Measurement Consistency from Magnetic Resonance Images[1]

Dongjun Chung, MA, Moo K. Chung, PhD, Reid B. Durtschi, MS, Lindell R. Gentry, MD, Houri K. Vorperian, PhD

Rationale and Objectives. In quantifying medical images, length-based measurements are still obtained manually. Due to possible human error, a measurement protocol is required to guarantee the consistency of measurements. In this work, we review various statistical techniques that can be used in determining measurement consistency. The focus is on detecting a possible measurement bias and determining the robustness of the procedures to outliers.

Materials and Methods. We review correlation analysis, linear regression, Bland-Altman method, paired *t*-test, and analysis of variance (ANOVA). These techniques were applied to measurements, obtained by two raters, of head and neck structures from magnetic resonance images.

Results. The correlation analysis and the linear regression were shown to be insufficient for detecting measurement inconsistency. They are also very sensitive to outliers. The widely used Bland-Altman method is a visualization technique, so it lacks the numeric quantification. The paired *t*-test tends to be sensitive to small measurement bias. In contrast, ANOVA performs well even under small measurement bias.

Conclusions. In almost all cases, using only one method is insufficient and it is recommended that several methods be used simultaneously. In general, ANOVA performs the best.

Key Words. Measurement consistency; bias; outlier; head; neck; Bland-Altman.

We were motivated in part by the need to establish a reliable measurement protocol of head and neck structures involving both bony and soft tissue structures from magnetic resonance (MR) images collected for the purpose of quantifying the growth pattern of various oral and pharyngeal structures or vocal tract structures (1,2). Figure 1 depicts a select set of such measurements obtained manually from MR imaging.

It is crucial to obtain accurate and reliable measurements, particularly in developmental studies, and to establish an accurate measurement protocol. Unfortunately, because the ground truth for manual measurements is never known, it is difficult to quantitatively determine if a given protocol produces consistent measurements. We have addressed this problem by placing reference landmarks and obtaining repeated measures from MR images by two trained raters. Next, using those paired measurements, we assessed the consistency of measurements of our measurement protocol. The purpose of this study is to determine the ideal analysis method to check for consis-

Acad Radiol 2008; 15:1322–1330

[1] From the Department of Statistics, University of Wisconsin-Madison, Madison, WI (D.C.); Department of Biostatistics and Medical Informatics, Waisman Laboratory for Brain Imaging and Behavior, University of Wisconsin-Madison, 1500 Highland Avenue, No. 437, Madison, WI 53705 (M.K.C.); Department of Radiology, University of Wisconsin Hospital and Clinics, Madison, WI (L.R.G.); and Waisman Center, University of Wisconsin-Madison, Madison, WI (R.B.D., H.K.V.). Received March 3, 2008; accepted April 24, 2008. This work was supported in part by National Institutes of Health research grants R03 DC4362 (Anatomic Development of the Vocal Tract: MRI Procedures) and R01 DC6282 (MRI and CT Studies of the Developing Vocal Tract) from the National Institute of Deafness and Other Communicative Disorders (NIDCD), as well as core grant P-30 HD03352 to the Waisman Center from the National Institute of Child Health and Human Development (NICHHD). **Address correspondence to:** M.K.C. e-mail: mkchung@wisc.edu

doi:10.1016/j.acra.2008.04.020

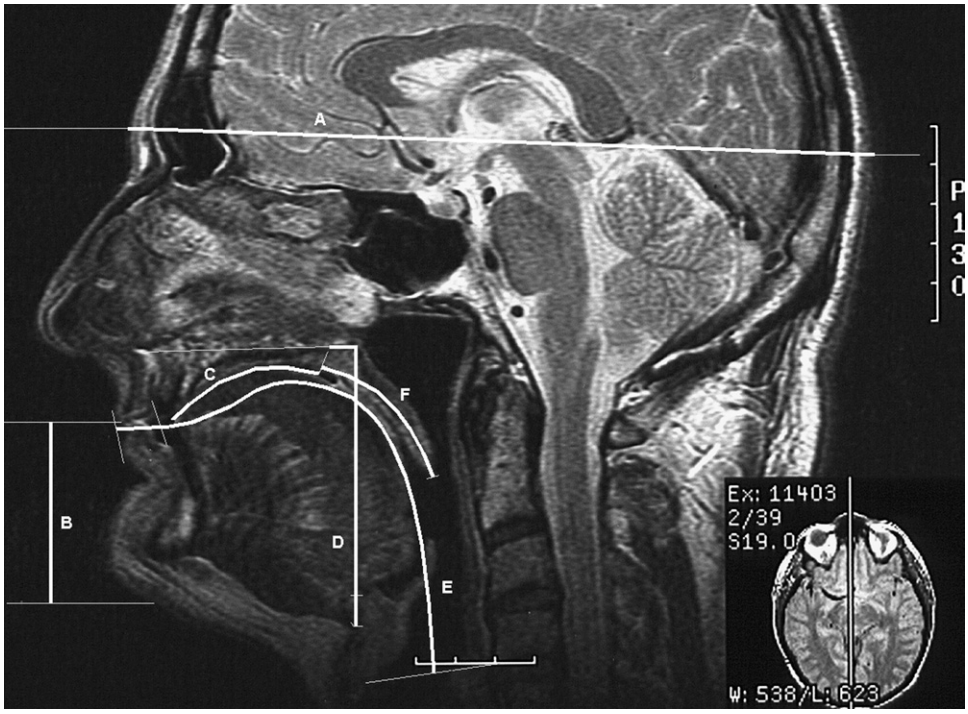

Figure 1. Mid-sagittal head and neck magnetic resonance images with the six measurements used for measurement consistency comparison: (**a**) Head length (HL); (**b**) lower anterior face height (LFH); (**c**) anterior tongue length (ATL); (**d**) hyoid vertical distance from posterior nasal spine (HVP); (**e**) vocal tract length (VTL); and (**f**) soft palate length (SP). See text for the definition of variables and tissue type and measurement type of each variable.

tency of measurements. We will refer to this problem as the "measurement consistency problem."

The measurement consistency problem occurs universally, and it is of broad interest to researchers in diverse medical imaging disciplines. There are several major statistical approaches that have been used to check measurement consistency. The most widely used methods are correlation analysis, linear regression, paired *t*-test, and the Bland-Altman method (3,4). A review of the measurement consistency problem can be found in Krummenauer and Doll (3). They (3) conclude that using only one method is insufficient and that several methods should be applied and compared. They also suggest making as many repeated measurements as time and cost permit for more accurate determination of measurement consistency.

Bland and Altman (4) found that the correlation analysis, which is a popular method in establishing measurement consistency (5–9), is not appropriate. They proposed a visualization technique called the "Bland-Altman method" based on the difference between measurements. A detailed discussion on this method can be found in Bland and Altman (10,11). Braždžionytė and Macas (12) claimed that the Bland-Altman method is more appropriate for assessing the measurement consistency compared to correlation analysis and linear regression. However, a shortcoming of the Bland-Altman approach is that it is a visualization technique and lacks numeric quantification.

Abate et al. (13) used the Bland-Altman method to analyze the measurement consistency between MR imaging and dissection for measuring adipose tissue mass. Powell et al. (7) used both a linear regression and the Bland-Altman method to analyze the measurement consistency between ultrasonic flowmeter measurements and phase-velocity cine MR imaging. Edvardsen et al. (5) used a paired *t*-test and the Bland-Altman method to compare the measurements from tissue Doppler echocardiography to the measurements from MR imaging. Liu et al. (6) used the correlation coefficient to analyze the measurement consistency between manual delineation and automated segmentation of thermal coagulation on three-dimensional elastographic images.

We review various quantitative techniques for determining measurement consistency and provide an MR imaging study that describes the strength and the weakness

of each technique. When comparing techniques, our main focus is on detecting the measurement bias and determining robustness to outliers. We provide further guidelines for using each technique.

MATERIALS AND METHODS

Description of Head and Neck Imaging Data

MR images from 10 male subjects (aged 0 to 4 years) were used for this study. The landmarks for making measurements were placed on the MR imaging slice independently by two trained raters, referred to as CC and RD. All landmarks and measurements were taken from the mid-sagittal slice of the MR images from the imaging database. To ensure unbiased placement of landmarks, RD and CC each placed landmarks on the image after suppressing the landmarks placed by the other. Thus, each rater landmarked and measured the selected image independently of the other. All landmarks and measurements were made using the Sigma Scan Pro version 5 (Systat Software, Inc., San Jose, CA), and data were recorded onto a hardcopy measurement sheet and entered into a measurement database for statistical analysis. All measurements were made in centimeters.

Both CC and RD obtained measurements from 10 MR images independently at three separate times, resulting in a total of 60 measurements. These measurements were classified into four different categories: consistent, less consistent, biased, and with outliers. Of the 38 variables measured in the head and neck region, the following 6 variables are used to illustrate each case: head length (HL), lower anterior facial height (LFH), anterior tongue length (ATL), hyoid vertical distance (HVP) from posterior nasal spine, vocal tract length (VTL), and soft palate length (SP). The definitions of those six variables are as follows (see Fig 1).

- HL (bony tissue—linear measurement): The maximum linear distance from the glabella to the opisthocranion.
- LFH (bony and soft tissue—linear measurement): The distance from the stomion to the gnathion. If the subject has an open mouth posture, the stomion was taken as the point at the anterosuperior edge of the mandibular lip.
- ATL (soft tissue—curvilinear measurement): The curvilinear distance along the dorsal superior contour of the tongue from the tongue tip to the intersection with the line dividing the hard palate and soft palate.
- HVP (bony tissue—linear measurement): The vertical distance from the inferior and anterior aspect of the hyoid bone to the level of the PNS.
- VTL (bony and soft tissue—curvilinear measurement): The curvilinear distance along the midline of the tract (ie, the distance along the midpoints of lines drawn between the inferior and superior boundaries of the vocal tract wall) starting at the level of the true vocal fold to the intersection with a line drawn tangentially to the lips.
- SP (bony and soft tissue—curvilinear measurement): The curvilinear distance from the posterior edge of the hard palate to the inferior edge of the uvula—a projection of variable length from the free inferior border of the soft palate. The criterion used to identify the end of the hard palate and the beginning of the soft palate is a line drawn at the beginning of the hard palate/soft palate overlap.

The measurement errors themselves are relatively small and measured by the average relative error (ARE), defined as:

$$ARE = \frac{1}{n}\sum_{i=1}^{n} \frac{|RD_i - CC_i|}{|RD_i + CC_i|/2}, \quad (1)$$

where RD_i and CC_i are the ith measurement of RD and CC, respectively, and $n = 30$, the number of measurements obtained by each rater. The average relative errors for HL, LFH, ATL, HVP, VTL, and SP are 0.016, 0.036, 0.041, 0.070, 0.046, and 0.1, respectively. The fairly large ARE of SP is caused by an outlier (Fig 2).

Figure 2 shows the scatterplot of the measurements of each head and neck structure. There are 30 data points on each scatterplot (three repeated measurements for 10 MR images). The solid line ($y = x$) indicates the perfect consistency between two raters. Two raters measured HL and LFH consistently and most points are placed near the $y = x$ line. ATL and HVP measurements are less consistent than for LFH. For VTL, most points are under the $y = x$ line and the measurements obtained by RD are biased against the measurements obtained by CC. For SP, there is an outlier caused by RD.

Correlation Analysis and Linear Regression

The correlation coefficient r measures the linear relationship between two variables, and ranges between -1 and 1. If measurements are consistent, we expect to have

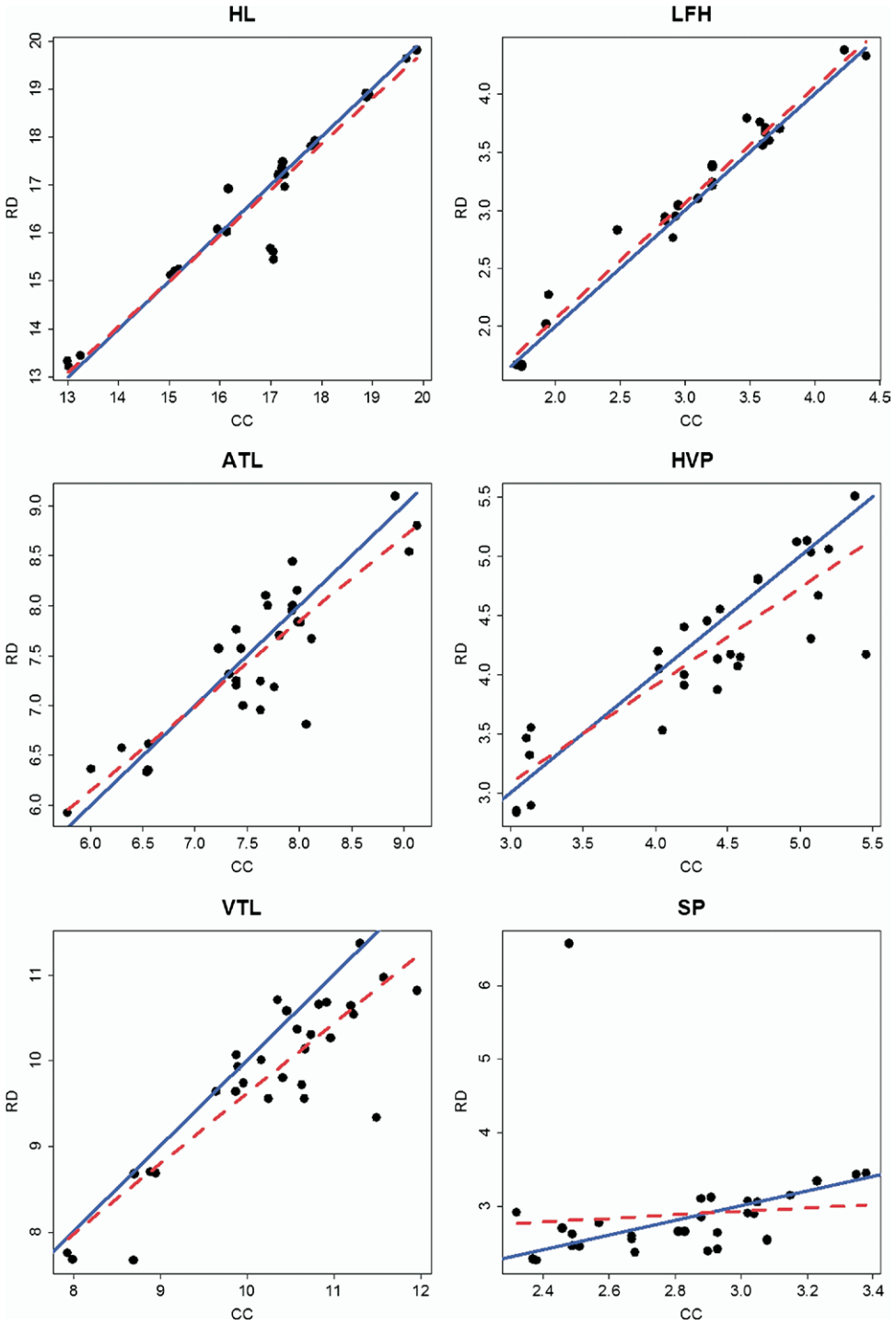

Figure 2. Scatterplots of head length (HL), lower anterior face height (LFH), anterior tongue length (ATL), hyoid vertical distance from posterior nasal spine (HVP), vocal tract length (VTL), and soft palate length (SP). The *solid lines* ($y = x$) indicate the perfect consistency between two raters. The *dotted lines* are the linear regression fit.

a strong linear relationship and, in turn, a correlation value close to 1. In contrast, if the measurements are less consistent, a correlation value close to 0 is expected. Under the null hypothesis of $r = 0$ (not consistent), the significance of correlation can be tested using a t-statistic with $n - 2$ degrees of freedom:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (2)$$

The correlation analysis has been previously used in measurement consistency (5–9). However, as we show in the Results section, it is not a proper procedure.

Alternately a linear regression can be used to determine the measurement consistency (7,12). The following regression model is used to fit measurements:

$$RD_i = \beta_0 + \beta_1 \times CC_i + \varepsilon_i.$$

When RD and CC are consistent, we expect the regression slope β_1 to be close to 1. By testing whether the slope is equal to 1, we can quantitatively determine the consistency. The regression fit is given in Figure 2. Because the slope is proportional to the correlation coefficient, both the correlation analysis and the linear regression are equivalent approaches, although this equivalence is not exploited previously (14). Similarly one can test whether the intercept β_0 is close to 0 for testing a bias if one rater is systematically obtaining larger or smaller measurements compared to the other rater.

Bland-Altman Method and Paired *t*-Test

Although the Bland-Altman method has been discussed in the literature (3–5,7,10–13), we briefly explain here for the completeness of this work. Let d_i be the measurement difference, ie, $d_i = CC_i - RD_i$. The measurement difference is the estimated bias of measurements between the two raters. Let $\bar{d}$ and S_d^2 be the mean and the variance of the difference. Bland and Altman plotted d_i versus the average of measurements of two raters, with the reference lines, $\bar{d}$, $\bar{d} - 1.96S_d$, and $\bar{d} + 1.96S_d$ (4). The range between $\bar{d} - 1.96S_d$ and $\bar{d} + 1.96S_d$ provides the "limit of agreement" (Fig 3).

The weakness of the Bland-Altman method is that the measurement consistency is mainly determined visually without statistical significance attached to the plot. To give the statistical significance to the Bland-Altman method procedure, a paired t-test can be used. We test whether the measurement difference is statistically small enough using the test statistic

$$T = \frac{\bar{d}}{\sqrt{S_d^2/n}}, \quad (3)$$

which is distributed as the t-distribution with $n - 1$ degrees of freedom.

ANOVA and Within-Rater Consistency

All the previous methods can determine consistency between a set of paired measurements. When there are more than two raters, the previous methods cannot be applied directly without significant modification. We propose to use the analysis of variance (ANOVA) approach for more general cases. The strength of ANOVA is that it can be used to determine both between- and within-rater measurement consistency. If we have information about how each rater measures the same MR image consistently, we can determine who is more consistent. This additional information can be used to further train less consistent raters.

Let X_{ijk} be the kth measurement on the jth MR image by the ith rater. Then, the two-way ANOVA model is given as

$$X_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}.$$

The usual measurement consistency between CC and RD can be determined by testing $\alpha_{CC} = \alpha_{RD}$. The interaction term $(\alpha\beta)_{ij}$ is used to determine the within-rater consistency for 10 MR images. The within-rater consistency can be determined by simultaneously testing $\alpha\beta_{CC,1} = \cdots = \alpha\beta_{CC,10} = \alpha\beta_{RD,1} = \cdots = \alpha\beta_{RD,10}$.

We can also visualize the within-rater consistency patterns using the box plot (15). The box plot is one of popular data visualization methods and it is drawn in the following way (16). First, we obtain the value corresponding to 25%, 50%, and 75% of the sorted observations. They are called the lower quantile q_1, the median q_2, and the upper quantile q_3, respectively. The median q_2 provides the information about the center, such that the half of the data are smaller than q_2 and the other half are larger than q_2. Then, we draw "the box" from q_1 to q_3 with the line of q_2 within the box. This box provides the range containing 50% of the data around q_2. Finally, we draw one line from q_1 to $q_1 - 1.5(q_3 - q_1)$ and another line from q_3 to $q_3 + 1.5(q_3 - q_1)$, which are called "the whiskers." In a

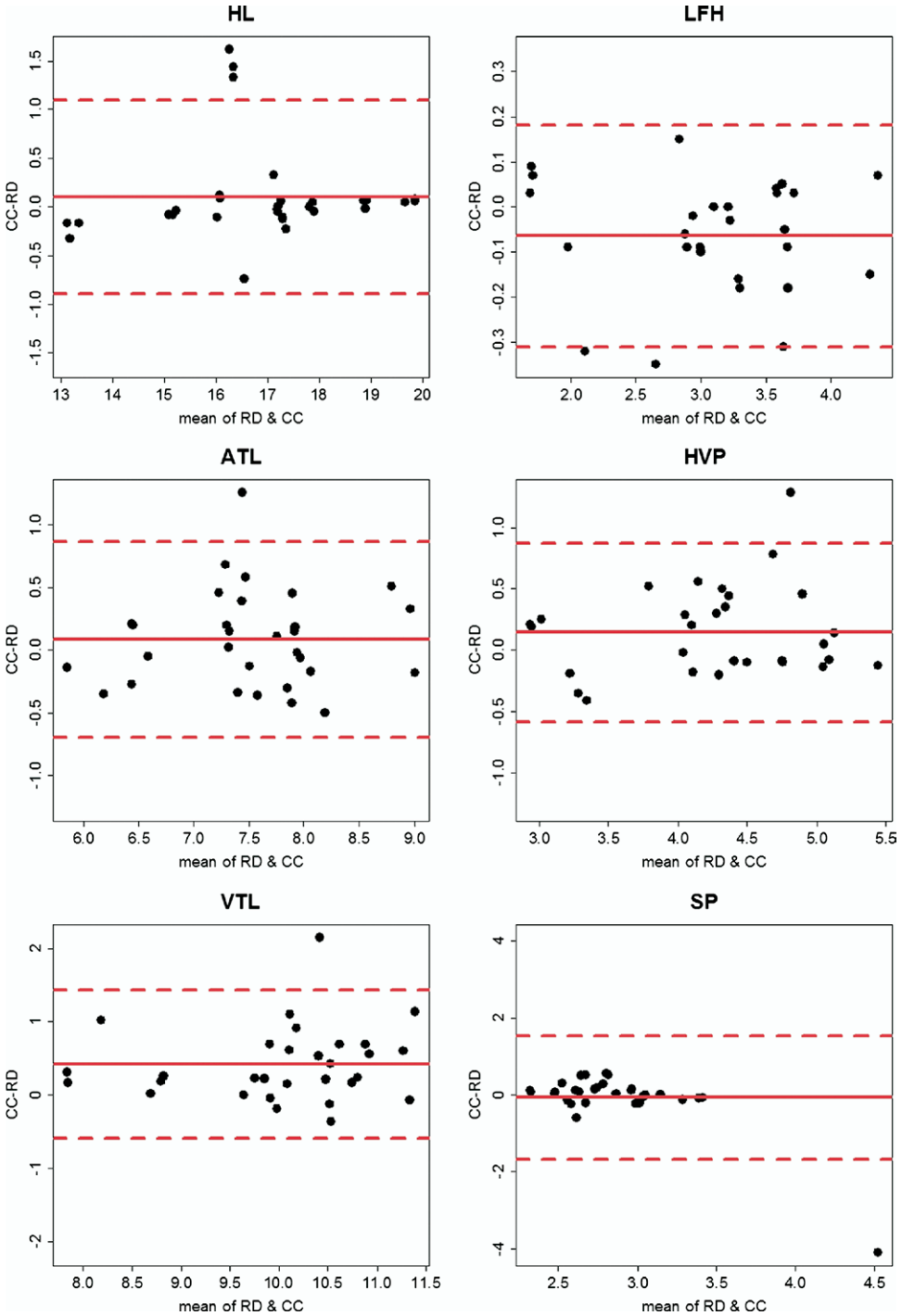

Figure 3. Bland-Altman plots of head length (HL), lower anterior face height (LFH), anterior tongue length (ATL), hyoid vertical distance from posterior nasal spine (HVP), vocal tract length (VTL), and soft palate length (SP). The *solid line* is the mean difference $\bar{d}$, and the *dotted lines* are $\bar{d} - 1.96S_d$ (*lower*) and $\bar{d} + 1.96S_d$ (*upper*).

box plot, the observations outside $q_1 - 1.5(q_3 - q_1)$ and $q_3 + 1.5(q_3 - q_1)$ are determined as potential outliers.

Let $d_{j,k}$ be the difference between the kth measurement of the jth MR image and the average measurements of the jth MR image by one fixed rater. The box plot of $d_{j,k}$ shows the diversity of measurements for each MR image. We can see how consistent each MR image is measured by a specific rater using the box plot of $d_{j,k}$. We can visu-

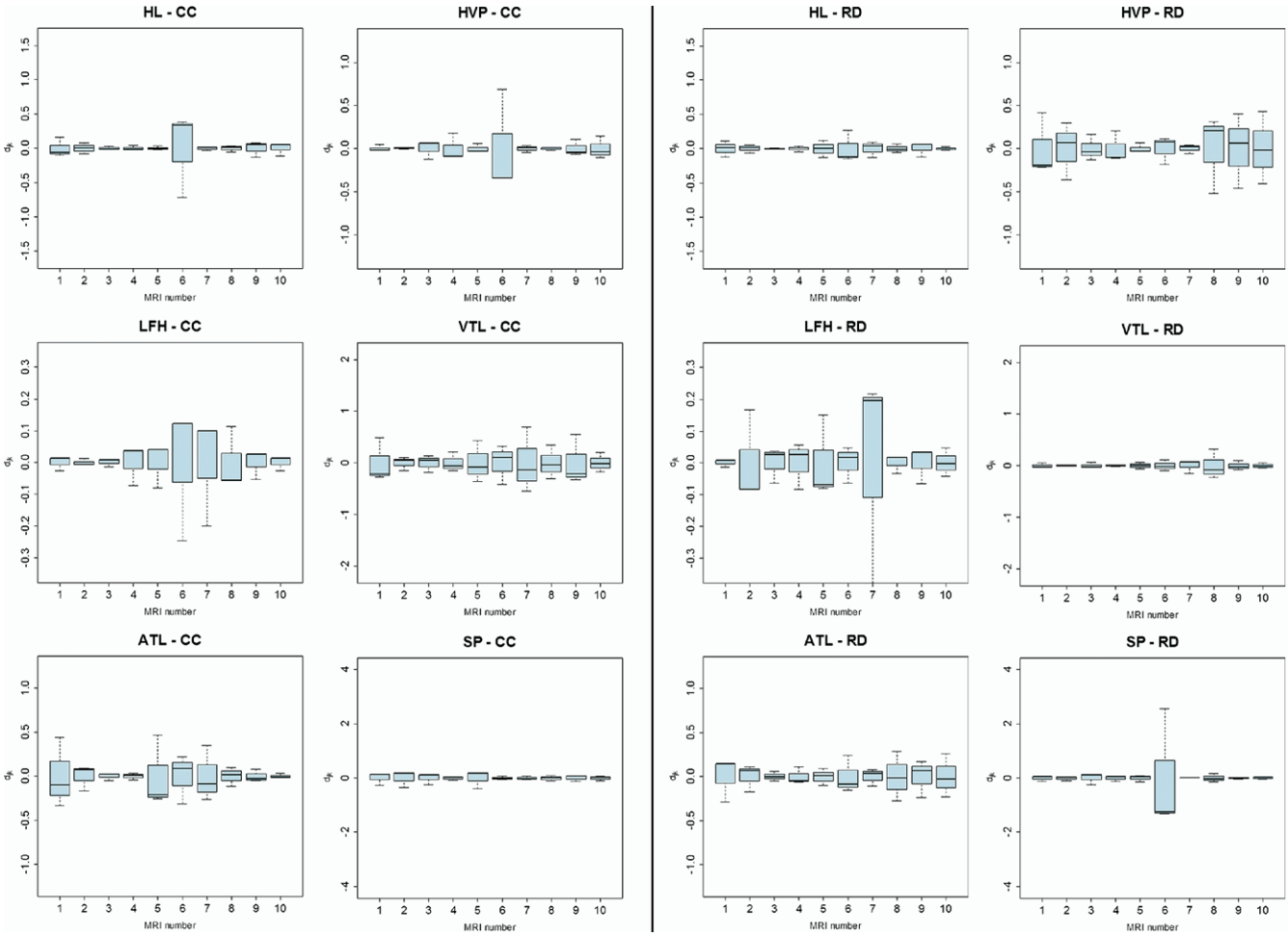

Figure 4. Within-rater consistency box plot of $d_{j,k}$ for 10 magnetic resonance images of head length (HL), lower anterior face height (LFH), anterior tongue length (ATL), hyoid vertical distance from posterior nasal spine (HVP), vocal tract length (VTL), and soft palate length (SP) for raters CC (*left*) and RD (*right*).

ally compare within-rater consistency by comparing the box plots between the raters CC and RD (Fig 4).

RESULTS

Correlation Analysis and Linear Regression

The linear regression fitting line for each head and neck structure appears as the dotted line in Figure 2. The measurements are more consistent when the dotted line is close to the solid line ($y = x$). Two lines were very close in HL, LFH, ATL, HVP, and SP. In contrast, the dotted line was far from the solid line in VTL. The correlation coefficients of HL, LFH, ATL, and HVP were 0.963, 0.987, 0.880, and 0.871, respectively ($P < .001$ in all cases). This implies the measurements are consistent for HL, LFH, ATL, and HVP, and this coincides with what we observe in Figure 2.

In contrast, the correlation coefficient was 0.875 ($P < .001$) for VTL, and this seems to contradict Figure 2 because there was a clear systematic bias in VTL. We can infer from this that the correlation coefficient cannot detect the measurement inconsistency. The correlation coefficient of SP was 0.089 ($P = .639$). Despite existing consistency between CC and RD, an outlier made the correlation coefficient close to 0. After removing the outlier, correlation coefficient of SP becomes 0.673 ($P < .001$). This implies that the correlation coefficient is very sensitive to outliers.

Table 1
Summary of Statistical Method Used in Determining the Measurement Consistency

Method	Strength	Weakness	Agreement	Disagreement
Correlation and regression	Show degree of consistency Simple procedure	Cannot easily detect inconsistency Sensitive to outliers	HL, LFH, ATL, HVP	VTL, SP
Bland-Altman Method	Visualization technique	Lacks statistical significance Not easy to quantify the degree of consistency	The method does not provide a decision.	
Paired *t*-test	Detect bias fairly well Simple procedure	Fails under systematic bias	HL, ATL, VTL, SP	LFH, HVP
ANOVA	Best performance Provide additional information of the within-rater consistency Applicable for more than two raters	Complicated procedure	HL, LFH, ATL, HVP, VTL, SP	

ANOVA: analysis of variance.
The last two columns show whether the method agrees with the ANOVA result for the six variables: head length (HL), lower anterior face height (LFH), anterior tongue length (ATL), hyoid vertical distance from posterior nasal spine (HVP), vocal tract length (VTL), and soft palate length (SP).

In summary, the correlation analysis has difficulty detecting the inconsistency between measurements. This is due to the fact that the correlation coefficient shows the degree of association, not the degree of consistency. The correlation analysis is very sensitive to outliers. As a result, the correlation analysis is not appropriate as the measurement consistency analysis.

Bland-Altman Method and Paired *t*-Test

Figure 3 shows the Bland-Altman plots for head and neck structures. Although these plots provide the degree of bias, it is not easy to infer about the measurement consistency based on these plots. This is because the Bland-Altman method lacks statistical significance attached to the plot. Moreover, in measuring SP, one outlier severely increases the limit of agreement. In summary, the Bland-Altman method is not appropriate as a technique for determining measurement consistency.

The paired *t*-test indicates that there is significant inconsistency in measuring LFH ($P = .008$) and HVP ($P = .038$), although the scatterplots of LFH and HVP in Figure 2 show measurement consistency. This contradiction can happen if one rater's measurements are systematically either larger or smaller than those of the other rater. When this systematic bias becomes larger than the measurement variance, this contradiction will happen.

In summary, the paired *t*-test can detect measurement bias between raters fairly well in most cases. However, it may fail when one rater systematically makes either larger or smaller measurements than the other rater.

ANOVA and Within-Rater Consistency

ANOVA results show that measurements are consistent between raters in measuring HL ($P = .110$), LFH ($P = .517$), ATL ($P = .576$), HVP ($P = .937$), and SP ($P = .279$) but not in measuring VTL ($P = .029$). This finding exactly coincides with what we found in Figure 2. The box plots in the Figure 4 and the interaction term in ANOVA show which rater performs better. RD is significantly more consistent than CC in measuring HL (the first row in the Fig. 4; $P < .001$). CC is more consistent than RD in measuring LFH (the second row in the Fig. 4) but the difference was not significant ($P = .770$). RD is significantly more consistent than CC in measuring ATL (the third row in the Fig 4; $P = .008$). CC is more consistent than RD in measuring HVP (the fourth row in the Fig 4) but the difference was not significant ($P = .152$). RD is significantly more consistent than CC in measuring VTL (the fifth row in the Fig. 4; $P = .016$). CC is more consistent than RD in measuring SP (the sixth row in the Fig. 4) but this difference was not significant ($P = .115$).

In summary, ANOVA extends the paired *t*-test method by considering the within-rater consistency. ANOVA analysis shows a good performance in detecting measurement bias.

DISCUSSION

In this work, we reviewed five techniques for determining measurement consistency of structures measured

from head and neck MR images: the correlation analysis, the linear regression, the Bland-Altman method, the paired *t*-test, and the ANOVA. We showed the strength and weakness of each technique in detecting the measurement bias and determining the robustness to outliers. Table 1 provides the summary of the strength and weakness of each technique.

A correlation analysis cannot detect the measurement inconsistency between raters and it is sensitive to outliers. It is inappropriate to use the correlation analysis for determining measurement consistency. A linear regression should not be used either because it is equivalent to the correlation analysis.

It is not easy to make a quantitative decision using the Bland-Altman method. This is mainly because the Bland-Altman plot does not have statistical significance attached to it. The paired *t*-test provides quantification for the Bland-Altman method, and it has a good performance in detecting measurement bias. However, when most of the measurements of one rater are consistently larger or smaller than those of the other rater, the paired *t*-test tends to fail.

ANOVA provides the best performance in all cases studied and showed accurate analysis results in determining the measurement consistency. In addition, it provides the additional information of within-rater consistency.

As suggested by Krummenauer and Doll (3), a good rule to follow is not to limit measurement consistency assessment on only one method but rather to apply and compare several methods. We also recommend making as many repeated measurements as time and cost permit for more accurate determination of measurement consistency.

ACKNOWLEDGMENTS

We thank Celia Choih for assistance with placement of the anatomic landmarks and for making the necessary measurements.

REFERENCES

1. Vorperian HK, Kent RD, Lindstrom MJ, Kalina CM, Gentry LR, Yandell BS. Development of vocal tract length during early childhood: A magnetic resonance imaging study. J Acoust Soc Am 2005; 117:338–350.
2. Vorperian HK, Durtschi RB, Wang S, Chung MK, Ziegert AJ, Gentry LR. Estimating head circumference from pediatric imaging studies: An improved method. Acad Radiol 2007; 14:1102–1107.
3. Krummenauer F, Doll G. Statistical methods for the comparison of measurements derived from orthodontic imaging. Eur J Orthod 2000; 22:257–269.
4. Bland JM, Altman DG, Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307–310.
5. Edvardsen T, Gerber BL, Garot J, Bluemke DA, Lima JAC, Smiseth OA. Quantitative assessment of intrinsic regional myocardial deformation by Doppler strain rate echocardiography in humans: Validation against three-dimensional tagged magnetic resonance imaging. Circulation 2002; 106:50–56.
6. Liu W, Zagzebski JA, Varghese T, Dyer CR, Techavipoo U, Hall TJ. Segmentation of elastographic images using a coarse-to-fine active contour model. Ultrasound Med Biol 2006; 32:397–408.
7. Powell AJ, Maier SE, Chung T, Geva T. Phase-velocity cine magnetic resonance imaging measurement of pulsatile blood flow in children and young adults: In vitro and in vivo validation. Pediatr Cardiol 2000; 21:104–110.
8. Vallejo E, Dione DP, Bruni WL, et al. Reproducibility and accuracy of gated SPECT for determination of left ventricular volumes and ejection fraction: Experimental validation using MRI. J Nucl Med 2000; 41:874–882.
9. Van Oosterhout MFM, Willigers HMM, Reneman RS, Prinzen FW. Fluorescent microspheres to measure organ perfusion: Validation of a simplified sample processing technique. Am J Physiol 1995; 269: H725–H733.
10. Bland JM, Altman DG. Comparing methods of measurement: Why plotting difference against standard method is misleading. Lancet 1995; 346:1085–1087.
11. Bland JM, Altman DG. Applying the right statistics: Analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22:85–93.
12. Bražd̆ionytė J, Macas A. Bland-Altman analysis as an alternative approach for statistical evaluation of agreement between two methods for measuring hemodynamics during acute myocardial infarction. Medicina 2007; 43:208–214.
13. Abate N, Burns D, Peshock RM, Garg A, Grundy SM. Estimation of adipose tissue mass by magnetic resonance imaging: Validation against dissection in human cadavers. J Lipid Res 1994; 35:1490–1496.
14. Chatterjee S, Hadi AS, Price B. Regression analysis by example, 3rd ed. New York: John Wiley & Sons, Inc., 2000.
15. Tukey JW. Exploratory data analysis. New York: Addison-Wesley, 1977.
16. Martinez WL, Martinez AR. Exploratory data analysis With MATLAB. London: Chapman & Hall/CRC, 2005.