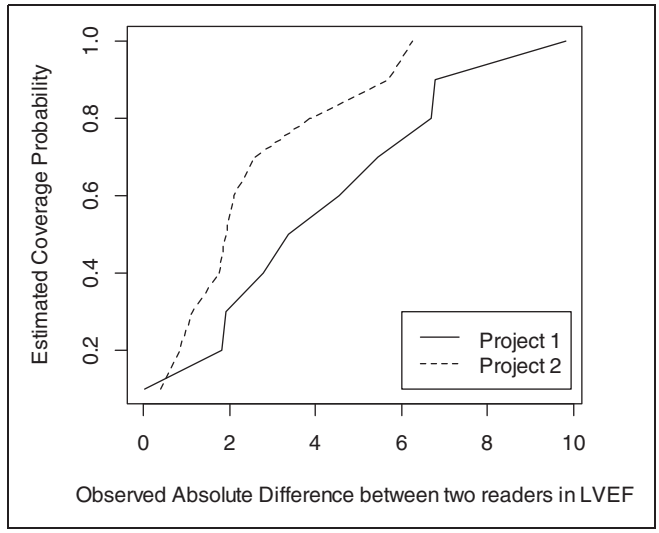

Figure 5. Coverage probability curves for LVEF example.

Figure 5 illustrates such coverage curves that indicate readers in project 2 agree better than readers in project 1. If we want to use a summary measure based on the CP curve to evaluate the extent of agreement, one can consider the new approach by looking at the area under the CP curve.[19] Barnhart proposed to use a relative area under the CP curve (RAUCPC) to assess agreement where multiple acceptable differences are specified or to compare different CP curves. This approach evaluates agreement based on all possible differences up to a pre-specified maximum difference, where we expect 100% CP. Further research is needed in this area.

This paper examines QA and QC in terms of reader reproducibility/agreement where true value or gold standard is not available. One limitation of such QA and QC process is that the readers may agree very well but may agree to the wrong value. Without true value or gold standard, accuracy cannot be truly evaluated.

Funding

This work was supported in part (except for Robert Bigelow, Michael Pencina) by the American Society of Echocardiography (ASE) Education and Research Foundation (grant number 12-G-10-ASE). The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Society of Echocardiography or the ASE Education and Research Foundation.

Conflict of interest

None declared.

References

1. Chung ES, Leon AR, Tavazzi L, et al. Results of the predictors of response to CRT (PROSPECT) trial. *Circulation* 2008; **117**: 2608–2616.

2. Gottdiener JS, Bednarz J, Devereux R, et al. American society of echocardiography recommendations for use of echocardiography in

clinical trials. *J Am Soc Echocardiogr* 2004; **17**: 1086–1119.
3. Douglas PS, DeCara JM, Devereux RB, et al. Echocardiographic imaging in clinical trials: American society of echocardiography standards for echocardiography core laboratories: endorsed by the American College of Cardiology Foundation. *J Am Soc Echocardiogr* 2009; **22**: 755–765.
4. Barnhart HX, Haber MJ and Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat* 2007; **17**: 529–569.
5. Atkinson G and Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics* 1997; **53**: 775–777.
6. Shoukri MM. *Measures of interobserver agreement and reliability*, 2nd ed. New York, NY: CRC Press, Taylor & Francis Group, 2011.
7. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; **64**: 96–106.
8. Chen CC and Barnhart HX. Comparison of ICC and CCC for assessing agreement for data without and with replications. *Comput Stat Data Anal* 2008; **53**: 554–564.
9. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
10. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**: 225–268.
11. Barnhart HX, Haber M and Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2002; **58**: 1020–1027.
12. Barnhart HX, Haber M and Kosinski AS. Assessing individual agreement. *J Biopharm Stat* 2007; **17**: 697–719.
13. Haber M, Gao J and Barnhart HX. Assessing observer agreement in studies involving replicated binary observations. *J Biopharm Stat* 2007; **17**: 757–766.
14. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160.
15. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Stat Med* 2000; **19**: 255–270.
16. Lin LI, Hedayat AS, Sinha B, et al. Statistical methods in assessing agreement: models, issues and tools. *J Am Stat Assoc* 2002; **97**: 257–270.
17. Lin L, Hedayat AS and Wu W. *Statistical tools for measuring agreement*. New York, NY: Springer, 2012.
18. Daubert MA, Yow E, Barnhart HX, et al. A practical approach to continuous quality improvement implementation: improving reproducibility in the clinical echocardiography laboratory. *Circulation* 2012; **126**: A14418.
19. Barnhart HX. Assessing agreement with relative area under the coverage probability curve. In: *Abstract and oral presentation at the 2013 Joint Statistical Meetings*, Montreal, Canada, 3–8 August 2013.