Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Novel Graphical Evaluation of Agreement

Jongphil Kim (jongphil.kim@moffitt.org)
H Lee Moffitt Cancer Center and Research Institute Department of Molecular Oncology
https://orcid.org/0000-0002-7430-2226
Ji-Hyun Lee
University of Florida

Research article

Keywords: Agreement, Bland-Altman plot, Concordance Correlation Coefficient, Graphical Evaluation, Limit of Agreement, Reference Band

Posted Date: December 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-122165/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on February 20th, 2022. See the published version at https://doi.org/10.1186/s12874-022-01532-w.

A Novel Graphical Evaluation of Agreement

Jongphil Kim[1,2*] and Ji-Hyun Lee[3,4]

[1]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center & Research Institute
[2]Department of Oncologic Sciences, University of South Florida
[3]Department of Biostatistics, University of Florida
[4]Division of Quantitative Sciences, University of Florida Health Cancer Center

*** Corresponding author**:
Jongphil Kim, PhD
Moffitt Cancer Center & Research Institute
12902 Magnolia Drive, MRC-Bio2, Tampa, FL 33612
Phone: 813-745-6908; Fax: 813-745-6107
Email: Jongphil.Kim@moffitt.org

Running Title: A Novel Graphical Evaluation of Agreement

Text word count: 2861
Abstract word count: 226
Figures: 4
Tables: 2
Supplemental Figures: 1
Supplemental Tables: 1
References: 15

Abstract

Background: The Bland-Altman plot with limit of agreement has been widely used as a visual tool for assessing test-retest reliability or reproducibility between two measurements. We have observed, however, that in certain circumstances the limit of agreement approach may mislead practitioners. Particularly, if the acceptable difference is not available and two readers are highly concordant but the common variance of the data is large, the broad width of the limit of agreement plot may incorrectly indicate a lack of agreement.

Methods: This paper proposes a novel, scaled index-based guidance for graphical evaluation of reproducibility or reliability. We create a reference band from two measurements, which is based on the concordance correlation coefficient.

Results: Simulation studies have been carried out to demonstrate the benefits of our method over the limit of agreement. We also consider the application to the real examples, including the peak expiratory flow rate data in Bland and Altman's paper and the test-retest reproducibility data of Radiomics study.

Conclusions: In absence of acceptable difference, we found that the limit of agreement seems to derive subjective inference and may not be consistent with concordance correlation coefficient. Our simulation study results and real data application show that the proposed method can provide practitioners with a novel graphical evaluation method which is consistent with results from concordance correlation coefficient approach than the limit of agreement approach.

Keywords: Agreement, Bland-Altman plot, Concordance Correlation Coefficient, Graphical Evaluation, Limit of Agreement, Reference Band

Background

In the process of development of new predictors or features in clinical studies, it is essential to assess how reliable or reproducible they are. The reliability or reproducibility of the features are evaluated by either unscaled summary indices based on absolute difference of measurements such as limit of agreement (LoA) (1-3), coverage probability (CP) and total deviation index (TDI) (4, 5) or scaled summary indices such as the concordance correlation coefficient (CCC) by Lin (6) or intraclass correlation coefficient (ICC). If the difference is interpretable and acceptable difference is available between measurements (e.g., blood pressure, peak expiratory flow rate in Bland and Altman(1), etc.), unscaled indices should be selected for assessing reliability or reproducibility. However, if the difference may not be interpretable and acceptable difference is not available, ICC or CCC are commonly selected as scaled indices for two or more continuous measurements. For example, Balagurunathan et al. (7) developed 219 quantitative 3D imaging features derived from computed tomographic (CT) images, which may be useful as prognostic biomarkers in non-small cell lung cancer studies. These imaging features include texture features such as pixel histogram, run length, co-occurrence, 3D-Laws, and the difference of these features are hard to be clinically interpreted and thus the acceptable difference for such feature cannot be predetermined. The CCC was selected to evaluate the reproducibility or reliability of imaging features. More details regarding definition of repeatability, reproducibility, validity, reliability and agreement indices for continuous measurements are available in Barnhart et al. (8). We use agreement, reliability and reproducibility interchangeably in this paper since we seek to propose a novel visual tool for assessing agreement between two measurements. The pros and cons of different agreement indices are well compared in Barnhart et al. (9).

The Bland-Altman (B-A) plot with the LoA has been widely used as a visual tool for assessing agreement due to simplicity and intuitive appeal and was reported as one of the top 100 most cited papers of all time (Van Noorden et al. (10)). Suppose that n pairs of samples (X_{1j}, X_{2j}), $j = 1, \dots, n$ are collected independently from a bivariate normal distribution $X = (X_1, X_2)^T$ with mean $\mu = (\mu_1, \mu_2)^T$ and variance-covariance matrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, $|\rho| < 1$. The CCC, ρ_c, is expressed as the product of two terms;

$$\rho_c = \rho C_b, \; 0 < C_b = \frac{2}{\frac{\sigma_1}{\sigma_2} + \frac{\sigma_2}{\sigma_1} + \frac{(\mu_1 - \mu_2)^2}{\sigma_1\sigma_2}} \leq 1,$$

where ρ is the correlation coefficient and the term C_b measures how far the best-fit line deviates from the perfect concordance line $X_1 = X_2$. Bland and Altman (1) proposed a residual type plot of the observed pairs of data for graphical evaluation of agreement. The LoA is defined as

$$\bar{d} \pm t_{n-1,0.025} S_d,$$

where $d_j = X_{2j} - X_{1j}$, $\bar{d} = \frac{1}{n}\sum d_j$, $S_d^2 = \frac{1}{n-1}\sum(d_j - \bar{d})^2$, $t_{n-1,0.025}$ is 100×(1-0.025) percentile of the t-distribution with $n-1$ degrees of freedom. The LoA contains nearly 95% of the data, and the inference is made by comparing the width of the LoA with the predetermined acceptable difference. The approximate and exact 95% confidence intervals for the LoA were investigated by Bland and Altman (2) and Carkeet (3), respectively. Assuming $C_b = 1$ (i.e., $\mu_1 = \mu_2$ and $\sigma = \sigma_1 = \sigma_2$), S_d^2 is an estimator of $2\sigma^2(1-\rho)$ since the variance of d_j is $2\sigma^2 - 2\rho\sigma^2$, and the LoA is then directly proportional to the between-subject variability (i.e., common standard deviation σ) and the square root of $1-\rho$. As in biomarker studies including Balagurunathan et al.'s Radiomics study, the reproducibility of feature was evaluated by CCC but the B-A plot with LoA was presented as graphical illustration of reproducibility. However, the LoA may not be consistent with the CCC values since LoA is an unscaled index while CCC is a scaled index. Thus, a novel, scaled-based graphical evaluation of agreement is needed to address this inconsistency.

In this paper, we seek to present a CCC-based visual tool for assessing agreement in cases where no acceptable difference is available and the scaled index are used for evaluating the reliability or reproducibility. We believe that the proposed method provides practitioners with not only guidelines for a descriptive graphical evaluation of agreement but also useful information such as recognition of patterns and identification of outliers in the data. Methods section of this paper shows how a reference band (RB) as a descriptive visual tool is derived from the CCC. Results section presents simulation results with comparisons to the LoA. A peak expiratory flow rate study data in Bland and Altman's paper and the Radiomics features extracted from 3D CT images in Balagurunathan et al. are considered as examples in Results section. The paper concludes with final comments in Conclusions section.

Methods

Unlike total deviation index (TDI) and coverage probability (CP) by Lin (4), Lin et al.(5), and Escaramis et al. (11), we assume that C_b = 1 (i.e., $\mu_1 = \mu_2$ and $\sigma = \sigma_1 = \sigma_2$). Then, $X_2 - X_1$ is normally distributed with mean 0 and variance $2\sigma^2(1-\rho)$, and the pooled estimator S^2 for σ^2,

$$S^2 = \frac{1}{2(n-1)}\sum_{i=1}^{2}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2, \bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij},$$

is distributed as $\sigma^2\chi_\nu^2/\nu$ with degrees of freedom $\nu = 2(n-1)$. For a given the correlation coefficient ρ, the variable $t = \frac{X_2 - X_1}{S\sqrt{2(1-\rho)}}$ is distributed as a central t-distribution with degrees of freedom ν. Thus, the probability over a band in (X_1, X_2) plane defined as

$$\Re(X) = (X_1, X_2), -\infty < X_1 < \infty, X_1 - \omega_r \leq X_2 \leq X_1 + \omega_r,\ \omega_r = t_{\nu,\alpha/2}\, S\sqrt{2(1-\rho)}$$

is exactly 1 – α under the assumption that C_b = 1.

For a graphical tool for assessing agreement, we introduce 100(1 - α) % "reference band (RB)" $\Re(X)$ with width ω_r for a given confidence level 1 - α and degrees of freedom ν. The two lines, $X_2 - X_1 = \pm\,\omega_r$, in $\left(\frac{X_1+X_2}{2}, X_2 - X_1\right)$ plane, are the boundary lines of the RB, as illustrated in Figure 1. The benefit of the difference against the average plot over X_1 against X_2 plot with 45° line is that it allows us to better investigate any possible relationship between the discrepancies and the average values (Bland and Altman (2)). If the absolute value of the difference $|X_2 - X_1|$ exceeds the width ω_r, those data can then be viewed as outliers from the RB.

Practitioners may choose different values of the CCC for a lower bound of excellent concordance, depending on their practical interpretation of the CCC. In this paper, we employ the lower bound of the CCC of 0.75 for excellent concordance, and this threshold has been well accepted in Nickerson (12) and Rosner (13). Assuming C_b = 1, the width of the RB is

$$\omega_r = \frac{1}{\sqrt{2}}\, t_{\nu,\alpha/2}\, S.$$

It is clear that excellent concordance would not be achieved if ρ is lower than 0.75 since $0 < C_b \leq 1$. Thus, nearly 95% of data should be located within the RB if the CCC is at least 0.75 and $C_b = 1$. Note that the random samples from a bivariate normal variable are distributed to the line

$$X_2 = \frac{\sigma_2}{\sigma_1}\,(X_1 - \mu_1) + \mu_2$$

in (X_1, X_2) plane, and that the slope of the best-fit line would be negative for $\sigma_2 < \sigma_1$, positive for $\sigma_2 > \sigma_1$, and 0 for $\sigma_2 = \sigma_1$ in $\left(\frac{X_1+X_2}{2}, X_2 - X_1\right)$ plane. Thus, the vertical shift of the mean difference from 0 and the slope of the best-fit line indicate the degree of heterogeneity of the two means and variances. We will investigate this in Section 3.

Results

Simulation Studies

This section considers four different scenarios to illustrate the performance of our approach; the results are compared with those obtained by the LoA approach by using 50 simulated data. The RB in $\left(\frac{X_1+X_2}{2},\ X_2 - X_1\right)$ plane is constructed by using $\alpha = 0.05$ and the CCC of 0.75 as the lower limit of excellent concordance. To simulate data, the normally distributed bivariate random numbers with ρ were

generated by the method by Kim (14). Scenario I evaluates the number of outliers detected by proposed method when $\rho_c = 0.75$ and $C_b = 1$. In scenario II and III, the proposed method is compared with the LoA of B-A plot when data are highly concordant ($\rho = 0.95$ and $C_b = 1$) and the common variance is relatively small ($\sigma_1 = \sigma_2 = 1$) or large ($\sigma_1 = \sigma_2 = 2$). The effect of heterogeneity of both two variances and means is investigated in Scenario IV. Table 1 summarizes the parameters used for the scenarios, the estimates of $\rho_c, \rho, C_b, \omega_r, S, S_d$, the number of outliers of the proposed method, and the width of the LoA of B-A plot. The graphical comparisons are provided in Figure 2.

Scenario I ($\mu_1 = \mu_2 = 1, \sigma_1 = \sigma_2 = 1; C_b = 1, \rho = 0.75; \rho_c = 0.75$) The data are randomly distributed to the line $X_2 - X_1 = 0$ in $\left(\frac{X_1+X_2}{2}, X_2 - X_1\right)$ plane (top and left panel of Figure 2). No pattern is detected and nearly 4% of data deviates from the RB, which implies that data is close to CCC of 0.75. Note that the width of the RB (ω_r = 1.353) is close to that of the LoA (width = 1.387) since S_d = 0.69.

Scenario II ($\mu_1 = \mu_2 = 1, \sigma_1 = \sigma_2 = 1; C_b = 1, \rho = 0.95; \rho_c = 0.95$) As in scenario I, no pattern is detected, which indicates that the bias correction factor, C_b, would be close 1. Compared with scenario I, no data is deviated from the RB, while approximately 95% of data is located within the LoA as depicted in the top right panel of Figure 2. Based on the proposed approach, it is apparent that the agreement of the data is considerably higher than 0.75 since all data are clustered near 0 within the RB and the slope of best-fit line seems to be near 0.

Scenario III ($\mu_1 = \mu_2 = 1, \sigma_1 = \sigma_2 = 2; C_b = 1, \rho = 0.95; \rho_c = 0.95$) Compared with scenario II, the only difference is that both σ_1 and σ_2 are increased to 2, and the RB and the LoA are almost two-folds of scenario II (lower left panel of Figure 2 and Table 1). It appears less concordant than scenario II based on the width of LoA, despite of the fact that, the CCC of scenario III is the same as scenario II. Indeed, the degree of concordance of scenario III appears similar to that of scenario II. The proposed RB method correctly reflects its concordant level with no deviates of the data point from the RB.

Scenario IV ($\mu_1 = 1, \mu_2 = 2, \sigma_1 = 1, \sigma_2 = 2; C_b = 2/3, \rho = 0.9; \rho_c = 0.6$) In $\left(\frac{X_1+X_2}{2}, X_2 - X_1\right)$ plane, the data are vertically shifted (lower right panel), and the slope of best-fit line is positive, showing $\sigma_1 < \sigma_2$. Thus, it is anticipated that the bias correction factor, C_b, is much smaller than 1. Nearly 16% of data deviates from the RB, which implies that the value of CCC seems much lower than 0.75. However, the centerline of the LoA moves up by the mean of the differences, $\bar{d}$, while about 95% of the data remains within the LoA. The width of the LoA (width = 2.347) is larger than that of the proposed method (ω_r =

2.202). Thus, the proposed method is more consistent with the CCC and provides a better visual tool for evaluating the agreement in comparison with the LoA.

In summary, nearly 95% of the data lie in the LoA for all scenarios, and the visual evaluation on agreement depends on the width of the LoA and the predetermined acceptable difference. If the same acceptable difference is applied to all scenarios, the rank by the LoA approach is that scenario II > scenario I & III > scenario IV, while the proposed method ranks the concordance as scenario II & III > scenario I > scenario IV, based on the number of outliers from the RB. We observe that the proposed method is consistent with the CCC values and is considerably less sensitive to between-subject variability since the CCC is a scaled index. The similar results are obtained for the sample size of 500 (plots and summary tables are omitted).

A graphical comparisons of simulation results for n = 500 were presented in Supplementary Figure 1 when the data X_1 and X_2 are generated from uniform distribution and correlation of 0.65, 0.75, 0.85, and 0.95 by using Demirtas method (15). The numbers of outliers of RB are 53 (10.6%), 42 (8.4%), 23 (4.6%), and 9 (1.8%), respectively. The width of RB is not dependent to correlation ρ while the width of LoA is inversely associated with ρ (the width of LoA = 0.47, 0.40, 0.30, and 0.18). Thus, we observed that the proposed method is not sensitive to the sample size and robust to the normality assumption.

Applications to Real Data

A peak expiratory flow rate (PEFR) study data in Bland and Altman's paper (1) and the Radiomics features extracted from 3D CT images in Balagurunathan et al. (7) are investigated as real examples below.

Example 1 (PEFR data): PEFR was measured using two different types of equipment: a large Wright peak flow meter and a mini Wright peak flow meter. There are two measurements for each meter, and data are shown in Supplementary Table 1. Only the first measurement by each meter is used for the comparison of our proposed method with the LoA which is obtained as

$$\bar{d} \pm t_{16,0.025}\, S_{1-2} = -2.12 \pm 82.18\ (l/min).$$

The boundary lines of the RB are

$$X_2 - X_1 = \pm \frac{1}{\sqrt{2}} \times t_{32,0.025} \times 118.26 = \pm 170.33\ (l/min),$$

in $\left(\frac{X_1+X_2}{2},\ X_2 - X_1\right)$ plane. As depicted in Figure 3, the width of LoA is approximately two-folds of the RB. All data are clustered in the RB, implying that the CCC value would be greater than 0.75 and that the two meters have an excellent concordance based on the scaled index. Note that estimates of the CCC, the

Pearson correlation coefficient, and the bias correction factor are 0.943, 0.943, and 0.999, respectively, due to the large between-subject variability. However, the mini meter would be unacceptable for clinical purposes because the width of LoA (± 82.18) is too wide that considered as the evidence of the lack of reproducibility.

Example 2 (Radiomics Data): In Balagurunathan et al. study (7), they developed and identified a set of features extracted from CT images that can be converted into quantifiable and minable data as a potential prognostic and predictive biomarker of clinical outcomes. The unenhanced thoracic CT images for 32 patients in test-retest settings were acquired within 15 minutes of each other, using the same CT scanner. All patients had a primary pulmonary tumor of 1cm or larger. A total of 64 lesions (2 per patient) were segmented, and a total of 219 3D features were extracted from CT scans. Two segmentation methods, manual and automatic single-click ensemble segmentation developed by Balagurunathan et al., were used to get the correct segmentation boundaries of tumors. These 219 features can be broadly divided into two classes; non-texture and texture features. Non-texture features include tumor size, shape, and location description, while texture features include pixel histogram, run length, co-occurrence, Laws, and wavelet-based features (see details in Balagurunathan et al. (7)). The first step of the process is to screen out less reproducible features. Unlike PEFR study, it is impractical to determine the acceptable difference for assessing the agreement between two observations. Thus, the scaled index such as CCC would be a reasonable measure for assessing agreement.

In this paper, two non-texture features, shortest × longest diameter and volume, out of 219 features are considered. The log-transformation is taken to improve the normality. The estimated CCC values, $\hat{\rho}_c$, of two features obtained by two segmentation methods are very close to 1 (Table 2), and the graphical evaluation of agreement is presented in Figure 4. The CCC value of 0.75 is selected as the lower limit of excellent concordance. As shown in Figure 4, all data are clustered near 0 within the RB, all CCC values are considerably larger than 0.75, and it is anticipated from visual evaluation that the agreement of volume by manual segmentation (lower left panel) is the highest while shortest × longest diameter by ensemble segmentation (upper right panel) is the lowest among them, which is consistent with the CCC values, $\hat{\rho}_c$ (Table 2).

Conclusions

The Bland-Altman (B-A) plot with the limit of agreement (LoA) has been widely used as a visual tool for assessing agreement. The agreement is evaluated by comparing acceptable difference with the LoA, an unscaled index. If acceptable difference cannot be determined or difference of data may not be

interpretable, scaled indices such as ICC or CCC may be used to assess the agreement. Despite of the popularity, the B-A plot may mislead practitioners since it may not be consistent with scaled indices, especially when the common variance is large but two measurement are highly concordant (scenario III in comparison with scenario II). In this paper, we propose a novel, CCC-based visual tool for assessing agreement. We propose use of the reference band (RB), using the $\alpha/2$ upper critical point of the $t-$distribution for a given a confidence level α and degrees of freedom v. In absence of acceptable difference, we evaluate the LoA in B-A plot by simulation studies and found that the LoA seems to derive subjective inference from between-subject variability and may not be consistent with CCC. Our simulation studies show that our visual tool is consistent with CCC than the LoA. Unlike the LoA, the RB is a descriptive visual tool which is developed from CCC under normality assumption. Although the proposed method appears robust when data are uniformly distributed, the number of outliers from the RB may not be associated with CCC if data are not normally distributed. We also hope that the proposed method can provide practitioners with additional useful information such as recognition of patterns and identification of outliers in data. *Maltab* programs and R shiny app used for this paper is available upon request.

Abbreviations

B-A: Bland-Altman
CCC: concordance correlation coefficient
CP: coverage probability
ICC: intraclass correlation coefficient
LoA: limit of agreement
PEFR: peak expiratory flow rate
RB: reference band
TDI: total deviation index

Declarations

Ethics approval and consent to participate: Not Applicable

Consent for publication: Not Applicable

Availability of data and material

PEFR Data are presented in Supplementary Table 1. Radiomic data will be available, depending upon Dr. Robert Gillies' approval.

Competing Interests

All authors declare that they have no competing interest and have nothing to disclose any conflict of interest.

Funding

This work has been supported in part (i.e., for conducting simulations and manuscript writing) by the Biostatistics and Bioinformatics Shared Resource at the H. Lee Moffitt Cancer Center & Research Institute, an NCI designated Comprehensive Cancer Center (P30-CA076292). No support was received for the design of study and collection, analysis, and interpretation of data.

Authors' Contribution

JK and JL develop the method and are responsible of interpretation of results and drafting the paper. All authors read and approved the final manuscript.

Acknowledgments

The authors express many thanks to Dr. Robert Gillies and his team for providing Radiomics data.

References

1. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet. 1986;1(8476):307-10. PubMed PMID: WOS:A1986AYW4000013.
2. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135-60. doi: Doi 10.1177/096228029900800204. PubMed PMID: WOS:000083700100004.
3. Carkeet A. Exact Parametric Confidence Intervals for Bland-Altman Limits of Agreement. Optometry Vision Sci. 2015;92(3):E71-E80. doi: Doi 10.1097/Opx.0000000000000513. PubMed PMID: WOS:000350314000001.
4. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Stat Med. 2000;19(2):255-70. Epub 2000/01/21. doi: 10.1002/(sici)1097-0258(20000130)19:2<255::aid-sim293>3.0.co;2-8. PubMed PMID: 10641028.
5. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. J Am Stat Assoc. 2002;97(457):257-70. PubMed PMID: WOS:000173997500028.
6. Lin LI. A Concordance Correlation-Coefficient to Evaluate Reproducibility. Biometrics. 1989;45(1):255-68. PubMed PMID: WOS:A1989U124500022.
7. Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O, Hawkins S, Kim J, Goldgof DB, Hall LO, Gatenby RA, Gillies RJ. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational Oncology. 2014;7(1):72-87. PubMed PMID: WOS:000342684300010.
8. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat. 2007;17(4):529-69. Epub 2007/07/07. doi: 10.1080/10543400701376480. PubMed PMID: 17613641.
9. Barnhart HX, Yow E, Crowley AL, Daubert MA, Rabineau D, Bigelow R, Pencina M, Douglas PS. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res. 2016;25(6):2939-58. Epub 2014/05/17. doi: 10.1177/0962280214534651. PubMed PMID: 24831133.
10. Van Noorden R, Maher B, Nuzzo R. The Top 100 Papers. Nature. 2014;514(7524):550-3. doi: DOI 10.1038/514550a. PubMed PMID: WOS:000343801500021.
11. Escaramis G, Ascaso C, Carrasco JL. The total deviation index estimated by tolerance intervals to evaluate the concordance of measurement devices. BMC Med Res Methodol. 2010;10:31. Epub 2010/04/10. doi: 10.1186/1471-2288-10-31. PubMed PMID: 20377875; PMCID: PMC2859350.
12. Nickerson CAE. A note on "A concordance correlation coefficient to evaluate reproducibility". Biometrics. 1997;53(4):1503-7. PubMed PMID: WOS:000071147700028.
13. B R. Fundamentals of Biostatistics. 6th edition ed: Thomson, USA; 2006.
14. Kim J. The computation of bivariate normal and t probabilities, with application to comparisons of three normal means. Comput Stat Data An. 2013;58:177-86. doi: 10.1016/j.csda.2012.08.015. PubMed PMID: WOS:000312359100015.
15. Demirtas H. Generating Bivariate Uniform Data with a Full Range of Correlations and Connections to Bivariate Binary Data. Commun Stat-Theor M. 2014;43(17):3574-9. doi: 10.1080/03610926.2012.700373. PubMed PMID: WOS:000340366200002.

Figure 1. Definition of the reference band (RB) in $\left(\frac{X_1+X_2}{2},\ X_2-X_1\right)$ plane, assuming $C_b = 1$.

Figure 2. Comparisons with the limit of agreement for 4 different scenarios; scenario I ($\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$; $C_b = 1$, ρ $= 0.75$; $\rho_c = 0.75$), scenario II ($\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$; $C_b = 1$, ρ $= 0.95$; $\rho_c = 0.95$), scenario III ($\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 2$; $C_b = 1$, ρ $= 0.95$; $\rho_c = 0.95$) and scenario IV ($\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 2$; $C_b = 2/3$, ρ $= 0.9$; $\rho_c = 0.6$). The CCC of 0.75 is selected as a lower bound of excellent concordance. The sample size is 50.

Figure 3. Peak Expiratory Flow Rate (PEFR) data analysis; the comparison of the reference band (RB) with the limit of agreement. CCC was set up at 0.75.

Figure 4. Radiomics Data Analysis; the comparison of the reference band (RB) with the limit of agreement (LoA). Shortest × longest diameter from manual segmentation (upper left panel), shortest × longest diameter from ensemble segmentation (upper right panel), volume manual segmentation (lower left panel), and volume from ensemble segmentation (lower right panel). All data were log-transformed.

Figure 1.

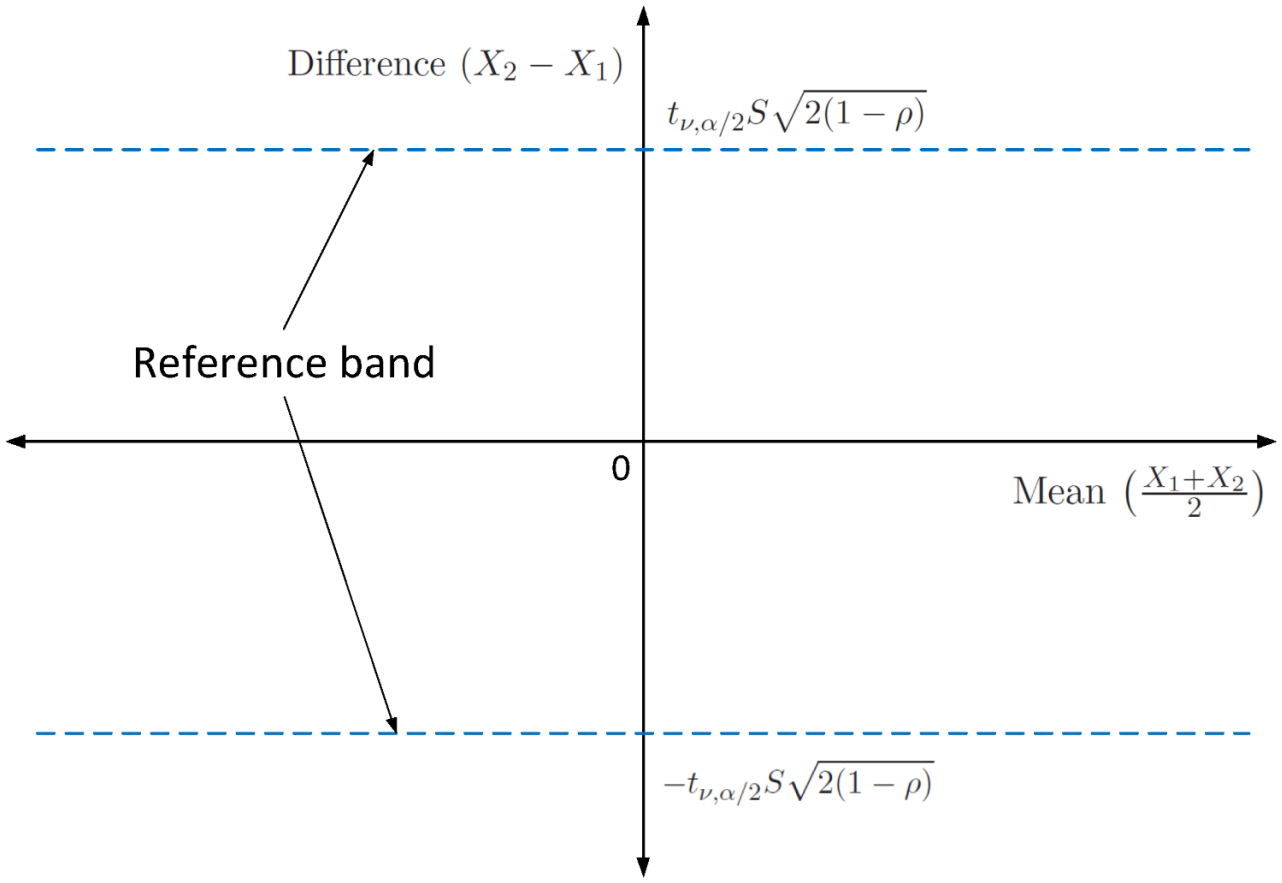

Difference $(X_2 - X_1)$
$t_{\nu,\alpha/2} S\sqrt{2(1-\rho)}$
Reference band
0
Mean $\left(\frac{X_1+X_2}{2}\right)$
$-t_{\nu,\alpha/2} S\sqrt{2(1-\rho)}$

Figure 2.

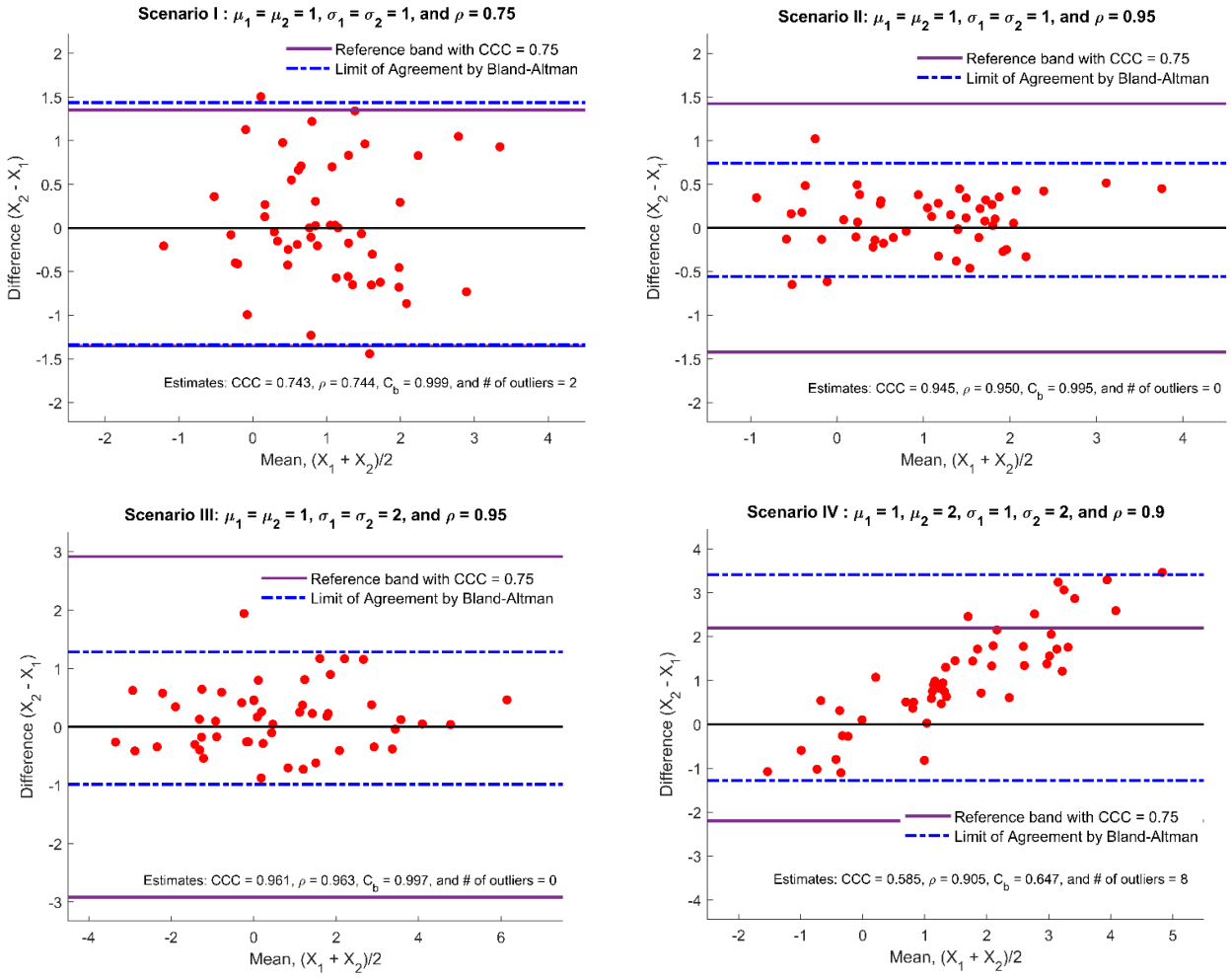

Scenario I : $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.75$
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.743, ρ = 0.744, C_b = 0.999, and # of outliers = 2
Scenario II: $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.95$
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.945, ρ = 0.950, C_b = 0.995, and # of outliers = 0
Scenario III: $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 2$, and $\rho = 0.95$
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.961, ρ = 0.963, C_b = 0.997, and # of outliers = 0
Scenario IV : $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 2$, and $\rho = 0.9$
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.585, ρ = 0.905, C_b = 0.647, and # of outliers = 8
Difference $(X_2 - X_1)$
Mean, $(X_1 + X_2)/2$

Figure 3.

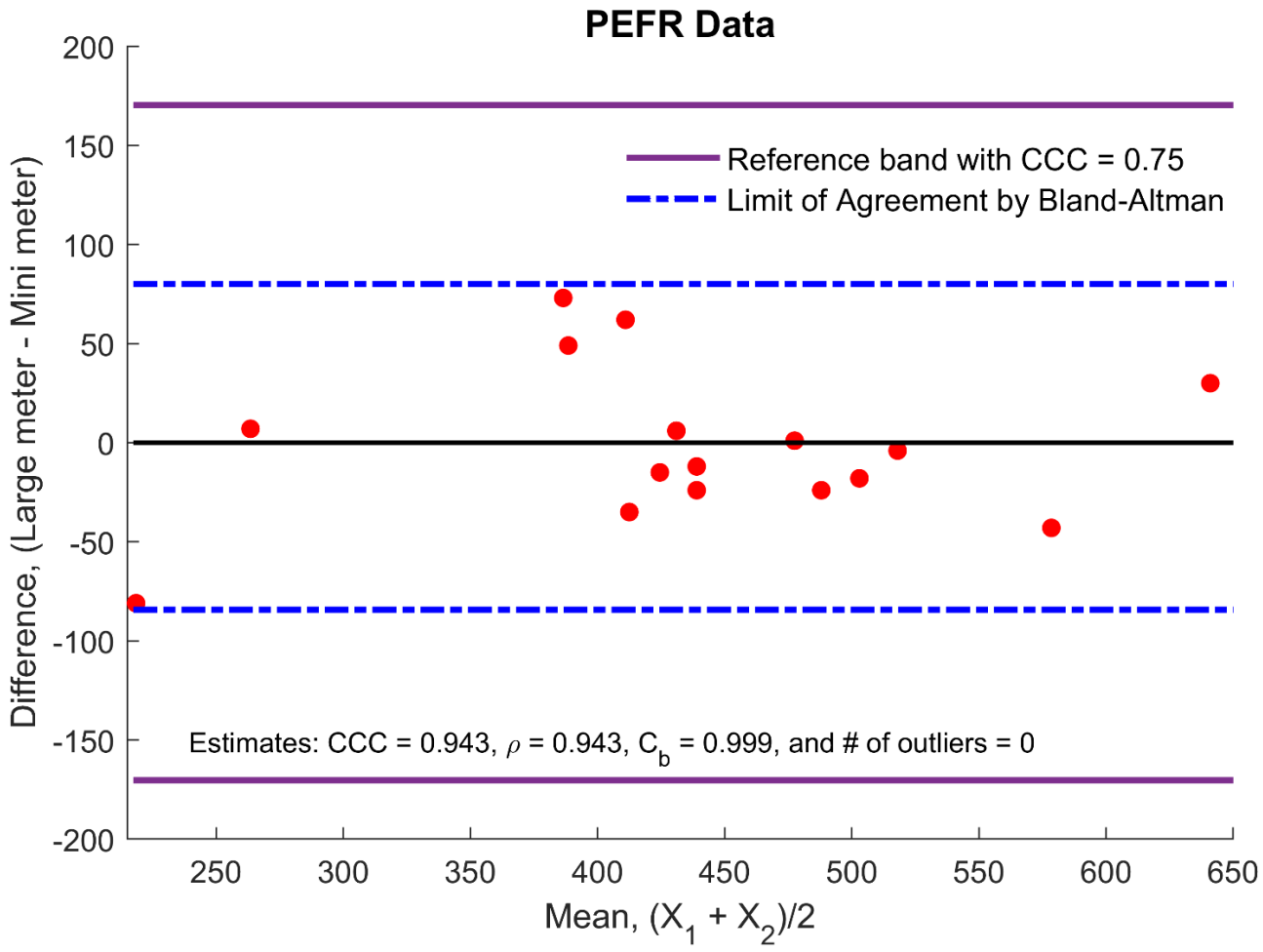

PEFR Data
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.943, ρ = 0.943, C_b = 0.999, and # of outliers = 0
Difference, (Large meter - Mini meter)
Mean, (X1 + X2)/2
200
150
100
50
0
-50
-100
-150
-200
250
300
350
400
450
500
550
600
650

Figure 4.

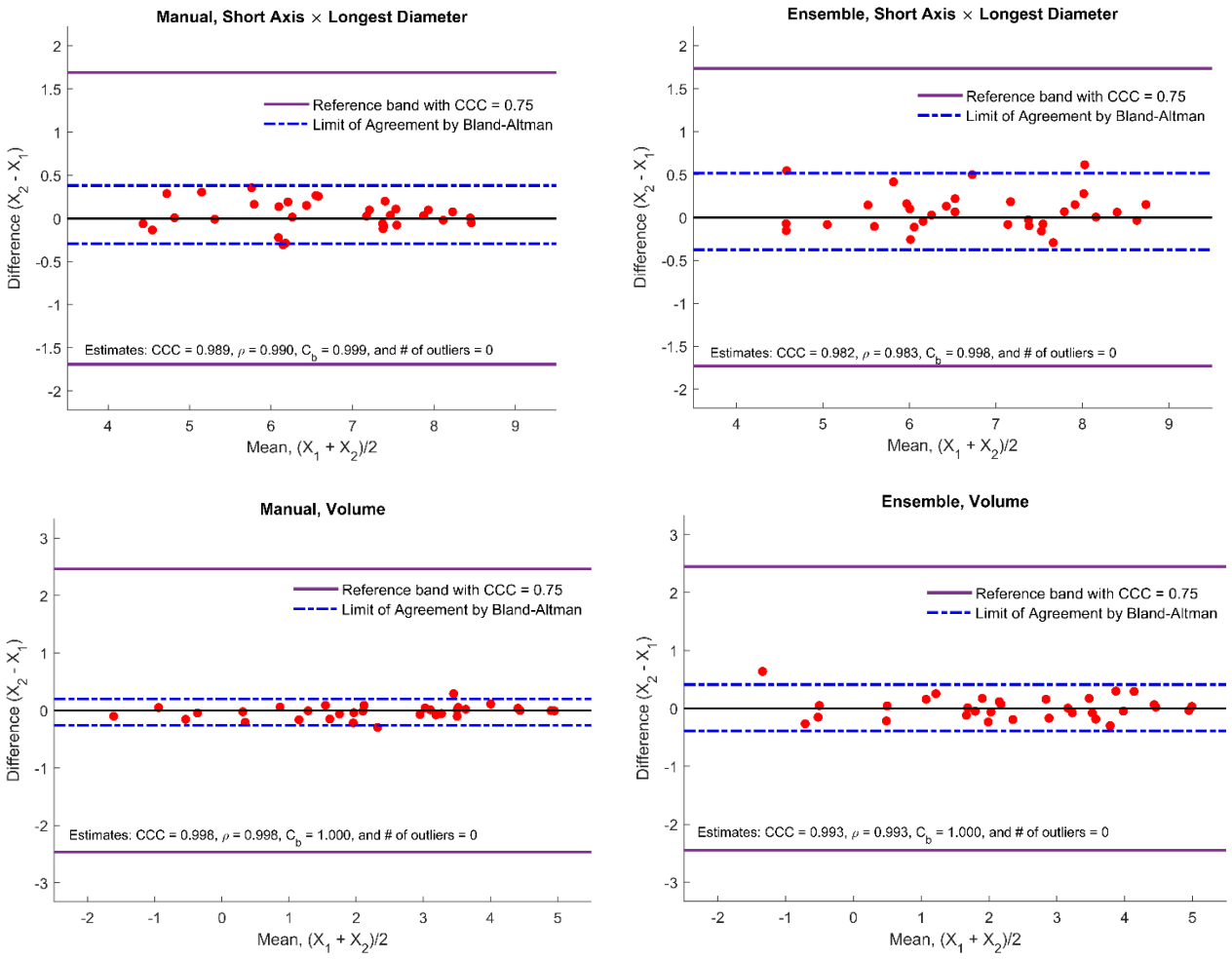

Manual, Short Axis × Longest Diameter
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.989, ρ = 0.990, C_b = 0.999, and # of outliers = 0
Difference (X_2 - X_1)
Mean, (X_1 + X_2)/2
Ensemble, Short Axis × Longest Diameter
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.982, ρ = 0.983, C_b = 0.998, and # of outliers = 0
Difference (X_2 - X_1)
Mean, (X_1 + X_2)/2
Manual, Volume
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.998, ρ = 0.998, C_b = 1.000, and # of outliers = 0
Difference (X_2 - X_1)
Mean, (X_1 + X_2)/2
Ensemble, Volume
Reference band with CCC = 0.75
Limit of Agreement by Bland-Altman
Estimates: CCC = 0.993, ρ = 0.993, C_b = 1.000, and # of outliers = 0
Difference (X_2 - X_1)
Mean, (X_1 + X_2)/2

Table 1. Simulation study results (n=50 simulated data from bivariate normal distribution)

Case	Parameters						Estimates								
	μ_1	μ_2	σ_1	σ_2	ρ	C_b	$\hat{\rho}_c$	$\hat{\rho}$	$\hat{C}_b$	ω_r	S	S_d	# outliers[1]		Width of LoA[2]
I	1	1	1	1	0.75	1	0.743	0.744	0.999	1.353	0.964	0.690	2	4%	1.387
II	1	1	1	1	0.95	1	0.945	0.950	0.995	1.423	1.014	0.323	0	0%	0.650
III	1	1	2	2	0.95	1	0.961	0.963	0.997	2.919	2.081	0.565	0	0%	1.135
IV	1	2	1	2	0.9	2/3	0.585	0.905	0.647	2.202	1.569	1.168	8	16%	2.347

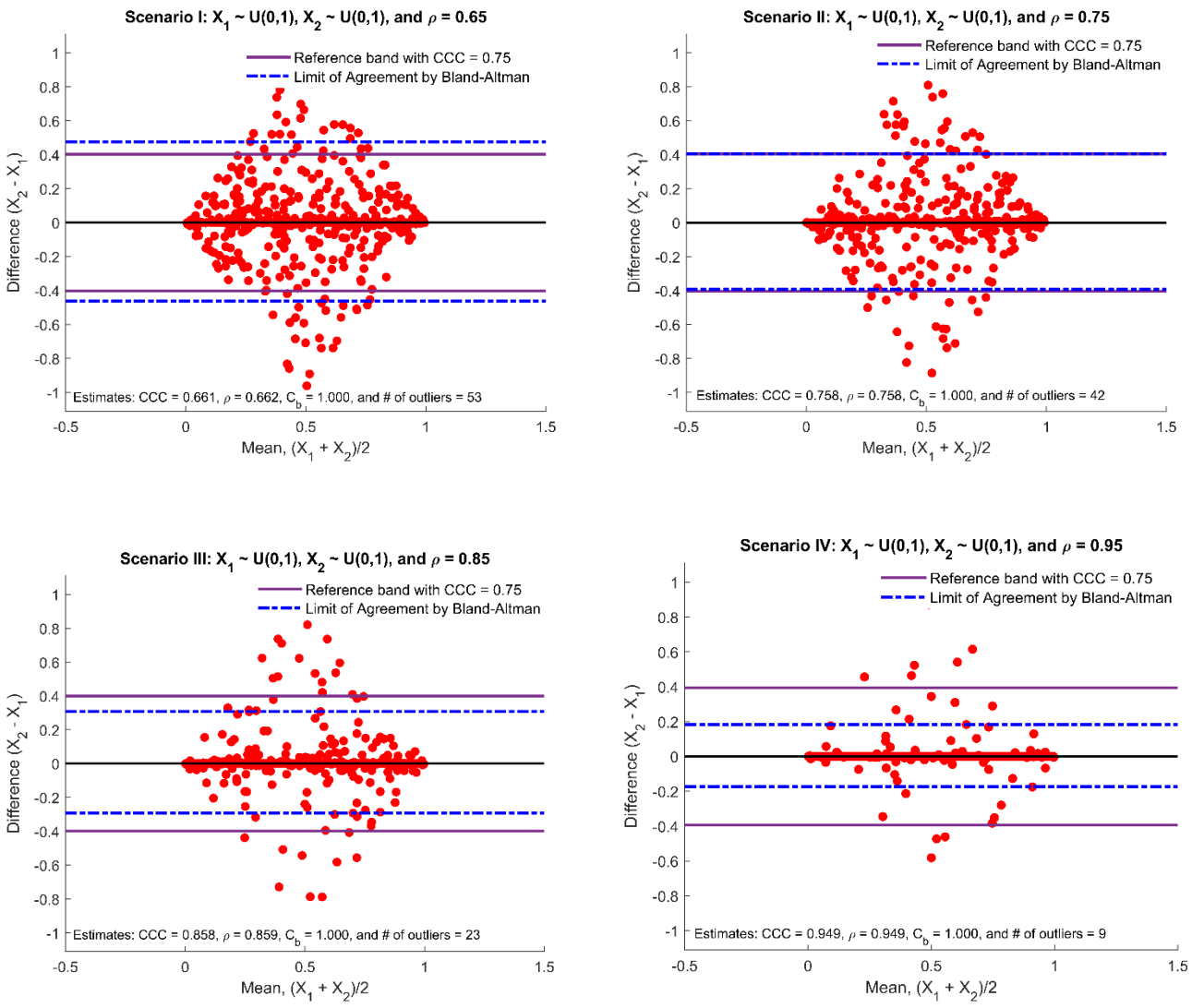
1. # of outliers indicate the number of outliers of the reference band
2. Width of LoA indicates the width of the limit of agreement.

Table 2. Radiomics Data Analysis; agreement of the features obtained from manual and ensemble segmentation. Data were all log-transformed.

Features	Manual			Ensemble		
	$\hat{\rho}_c$	$\hat{\rho}$	$\hat{C}_b$	$\hat{\rho}_c$	$\hat{\rho}$	$\hat{C}_b$
Short Axis × Longest Diameter [mm^2]	0.9895	0.9902	0.9992	0.9818	0.9835	0.9983
Volume [cm^3]	0.9977	0.9981	0.9997	0.9933	0.9934	0.9999

Supplementary Table 1. Peak Expiratory Flow Rate (PEFR; l/min) measured with Wright peak flow and Mini Wright flow meters

Subject	Large Wright Peak Flow Meter		Mini Wright Peak Flow Meter		Large meter - Mini meter
	1st PEFR	2nd PEFR	1st PEFR	2nd PEFR	
1	494	490	512	525	-18
2	395	397	430	415	-35
3	516	512	520	508	-4
4	434	401	428	444	6
5	476	470	500	500	-24
6	557	611	600	625	-43
7	413	415	364	460	49
8	442	431	380	390	62
9	650	638	658	642	-8
10	433	429	445	432	-12
11	417	420	432	420	-15
12	656	633	626	605	30
13	267	275	260	227	7
14	478	492	477	467	1
15	178	165	259	268	-81
16	423	372	350	370	73
17	427	421	451	443	-24
Mean	450.35	445.41	452.47	455.35	-2.12
SD	116.31	119.61	113.12	111.32	38.77



Supplementary Figure 1. Comparisons with the limit of agreement for 4 different scenarios; scenario I ($C_b = 1$, $\rho = 0.65$), scenario II ($C_b = 1$, $\rho = 0.75$), scenario III ($C_b = 1$, $\rho = 0.85$), and scenario IV ($C_b = 1$, $\rho = 0.95$). The CCC of 0.75 is selected as a lower bound of excellent concordance. The sample size is 500. X_1 and X_2 are generated from uniform distribution. The numbers of outliers of RB are 53 (10.6%), 42 (8.4%), 23 (4.6%), and 9 (1.8%), respectively. The width of RB is not dependent to correlation ρ while the width of LoA is inversely associated with ρ (the widths of LoA are 0.47, 0.40, 0.30, and 0.18, respectively).

Figures

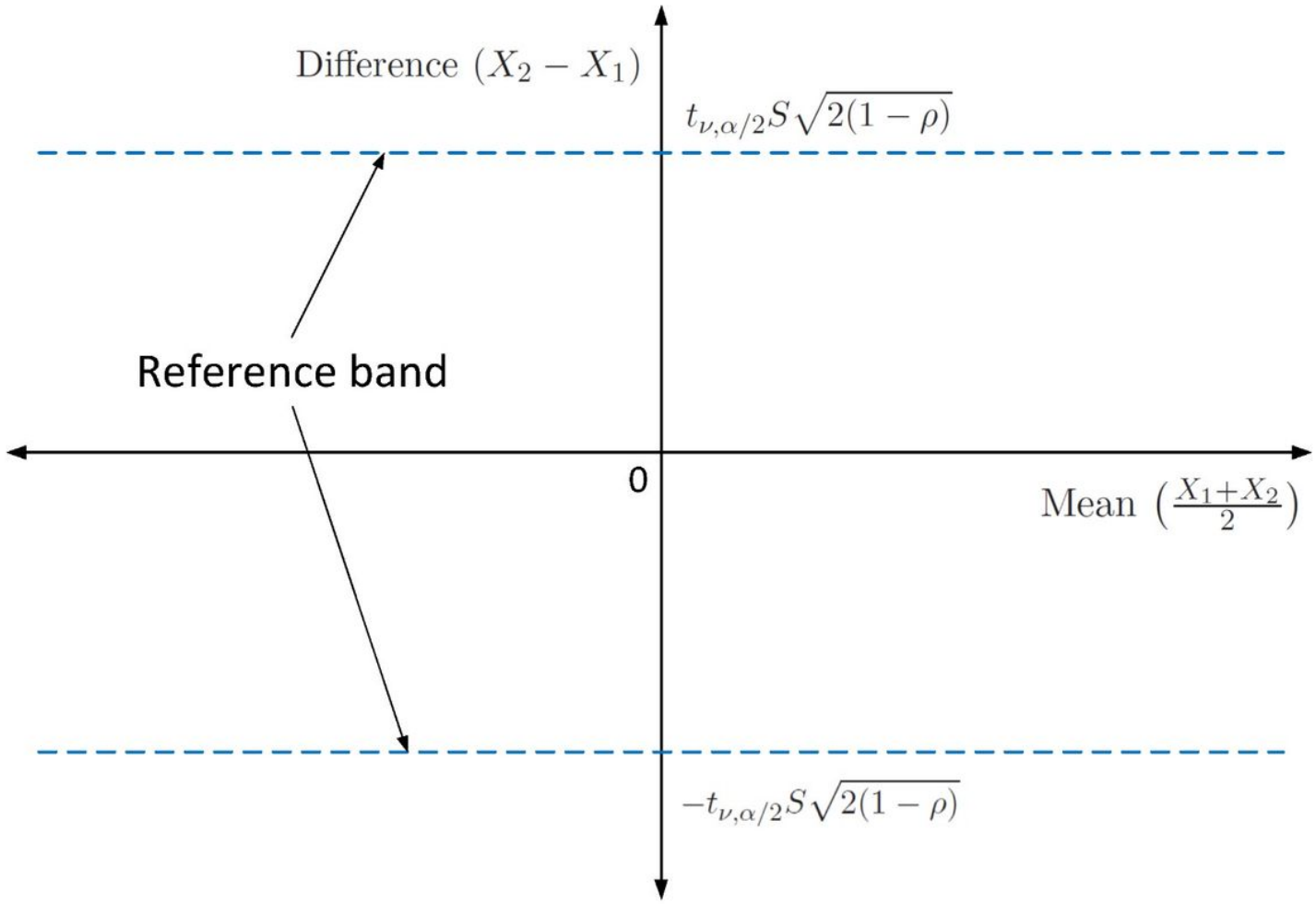


Figure 1

Please see the Manuscript PDF file for the complete figure caption.

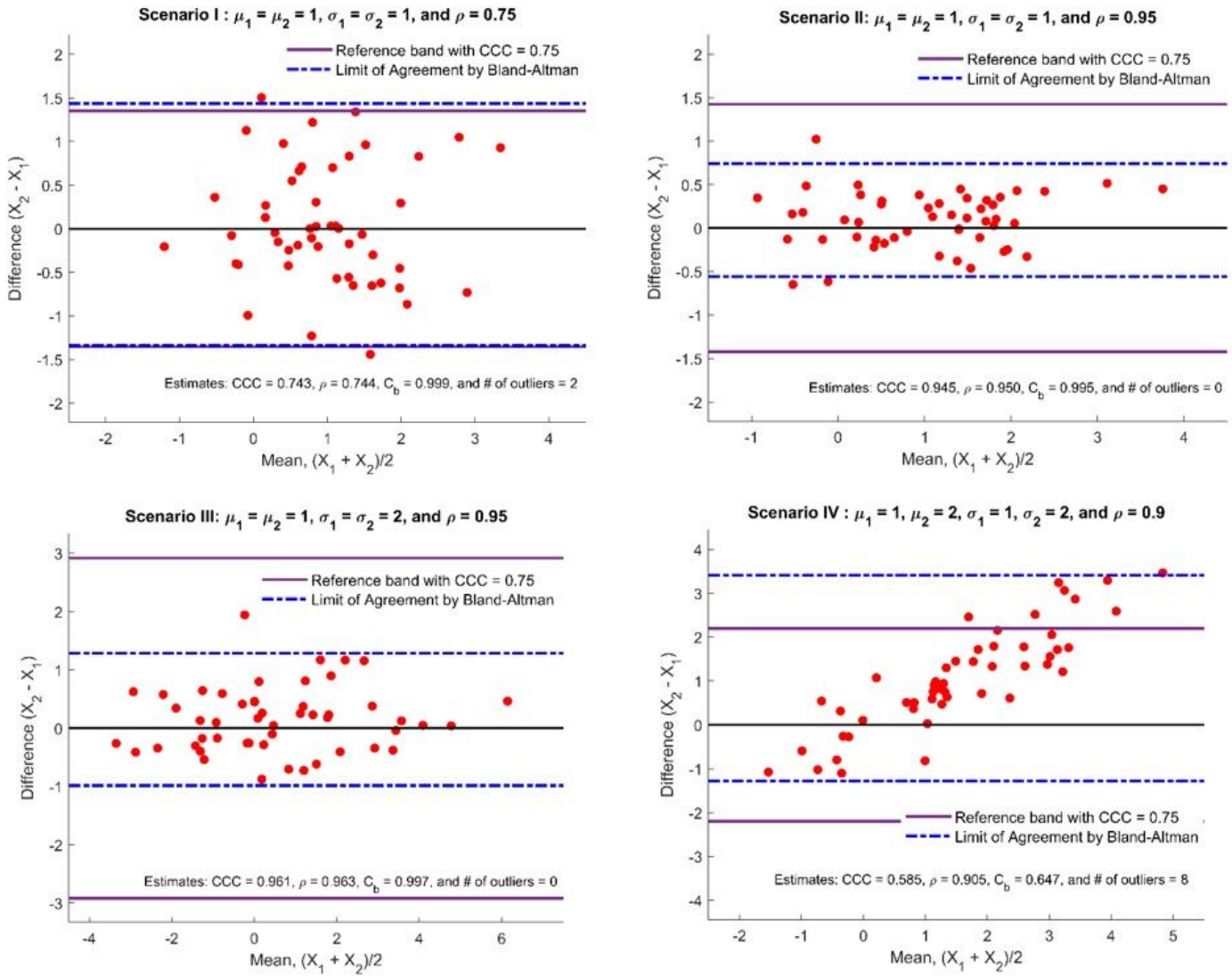


Figure 2

Please see the Manuscript PDF file for the complete figure caption.

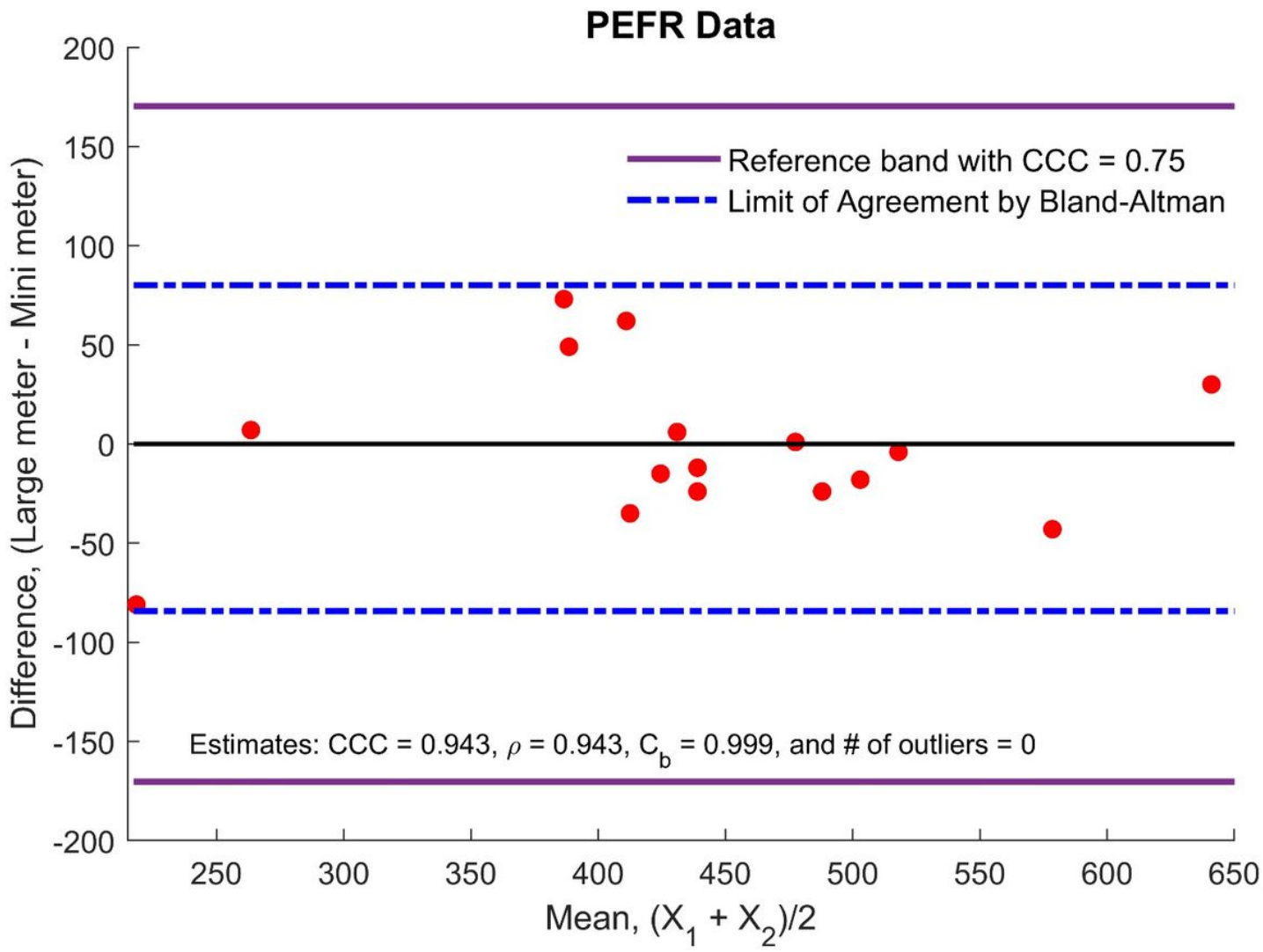


Figure 3

Peak Expiratory Flow Rate (PEFR) data analysis; the comparison of the reference band (RB) with the limit of agreement. CCC was set up at 0.75.

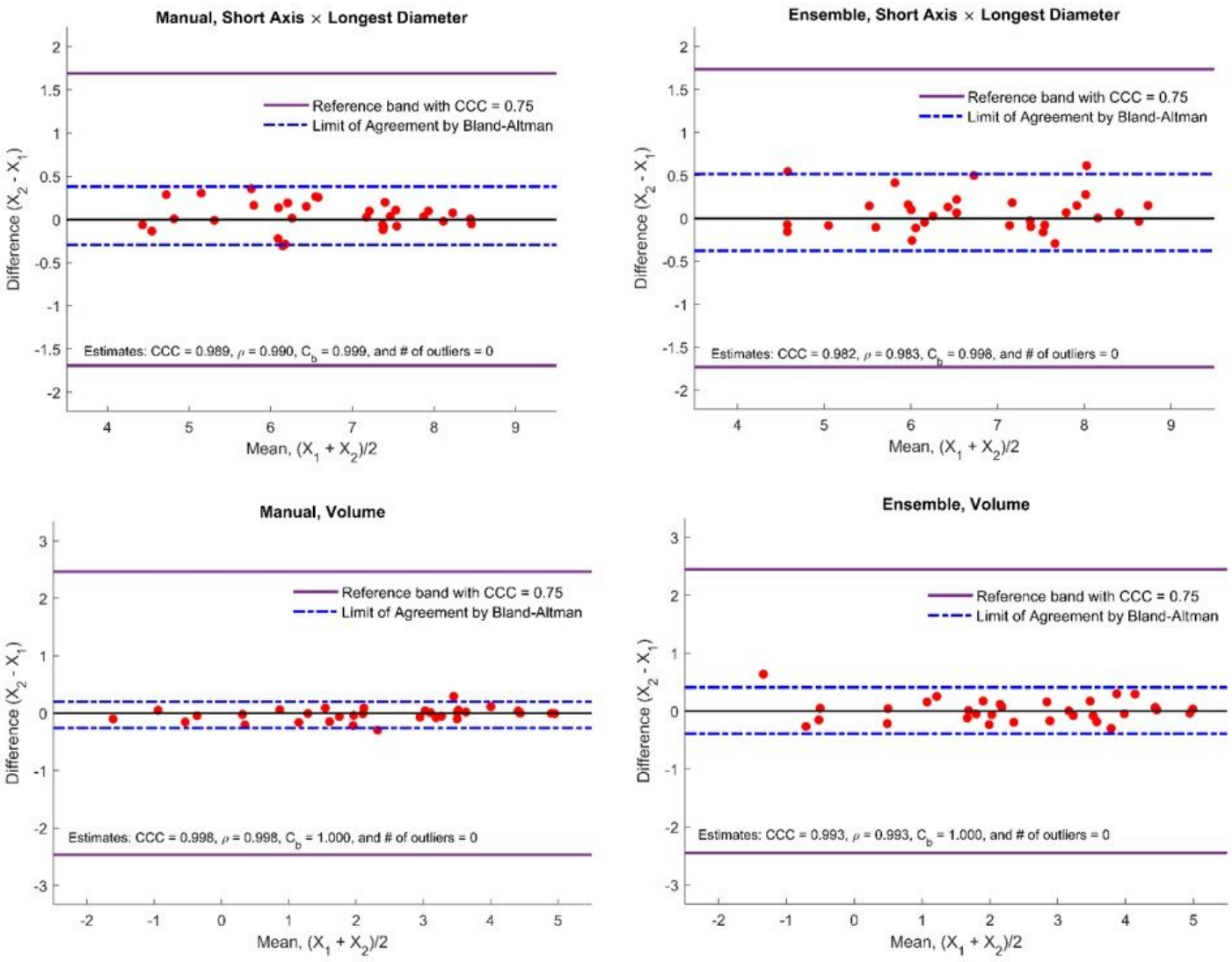


Figure 4

Radiomics Data Analysis; the comparison of the reference band (RB) with the limit of agreement (LoA). Shortest × longest diameter from manual segmentation (upper left panel), shortest × longest diameter from ensemble segmentation (upper right panel), volume manual segmentation (lower left panel), and volume from ensemble segmentation (lower right panel). All data were log-transformed.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableandFigure.docx