A practical guide to assess reproducibility of echocardiographic measurements

Bunting, Karina V.; Steeds, Richard P.; Slater, Luke T.; Rogers, Jennifer K.; Gkoutos, Georgios V.; Kotecha, Dipak

DOI:
10.1016/j.echo.2019.08.015
10.1016/j.echo.2019.08.015

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Bunting, KV, Steeds, RP, Slater, LT, Rogers, JK, Gkoutos, GV & Kotecha, D 2019, 'A practical guide to assess reproducibility of echocardiographic measurements', *Journal of the American Society of Echocardiography*, vol. 32, no. 12, pp. 1505-1515. https://doi.org/10.1016/j.echo.2019.08.015, https://doi.org/10.1016/j.echo.2019.08.015

A Practical Guide to Assess the Reproducibility of Echocardiographic Measurements

Karina V. Bunting, BSc, MSc, Richard P. Steeds, MBBS, MA, MD, FESC, Luke T. Slater, BSc, Jennifer K. Rogers, BSc, PhD, CStat, AFHEA, Georgios V. Gkoutos, PhD, DIC, and Dipak Kotecha, MBChB, PhD, MSc, FESC, FHEA, *Birmingham, United Kingdom*

Echocardiography plays an essential role in the diagnosis and assessment of cardiovascular disease. Measurements derived from echocardiography are also used to determine the severity of disease, its progression over time, and to aid in the choice of optimal therapy. It is therefore clinically important that echocardiographic measurements be reproducible, repeatable, and reliable. There are a variety of statistical tests available to assess these parameters, and in this article the authors summarize those available for use by echocardiographers to improve their clinical practice. Correlation coefficients, linear regression, Bland-Altman plots, and the coefficient of variation are explored, along with their limitations. The authors also provide an online tool for the easy calculation of these statistics in the clinical environment (www.birmingham.ac.uk/echo). Quantifying and enhancing the reproducibility of echocardiography has important potential to improve the value of echocardiography as the basis for good clinical decision-making. (J Am Soc Echocardiogr 2019;32:1505-15.)

Keywords: Echocardiography, Reproducibility, Repeatability, Reliability

Echocardiography is a key cardiac investigation that has contributed to improvements in the diagnosis and management of cardiovascular disease.[1,2] The use of echocardiography continues to grow, not only in number but also in the types of measurements, from M-mode to two-dimensional imaging, Doppler echocardiography, three-dimensional imaging, and speckle-tracking. However, inpatient hospital data suggest that echocardiography continues to be underused in critical cardiovascular conditions, and it is an operator-dependent technique that is prone to variable reproducibility.[3] Defining a reproducible measurement from echocardiography is challenging because of intrinsic biologic variation and the difference in measurement and interpretation between operators.

The term *reproducibility* covers many overlapping concepts. It is explicitly defined as the variation of the same measurement made on a subject under changing conditions, but in real-life practice it also includes changes in measurement method, observer, time frame, instrumentation, location, and/or environment. Repeatability can be separately considered as the variation in repeat measurements made on the same subject under identical conditions, whereas reliability is the magnitude of error between measurements.[4] It is inevitable that there will be some degree of error in clinical measurements, and the acceptable amount will depend on particular circumstances.[5-7] The correct statistical tests to determine these forms of reproducibility are often poorly considered, with the potential to mislead and confound clinical decision-making.[7]

As clinical indications for echocardiography increase, it is essential that these measurements can be relied upon for accurate diagnosis and serial assessment of cardiac function.[8,9] In this article, we

From the Institute of Cardiovascular Sciences, University of Birmingham (K.V.B., R.P.S., D.K.); University Hospitals Birmingham NHS Foundation Trust (K.V.B., R.P.S., D.K.); the Institute of Cancer and Genomic Sciences, University of Birmingham (L.T.S., G.V.G.); Biomedical Research Centre, National Institute for Health Research (G.V.G.); PHASTAR Statistical Research (J.K.R.); and the Medical Research Council Health Data Research UK, Birmingham (G.V.G.), United Kingdom.

Dr. Kotecha is supported by a National Institute for Health Research (NIHR) Career Development Fellowship (CDF-2015-08-074). Ms. Bunting is funded through this fellowship as a research assistant and PhD student. Dr. Kotecha and Ms. Bunting are supported through an Accelerator Award of the British Heart Foundation awarded to the University of Birmingham Institute of Cardiovascular Sciences (AA/18/2/34218). Dr. Gkoutos acknowledges support from H2020-EINFRA (731075) and the National Science Foundation (IOS:1340112) as well as support from the NIHR Birmingham Experimental Cancer Medicine Centre, NIHR Birmingham Surgical Reconstruction and Microbiology Research Centre, and NIHR Birmingham Biomedical Research Centre and Medical Research Council Health Data Research UK (HDRUK/CFC/01). The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the NIHR, the Medical Research Council, or the Department of Health. Dr. Kotecha is chief investigator of the RATE-AF clinical trial (NCT02391337); is Steering Committee lead for the Beta-blockers in Heart Failure Collaborative Group (NCT00832442); has received a Career Development Fellowship from the NIHR (CDF-2015-08-074); is the recipient of a British Heart Foundation project grant (PG/17/55/33087); collaborates on a European Union/EFPIA Innovative Medicines Initiative Grant (BigData@Heart; #116074); and has received fees from Bayer and Atricure. Dr. Rogers has received grants from the NIHR and consultancy fees from Bayer, Quintiles-IMS, and the Institute of Cancer Research, all outside the submitted work.

Conflicts of Interest: None.

Reprint requests: Dipak Kotecha, MBChB, PhD, MSc, FESC, FHEA, University of Birmingham, Institute of Cardiovascular Sciences, Medical School, Vincent Drive, Birmingham B15 2TT, United Kingdom (E-mail: *d.kotecha@bham.ac.uk*).

0894-7317

https://doi.org/10.1016/j.echo.2019.08.015

Abbreviations

ICC = Intraclass correlation coefficient

LVEF = Left ventricular ejection fraction

review the literature on reproducibility, repeatability, and reliability in the practical context of echocardiography. In Table 1, we provide a summary of statistical tests to assess reproducibility, repeatability, and reliability. Table 2 highlights the application of these tests, giving examples specific to echocardiography. Our aim is to provide echocardiographers and clinicians with the tools to appraise their own measurements, reduce inconsistencies within and between operators, and improve the reliability of echocardiography in clinical practice. To enable assessment in routine clinical care, we provide an online calculator for key statistical tests and graphs, allowing users to input measurements and easily assess their own reproducibility: www.birmingham.ac.uk/echo.

REPRODUCIBILITY

Reproducibility assesses the degree of variation in a measurement when conditions are changed. In echocardiography, this could be used to assess the variability in different operators, between different echocardiography sessions, or across separate patients with the same condition. When assessing reproducibility, statistical tests can assess correlation, bias, and agreement; together these are used to form a conclusion as to whether a study is reproducible (see Table 1 for a summary of terms). Correlation is defined as how well one variable can be used to predict the other, bias is whether there is a systematic difference from the expected value (either under- or overestimation), and agreement is defined as how close two measurements are from one another when on the same scale.[5,21]

We explore these three aspects in detail below, along with their limitations, using the example of Simpson's biplane left ventricular ejection fraction (LVEF) assessed by two operators (Figure 1).

Association

Association assesses the relationship between groups of data, with higher values (either positive or negative) suggesting a closer association. The choice of association statistic depends on the type of data available, and below we discuss four main options.

Table 1 Terms of reference

Term	Explanation	Examples of practical application in echocardiography	Most valuable statistical tests
Reproducibility	Variation of the same measurement made on a subject under changing conditions or different operators	Comparing measurements of aortic valve peak velocity on the same patient by two different echocardiographers.	Correlation coefficients for association: Pearson or Spearman correlation (r) and linear regression (percentage of variation explained = r^2).
		Comparing the grade of mitral regurgitation by two echocardiographers as none, mild, moderate, or severe.	Measure of the agreement between two operators: Cohen's κ (or weighted κ for degree of disagreement). κ = (total number of agreements − total agreements due to chance)/(total observations − total agreements due to chance).
		Comparing the difference in measurement of ejection fraction made by 2D Simpson's biplane and 3D volumes.	Agreement: Bland-Altman limits of agreement = bias (average difference between measurements) ± 1.96 × SD.
		Assessing the correlation of left ventricular outflow tract diameter measured by multiple operators in the echocardiography department.	Intraclass correlation coefficient (requires complex computation).
Repeatability	Variation in repeat measurements made on the same subject under identical conditions	Assessing the difference in consecutive beats by the same operator for tissue Doppler E/e′ ratio.	Repeatability coefficient = within-subject SD × √2 × 1.96.
Reliability	Magnitude of error between repeated measurements	Assess within a department the variation in mitral regurgitation effective regurgitant orifice area between operators.	Minimal detectable change = 1.96 × √2 × SEM. Coefficient of variation = SD/mean × 100. Percentage change = (second measurement − first measurement)/average × 100.

2D, Two-dimensional; *3D*, three-dimensional.

HIGHLIGHTS

- Good reproducibility, repeatability, and reliability are essential for echo studies.
- Straightforward statistical evaluation can improve echocardiography practice.
- A free online tool is available at www.birmingham.ac.uk/echo.

Correlation Coefficient. The correlation coefficient (*r*) simply measures the linear relationship between two variables. The most commonly used methods are the Pearson correlation coefficient (for normally distributed variables) and the Spearman correlation coefficient (for skewed variables, using a ranking of the measurements). An *r* value of 0 implies no correlation between the variables at all. If there were a perfect correlation between two variables, the *r* value would equal 1 (or −1 if perfectly and inversely correlated).[22,23] In reality, no clinical variables could attain this level of correlation, but an *r* value above 0.8 shows a very strong correlation and between 0.6 and 0.8 a strong correlation.[24] It should be emphasized that correlation is not a good measure of agreement (discussed later) and will depend on the range of the measurement in question.[25] Statistical tests can be used to determine if these correlations are likely due to chance; *P* values in this context do not refer to the strength of correlation but instead indicate whether we have confidence in the correlation coefficient. Figures 1A, 1B, and 1C demonstrate strong or very strong correlations between the two operators for LVEF, in contrast to Figure 1D, in which correlation is relatively weak. These methods are best used for paired parameters, such as intraobserver variability (the same operator taking an echocardiographic measure twice on the same subject and assessing the variation) or interobserver variability (two operators taking the same measure on the same subject and assessing their variation).

Linear Regression. Regression analysis describes how well one variable can be used to predict the value of another, or the strength of their relationship. With enough data points, a "line of best fit" can be created on the basis of the regression equation: variable 1 = constant value + coefficient × variable 2. Given the two variables, the regression model provides the constant value and the coefficient by trying to minimize the difference between the true observed value and the value predicted from the model (also known as the residual). This method requires data that are normally distributed (not skewed) and can be affected by outlying values. It only measures to what extent two variables are linearly related (in many cases, the association can be more complex).[9,26] The value of linear regression is limited, but it can be useful to visualize the association of paired data before other statistical tests. As with any statistical measurement, there is some variability. The 95% CI gives us an idea of the bounds of uncertainty around the calculated estimate.

Figure 1A shows a very close relationship between the values from the two operators, with a regression coefficient of 0.90. This means that for every 1.0% increase in LVEF ratings in the future by operator 1, we would expect that the corresponding average LVEF measurement of operator 2 would increase by 0.9%. The CI suggests that if repeated samples are taken, there is a 95% chance that the true regression coefficient will lie in the interval between 0.45 and 1.34. Conversely, Figure 1D shows a very variable relationship, with a regression coefficient of 0.15 and a broad CI (from −0.28 to 0.60). This includes the value of 0, meaning that there may be no association between the operators at all.

Intraclass Correlation Coefficient. The intraclass correlation coefficient (ICC) is often used to determine the reproducibility and reliability of numeric measurements organized into groups beyond a simple pairing, for example, different operators measuring the same variable in different patients. The formulas for ICC are complex, but essentially they pool data and compare within and across operators on the basis of an analysis of variance. This divides the total variability into actual difference and error, and the ICC is an average of all the correlations on the basis of all the possible pairs of data.[27] In essence, the ICC is giving us confidence about how closely the values are alike in different groups of data. The ICC can be used to assess variability within a single operator (intraobserver), between different operators (interobserver), or across different time points. It possesses the advantage of being able to compare more than two groups of variables (more than two operators) and may be superior to Pearson and Spearman correlation coefficients because it considers systematic differences.[28] The disadvantage of ICC is that it has limited value for comparing reproducibility of results in different populations. Because the ICC is a dimensionless value, the outcome will vary according to the dependent variables in the population sampled; data with a wide range of values will generate a high ICC value, whereas data with a narrow range of values will result in a low ICC.[22,29,30] ICC is unhelpful if the indices show poor agreement, as there is no indication as to the source of the error. Although there is no strict ICC value that marks the cutoff for appropriate correlation,[7,23] values between 0.75 and 1.00 suggest excellent correlation, between 0.60 and 0.74 good correlation, and <0.4 poor correlation.[31] Interpretation of the ICC is demonstrated when comparing Figures 1A and 1B. Whereas the standard correlation coefficients are similar (0.82 and 0.70), the ICCs are considerably different (0.90 and 0.48) because of greater variance between the two operators in example B. However both are statistically significant, in contrast to Figure 1D.

With the online calculator, this method can be more widely used, as it allows the assessment of reproducibility in more than two groups. For example, it can be used to assess interobserver variability across all members of the echocardiography department. For more complex scenarios, or for readers planning to calculate the ICC, we would recommend review of guidance on the correct selection and reporting of the ICC.[32]

Bias

Bias indicates to what extent there is a true difference in two data points that has not resulted from chance. These statistical tests can help determine if there are significant differences in paired data, for example, two recordings of left ventricular outflow tract diameter. A small probability (often $P < .05$, which is less than one in 20) suggests that there is evidence for a difference in the two measurements (i.e., we reject the null hypothesis, which is that there is no difference in values).[33] It is important to note that the strength of statistical significance is not related to the extent of bias but rather to whether there is confidence in the rejection of a chance effect. Paired *t* tests are used for normally distributed data, and the Wilcoxon test is used for skewed data. In our example, the bias assessment is not significant for Figures 1A and 1D, whereas there is a systematic bias in Figure 1B that is highly statistically significant at $P = .004$ and gives evidence for a true difference between the

Table 2 Statistical methods useful in echocardiography

Statistical method	Strengths of method	Weakness of method	Examples from published literature
Association			
Correlation coefficient	Options for normally distributed data (Pearson) and skewed data (Spearman).	Can only account for linear relationships. Sensitive to outlying values.	n = 17 with heart failure or dilated cardiomyopathy. Very strong interoperator association between LVEF and GLS: Pearson's correlation coefficient r = 0.89 for LVEF and r = 0.97 for GLS.[10]
Linear regression	The regression line can be used to predict the value of one variable from another. Analysis of the difference between the observed and predicted values (residuals).	Can be used only if the data are normally distributed. Assumes the same degree of variance across the whole variable. Sensitive to outlying values.	n = 31 patients clinically indicated for cardiac CT. Strong correlations seen between different imaging modalities when measuring volumes and LVEF. For cardiac CT vs CMR, linear regression r^2 = 0.85; regression equation $y = 0.97x - 1.3$. For 3D TTE vs CMR, r^2 = 0.93; regression equation $y = 0.87x + 6.3$.[11]
Intraclass correlation coefficient	Assesses how closely variables are related to each other. Best for a large number of observations. Accounts for a change in the mean over time. Independent of the scale of measurement and size of error.	As with other measures above, shows correlation, not causation. No fixed clinical interpretation for level of agreement. Cannot be used to compare reliability of measurements among different studies. Affected by the size of the range of data.	n = 183 patients with hypertension, comparing two measurements 45 days apart. Excellent correlation between first and second study: ICCs of 0.90 for indexed LV mass and 0.85 for septal diameter.[12]
Bias			
t test	Provides a P value for paired data sets.	For normally distributed data only.	n = 88 patients before chemotherapy. Differences in intra- and interobserver variability of LVEF and volumes were assessed using a t test, with $P < .001$ considered to indicate statistical significance. Noncontrast 3D echocardiography had significantly lower variability than 2D Simpson's method, 2D triplane, or studies using contrast.[13]
Wilcoxon signed rank/Mann-Whitney U test	Can be used for skewed data.	Uses ranking, so assessment of raw data is needed to interpret the P value.	n = 284 children with evaluation of MAPSE using B-mode and M-mode imaging. M-mode MAPSE had significantly lower variability than B-mode lateral MAPSE for both interobserver ($P < .001$) and intraobserver ($P < .001$) observer variability (using Wilcoxon signed rank test).[14]

Agreement			
Bland-Altman plot	Demonstrates degree of agreement and depicts outliers. Demonstrates systematic bias.	Unable to detect proportional bias. Assumes normal distribution. Numeric data only. A clinical decision needs to be made as to whether there is good agreement on the basis of the width of confidence limits.	n = 50 trastuzumab patients comparing two scans a minimum of 14 days apart by the same operator, showing better agreement for GLS than Simpson's biplane LVEF. For GLS, bias = −0.1 between the two time periods; limits of agreement = −1.8 to 1.7. For LVEF, bias = 0.5; limits of agreement = −11.2 to 12.1.[15]
Cohen's κ	Measures agreement and takes into account the amount of agreement which is there by chance.	Dependent on the prevalence of a condition. Doesn't account for degree of disagreement.	n = 146 enrolled in the Multi-Ethnic Study of Atherosclerosis with echocardiography and CMR on the same day. For classification of hypertrophy (normalized for body surface area), there was weak agreement between modalities, albeit statistically significant (Cohen's κ = 0.37, $P < .001$).[16]
Weighted κ	Weights the degree of agreement and disagreement between data sets.	Requires a predefined table of weights.	n = 80 with clinical aortic stenosis undergoing cardiac CT and TEE. Weak agreement between modalities for grading aortic valve calcification: weighted κ = 0.34.[17]
Repeatability			
RC	Uses the units of the variable.	Assumes normally distributed data. Unsuitable if the extent of agreement depends on the value of the measurement.	n = 67 pregnant women with measurement of transabdominal Doppler ultrasound of the ductus venosus at 10–14 weeks of gestation. Intraobserver repeatability was better for pulsatility index for veins (RC = 1.27) compared with end-diastolic velocity (RC = 2.03).[18]
Reliability			
MDC	To assess reliability of measurements. Simple method to detect change.	Suited more for short intervals between repeated measurements.	n = 56 patients referred for echocardiography before beginning trastuzumab treatment. Lowest intra- and interobserver variability for assessing LVEF shown with 3D without contrast (MDC = 0.048% and 0.075%) vs other echocardiographic methods with and without contrast.[13]

(*Continued*)

Table 2 (*Continued*)

Statistical method	Strengths of method	Weakness of method	Examples from published literature
CV	Optimal method if the SD is proportional to the mean.	Suboptimal method if there is a large difference between the highest and lowest possible values. Limited if the degree of error is not associated with the value of the measurement. Cannot be used if there are both positive and negative values.	$n = 60$ ($n = 20$ with heart failure, $n = 20$ with LVH, and $n = 20$ with normal structure). CV values comparing CMR vs TTE show lower variation with CMR: for LVEF, 2.4%–7.3% using CMR vs 8.6%–19.4% with TTE; for LVM, 2.8%–4.8% vs 11.6%–15.7%, respectively.[19]
Percentage change	Scale independent. Simple to interpret.	Low statistical power compared with other methods. Does not correct for imbalance between groups.	$n = 608$ with Marfan syndrome assessing interobserver assessment of aortic root dimensions. Measurements using a single beat were less reliable than taking the average of three beats (percentage error 3.9 ± 3.0% vs 3.6 ± 2.6%, $P = .0002$).[20]

2D, Two-dimensional; *3D*, Three-dimensional; *CT*, Computed tomography; *CMR*, cardiac magnetic resonance; *CV*, coefficient of variation; *GLS*, global longitudinal strain; *LV*, left ventricular; *LVH*, left ventricular hypertrophy; *LVM*, left ventricular mass; *MAPSE*, mitral annular plane systolic excursion; *MDC*, minimal detectable change; *RC*, repeatability coefficient; *TEE*, transesophageal echocardiography; *TTE*, transthoracic echocardiography.

two operators. If the data points are clustered equally around the line of equality, this suggests that there is no systematic bias. Notably, Figure 1C shows a proportional bias, and so a *t* test is likely to be inaccurate in this case. Proportional bias occurs when the difference between measurements is dependent on the value of the measurement taken.[34]

Agreement

Agreement defines the degree of consensus between different measurements, and different statistical comparisons are available according to whether the data are continuous or categorical.

Bland-Altman Plot. The Bland-Altman plot is widely used to visualize the difference in two continuous measurements from the same individual, graphed according to the average value of the two measures. In terms of echocardiography, this is highly valuable to assess measurements taken on the same patient by two different echocardiographers. This method can also be used for assessing two measurements made by the same operator or two measurements using different techniques or in different environments.

Creating the Bland-Altman plot is straightforward and requires (1) plotting the difference in the pair of measurements against their mean, (2) calculating and plotting the bias (the mean of the differences), and (3) calculating and plotting the upper and lower limits of agreement (bias ± 1.96 × SD of difference). The limits of agreement indicate where the true mean (and future measurements) are likely to lie, and interpretation will depend on the clinical magnitude of the limits.[23,35] If values are consistently outside the confidence limits, it may indicate a lack of agreement or a true biologic difference that is not due just to sampling error.[5,35]

Figure 2 shows the examples from Figure 1 constructed into Bland-Altman plots. Figure 2A shows a small degree of bias (−0.67) and narrow limits of agreement (−8.3 to 6.9), whereas Figure 2B shows a systematically higher LVEF from operator 2 for each measurement. In Figure 2C, there is evidence of a proportional error, with increasing difference between measurements at both extremes of LVEF. Figure 2D shows very wide limits of agreement (−35.6 to 36.8) that are likely to be highly clinically relevant.

Kappa Statistics. Cohen's κ is used to assess the agreement between categorical data (measurements with different levels, such as the severity of valve disease or categories of left ventricular dysfunction). The result ranges from 0 (no agreement) to 1 (perfect agreement), with values < 0.6 indicating weak agreement and >0.8 strong agreement. Cohen's κ takes into account disagreement between the two operators and also agreement by chance. A modified approach, the weighted κ, can be used to determine the degree of disagreement using a predefined table of weights.[7] Figure 3A demonstrates strong agreement between the two observers ($\kappa = 0.89$) as for each case they made a similar grading for the severity of mitral regurgitation in the same patients. In contrast, Figure 3B shows almost no agreement between the two observers across the severity of mitral regurgitation ($\kappa = 0.05$). Figure 3C shows that there is good agreement between the two observers for cases with "severe" disease, but overall the agreement is weak because of a lack of consistent results with other categories of mitral regurgitation ($\kappa = 0.27$). Figure 2D shows overall moderate agreement across all cases ($\kappa = 0.65$) despite there being 100% agreement for patients with no mitral regurgitation. Note that Cohen's κ is not the only statistic that can be used to assess agreement for categorical data, and other measures are available to address some of its assumptions and shortcomings. These include

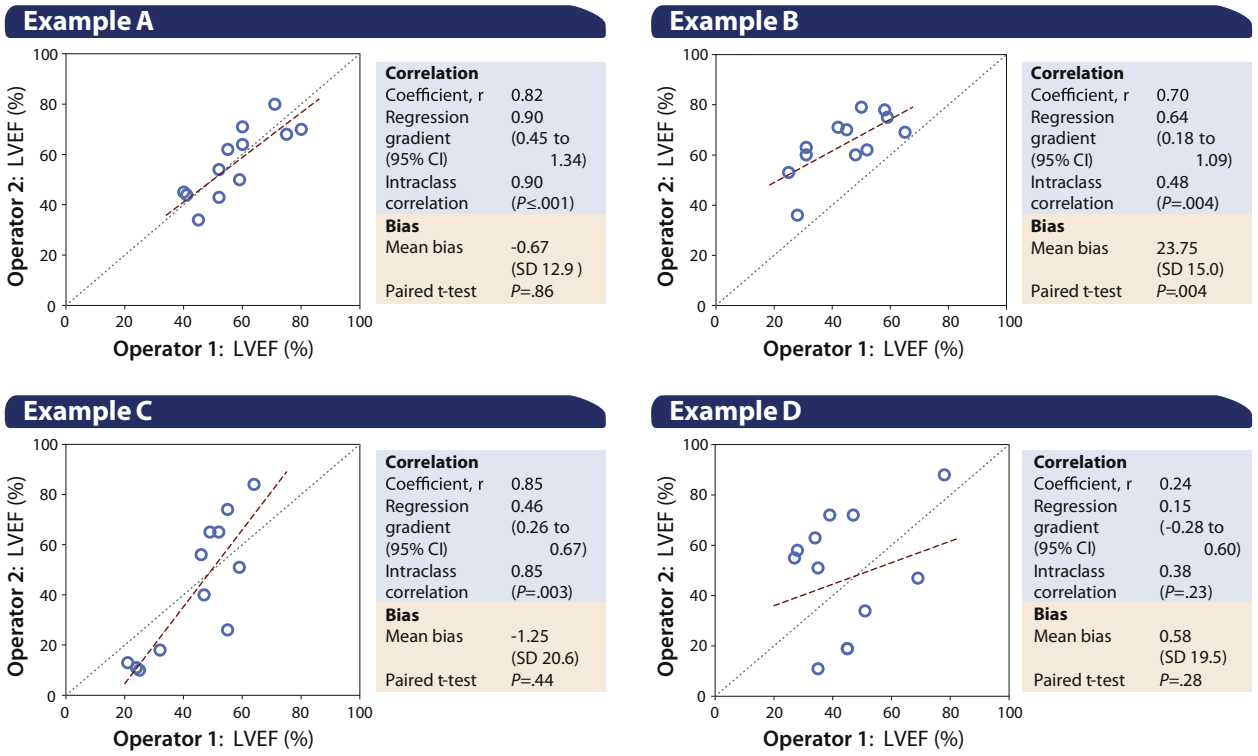

Figure 1 Reproducibility assessment between two operators. Four examples of scatterplots showing interoperator reproducibility and related statistical tests for LVEF. *Black dotted line* is the line of equality, and *red dashed line* is the line of best fit (linear regression line). **(A)** Very strong correlation with no evidence of bias. **(B)** Strong correlation but with significant bias, as operator 1 is consistently measuring LVEF at a lower value than operator 2. **(C)** Very strong correlation, but the difference in the two operators changes according to ejection fraction (proportional bias). **(D)** Weak correlation but no bias. Note that *P* values provide an assessment of statistical significance (whether due to chance) but do not imply any strength of correlation.

Fleiss's generalized κ, Scott's π coefficient, and Gwet's agreement coefficient, among others.[36]

REPEATABILITY

Repeatability studies are used to ensure that minimal variation exists when the measurement is retested on the same subject or group: the smaller the variation, the more reliable the results. When carrying out test-retest procedures, if the conditions in which the measurement are taken are kept exactly the same, then any variation detected can be attributed to the accuracy of the measurement. The time interval between repetitions should be short enough to exclude any biological change but long enough to prevent any interference from the preceding test. The appropriate time interval will vary depending on the situation,[23] but for echocardiography, repeating measurements on the same day with an interval of at least a few minutes would seem appropriate. For calculations that require multiple echocardiographic measurements for calculation, such as aortic valve area, it is important to obtain measurements under similar hemodynamic conditions. Therefore, in the context of any cardiac arrhythmias, similar cardiac cycle lengths should be selected for measurement.[37] Variability of results within a single patient can be assessed statistically by the repeatability coefficient (using the SD of differences), the coefficient of variation (discussed further below), or an ICC. The advantage of the repeatability coefficient is that its value is in the same units as the measurement, allowing easier interpretation to guide decision-making.[6,30]

Figure 4 shows an example for peak aortic valve velocity in four patients undergoing 10 consecutive measurements by the same operator for possible aortic stenosis. Figure 4A demonstrates a repeatability coefficient of 13 cm/sec, meaning that the variation in future measurements for aortic valve peak velocity are small (by that echocardiographer on that particular patient). Figure 4B shows proportional bias for velocity to increase in value as the observer takes more measurements, whereas Figure 4C displays clinically relevant variation (perhaps due to a patient factor such as atrial fibrillation). In Figure 4D we see major issues in repeatability (e.g., because of equipment problems).

RELIABILITY

To be a reliable measurement, the magnitude of the difference between repeated measurements should be within a clinically acceptable limit. The test should be precise enough to give us confidence that we can differentiate between normal or abnormal in a given population or between different patients or populations. The minimal detectable change can be used to assess reliability when measurements are repeated over a short time interval.[38] It is expressed as a percentage and represents the minimal change required to be sure that the differences observed reflect a real change rather than measurement error (with higher percentages suggesting a less reliable method).[39] The coefficient of variation is a common method to compare reliability between tests. It is calculated as the ratio of SD to mean, with a smaller percentage indicating a more precise method.[9,40] This would be useful in echocardiography for assessing

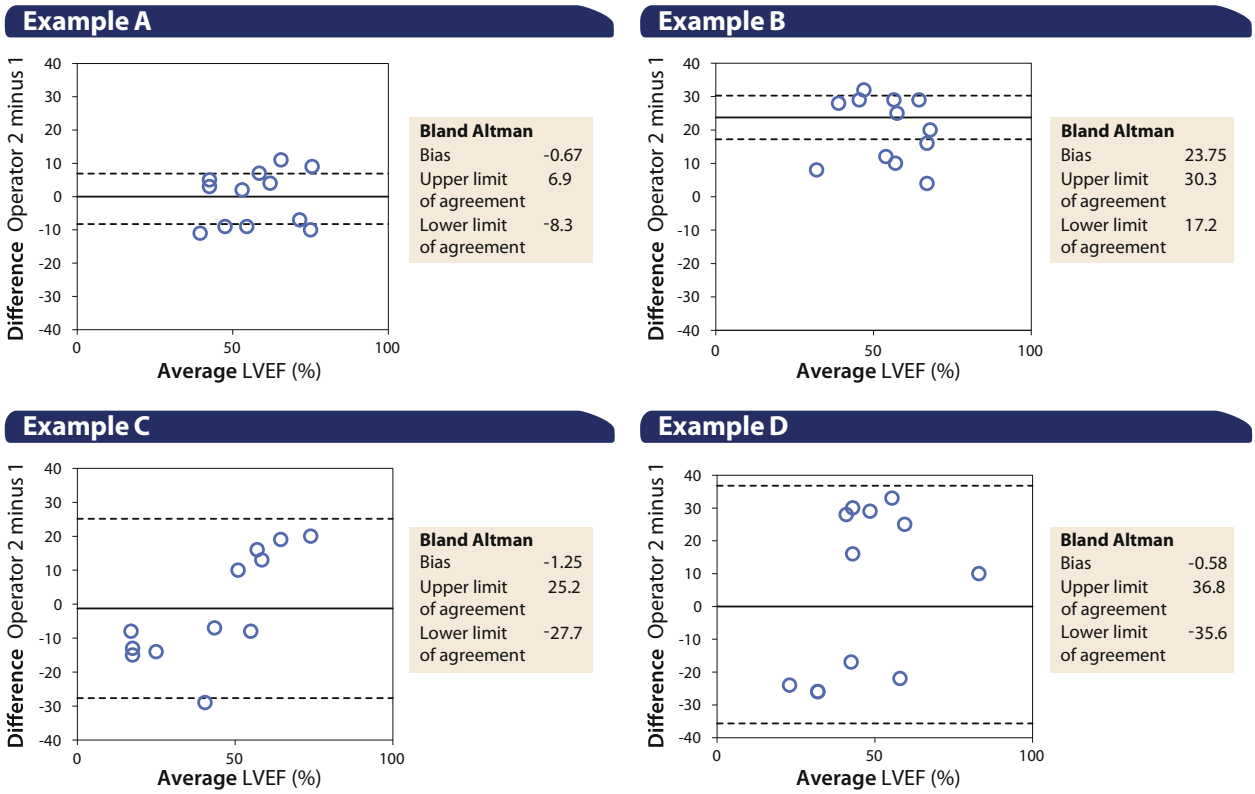

Figure 2 Bland-Altman plots for agreement between two tests or operators. Bland-Altman plots showing assessment of agreement for the four examples in Figure 1. **(A)** No bias between measurements or assessors and narrow limits of agreement. **(B)** Substantial bias (observation 2 is systematically higher than observation 1) but with narrow limits of agreement. **(C)** Low levels of bias but broad limits of agreement. **(D)** Low levels of bias with even broader limits of agreement.

the variation in parameters within a certain patient population; for example, different measures of left atrial dilatation in the same patients with hypertension or the reliability of averaging different numbers of cardiac cycles in those with atrial fibrillation. Interpretation of minimal detectable change and the coefficient of variation is depicted in Figure 4.

Simple assessment of reliability can also be calculated, such as the absolute or percentage change in two measurements. However, these tests have limited statistical power to determine differences, are unable to account for inherent variation, and the results are highly dependent on the value at baseline.[41,42] Whatever method is used, echocardiographers need to consider whether the change in measurement is due to the reliability of the test or if a biological change in the patient could explain the difference (e.g., worsening of valve disease over the time period).

DISCUSSION

To ensure that the methods we use in echocardiography are useful for clinical decisions, reproducibility, repeatability, and reliability should be assessed. Unreliable estimates have the potential to impact patient management and outcomes, as well as leading to a waste of time and resources. The challenge for the echocardiographer is not only to identify a change in a biologic parameter, but then to know whether that change is real or clinically significant. For example, is there a true change in cardiac structure and/or function that would require additional treatment, or is the change inconsistent or accounted for by changes to environment, operator, or other factors?[43]

To identify any significant variability in echocardiographic parameters within a department, intra- and interobserver variability can be measured.[9,44] The possibility of measurement error should be minimized as much as possible by ensuring that all equipment is accurately calibrated, adequate training is given to echocardiographers, and standardized guidelines are followed.[7] In clinical practice, a patient being serially assessed will likely be scanned by different echocardiographers on each occasion, hence the importance of ensuring no significant variation between operators. For numeric data (such as LVEF or Doppler values), the degree of agreement can be assessed using either the Bland-Altman plot or the ICC. Pure measures of association (such as correlation and linear regression) provide limited information but are essential components of understanding and visualizing data to assess for outliers and points that influence the trend of the data. When assessing a categorical result (such as quantifying the severity of mitral regurgitation), a κ test can be used.[7,9]

Repeatability and reliability measurements are as important and give confidence that the values obtained can be used to make clinical decisions. Repeatability coefficients and the coefficient of variation are commonly used and can be calculated without difficulty.[7] For assessment of within-subject variation, three repeat measurements are usually considered appropriate,[7,45] translated into echocardiography as the average of three Doppler indices. This is probably appropriate for sinus rhythm, but in the case of atrial fibrillation, the assessment of reproducibility is even more challenging because of the variation

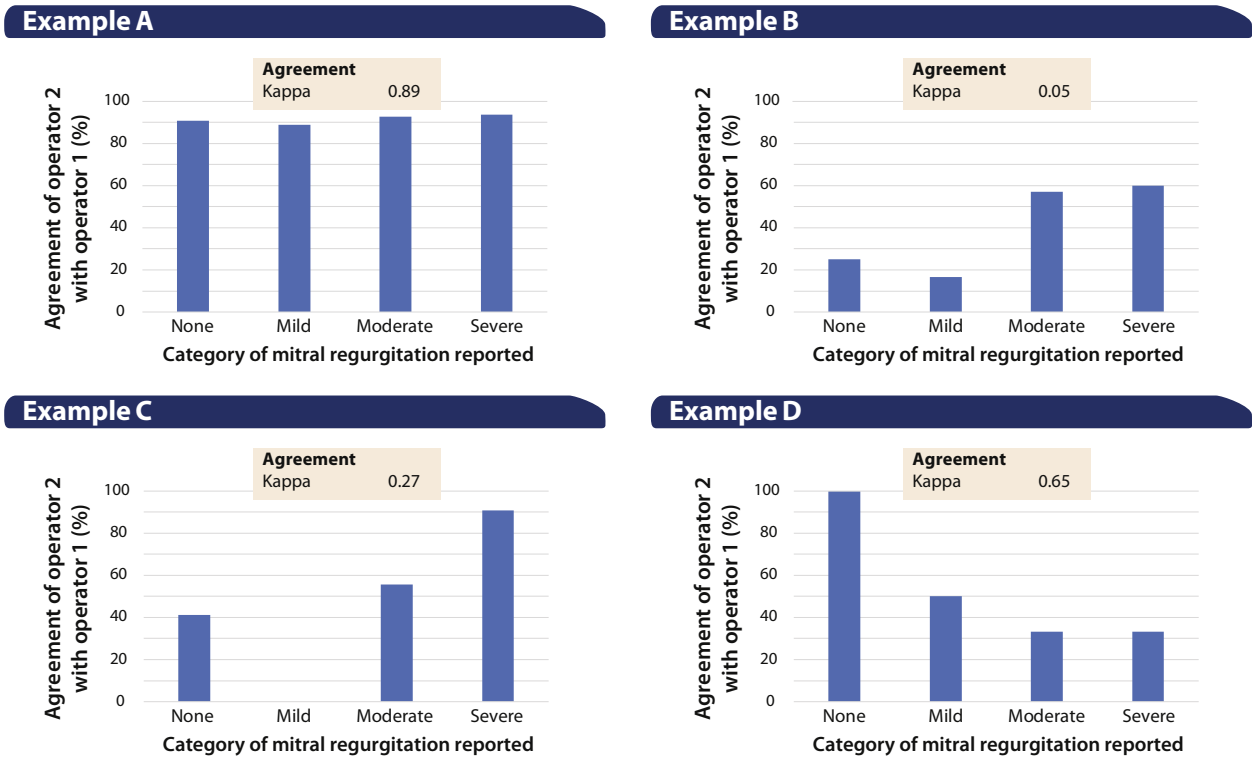

Figure 3 Cohen's κ to assess agreement of two observers for a graded result. Cohen's κ for the severity of mitral regurgitation in the same patients by two echocardiographers. **(A)** Strong agreement demonstrated across all severity categories. **(B)** Minimal agreement between the two echocardiographers. **(C)** Overall weak agreement between observers, with evidence of difference according to the severity of mitral regurgitation. **(D)** Overall moderate agreement across all categories, despite complete agreement for the absence of mitral regurgitation.

in ejection time and volume between consecutive heart beats. Loss of atrial contraction and irregular ventricular contraction lead to beat-to-beat changes in preload and hence variation in load-dependent echocardiography variables.[46] Although echocardiographers are recommended to average multiple consecutive beats in patients with atrial fibrillation, a systematic review by our group showed that isolating and averaging beats with similar cardiac cycle length (the index beat approach) could improve the overall reproducibility of measurement in atrial fibrillation.[47]

Systematic assessment of reproducibility, repeatability, and reliability is rare in routine clinical practice because of the existing demands on clinical services, the extra time required, and the lack of knowledge and tools to perform statistical evaluation. To facilitate this process, we provide a simple online calculator that offers key statistical tests and can automatically create graphical plots by entering echocardiographic data. The tool is available free of charge at www.birmingham.ac.uk/echo. Users are encouraged to provide feedback to improve this open-source tool and can elect to submit anonymous data to help build up a picture of global echocardiographic reproducibility.

In contrast to clinical practice, there is already awareness in imaging research for the need to quantify intra- and interoperator reproducibility, thereby providing some idea of generalizability to routine care. Design of research studies that formally evaluate reproducibility, repeatability, or reliability should clearly delineate what variation is specifically being assessed, with clear use of terminology to avoid confusion. To accurately measure reproducibility, these data should not be gathered as an accessory to other data but with a distinct study plan. Similar to other study outcomes, prior ascertainment of required sample size is vital so that a sufficient number of observations are obtained for quantification beyond the play of chance. Although outside the scope of this report, resources are available to help plan appropriate study size.[48,49] Other important considerations are the method of subject sampling and whether this is consecutive, random, or by convenience (with implications on statistical method and potential inclusion of bias), the degree of blinding possible, and appropriate reporting of all facets of the study.[50]

Our discussion of reproducibility, repeatability, and reliability is limited to common statistical tests that can easily be performed in the clinical environment by echocardiographers without much statistical knowledge. However, there are many other useful statistical analyses that can be performed, and we would always recommend working with a medical statistician to properly interpret results. Quantifying these values in individuals, within echocardiography departments, and across different cardiac centers has the potential to improve and enrich the clinical value of echocardiography. In the future, artificial intelligence algorithms may be able to automatically calculate and demonstrate the reproducibility of our techniques, but for the foreseeable future, we still require human

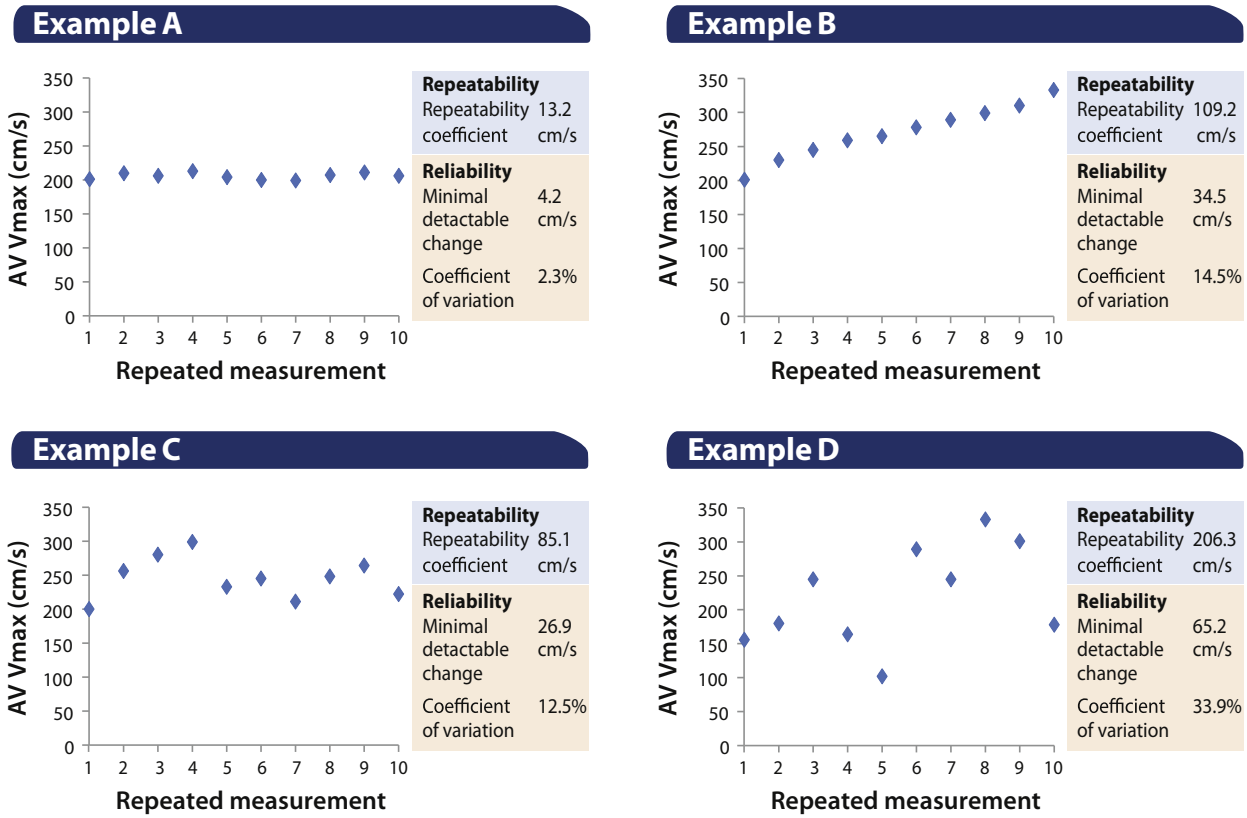

Figure 4 Repeatability and reliability assessment. Test-retest assessment for aortic valve peak velocity in the same patient with the same operator. **(A)** Repeatable and reliable measurements, with low values for repeatability coefficient and coefficient of variation, and small minimal detectable change. **(B)** Proportional bias with velocity increasing with more measurements taken, affecting both repeatability and reliability. **(C)** Example of intermediate levels of repeatability and reliability. **(D)** Clear issue with repeatability and reliability (possible equipment problem).

effort to ensure the validity of physiological assessment through echocardiography.

CONCLUSION

Echocardiography continues to grow in number, complexity, and clinical importance. Ensuring reproducible, repeatable, and reliable measurements is vital to base clinical decisions on echocardiographic parameters. More frequent and better use of appropriate (and straightforward) statistical tests has the potential to improve echocardiography at the level of individual echocardiographers and across imaging departments. An online tool for easy calculation of these statistics in the clinical environment is available at www.birmingham.ac.uk/echo.

ACKNOWLEDGMENT

We thank James Hodson (Institute of Clinical Sciences, University of Birmingham, United Kingdom).

REFERENCES

1. Steeds RP. Echocardiography: frontier imaging in cardiology. Br J Radiol 2011;84:S237-45.
2. Sonderegger-Iseli K, Burger S, Muntwyler J, Salomon F. Diagnostic errors in three medical eras: a necropsy study. Lancet 2000;355:2027-31.
3. Papolos A, Narula J, Bavishi C, Chaudhry FA, Sengupta PP. U.S. hospital use of echocardiography: insights from the Nationwide Inpatient Sample. J Am Coll Cardiol 2016;67:502-11.
4. Plana JC, Galderisi M, Barac A, Ewer MS, Ky B, Scherrer-Crosbie M, et al. Expert consensus for multimodality imaging evaluation of adult patients during and after cancer therapy: a report from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging 2014;15:1063-93.
5. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008;31:466-75.
6. Shammas HJ, Hoffer KJ. Repeatability and reproducibility of biometry and keratometry measurements using a noncontact optical low-coherence reflectometer and keratometer. Am J Ophthalmol 2012;153:55-61.e2.
7. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010;73:1167-79.
8. Galderisi M, Henein MY, D'Hooge J, Sicari R, Badano LP, Zamorano JL, et al. Recommendations of the European Association of Echocardiography: how to use echo-Doppler in clinical trials: different modalities for different purposes. Eur J Echocardiogr 2011;12:339-53.
9. Barnhart HX, Yow E, Crowley AL, Daubert MA, Rabineau D, Bigelow R, et al. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res 2016;25:2939-58.
10. Kengne C, Ouattara B, Angoran I, Anzouan C. Interobserver variability in estimation of the left ventricular systolic function of dilated cardiopathies: ejection fraction and global longitudinal strain. Arch Cardiovasc Dis 2018;10:36.

11. Sugeng L, Mor-Avi V, Weinert L, Niel J, Ebner C, Steringer-Mascherbauer R, et al. Quantitative assessment of left ventricular size and function: side-by-side comparison of real-time three-dimensional echocardiography and computed tomography with magnetic resonance reference. Circulation 2006;114:654-61.
12. Palmieri V, Dahlof B, DeQuattro V, Sharpe N, Bella JN, de Simone G, et al. Reliability of echocardiographic assessment of left ventricular structure and function: the PRESERVE study. Prospective Randomized Study Evaluating Regression of Ventricular Enlargement. J Am Coll Cardiol 1999;34:1625-32.
13. Thavendiranathan P, Grant AD, Negishi T, Plana JC, Popovic ZB, Marwick TH. Reproducibility of echocardiographic techniques for sequential assessment of left ventricular ejection fraction and volumes: application to patients undergoing cancer chemotherapy. J Am Coll Cardiol 2013;61:77-84.
14. Hensel KO, Roskopf M, Wilke L, Heusch A. Intraobserver and interobserver reproducibility of M-mode and B-mode acquired mitral annular plane systolic excursion (MAPSE) and its dependency on echocardiographic image quality in children. PLoS ONE 2018;13:e0196614.
15. King A, Thambyrajah J, Leng E, Stewart MJ. Global longitudinal strain: a useful everyday measurement? Echo Res Pract 2016;3:85-93.
16. Armstrong AC, Gjesdal O, Almeida A, Nacif M, Wu C, Bluemke DA, et al. Left ventricular mass and hypertrophy by echocardiography and cardiac magnetic resonance: the multi-ethnic study of atherosclerosis. Echocardiography 2014;31:12-20.
17. LaBounty TM, Sundaram B, Agarwal P, Armstrong WA, Kazerooni EA, Yamada E. Aortic valve area on 64-MDCT correlates with transesophageal echocardiography in aortic stenosis. AJR Am J Roentgenol 2008;191:1652-8.
18. Mavrides E, Holden D, Bland JM, Tekay A, Thilaganathan B. Intraobserver and interobserver variability of transabdominal Doppler velocimetry measurements of the fetal ductus venosus between 10 and 14 weeks of gestation. Ultrasound Obstet Gynecol 2001;17:306-10.
19. Grothues F, Smith GC, Moon JC, Bellenger NG, Collins P, Klein HU, et al. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. Am J Cardiol 2002;90:29-34.
20. Selamet Tierney ES, Levine JC, Chen S, Bradley TJ, Pearson GD, Colan SD, et al. Echocardiographic methods, quality review, and measurement accuracy in a randomized multicenter clinical trial of Marfan syndrome. J Am Soc Echocardiogr 2013;26:657-66.
21. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255-68.
22. Bland. An introduction to medical statistics. Oxford, UK: Oxford University Press; 2015.
23. Liao JJ, Capen RC, Schofield TL. Assessing the reproducibility of an analytical method. J Chromatogr Sci 2006;44:119-22.
24. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8:23-34.
25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
26. Motulsky H, Christopoulos A. Fitting models to biological data and linear and nonlinear regression. New York: Oxford University Press; 2004.
27. Popovic ZB, Thomas JD. Assessing observer variability: a user's guide. Cardiovasc Diagn Ther 2017;7:317-24.
28. van Stralen KJ, Jager KJ, Zoccali C, Dekker FW. Agreement between methods. Kidney Int 2008;74:1116-20.
29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
30. Vaz S, Falkmer T, Passmore AE, Parsons R, Andreou P. The case for using the repeatability coefficient when calculating test-retest reliability. PLoS ONE 2013;8:e73990.
31. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284-90.
32. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155-63.
33. Hedberg EC, Ayers S. The power of a paired *t*-test with a covariate. Soc Sci Res 2015;50:277-91.
34. Ludbrook J. Linear regression analysis for comparing two measurers or methods of measurement: but which regression? Clin Exp Pharmacol Physiol 2010;37:692-9.
35. Giavarina D. Understanding Bland Altman analysis. Biochemia Medica 2015;25:141-51.
36. Gwet KL. Testing the difference of correlated agreement coefficients for statistical significance. Educ Psychol Meas 2016;76:609-37.
37. Esquitin KA, Khalique OK, Liu Q, Kodali SK, Marcoff L, Nazif TM, et al. Accuracy of the single cycle length method for calculation of aortic effective orifice area in irregular heart rhythms. J Am Soc Echocardiogr 2019;32:344-50.
38. Naylor JM, Hayen A, Davidson E, Hackett D, Harris IA, Kamalasena G, et al. Minimal detectable change for mobility and patient-reported tools in people with osteoarthritis awaiting arthroplasty. BMC Musculoskelet Disord 2014;15:235.
39. Powden CJ, Hoch JM, Hoch MC. Reliability and minimal detectable change of the weight-bearing lunge test: a systematic review. Man Ther 2015;20:524-32.
40. Albert A, Zhang L. A novel definition of the multivariate coefficient of variation. Biom J 2010;52:667-75.
41. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol 2001;1:6.
42. Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 2016;32:669-79.
43. Rijsterborgh H, Mayala A, Forster T, Vletter W, van der Borden B, Sutherland GR, et al. The reproducibility of continuous wave Doppler measurements in the assessment of mitral stenosis or mitral prosthetic function: the relative contributions of heart rate, respiration, observer variability and their clinical relevance. Eur Heart J 1990;11:592-600.
44. Douglas PS, DeCara JM, Devereux RB, Duckworth S, Gardin JM, Jaber WA, 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;22:755-65.
45. Frikha Z, Girerd N, Huttin O, Courand PY, Bozec E, Olivier A, et al. Reproducibility in echocardiographic assessment of diastolic function in a population based study (the STANISLAS Cohort study). PLoS ONE 2015;10:e0122336.
46. Muntinga H, Gosselink A, Blanksma P, De Kam PJ, Wall E, Crijns H. Left ventricular beat to beat performance in atrial fibrillation: dependence on contractility, preload, and afterload. Heart 1999;82:575-80.
47. Kotecha D, Mohamed M, Shantsila E, Popescu BA, Steeds RP. Is echocardiography valid and reproducible in patients with atrial fibrillation? A systematic review. Europace 2017;19:1427-38.
48. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17:101-10.
49. Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Stat Med 2001;20:3205-14.
50. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96-106.