Critical Review of Current Approaches for Echocardiographic Reproducibility and Reliability Assessment in Clinical Research

Anna Lisa Crowley, MD, Eric Yow, MS, Huiman X. Barnhart, PhD, Melissa A. Daubert, MD, Robert Bigelow, PhD, Daniel C. Sullivan, MD, Michael Pencina, PhD, and Pamela S. Douglas, MD, *Durham, North Carolina*

Background: There is no broadly accepted standard method for assessing the quality of echocardiographic measurements in clinical research reports, despite the recognized importance of this information in assessing the quality of study results.

Methods: Twenty unique clinical studies were identified reporting echocardiographic data quality for determinations of left ventricular (LV) volumes (n = 13), ejection fraction (n = 12), mass (n = 9), outflow tract diameter (n = 3), and mitral Doppler peak early velocity (n = 4). To better understand the range of possible estimates of data quality and to compare their utility, reported reproducibility measures were tabulated, and de novo estimates were then calculated for missing measures, including intraclass correlation coefficient (ICC), 95% limits of agreement, coefficient of variation (CV), coverage probability, and total deviation index, for each variable for each study.

Results: The studies varied in approaches to reproducibility testing, sample size, and metrics assessed and values reported. Reported metrics included mean difference and its SD (n = 7 studies), ICC (n = 5), CV (n = 4), and Bland-Altman limits of agreement (n = 4). Once de novo estimates of all missing indices were determined, reasonable reproducibility targets for each were identified as those achieved by the majority of studies. These included, for LV end-diastolic volume, ICC > 0.95, CV < 7%, and coverage probability > 0.93 within 30 mL; for LV ejection fraction, ICC > 0.85, CV < 8%, and coverage probability > 0.85 within 10%; and for LV mass, ICC > 0.85, CV < 10%, and coverage probability > 0.60 within 20 g.

Conclusions: Assessment of data quality in echocardiographic clinical research is infrequent, and methods vary substantially. A first step to standardizing echocardiographic quality reporting is to standardize assessments and reporting metrics. Potential benefits include clearer communication of data quality and the identification of achievable targets to benchmark quality improvement initiatives. (J Am Soc Echocardiogr 2016;29:1144-54.)

Keyword: Reproducibility in echocardiography

Cardiac ultrasound has the potential to contribute critically important imaging end points in clinical cardiovascular research. However, its use is limited by real and perceived suboptimal measurement reproducibility. One notable example of poor data reliability is the Predictors of Response to Cardiac Resynchronization Therapy trial, which reported a large percentage of nonassessable echocardiographic data (up to 50% for Doppler tissue imaging parameters) and low interobserver reproducibility results (coefficient of variation up to 72%), making the trial findings difficult to interpret.[1]

Additionally, comparison among studies is difficult, because a variety of statistical approaches to assess variability have been reported in the literature.[2] Some studies report coefficient of variation (CV),[1,3] while others report the Pearson correlation coefficient,[3,4] percentage error,[5] or the intraclass correlation coefficient (ICC).[2,6-8] Of note, the American Society of Echocardiography Cardiovascular Technology and Research Summit developed a roadmap for 2020 that included the goals of (1) documenting the reproducibility of quantitative echocardiographic biomarkers and (2) developing reproducibility standards for echocardiography core laboratories.[9]

In response to this roadmap, the aims of the present study were to (1) determine the range of reproducibility methods in use, (2) calculate not-reported reproducibility metrics for selected variables across representative studies to compare these results, and (3) determine if these findings hold implications for establishing a standard of precision for echocardiography in clinical research.

From the Department of Medicine (A.L.C., M.A.D., P.S.D.) and the Department of Radiology (D.C.S.), Duke University, Durham, North Carolina; and the Duke Clinical Research Institute, Durham, North Carolina (A.L.C., E.Y., H.X.B., M.A.D., R.B., D.C.S., M.P., P.S.D.).

This work was supported in part by the American Society of Echocardiography (ASE) Education and Research Foundation, Award No. 12-G-10-ASE (A.L.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the ASE or the ASE Education and Research Foundation.

Reprint requests: Anna Lisa Crowley, MD, Box 2805, DUMC, Durham, NC 27705 (E-mail: annalisa.crowley@duke.edu).

0894-7317/$36.00

http://dx.doi.org/10.1016/j.echo.2016.08.006

Abbreviations
ANOVA = Analysis of variance
CP = Coverage probability
CV = Coefficient of variation
ICC = Intraclass correlation coefficient
LOA = Limits of agreement
LV = Left ventricular
TDI = Total deviation index

METHODS

Overview

A representative cohort of studies was identified to capture the range of metrics used to assess reproducibility and their corresponding results for six variables: left ventricular (LV) end-diastolic volume, LV end-systolic volume, LV ejection fraction, LV mass, LV outflow tract diameter, and mitral diastolic inflow peak early velocity. Of these, studies with sufficient reported data were selected to calculate missing but commonly reported reproducibility metrics with reasonable assumptions. On the basis of these findings, estimates of reasonably achievable targets for these metrics are proposed for each variable.

Cohort Selection

To identify a sample of studies reporting metrics for echocardiographic reproducibility, a PubMed search was performed using the following search terms: "echocardiographic reproducibility," "LV ejection fraction," "interstudy echocardiography reproducibility," "echocardiography core lab," and "echocardiographic LV mass correlates" (see Supplemental Table 1 and Figure 1). Only human studies published in English with data providing quantitative assessments of the reproducibility of continuous variables were included. In addition, a review of each study's cited references was performed to identify additional relevant studies. Finally, unpublished results from our own core laboratory reproducibility testing were included. Each study was reviewed in detail to identify the study population, types and variables of reproducibility assessed, number of echocardiograms analyzed, and reported reproducibility metrics. If other commonly used reproducibility metrics were not reported, we determined if sufficient data were provided in the articles to make reasonable assumptions to calculate the other de novo estimates of data quality. Only those studies with sufficient data for us to report all commonly used reproducibility metrics for a given variable were included in the cohort for analyses.

Statistical Analysis

Descriptive statistics including mean and SD of the overall population were extracted from the selected published results. If only the reproducibility population was reported, then the descriptive statistics for the reproducibility population were used for the overall population. We also extracted any reported mean and SD of paired differences of measurements on the same subject (and assumed that the mean difference was zero if not reported) and all data presented on reproducibility methods, including (1) ICC, (2) 95% limits of agreement (LOA), (3) CV, (4) coverage probability (CP), and (5) total deviation index (TDI). These indices were selected on the basis of a previous review of agreement indices for assessing and improving measurement reproducibility in an echocardiography core laboratory setting.[2]

Any indices not reported were calculated with specific approaches detailed in the Appendix, such that a complete range of indices were available for each variable in each study. Although the specific formula for each of the indices has been reported previously[2] and is provided in the Appendix, we provide below a brief description and interpretation of these indices.

ICC. ICC has been the most popular index used to report reliability in the medical literature. Although there are different versions of ICC depending on different assumed analysis of variance (ANOVA) models, the original ICC based on a one-way ANOVA model with subject effect is still the most commonly reported index and is the ICC used in this study. It is defined as the ratio of between-subject variability to total variability (sum of between-subject variability and error variability). For the case of two observers, the error variability is equal to half of the variance of the differences of measurements by the two observers. ICC values range from −1 to 1. Interpretation of the ICC is that the larger the ICC value, the better the reproducibility. However, the definition of adequate and inadequate reproducibility on the basis of ICC is controversial. Landis and Koch provided adjectives of "substantial" for values between 0.6 and 0.8 and of "almost perfect" for values between 0.81 and 1.0, although these cut points are arbitrary and subjective. Intuitively, if the error variability is small relative to the between-subject variability, then the ICC value will be high (close to 1).

Because of the relativeness of this index, an artificially high or low ICC value can be obtained if the between-subject variability is small or high, respectively, even though the error variability is the same. This drawback can be visualized in Figure 2 with two examples of 20 pairs of hypothetically generated LV ejection fractions. In Figure 2A, the first observer's readings were randomly generated with values from 10% to 90%. The second observer's readings were obtained by adding or subtracting 10% to the first reader's readings so that the difference between the two observers is always equal to ±10%. The ICC for these data is estimated to be very high (ICC = 0.91; 95% CI: 0.79 to 0.96). In Figure 2B, 20 pairs of hypothetical LV ejection fraction readings were also generated randomly. However, in this case, the first observer's readings were randomly generated between values of 55% and 65%. The second observer's readings were also obtained by either adding or subtracting 10 from the first observer's readings. As a result, the difference between the two readers again is always equal to ±10%. However, the ICC for the second example is estimated to be very low (ICC = 0.15; 95% CI: −0.29 to 0.54). On the basis of ICC, one would conclude that there is excellent reproducibility in the first data set but poor reproducibility in the second data set. In fact, the reproducibility is the same for both data sets, because the differences between the two readers are always equal to ±10%. The low ICC value in Figure 2B is due to the relatively small between-subject variability in the data. Even though this drawback has been recognized in the statistical literature, there is insufficient recognition of this drawback in the medical literature, and it continues to be a popular index for reporting reliability. We direct the interested reader to a previous publication[2] in which different reproducibility metrics were explored more thoroughly.

Ninety-Five Percent LOA. The 95% LOA of Altman and Bland[29] are a popular tool for examining agreement between two measurements on the same subject because of their intuitive appeal. The 95% LOA are centered on the mean difference; assuming that the differences are normally distributed, 95% of the differences would fall within these limits. Asymmetric limits imply some bias, and the magnitude of the limits indicates the magnitude of disagreement for 95% of the subjects. Smaller limits imply better reproducibility. A Bland and Altman plot, such as Figure 2A or 2B, plotting average versus difference, is used to display the data visually. For both the two hypothetical data sets in Figure 2, the estimated 95% LOA are −9.81% and 7.81%. This also implies that 95% of differences are within ±10%.

Within-Subject CV. Only the within-subject CV is used for assessing reproducibility. The CV is defined as the within-subject SD (i.e., the square root of error variability or half of the SD of differences)

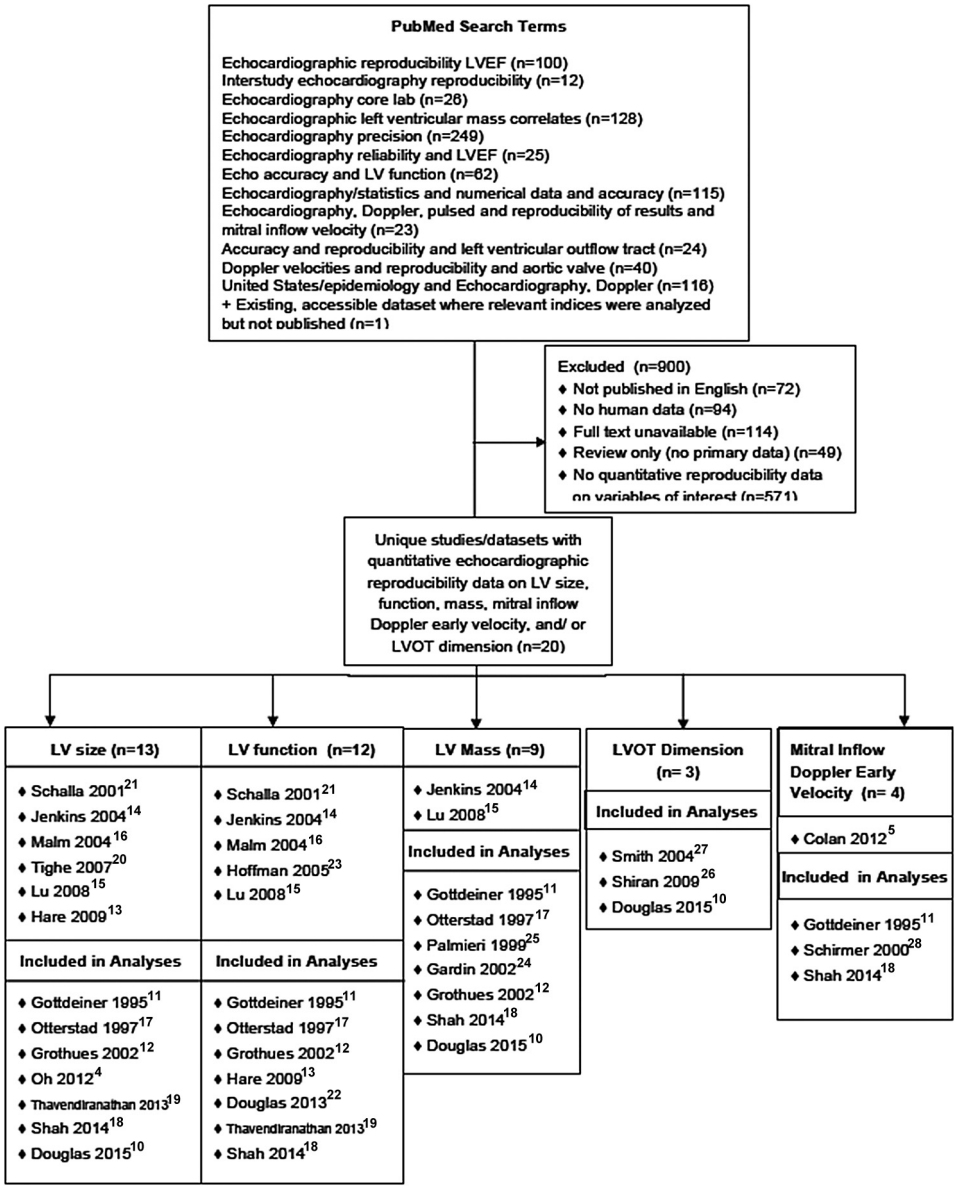

Figure 1 PubMed search strategy to identify studies with quantitative echocardiographic reproducibility data on LV end-diastolic and end-systolic volumes (n = 13),[4,10-21] LV ejection fraction (LVEF) (n = 12),[11-19,21-23] LV mass (n = 9),[10-12,14,15,17,18,24,25] LV outflow tract (LVOT) dimension (n = 3),[10,26,27] and mitral inflow Doppler early velocity (n = 4).[5,11,18,28] Of these, studies with sufficient reported data to enable further calculations were selected to calculate missing but commonly reported reproducibility metrics with reasonable assumptions.

divided by the population mean. The smaller the within-subject CV, the better the reproducibility. By definition, the CV depends on the population mean, and its value could be artificially small or large for a study with a large or small population mean, even though the error variability is the same. For the two generated data sets in Figure 2, the estimated within-subject SD is equal to 2.25% because of the same differences of the two observes, but the estimated population means are 43.3% and 59.5%, respectively, resulting in a larger within-subject CV (0.052 or 5.3%) for the first data set than the within-subject CV (0.038 or 3.8%) for the second data set.

CP and TDI. CP and TDI are two agreement indices, with equivalent concepts, to measure the proportion of cases within a boundary for allowed differences. Intuitively speaking, a reasonable criterion to judge whether reproducibility is satisfactory would be to require that an overwhelming majority (e.g., 95% or 80%) of the absolute differences be within a preset acceptable difference (i.e., prespecified acceptable measurement error). The probability of the observed absolute differences falling within the acceptable difference is the CP for the given acceptable difference. The higher the CP, the better the reproducibility. Estimation of the CP can be accomplished simply by using the proportion of paired absolute differences falling within limits that are considered acceptable. In the hypothetical data in Figure 2, the CP for the acceptable difference of 10 is equal to 100% because the paired absolute differences are all equal to 10. If data are available, a CP curve can be constructed by plotting the observed absolute differences versus the corresponding estimated CPs connected with lines. One can then choose an acceptable difference and find the corresponding estimated CP in the curve.

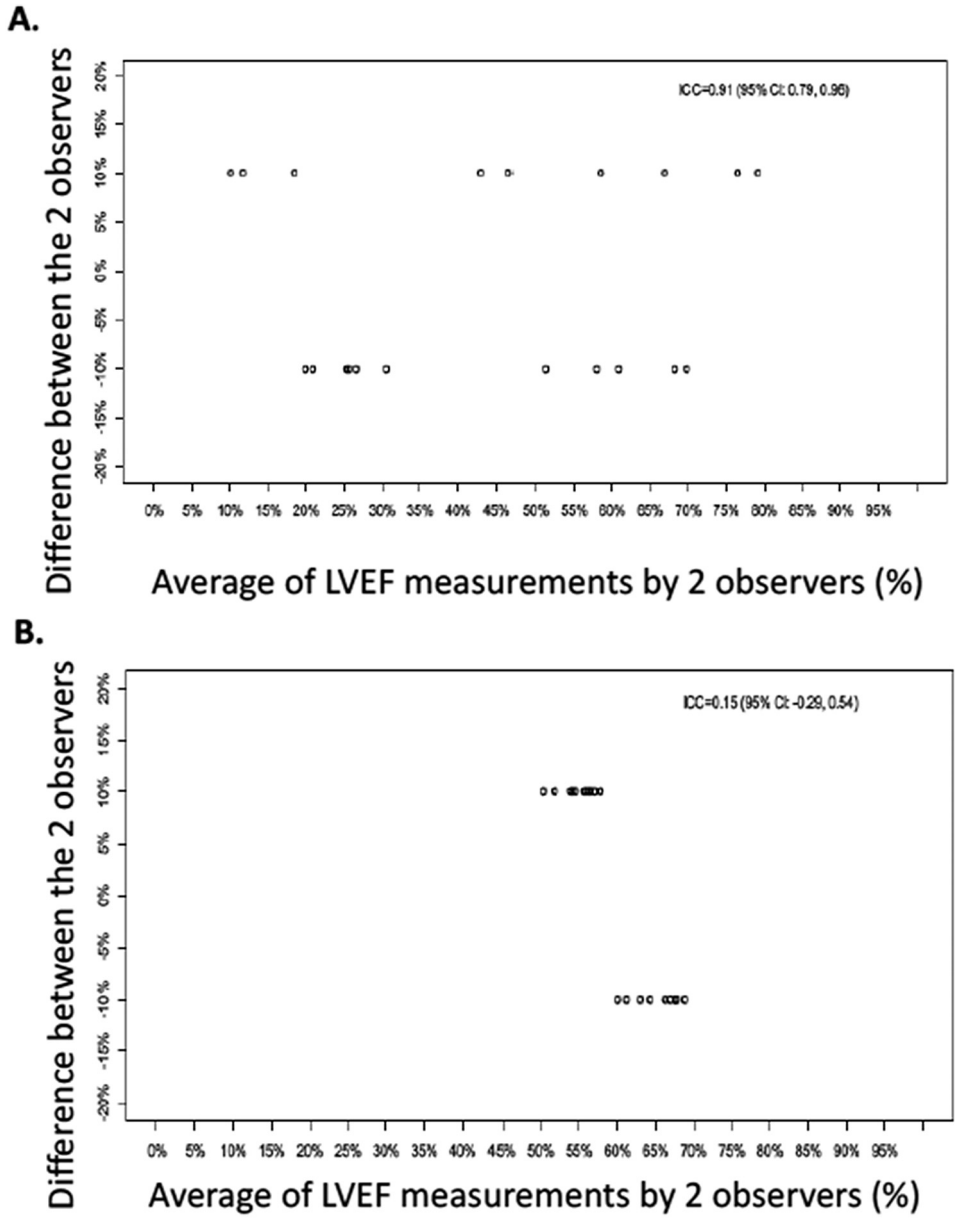

Figure 2 **(A)** The difference in LV ejection fraction (LVEF) measurements between 20 pairs of randomly generated LVEF measurements is displayed on the y axis, and the average between the two LVEF measurement is displayed on the x axis. The first measurement has values ranging from 10% to 90%. The second measurement is always equal to ±10 of the first measurement. The ICC in this example is 0.91. **(B)** The difference in LVEF measurements between 20 pairs of randomly generated LVEF measurements is displayed on the y axis, and the average between the two LVEF measurement is displayed on the x axis. The first measurement has values ranging from 55% to 65%. The second measurement is always equal to ±10 of the first measurement. The ICC in this example is 0.15. However, the data in both **(A)** and **(B)** represent the same the reproducibility, because the differences between the two readers are always equal to ±10%. The low ICC value in **(B)** is due to the relatively small between-subject variability in the data.

Therefore, the CP curve provides a visual spectrum of measurement error for different acceptable differences. For the hypothetical data in Figure 2, the estimated CP curve is the straight line connecting two points: (0, 0) and (10%, 100%). If data are not available, the CP curve can be constructed, assuming that the differences are normally distributed, by using reported or imputed mean and SD. This latter approach is used for the computed CP curves presented in the "Results" section for the chosen studies, because the original data were not available. Comparisons of CP curves would be available in the future by using relative area under the curve.[30]

Sometimes it may be difficult to prespecify the acceptable difference. If we believe that the observers are making quality measurements, and they represent the best achievable measurements for a given parameter, then it is of interest to know the range within which an overwhelming majority (e.g., 95% or 80%) of the observed absolute differences would be expected to fall. This expected absolute difference is called TDI (for a given probability as the measure of majority). We used 80% and 95% as the given probability in this study. In Figure 2, the estimated TDI would 10 for both cases.

Table 1 Reported and calculated measures of echocardiographic reproducibility of LV end-diastolic volume measurements

Data source Population studied; Reproducibility type[reference]	Overall population		Reported and calculated results for reproducibility agreement indices						
	n	Mean mL (SD)	n	Mean difference mL (SD)	ICC	95% LOA	CV (%)	CP*	TDI† 80%, 95%
A. Hypertension; intraobserver/interstudy[11]	**81**	**132.5 (40.0)**	**81**	0.0 (21.2)	**0.86**	−41.5, 41.5	11.3	0.844	27.1, 41.5
B. Myocardial infarction and normal; interobserver[17]	**24**	**184.8 (45.6)**	**24**	0.0 (12.8)	0.96	−25.0, 25.0	**4.9**	0.981	16.4, 25.0
C. Mixed LV hypertrophy, heart failure, and normal volunteers; intraobserver/interstudy[12]	**60**	**180.0 (65.0)**	**60**	**0.9 (13.5)**	0.98	−25.6, 27.4	5.3	0.973	17.3, 26.5
D. Heart failure; interobserver[4]	**1,460**	**222.4 (68.8)**	**67**	**8.9 (24.8)**	**0.94**	−39.7, 57.5	7.9	0.744	33.8, 51.6
E. Cancer; interobserver[19]	**56**	**88.0 (25.0)**	**56**	0.0 (23.3)	0.56	−45.7, 45.7	18.8	0.988	29.9, 45.7
F. Elderly community population; intraobserver[18]	**40**	**95.7 (39.0)**	**20**	**−4.3 (7.8)**	**0.98**	−19.6, 11.0	5.8	0.801	11.5, 17.4
G. Mixed normal volunteers, heart failure, aortic stenosis; interobserver[10]	10	198.3 (61.3)	10	0.31 (17.1)	0.96	−28.3, 28.9	6.1	1.000	22.9, 38.0

Reported data are in boldface type, and extrapolated data are in regular type. General guidelines for statistical test interpretation are as follows: the ICC ranges between −1 and 1, with higher values indicating better reproducibility; the CV ranges from 0 to infinity, with smaller values indicating better reproducibility; the CP ranges from 0 to 1, with higher values indicating better reproducibility; and the TDI ranges from 0 to infinity, with smaller values indicating better reproducibility.

*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 30 mL.

†The TDI is the absolute paired difference with the desired CP of 80% and 95%.

Barnhart *et al.*,[2] who compared the pros and cons of various reproducibility metrics, concluded that the CP is the preferred index for assessing reproducibility. This conclusion was based on the computational simplicity of this index, its ability to identify discordant measurements to provide guidance for review and retraining, and its consistent evaluation of data quality across multiple reviewers, populations, and continuous as well as categorical data.

All statistical analyses were performed using SAS version 9.4 (SAS Institute, Cary, NC).

RESULTS

Cohort Selection

Figure 1 details the search strategy used to identify 20 unique studies reporting reproducibility data for LV end-diastolic volume, LV end-systolic volume, LV ejection fraction, LV mass, LV outflow tract diameter, and mitral diastolic inflow peak early velocity. Individual study details regarding reproducibility types, number and type of observer, sample size, and statistical indices used are listed in Supplemental Table 1.

Interestingly, there were a variety of different approaches to reproducibility assessments (interacquisition [$n = 1$], interobserver only [$n = 1$], both interobserver and intraobserver [$n = 6$], intersite variability [$n = 1$], intrasubject-interstudy variability [$n = 5$], and temporal drift [$n = 2$]), with different numbers of subject echocardiograms reviewed for reproducibility testing, ranging from 10 to 83 (or ranging from <1% to 100% of the overall study sample size), and different types of observers (sonographers, physicians, core laboratory, and site) (Supplemental Table 1). Reported metrics included mean difference and its SD ($n = 7$ articles), ICC ($n = 5$), CV ($n = 4$), and LOA ($n = 4$). Finally, none of the 20 studies cited a benchmark to allow the reader to interpret their findings regarding the quality of the reproducibility.

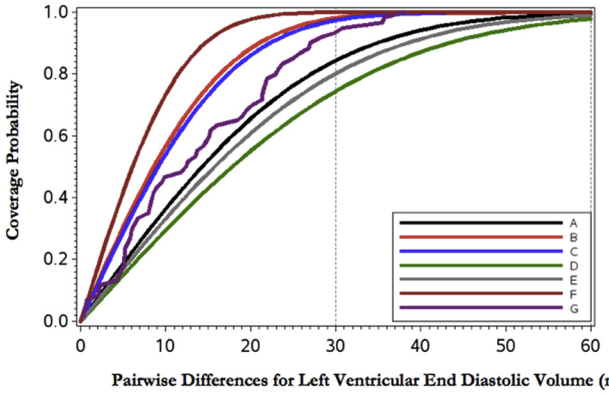

Figure 3 A comparative graphical display of the reproducibility data for echocardiographically determined LV end-diastolic volume is shown for the studies identified in Table 1.[4,10-12,17-19] The computed CP (in percentage on the y axis) curves for a range of pairwise differences (in milliliters on the x axis) for echocardiography-determined LV end-diastolic volume is based on an assumed normal distribution of the difference (except for study G, for which the curve was estimated from the data). If an acceptable pairwise difference of 60 mL is selected, then all seven studies achieved a CP of almost 100%. However, if an acceptable pairwise difference of 10 mL is selected, then the CPs ranged from about 25% to about 75%. The dotted vertical line represents an acceptable pairwise difference of 30 mL, a value for which the majority of studies achieved a CP of >80%.

Comparison of Reproducibility Results

Of the 20 unique studies identified in Figure 1, sufficient quantitative echocardiographic data on reproducibility were available for us to compute the missing reproducibility metrics on LV volumes ($n = 6$), LV ejection fraction ($n = 7$), LV mass ($n = 7$), LV outflow

Table 2 Reported and calculated measures of echocardiographic reproducibility of LV end-systolic volume measurements

Data source Population studied; reproducibility type[reference]	Overall population		Reported and calculated results for reproducibility agreement indices						
	n	Mean mL (SD)	*n*	Mean difference mL (SD)	ICC	95% LOA	CV (%)	CP*	TDI† 80%, 95%
A. Hypertension; intraobserver/interstudy[11]	**75**	**31.3 (15.8)**	**75**	0.0 (11.6)	**0.73**	−22.7, 22.7	26.2	0.990	14.8, 22.7
B. Myocardial infarction and normal; interobserver[17]	**24**	**100.4 (46.7)**	**24**	0.0 (4.5)	1.00	−8.9, 8.9	**3.2**	1.000	5.8, 8.9
C. Mixed LV hypertrophy, heart failure, and normal volunteers; intraobserver/interstudy[12]	**60**	**87.0 (70.0)**	**60**	**0.9 (14.0)**	0.98	−26.5, 28.3	11.4	0.968	18.0, 27.5
D. Heart failure; interobserver[4]	**1,460**	**160.7 (60.4)**	**67**	**7.6 (23.3)**	**0.93**	−38.1, 53.3	10.3	0.779	31.5, 48.0
E. Cancer; interobserver[19]	**56**	**37.0 (17.0)**	56	0.0 (10.5)	0.81	−20.5, 20.5	20.0	0.996	13.4, 20.5
F. Elderly community population; intraobserver[18]	**40**	**41.4 (36.8)**	**20**	**0.4 (5.2)**	**0.99**	−9.8, 10.6	8.9	1.000	6.7, 10.2
G. Mixed normal volunteers, heart failure, aortic stenosis; interobserver[10]	10	133.4 (62.6)	10	2.6 (12.2)	0.98	−19.7, 24.8	6.5	1.000	13.7, 21.2

Reported data are in boldface type, and extrapolated data are in regular type. General guidelines for statistical test interpretation are as follows: the ICC ranges between −1 and 1, with higher values indicating better reproducibility; the CV ranges from 0 to infinity, with smaller values indicating better reproducibility; the CP ranges from 0 to 1, with higher values indicating better reproducibility; and the TDI ranges from 0 to infinity, with smaller values indicating better reproducibility.

*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 30 mL.

†The TDI is the absolute paired difference with the desired CP of 80% and 95%.

Table 3 Reported and calculated measures of echocardiographic reproducibility of LV ejection fraction

Data source Population studied; reproducibility type[reference]	Overall population		Reported and calculated results for reproducibility agreement indices						
	n	Mean % (SD)	*n*	Mean difference % (SD)	ICC	% LOA	CV (%)	CP*	TDI† 80%, 95%
A. Hypertension; intraobserver/interstudy[11]	**78**	**77.3 (7.8)**	**78**	0.0 (3.3)	**0.91**	−6.4, 6.4	3.0	0.998	4.2, 6.5
B. Myocardial infarction and normal; interobserver[17]	**24**	**47.3 (11.2)**	**24**	0.0 (3.8)	0.94	−7.4, 7.4	**5.7**	0.992	4.9, 7.4
C. Mixed LV hypertrophy, heart failure, and normal volunteers; intraobserver/interstudy[12]	**60**	**57.0 (20.0)**	**60**	**−0.3 (6.1)**	0.95	−12.3, 11.7	7.6	0.898	7.8, 12.0
D. Oncology; interobserver[13]	**35**	**59.6 (7.3)**	**35**	**−0.2 (9.5)**	**0.79**	−18.8, 18.4	11.3	0.707	12.2, 18.6
E. Aortic stenosis; interobserver[22]	**592**	**52.6 (13.6)**	**30**	0.0 (6.4)	**0.89**	−12.5, 12.5	8.6	0.883	8.2, 12.5
F. Cancer; interobserver[19]	**56**	**61.0 (6.0)**	**56**	0.0 (5.7)	0.56	−11.1, 11.1	6.6	0.923	7.3, 11.1
G. Elderly community population; intraobserver[18]	**40**	**59.5 (13.9)**	**20**	**−2.3 (5.2)**	**0.93**	−12.5, 7.9	6.2	0.922	7.3, 11.1

Reported data are in boldface type, and extrapolated data are in regular type. General guidelines for statistical test interpretation are as follows: the ICC ranges between −1 and 1, with higher values indicating better reproducibility; the CV ranges from 0 to infinity, with smaller values indicating better reproducibility; the CP ranges from 0 to 1, with higher values indicating better reproducibility; and the TDI ranges from 0 to infinity, with smaller values indicating better reproducibility.

*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 10%.

†The TDI is the absolute paired difference with the desired CP of 80% and 95%.

tract diameter ($n = 3$), and mitral Doppler peak early velocity ($n = 3$).

The range of reproducibility values reported or computed for LV end-diastolic volume is shown in Table 1. Reported values are in boldface type, while computed values based on reported data are in regular type. We also reported unpublished results (in regular type) on the basis of the actual data from the Duke Clinical Research Institute Imaging Core Laboratory. The latter data are included because they are readily available, represent an in-depth assessment of intraobserver and interobserver reproducibility, and, unlike published studies, provide a complete data set that allows the calculation of all indices without any assumptions or extrapolations. For example,

Table 4 Reported and calculated measures of echocardiographic reproducibility of LV mass

Data source Population studied; reproducibility type[reference]	Overall population		Reported and calculated results for reproducibility agreement indices						
	n	Mean g (SD)	*n*	Mean difference g (SD)	ICC	95% ICC	CV (%)	CP*	TDI† 80%, 95%
A. Hypertension; intraobserver/interstudy[11]	**74**	**321.5 (79.5)**	**74**	0.0 (42.3)	**0.86**	−83.0, 83.0	9.3	0.363	54.3, 83.0
B. Myocardial infarction and normal; interobserver[17]	**24**	**201.8 (44.7)**	**24**	0.0 (23.7)	0.86	−46.5, 46.5	**8.3**	0.600	30.4, 46.5
C. LV hypertrophy; intraobserver/interstudy[25]	**366**	**242.0 (53.5)**	**183**	**−1.7 (19.8)**	**0.93**	−40.5, 37.1	5.8	0.686	25.5, 39.0
D. Young adults; interobserver[24]	**1,189**	**138.5 (38.6)**	**NR**	0.0 (20.8)	0.85	−40.7, 40.7	**10.6**	0.664	26.6, 40.7
E. Mixed LV hypertrophy, heart failure, and normal volunteers; intraobserver/interstudy[12]	**60**	**195.0 (51.0)**	**60**	**8.7 (25.0)**	0.88	−40.3, 57.7	9.1	0.549	34.0, 51.8
F. Elderly community population; intraobserver[18]	**40**	**156.7 (82.5)**	**20**	**−3.2 (20.2)**	**0.97**	−42.8, 36.4	9.1	0.672	26.2, 40.1
G. Mixed normal volunteers, heart failure, aortic stenosis, interobserver[10]	10	257.1 (46.0)	10	−2.27 (24.9)	0.85	−45.1, 40.6	6.8	0.621	33.2, 53.3

NR, Not reported.
Reported data are in boldface type, and extrapolated data are in regular type. General guidelines for statistical test interpretation are as follows: the ICC ranges between −1 and 1, with higher values indicating better reproducibility; the CV ranges from 0 to infinity, with smaller values indicating better reproducibility; the CP ranges from 0 to 1, with higher values indicating better reproducibility; and the TDI ranges from 0 to infinity, with smaller values indicating better reproducibility.
*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 20 g.
†The TDI is the absolute paired difference with the desired CP of 80% and 95%.

Table 5 Reported and calculated measures of echocardiographic reproducibility of LV outflow tract dimension

Data source Population studied; reproducibility type[reference]	Overall population		Reported and calculated results for reproducibility agreement indices						
	n	Mean cm (SD)	*n*	Mean difference cm (SD)	ICC	95% LOA	CV (%)	CP*	TDI† 80%, 95%
A. Aortic stenosis; interobserver[27]	**20**	**2.1 (0.2)**	**20**	**0.0 (0.1)**	0.83	−0.2, 0.2	3.4	0.953	0.1, 0.2
B. Mixed aortic stenosis and normal volunteers; interobserver[26]	**50**	**2.1 (0.2)**	**50**	**0.1 (0.2)**	0.74	−0.2, 0.4	5.0	0.706	0.3, 0.4
C. Aortic stenosis; interobserver[10]	10	2.1 (0.1)	10	0.0 (0.1)	0.62	−0.2, 0.2	3.0	0.967	0.1, 0.2

Reported data are in boldface type, and extrapolated data are in regular type. General guidelines for statistical test interpretation are as follows: the ICC ranges between −1 and 1, with higher values indicating better reproducibility; the CV ranges from 0 to infinity, with smaller values indicating better reproducibility; the CP ranges from 0 to 1, with higher values indicating better reproducibility; and the TDI ranges from 0 to infinity, with smaller values indicating better reproducibility.
*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 0.2 cm.
†The TDI is the absolute paired difference with the desired CP of 80% and 95%.

with the complete data from Duke, the actual CP curves are displayed rather than the extrapolated, smoothed versions due to best-case scenario assumptions we made from the available data in the published articles. Mean difference ranged from −4.3 to 8.9 mL, while the SD of differences ranged from 7.8 to 24.8 mL and ICC values ranged from 0.56 to 0.98. The CV ranged from 4.88% to 18.8%, while the CP (for an acceptable paired difference of 30 mL) ranged from 0.74 to 1.0.

Figure 3 displays the computed CP curves for a range of pairwise differences for echocardiographically determined LV end-diastolic volume. If an acceptable pairwise difference of 60 mL is selected, then all six studies achieved a CP of almost 100%. However, if an acceptable difference of 10 mL is selected, then the CPs ranged from about 25% to about 75%. The dotted vertical line represents an acceptable pairwise difference of 30 mL, a value for which the majority of studies achieved a CP of >80%.

Similarly, the range of reproducibility values reported or computed for echocardiographically determined LV end-systolic volume, LV ejection fraction, LV mass, LV outflow tract diameter, and mitral valve Doppler peak early diastolic velocity are shown in Tables 2–6, respectively. Reported values are in boldface type, while computed values are not.

Correspondingly, Figures 4–8 display the computed CP curves (y axis) for a range of pairwise differences (x axis) for echocardiographically derived LV end-systolic volume, LV ejection fraction, LV mass, LV outflow tract diameter, and mitral valve Doppler peak early diastolic velocity,

Table 6 Reported and calculated measures of echocardiographic reproducibility of mitral Doppler peak early diastolic velocity

Data source Population studied; reproducibility type[reference]	Overall population		Reported and calculated results for reproducibility agreement indices						
	n	Mean cm/sec (SD)	*n*	Mean difference cm/sec (SD)	ICC	95% LOA	CV (%)	CP*	TDI† 80%, 95%
A. Hypertension; intraobserver/ interstudy[11]	**88**	**53.4 (15.1)**	**88**	0.0 (12.9)	**0.64**	−25.3, 25.3	17.1	0.302	16.5, 25.3
B. General population; interobserver[28]	**3,022**	**68.1 (15.7)**	**58**	**0.0 (7.8)**	0.88	−15.3, 15.3	8.1	0.478	10.0, 15.3
C. Elderly community population; intraobserver[18]	**40**	**67.0 (11.0)**	**20**	**2.1 (2.2)**	**0.98**	−2.2, 6.4	2.3	0.906	4.0, 5.7

Reported data are in boldface type, and extrapolated data are in regular type. General guidelines for statistical test interpretation are as follows: the ICC ranges between −1 and 1, with higher values indicating better reproducibility; the CV ranges from 0 to infinity, with smaller values indicating better reproducibility; the CP ranges from 0 to 1, with higher values indicating better reproducibility; and the TDI ranges from 0 to infinity, with smaller values indicating better reproducibility.

*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 5 cm/sec.

†The TDI is the absolute paired difference with the desired CP of 80% and 95%.

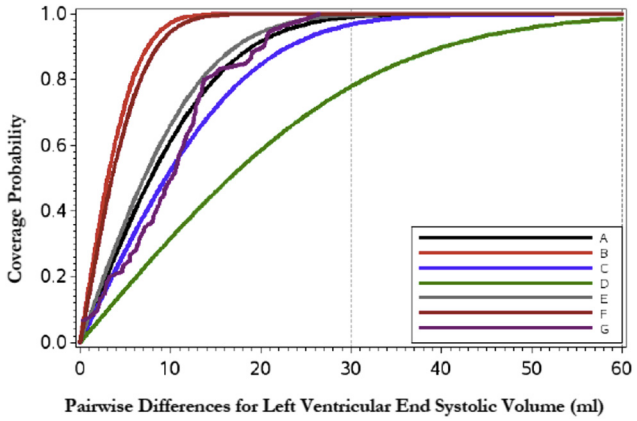

Figure 4 A comparative graphical display of the reproducibility data for echocardiographically determined LV end-systolic volume is shown for the studies identified in Table 2.[4,10-12,17-19] The computed CP (in percentage on the y axis) for a range of pairwise differences (in milliliters on the x axis) is shown for echocardiography-determined LV end-systolic volume. If an acceptable pairwise difference of 60 mL is selected, then all seven studies achieved a CP of almost 100%. However, if an acceptable pairwise difference of 10 mL is selected, then the CP ranged from about 25% to >90%. Similarly, a dotted vertical line represents an acceptable pairwise difference of 30 mL, a value for which the majority of studies achieved a CP of >80%.

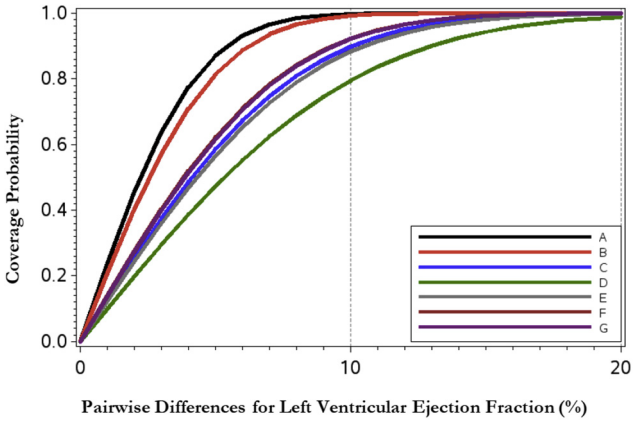

Figure 5 A comparative graphical display of the reproducibility data for echocardiographically derived LV ejection fraction is shown for the studies identified in Table 3.[11-13,17-19,22] The computed CP (in percentage on the y axis) for a range of pairwise differences (in percentage on the x axis) is shown for echocardiography-determined LV ejection fraction. If an acceptable pairwise difference of 20% is selected, then all seven studies achieved a CP of >90%. However, if an acceptable pairwise difference of 5% is selected, then the CP ranged 47% to 87%. A dotted vertical line represents an acceptable pairwise difference of 10%, a value for which the majority of studies achieved a CP of >80%.

respectively. A dotted vertical line represents an acceptable pairwise difference value for which the majority of studies achieved a CP of >80% for LV end-diastolic volume, LV end-systolic volume, LV ejection fraction, and LV outflow tract diameter, >50% for LV mass, and >45% for echocardiography-derived mitral valve Doppler peak early diastolic velocity.

Proposed Achievable Targets

Review of the reported and calculated data revealed precision metrics that were achieved by a majority of studies (Table 7). The best precision was achieved for LV end-diastolic volume (ICC > 0.95, CP > 93% for an acceptable pairwise difference of 30 mL), and the worst was achieved for mitral valve Doppler peak early diastolic velocity (ICC > 0.8, CP > 45% for an acceptable pairwise difference of 5 cm/sec) (Table 7).

DISCUSSION

We performed a critical review of current approaches for echocardiographic reproducibility and reliability assessment in clinical research. Our study has three important findings. First, data quality in echocardiographic clinical research was infrequently reported, and the statistical methods used to assess it varied substantially. Second, the range of reported values for each metric was large for each of the parameters examined, and even within a given study, the relative data quality varied depending on the metric examined.

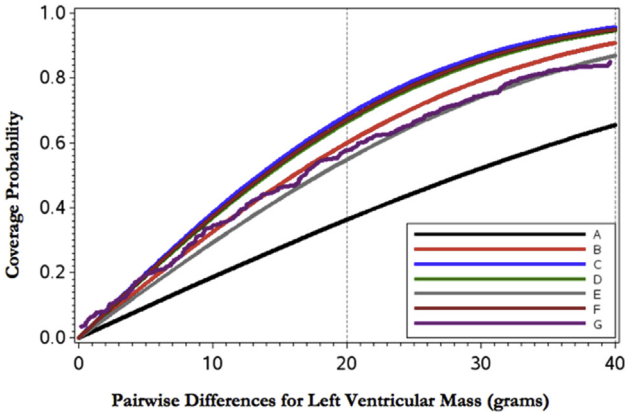

Figure 6 A comparative graphical display of the reproducibility data for echocardiographically derived LV mass is shown for the studies identified in Table 4.[10-12,17,18,24,25] The computed CP (in percentage on the y axis) for a range of pairwise differences (in grams on the x axis) is shown for echocardiography-derived LV mass. If an acceptable pairwise difference of 40 g is selected, then all seven studies achieved a CP of >60%. However, if an acceptable pairwise difference of 5 g is selected, then the CP ranged from 9% to 20%. A dotted vertical line represents an acceptable pairwise difference of 20 g, a value for which the majority of studies achieved a CP of >60%.

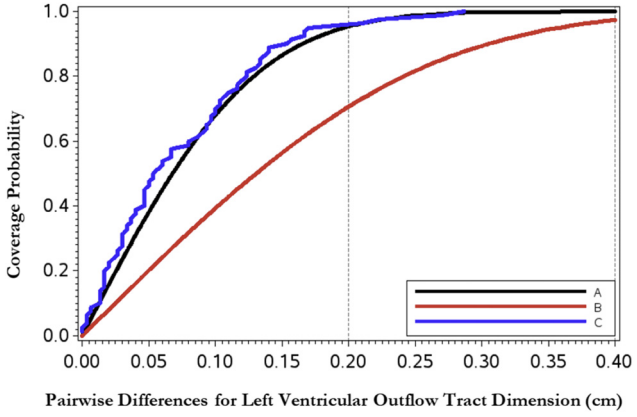

Figure 7 A comparative graphical display of the reproducibility data for echocardiographically derived LV outflow tract diameter is shown for the studies identified in Table 5.[10,26,27] The computed CP (in percentage on the y axis) for a range of pairwise differences (in centimeters on the x axis) is shown for echocardiography-derived LV outflow tract diameter. If an acceptable pairwise difference of 0.4 cm is selected, then all three studies achieved a CP of almost 100%. However, if an acceptable pairwise difference of 0.05 cm is selected, then the CP ranged from about 20% to 40%. A dotted vertical line represents an acceptable pairwise difference of 0.2 cm, a value for which the majority of studies achieved a CP of >80%.

Third, achievable precision metrics for the studies identified were listed for each variable.

Our literature review showed that 571 of 920 articles in echocardiographic clinical research did not report reproducibility metrics, and that when they were used, the statistical methods used to assess data quality varied substantially. In our sample, five separate approaches to reproducibility were assessed (intraobserver, interobserver, interstudy, interacquisition, and temporal drift). Furthermore, the sample

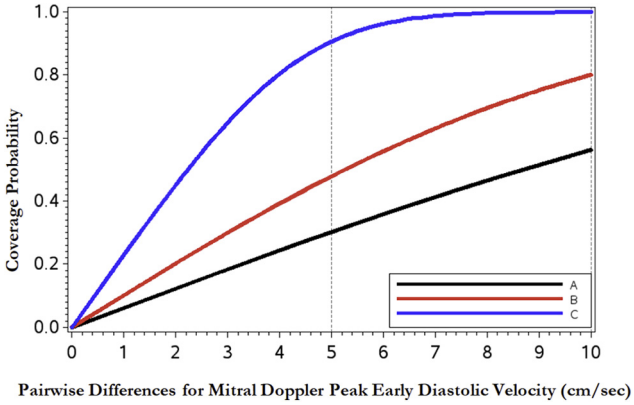

Figure 8 A comparative graphical display of the reproducibility data for echocardiographically derived mitral valve Doppler peak early diastolic velocity is shown for the studies identified in Table 6.[11,18,28] The computed CP (in percentage on the y axis) for a range of pairwise differences (in centimeters per second on the x axis) is shown for echocardiography derived mitral valve Doppler peak early diastolic velocity. If an acceptable pairwise difference of 10 cm/sec is selected, then all three studies achieved a CP of >50%. However, if an acceptable pairwise difference of 2 cm/sec is selected, then the CP ranged from <12% to 45%. A dotted vertical line represents an acceptable pairwise difference of 5 cm/sec, a value for which the majority of studies achieved a CP of >45%.

Table 7 Precision metrics achieved by a majority of studies

	ICC	CV	CP
LV end-diastolic volume	>0.95	<7	>93%*
LV end-systolic volume	>0.90	<12	>90%*
LV ejection fraction	>0.85	<8	>85%†
LV mass	≥0.85	<10	>60%‡
LV outflow tract diameter	>0.80	<5	>87%§
Mitral Doppler E velocity	>0.80	<12	>45%‖

*The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 30 mL.
†The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 10%.
‡The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 20 g.
§The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 0.2 cm.
‖The CP is the proportion of subjects or values that fall within the preset acceptable paired absolute difference of 5 cm/sec.

size was not always related to the overall cohort size and varied from <1% up to 100% of the overall cohort studied. Of note, the sample used to evaluate reproducibility should increase in size commensurate with increases in the expected variation. Last, ≥10 different metrics for reproducibility have been reported in the literature. Of note, some indices may be equivalent but may be named differently, for example, percentage error and CV. Consensus as to which types of reproducibility should be assessed, sample sizes needed, and metrics that should be reported would lead to a more consistent, uniform reproducibility process that could potentially improve data quality and therefore the utility of echocardiographic measurements. The first step toward this direction was a recent study that examined the

pros and cons of different agreement and reproducibility indices for assessing reproducibility in a core laboratory setting.[2] The investigators found that the CP is a preferred method because it is intuitive, computationally simple, and consistent across different settings.

Additionally, we found that the range of reproducibility metric values was large for each variable. The degree of variability in reproducibility metric values did not appear to be uniformly associated with the either the presence or absence of pathology. Although part of this variability is attributable to nonuniform processes, part is also due to the lack of widely accepted reproducibility standards for echocardiographic variables. In this setting, it is difficult to judge whether a study has good or acceptable levels of reproducibility if different indices were used for reporting on the same variable. By calculating multiple indices of reproducibility for each study, it became apparent that using different indices may lead to different conclusions regarding reproducibility. For example, a high ICC suggesting high reproducibility may be accompanied by a CP suggesting low reproducibility (Table 1, study D). Our findings suggest that reporting a panel of indices could more completely represent the echocardiographic reproducibility based on different indices for a given variable and study. Furthermore, by including CP metrics, actionable feedback can be obtained to target retraining efforts where indicated.[31]

Finally, by analyzing both reported and imputed metrics, we identified precision metrics that were achieved by a majority of studies for each variable. Although determining the standards for the echocardiography community is far beyond the scope of this report, we believe that these achievable targets may represent a starting point to discuss and develop standards. Consensus among echocardiographic investigators as to the preferred method and minimum acceptable values for reproducibility would enhance study comparability and provide benchmarked targets for quality improvement.

Limitations

Although this study has some important findings, some limitations exist. First, this study is a cross-section of available studies and not a comprehensive systematic review. As such, not all studies performed similar reproducibility assessments. Comparison among studies was limited to available data and restricted to intraobserver and interobserver assessments except as noted. Of note, the studies included in our analysis were performed or analyzed at highly reputable echocardiography laboratories. As a result, to the best of our knowledge, our analysis reflects a cross section of best-case practices.

Second, to calculate all indices in all studies, data and distributional assumptions were occasionally made in order to obtain approximations of unreported index values. For example, no mean differences between observers were assumed if these were not reported, in order to generate the least possible amount of variability, and to achieve a best-case scenario. These results are in normal (not boldface) type in the tables. Although these computed numbers are not directly estimated from the observed study data, they are reasonable approximations and allow us to compare reproducibility results across studies on the basis of different reproducibility indices. The ability to compare reproducibility results on the basis of different indices also allows a comprehensive evaluation of data quality in each study, an important advantage.

Third, we did not examine the relationship between the reproducibility of measurements and the future replicability of overall study results. Future studies will need to address this important point.

CONCLUSIONS

In conclusion, our study found that the statistical metrics used to assess data quality in echocardiographic clinical research vary substantially and are often not reported at all. Additionally, the range of reported values for each metric is large for each of the measurements examined. Even within a given study, the relative data quality varies depending on the metric examined. Despite these findings, we identified reasonable precision metrics that a majority of studies were able to achieve for the echocardiographic parameters examined and that most likely represent reasonable targets for echocardiography. These results are the first step in informing future larger scale studies and the eventual determination of a universal standard for echocardiographic reproducibility. Improving both the perception and reality of suboptimal echocardiographic measurement variability will benefit patients, clinical trialists, and echocardiography investigators.

REFERENCES

1. Chung ES, Leon AR, Tavazzi L, Sun JP, Nihoyannopoulos P, Merlino J, et al. Results of the Predictors of Response to CRT (PROSPECT) trial. Circulation 2008;117:2608-16.
2. 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 2014; http://dx.doi.org/10.1177/0962280214534651.
3. Foster E, Wasserman HS, Gray W, Homma S, Di Tullio MR, Rodriguez L, et al. Quantitative assessment of severity of mitral regurgitation by serial echocardiography in a multicenter clinical trial of percutaneous mitral valve repair. Am J Cardiol 2007;100:1577-83.
4. Oh JK, Pellikka PA, Panza JA, Biernat J, Attisano T, Manahan BG, et al. Core lab analysis of baseline echocardiographic studies in the STICH trial and recommendation for use of echocardiography in future clinical trials. J Am Soc Echocardiogr 2012;25:327-36.
5. Colan SD, Shirali G, Margossian R, Gallagher D, Altmann K, Canter C, et al. The Ventricular Volume Variability Study of the Pediatric Heart Network: study design and impact of beat averaging and variable type on the reproducibility of echocardiographic measurements in children with chronic dilated cardiomyopathy. J Am Soc Echocardiogr 2012;25:842-54.
6. Galderisi M, Benjamin EJ, Evans JC, D'Agostino RB, Fuller DL, Lehman B, et al. Intra- and interobserver reproducibility of Doppler-assessed indexes of left ventricular diastolic function in a population-based study (the Framingham Heart Study). Am J Cardiol 1992;70:1341-6.
7. 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.
8. Gottdiener JS, Bednarz J, Devereux R, Gardin J, Klein A, Manning WJ, et al. American Society of Echocardiography recommendations for use of echocardiography in clinical trials. J Am Soc Echocardiogr 2004;17:1086-119.
9. Pellikka PA, Douglas PS, Miller JG, Abraham TP, Baumann R, Buxton DB, et al. American Society of Echocardiography Cardiovascular Technology and Research Summit: a roadmap for 2020. J Am Soc Echocardiogr 2013;26:325-38.
10. Douglas PS. Duke echocardiography core laboratory inter-reader reproducibility, unpublished. 2015.
11. Gottdiener JS, Livengood SV, Meyer PS, Chase GA. Should echocardiography be performed to assess effects of antihypertensive therapy? Test-retest reliability of echocardiography for measurement of left ventricular mass and function. J Am Coll Cardiol 1995;25:424-30.
12. 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.
13. Hare JL, Brown JK, Leano R, Jenkins C, Woodward N, Marwick TH. Use of myocardial deformation imaging to detect preclinical myocardial dysfunction before conventional measures in patients undergoing breast cancer treatment with trastuzumab. Am Heart J 2009;158:294-301.
14. Jenkins C, Bricknell K, Hanekom L, Marwick TH. Reproducibility and accuracy of echocardiographic measurements of left ventricular parameters using real-time three-dimensional echocardiography. J Am Coll Cardiol 2004;44:878-86.
15. Lu X, Xie M, Tomberlin D, Klas B, Nadvoretskiy V, Ayres N, et al. How accurately, reproducibly, and efficiently can we measure left ventricular indices using M-mode, 2-dimensional, and 3-dimensional echocardiography in children? Am Heart J 2008;155:946-53.
16. Malm S, Frigstad S, Sagberg E, Larsson H, Skjaerpe T. Accurate and reproducible measurement of left ventricular volume and ejection fraction by contrast echocardiography: a comparison with magnetic resonance imaging. J Am Coll Cardiol 2004;44:1030-5.
17. Otterstad JE, Froeland G, St John Sutton M, Holme I. Accuracy and reproducibility of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and function. Eur Heart J 1997;18:507-13.
18. Shah AM, Cheng S, Skali H, Wu J, Mangion JR, Kitzman D, et al. Rationale and design of a multicenter echocardiographic study to assess the relationship between cardiac structure and function and heart failure risk in a biracial cohort of community-dwelling elderly persons: the Atherosclerosis Risk in Communities study. Circ Cardiovasc Imaging 2014;7:173-81.
19. 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.
20. Tighe DA, Rosetti M, Vinch CS, Chandok D, Muldoon D, Wiggin B, et al. Influence of image quality on the accuracy of real time three-dimensional echocardiography to measure left ventricular volumes in unselected patients: a comparison with gated-SPECT imaging. Echocardiography 2007;24:1073-80.
21. Schalla S, Nagel E, Lehmkuhl H, Klein C, Bornstedt A, Schnackenburg B, et al. Comparison of magnetic resonance real-time imaging of left ventricular function with conventional magnetic resonance imaging and echocardiography. Am J Cardiol 2001;87:95-9.
22. Douglas PS, Waugh RA, Bloomfield G, Dunn G, Davis L, Hahn RT, et al. Implementation of echocardiography core laboratory best practices: a case study of the PARTNER I trial. J Am Soc Echocardiogr 2013;26:348-3583.
23. Hoffmann R, von Bardeleben S, ten Cate F, Borges AC, Kasprzak J, Firschke C, et al. Assessment of systolic left ventricular function: a multicentre comparison of cineventriculography, cardiac magnetic resonance imaging, unenhanced and contrast-enhanced echocardiography. Eur Heart J 2005;26:607-16.
24. Gardin JM, Brunner D, Schreiner PJ, Xie X, Reid CL, Ruth K, et al. Demographics and correlates of five-year change in echocardiographic left ventricular mass in young black and white adult men and women: the Coronary Artery Risk Development in Young Adults (CARDIA) study. J Am Coll Cardiol 2002;40:529-35.
25. 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.
26. Shiran A, Adawi S, Ganaeem M, Asmer E. Accuracy and reproducibility of left ventricular outflow tract diameter measurement using transthoracic when compared with transesophageal echocardiography in systole and diastole. Eur J Echocardiogr 2009;10:319-24.
27. Smith LA, Cowell SJ, White AC, Boon NA, Newby DE, Northridge DB. Contrast agent increases Doppler velocities and improves reproducibility of aortic valve area measurements in patients with aortic stenosis. J Am Soc Echocardiogr 2004;17:247-52.
28. Schirmer H, Lunde P, Rasmussen K. Mitral flow derived Doppler indices of left ventricular diastolic function in a general population; the Tromso study. Eur Heart J 2000;21:1376-86.
29. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60.
30. Barnhart HX. Assessing agreement with relative area under the coverage proability curve. Stat Med 2016;35:3153-65.
31. Daubert MA, Yow E, Barnhart HX, Rabineau D, Crowley AL, Douglas PS. Quality improvement implementation: improving reproducibility in the echocardiography laboratory. J Am Soc Echocardiogr 2015;28:959-68.

SUPPLEMENTAL APPENDIX: COMPUTATIONAL DETAILS ON COMPUTED REPRODUCIBILITY VALUES

Specific computational details for Tables 1–6 and Figures 3–8 are provided in this appendix for reference. For the overall population description, the numbers were extracted from each article on the basis of the reported overall population or on the reproducibility population if the overall population was not presented. For some studies, the overall population r reproducibility population descriptions were only reported by subgroups, and these subgroup numbers were pooled to compute the overall population statistics. The calculated results for reproducibility were based on reported reproducibility values with certain assumptions. We elucidate the specific details and the assumptions employed. Let Y_{i1} and Y_{i2} be two measurements on subject i by the same observer or different observers. Then the following definitions were used for the reproducibility and agreement indices in Tables 2–7:

1. The mean difference is $\mu_D = E(Y_{i1} - Y_{i2})$ and SD of differences is $\sigma_D = \sqrt{Var(Y_{i1} - Y_{i2})}$.
2. The ICC is based on the one-way ANOVA model of $Y_{ij} = \mu + \alpha_i + e_{ij}$ with normal distribution of $\alpha_i \sim N(0, \sigma_S^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. The ICC is defined as $ICC = \sigma_S^2/\sigma_S^2 + \sigma_e^2$. Note that the overall population mean is $E(Y_{ij}) = \mu$ and the overall population variance is $Var(Y_{ij}) = \sigma_T^2 = \sigma_S^2 + \sigma_e^2$. The SD of difference has the following relationship with the error SD: $\sigma_D = \sqrt{2}\sigma_e$.
3. The LOA of 95% are defined as $LOAs = \mu_D \pm 1.96\sigma_D$.
4. The within-subject CV (wCV) is defined as $wCV = \sigma_e/\mu$.
5. The CP for a given acceptable difference of δ is defined as $\pi_\delta = Prob(|Y_{i1} - Y_{i2}| \leq \delta)$. Assume that $Y_{i1} - Y_{i2}$ is normally distributed as $N(\mu_D, \sigma_D^2)$. Then $\pi_\delta = \Phi(\delta - \mu_d/\sigma_D) - \Phi(-\delta - \mu_D/\sigma_D)$, where $\Phi(x)$ is the cumulative distribution function.
6. The TDI with CP of π is defined as the solution of the equation $\pi = Prob(|Y_{i1} - Y_{i2}| \leq TDI_\pi)$. With assumption that $Y_{i1} - Y_{i2}$ is normally distributed as $N(\mu_D, \sigma_D^2)$, then this equation becomes $\pi = \Phi(TDI_\pi - \mu_d/\sigma_D) - \Phi(-TDI_\pi - \mu_D/\sigma_D)$. We used $\pi = 0.80, 0.95$ in our computations.

At least two or three reported values in Tables 2–7 were required to compute the other missing reproducibility and agreement indices or population characteristics. If the study did not report the mean difference of paired measurements on the same subject, then 0.0 was assumed as the mean difference. The following scenarios were employed in our computations:

Scenario 1: If the error SD σ_e was reported but SD of differences was not reported, then SD of difference is computed as $\sigma_D = \sqrt{2}\sigma_e$.

Scenario 2: If the SD of differences was reported or computed as above, then the ICC is computed as $ICC = ((\sigma_T^2 - \sigma_D^2/2)/\sigma_T^2)$, by computing error variance as $\sigma_e^2 = (\sigma_D^2/2)$ and the reported overall population variance was used to estimate the total variability in the reproducibility study $\sigma_T^2 = \sigma_S^2 + \sigma_e^2$. The LOAs of 95% were computed as $LOAs = \sigma_D \pm 1.96\sigma_D$; the wCV was computed as $wCV = (\sigma_D/\mu\sqrt{2})$ with reported population mean μ; the CP was computed as $\pi_\delta = \Phi(\delta - \mu_d/\sigma_D) - \Phi(-\delta - \mu_D/\sigma_D)$ with given δ and normality assumption; the TDI as solution of $\pi = \Phi(TDI_\pi - \mu_d/\sigma_D) - \Phi(-TDI_\pi - \mu_D/\sigma_D)$ for $\pi = 0.80, 0.95$.

Scenario 3: If the ICC and overall population statistics were reported, then reported overall population variance was used for the total variability in the reproducibility study $\sigma_T^2 = \sigma_S^2 + \sigma_e^2$ to compute the error variability as $\sigma_e^2 = \sigma_T^2(1 - ICC)$. This leads to computation of the SD of differences as $\sigma_D = \sigma_T\sqrt{2(1 - ICC)}$. The computations of LOAs, wCV, CP and TDI follow as in scenario 2.

Scenario 4: If the wCV and the overall population statistics were reported only, then the error SD was computed as $\sigma_e = \mu * wCV$ where μ is the reported overall population mean. This leads to computation of the SD of differences as $\sigma_D = \sigma_e\sqrt{2}$. The other indices were then computed as in scenario 2.

Scenario 5: If the ICC and the SD of the differences were reported, but the overall population SD was not reported, then the overall SD was computed as the overall SD from the reproducibility study as $\sigma_T = \sigma_D/\sqrt{2(1 - ICC)}$.

For all figures, the CP curves were computed with normality assumption as $\pi_\delta = \Phi(\delta - \mu_d/\sigma_D) - \Phi(-\delta - \mu_D/\sigma_D)$ with δ representing the x axis and π_δ representing the y axis.

Supplemental Table 1 Details for the reproducibility methods used for the studies in Figures 1 and 3–8

Study (first author, population)	Method(s) for variable measurement	Overall study (*n*)	types of reproducibility assessed	No. MD readers	No. sonographers	No. echocardiograms read for reproducibility	Statistical methods reported for reproducibility
LV end-diastolic volume							
Schalla 2001[21]; Mixed CVD	2D biplane method of disks (modified Simpson's rule)	34	Interobserver Intraobserver	2 1	0	34	Pearson correlation coefficient
Jenkins 2004[14]; Referred for echocardiography	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	50	Interobserver Intraobserver Interstudy	0 0 NR	2 2 NR	20 20 50	Mean difference ± SD, Pearson correlation coefficient
Malm 2004[16]; Known or suspected heart disease	2D biplane method of disks (modified Simpson's rule)	110	Interobserver Intraobserver	2 1	0	30 NR	LOA
Tighe 2007[20]; Referred for SPECT	3D real-time with semiautomated summation of disks	64	Interobserver Intraobserver	>1 NR	1	10 10	Percentile difference
Lu 2008[15]; Normal children	Cube function of M-mode linear dimensions, 2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	20	Interobserver Intraobserver	NR	NR	20	Percentile difference
Hare 2009[13]; Oncology	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	35	Interobserver Intraobserver	NR	NR	10 10	Mean difference ± SD, 95% CI, ICC, repeatability coefficient
Studies with sufficient data to extrapolate additional indices and are included in further analyses							
Gottdeiner 1995[11]; Hypertension	Cube function of 2D targeted M-mode linear dimensions	96	Interstudy-intraobserver	1	≥1	81	Mean difference ± SD, SEM, ICC
Otterstad 1997[17]; MI and normal	2D biplane method of disks (modified Simpson's rule)	24	Interobserver Interstudy	2	0	24	ANOVA, CV
Grothues 2002[12]; LV hypertrophy, heart failure, normal	2D biplane method of disks (modified Simpson's rule)	60	Interstudy-intraobserver	1	NR	60	Mean difference ± SD, CV (indexed values only)
Oh 2012[4]; Heart failure	2D biplane method of disks (modified Simpson's rule)	1,460	Interobserver	NR	NR	67	Mean difference ± SD, mean percentage difference ± SD, Pearson correlation coefficient

Thavendiranathan 2013[19]; Oncology	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	20	Interobserver Intraobserver	2	NR	20 20	ANOVA, SEM ± 95% CI
Shah 2014[18]; Elderly community population	2D biplane method of disks (modified Simpson's rule) (ref ASE guidelines 2005) and 3D	>6,000	Intraobserver Temporal drift	Not specified	4	40	Bias ± SD, CV, ICC
Douglas 2015[10]; Unpublished Mixed population	2D biplane method of disks (modified Simpson's rule)		Interobserver Intraobserver	2	15	10 10	ICC, CP
LV end-systolic volume							
Schalla 2001[21]; Mixed CVD	2D biplane method of disks (modified Simpson's rule)	34	Interobserver Intraobserver	2 1	0	34	Pearson correlation coefficient
Jenkins 2004[14]; Referred for echocardiography	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	50	Interobserver Intraobserver Interstudy	0 0 NR	2 2 NR	20 20 50	Mean difference ± SD, Pearson correlation coefficient
Malm 2004[16]; Known or suspected heart disease	2D biplane method of disks (modified Simpson's rule)	110	Interobserver Intraobserver	2 1	0	30 NR	LOA
Tighe 2007[20]; Referred for SPECT	3D real-time with semiautomated summation of disks	64	Interobserver Intraobserver	>1 NR	1	10	Percentile difference
Lu 2008[15]; Normal children	Cube function of M-mode linear dimensions, 2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	20	Interobserver Intraobserver	NR	NR	20 20	Percentile difference
Hare 2009[13]; Oncology	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	35	Interobserver Intraobserver	NR	NR	10 10	Mean difference ± SD, 95% CI, ICC, repeatability coefficient
Studies with sufficient data to extrapolate additional indices and are included in further analyses							
Gottdeiner 1995[11]; Hypertension	Cube function of 2D targeted M-mode linear dimensions	96	Interstudy-intraobserver	1	≥1	75	Mean difference ± SD, SEM, ICC
Otterstad 1997[17]; MI and normal	2D biplane method of disks (modified Simpson's rule)	24	Interobserver Interstudy	2	0	24	ANOVA, CV

(*Continued*)

Supplemental Table 1 (*Continued*)

Study (first author, population)	Method(s) for variable measurement	Overall study (*n*)	types of reproducibility assessed	No. MD readers	No. sonographers	No. echocardiograms read for reproducibility	Statistical methods reported for reproducibility
Grothues 2002[12]; LV hypertrophy, heart failure, normal	2D biplane method of disks (modified Simpson's rule)	60	Interstudy-intraobserver	1	NR	20	Mean difference ± SD, CV (indexed values only)
Oh 201[22]; Heart failure	2D biplane method of disks (modified Simpson's rule)	1,460	Interobserver	NR	NR	67	Mean difference ± SD, mean percentage difference ± SD, Pearson correlation coefficient
Thavendiranathan 2013[19]; Oncology	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	20	Interobserver Intraobserver	2	NR	20 20	ANOVA, SEM ± 95% CI
Shah 2014[18]; Elderly community population	2D biplane method of disks (modified Simpson's rule) (ref ASE guidelines 2005) and 3D	>6,000	Intraobserver Temporal drift	Not specified	4	40	Bias ± SD, CV, ICC
Douglas 2015[10]; Unpublished Mixed population	2D biplane method of disks (modified Simpson's rule)	596	Interobserver Intraobserver	2	15	30 30	ICC, CP
LV ejection fraction							
Schalla 2001[21]; Mixed CVD	2D biplane method of disks (modified Simpson's rule)	34	Interobserver Intraobserver	2 1	0	34	Pearson correlation coefficient
Jenkins 2004[14]; Referred for echocardiography	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	50	Interobserver Intraobserver Interstudy	0 0 NR	2 2 NR	20 20 50	Mean difference ± SD, Pearson correlation coefficient, LOA
Malm 2004[16]; Known or suspected heart disease	2D biplane method of disks (modified Simpson's rule)	110	Interobserver Intraobserver	2 1	0	30 NR	LOA, mean difference
Hoffman 2005[23]; Known or suspected heart disease	2D biplane method of disks (modified Simpson's rule),	115	Interobserver	3	0	115	ICC, percentage error, 95% CI
Lu 2008[15]; Normal children	Cube function of M-mode linear dimensions, 2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	20	Interobserver Intraobserver	NR	NR	20 20	Percentile difference

Studies with sufficient data to extrapolate additional indices and are included in further analyses							
Gottdeiner 1995[11]; Hypertension	Cube function of 2D targeted M-mode linear dimensions	96	Interstudy-intraobserver	1	≥1	78	Mean difference ± SD, SEM, ICC
Otterstad 1997[17]; MI and Normal	2D biplane method of disks (modified Simpson's rule)	24	Interobserver Interstudy	2	0	24	ANOVA, CV
Grothues 2002[12]; LV hypertrophy, heart failure, normal	2D biplane method of disks (modified Simpson's rule) (ref ASE guidelines 2005)	60	Interstudy-intraobserver	1	NR	20	Mean difference ± SD, CV (indexed values only)
Hare 2009[13]; Oncology	2D biplane method of disks (modified Simpson's rule), 3D real-time	35	Interobserver Intraobserver	NR	NR	10 10	Mean difference ± SD, 95% CI, ICC, repeatability coefficient
Douglas 2013[22]; Aortic stenosis	2D biplane method of disks (modified Simpson's rule) (ref ASE guidelines 2005)	596	Interobserver Intraobserver	2	15	30 30	ICC, CP
Thavendiranathan 2013[19]; Oncology	2D biplane method of disks (modified Simpson's rule), 3D real-time with semiautomated summation of disks	20	Interobserver Intraobserver	2	NR	20 20	ANOVA, SEM ± 95% CI
Shah 2014[18]; Elderly community population	2D biplane method of disks (modified Simpson's rule) (ref ASE guidelines 2005) and 3D	>6,000	Intraobserver Temporal drift	Not specified	4	40	Bias ±SD, CV, ICC
LV mass							
Jenkins 2004[14]; Referred for echocardiography	M-mode, 2D, and 3D real-time (listed in accordance with ASE guidelines but methods not specified)	50	Interobserver Intraobserver Interstudy	0 0 NR	2 2 NR	20 20 50	Mean difference ± SD, Pearson correlation coefficient, LOA
Lu 2008[15]; Normal children	Cube function of M-mode linear dimensions, 2D area length method, 3D real-time	20	Interobserver Intraobserver	NR	NR	20 20	Percentile difference
Studies with sufficient data to extrapolate additional indices and are included in further analyses							
Gottdeiner 1995[11]; Hypertension	Cube function of 2D targeted M-mode linear dimensions (ref Troy 1972)	96	Interstudy-intraobserver	1	≥1	78	Mean difference ± SD, SEM, ICC
Otterstad 1997[17]; MI and normal	2D area-length method	24	Interobserver Interstudy	2	0	24	ANOVA, CV
Palmieri 1999[25]; LV hypertrophy	2D truncated ellipsoid formula	183	Interobserver Intraobserver	1–5	≥1	183	ANOVA, SEM ± 95% CI
Gardin 2002[24]; Normal	Cube function of 2D targeted M-mode linear dimensions	1,189	Interobserver Intraobserver intratechnician intertechnician	NR	NR	NR	CV

(*Continued*)

Supplemental Table 1 (*Continued*)

Study (first author, population)	Method(s) for variable measurement	Overall study (*n*)	types of reproducibility assessed	No. MD readers	No. sonographers	No. echocardiograms read for reproducibility	Statistical methods reported for reproducibility
Grothues 2002[12]; LV hypertrophy, heart failure, normal	2D truncated ellipsoid formula (ref Devereux *Am J Cardiol* 1986)	60	Interstudy-intraobserver	1	NR	20	Mean difference ± SD CV (indexed values only)
Shah 2014[18]; Elderly community population	Cube function of 2D linear dimensions	>6,000	Intraobserver Temporal drift	Not specified	4	40	Bias ± SD, CV, ICC
Douglas 2015[10]; Unpublished Mixed population	2D truncated ellipsoid formula	596	Interobserver Intraobserver	2	15	30 30	ICC, CP
LV outflow tract diameter							
Studies with sufficient data to extrapolate additional indices and are included in further analyses							
Smith 2004[27]; Aortic stenosis	2D linear dimension	20	Interobserver Intraobserver	2	1	20	Mean difference, coefficient of reproducibility (percentile difference)
Shiran 2009[26]; Aortic Stenosis and normal	2D linear dimension	50	Interobserver Intraobserver	2 readers, not specified if MDs	NR	10 10	Mean difference ± SD, LOA
Douglas 2015[10]; Unpublished Aortic stenosis	2D linear dimension	596	Interobserver Intraobserver	2	15	30 30	ICC, CP ICC, CP
Mitral inflow Doppler peak early diastolic velocity							
Colan 2012[5]; Dilated cardiomyopathy and normal	Not specified	173	Interacquisition Intraobserver Intraobserver drift Interobserver (ECL1-ECL2) Interobserver (site-ECL) Interstudy-intraobserver	1 1 1 2 2 Not specified	2 0 0 0 0 Not specified	171 171 171 171 171	Mean percentage error
Studies with sufficient data to extrapolate additional indices and are included in further analyses							
Gottdeiner 1995[11]; Hypertension	Pulsed-wave Doppler at mitral annular level in 4-chamber view	96	Interstudy-intraobserver	1	≥1	88	Mean difference ± SD, SEM, ICC
Schirmer 2000[28]; General population	Pulsed-wave Doppler at the level of and between the mitral leaflet tips in 4-chamber view	3,022	Interobserver Intraobserver	2 2	0	58 58	Mean difference ± SD

Shah 2014[18]; Elderly community population	Pulsed-wave Doppler at the level of the mitral leaflet tips in 4-chamber view	>6,000	Intraobserver Temporal drift	Not specified	4	40	Bias ± SD, CV, ICC

ASE, American Society of Echocardiography; *CVD*, cardiovascular disease; *ECL*, echocardiography core laboratory; *MD*, physician; *MI*, myocardial infarction; *NR*, not reported; *3D*, three-dimensional; *2D*, two-dimensional; *SPECT*, single-photon emission computed tomography.

Mixed CVD: coronary artery disease with and without myocardial infarction ($n = 15$), hypertension ($n = 15$), valvular heart disease ($n = 6$), dilated cardiomyopathy ($n = 4$), noncardiac chest pain ($n = 3$), and hypertrophic cardiomyopathy ($n = 2$).

Referred for echocardiography: regional wall motion abnormalities ($n = 41$ in the overall population and $n = 18$ in the reproducibility population), hypertension ($n = 2$ in the overall population and $n = 1$ in the reproducibility population), and normal ($n = 7$ in the overall population and $n = 1$ in the reproducibility population).

Referred for SPECT: coronary artery disease ($n = 24$), prior myocardial infarction ($n = 18$), prior coronary artery bypass surgery ($n = 14$), prior percutaneous coronary intervention ($n = 5$), hypertension ($n = 42$), diabetes mellitus ($n = 15$), hypercholesterolemia ($n = 39$), chest pain ($n = 42$), family history of coronary artery disease ($n = 18$), preoperative evaluation ($n = 11$), peripheral vascular disease ($n = 4$), dyspnea ($n = 9$), tobacco use ($n = 5$), cardiomyopathy ($n = 2$), heart failure ($n = 5$).

Percentile difference = (difference between the measurements/mean of the measurements) × 100%. Percentage error = (SD between 2 measurements/mean of 2 measurements) × 100%.