Published in final edited form as:

Acad Radiol. 2018 March ; 25(3): 391–396. doi:10.1016/j.acra.2017.09.010.

Essentials of Statistical Methods for Assessing Reliability and Agreement in Quantitative Imaging

Arash Anvari, MD, Elkan F. Halpern, PhD, and Anthony E. Samir, MD, MPH

Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114

Abstract

Quantitative imaging is increasing in almost all fields of radiological science. Modern quantitative imaging biomarkers measure complex parameters including metabolism, tissue microenvironment, tissue chemical properties or physical properties. In this paper, we focus on measurement reliability assessment in quantitative imaging. We review essential concepts related to measurement such as measurement variability and measurement error. We also discuss reliability study methods for intraobserver and interobserver variability, and the applicable statistical tests including: intraclass correlation coefficient, Pearson correlation coefficient, and Bland-Altman graphs and limits of agreement, standard error of measurement, and coefficient of variation.

Keywords

Measurement error; reliability; agreement; intraclass correlation coefficient; Bland-Altman graph; standard error of measurement; coefficient of variation

INTRODUCTION

Quantitative imaging technologies are increasingly used for the measurement of normal biological processes, pathologic processes, patient risk stratification, treatment response measurement in clinical care, and drug development (1–3). The goal of quantitative imaging is objective, accurate, and precise measurement of quantifiable features obtained from in vivo imaging studies, termed quantitative imaging biomarkers (QIBs). The simplest QIBs comprise measurement of the size of organs, vessels, or lesions. More complex QIBs measure parameters including metabolism, for example, the standardized uptake value in positron emission tomography imaging; tissue microenvironment, for example, diffusion or perfusion; tissue chemical properties, for example, spectroscopy; or physical properties, for example, tissue stiffness (4). QIBs are continuous variables, of which there are two subtypes: (1) ratio variables, such as shear wave velocity measured in meters per second (m/s) by shear wave sonoelastography methods for liver fibrosis assessment, or (2) interval variables, such as computed tomography (CT) densitometry measured in Hounsfield units for estimating emphysema severity. Ordinal variables are not QIBs. For example, the widely used Prostate Imaging Reporting and Data System (PI-RADS) classification system used for prostate

magnetic resonance imaging assessment has five numbered categories: PI-RADS 1 (very low probability) to PI-RADS 5 (very high probability) of prostate cancer (5). Although these categories are numbered, the numbers denote order, rather than quantity, and therefore PI-RADS is not a QIB.

To be clinically useful, QIBs must be reliably comparable to one another and to known reference measurements (6). The goal of this paper is to facilitate better understanding of QIB reliability measurement by imaging researchers new to the field, and to assist researchers to incorporate reliability study design principles into their own quantitative imaging studies.

In this review, we define relevant metrologic terminology and concepts including measurement, reliability, reproducibility, and agreement. We discuss common reliability studies, including intraobserver, interobserver, and method comparison studies. We introduce guidelines for reporting (7), reviewing (8), and critical appraisal (9) of reliability studies, and we review statistical measures of reliability for continuous variables, including intraclass correlation coefficient (ICC), Pearson correlation coefficient, and measures of agreement including Bland-Altman graph and limits of agreement, standard error of measurement (SEM), and coefficient of variation (CV).

DEFINITIONS AND STATISTICAL CONCEPTS

Measurements are central to biomedical research and clinical practice and are used to evaluate current disease status and change over time. In population studies, measurements permit useful comparison of health outcomes within or between patients. The Quantitative Imaging Biomarkers Alliance (QIBA) is an initiative by the Radiological Society of North America to promote the use of QIBs in clinical research and practice. QIBA working groups defined QIB concepts based on the Joint Committee for Guides in Metrology in 2012–2013 (3). For QIBs, *measurement* is the process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity. Quantity is a property of a variable, where the property has a magnitude and can be expressed as a number, which is called quantity value. The *measurand* is the quantity intended to be measured (4).

Mu (μ) or mean, is a measure of central tendency of a data set, and is computed as the sum of the data set values divided by the number of values. Sigma (σ) or standard deviation (SD) is a measure used to quantify the dispersion of a set of data values, and is computed as the square root of the average of the squared differences from the mean (μ) divided by number of values minus 1. A low SD implies that the data values tend to be close to the mean, whereas a high SD implies the data points are spread out over a wider value range.

MEASUREMENT UNCERTAINTIES AND MEASUREMENT VARIABILITY IN QIBS

Uncertainty is defined as the dispersion of the quantity values attributed to a measurand (4). Uncertainty in measurement (measurement error) generally has two components: (1) *systematic error (bias)*, which is the degree of closeness of a quantity value to the true value

and describes the difference between the average of measured values and average of true values in the same patients, and (2) *random error*, which is the degree of *measurement variability* in the same patients.

Measurement variability can arise at several levels: (1) biological variability occurs *within patients* based on diurnal, temporal, or situational changes, for example, postprandial state in liver stiffness, and also variation *between patients*; (2) technological variability, for example, different imaging acquisition algorithms within a single imaging modality, image acquisition protocol, or use of different imaging modalities in a method comparison study, for example, ultrasound vs CT in measurement of lesion size; and (3) observer variability, for example, between radiologists with different experience levels. Measurement variability sources act together and simultaneously during the measurement process. Measurement variability is a distinct concept from variance, which is a term that describes the true variability of values within a sample.

Error mitigation should include mitigation of measurement variability and *measurement bias*. Strategies to minimize measurement variability include (1) *restriction* – avoiding a specific variation source such as fasting, or acquiring images in a specific position; (2) *standardization* – defining nonvarying protocols for imaging acquisition or processing; (3) *observer training*; and (4) *averaging of repeated measurements*, which is particularly useful when the measurement has a wide range of random error. Measurement bias can be estimated only if the true value is known and is not mitigated by strategies that minimize random measurement error (4,10). Measurement bias mitigation requires improved measurement system calibration.

Reliability, agreement, precision, repeatability, and reproducibility, are terms often used interchangeably in general conversation, but are distinct concepts that are evaluated differently (4,11).

Reliability is defined as how well patients can be distinguished from each other despite observer measurement error (4).

Agreement is the degree of closeness between measurements made on the same patients by one or more observers or two methods of measurement, for example, the closeness of common bile duct diameter measurements in the same patients made by two radiologists using the same ultrasound machine.

Precision deals with variability; it is the closeness of agreement between measured quantity values obtained by replicate measurements on the same patients under specified and stable conditions, for example, common bile duct diameter in the same patients is repeatedly measured by a single radiologist using a single ultrasound machine in the same position and on a single occasion.

Repeatability refers to variability measured under a set of conditions that includes the same measurement procedure, the same observer, the same measuring system, the same operating conditions, and replicate measurements on the same patients over a short period of time (4,10,12). The *coefficient of repeatability* is twice the SD of the measurement differences.

Typically, the smallest detectable “real” change over time should be more than the coefficient of repeatability to be considered a “real” change in a pathophysiological process (12).

Reproducibility refers to the variation in measurements made on the same patient under “real world” conditions, which are circumstances analogous to clinical practice, where a variety of external factors cannot all be tightly controlled (13).

RELIABILITY STUDY DESIGN

In imaging science, we generally study two types of reliability: (1) *intraobserver reliability*, and (2) *interobserver reliability*. Intraobserver reliability studies evaluate the same observer using the same measurement instrument from the same set of images on the same patients on different occasions. For intraobserver reliability study, recall bias related to memory of prior measurements is minimized by requiring a reasonable time interval between measurements (usually several weeks) and image set de-identification and randomization. Interobserver reliability studies assess different observers (two or more) using the same measurement instrument to measure QIB on the same image(s) from the same patients. If interobserver reliability is high, it is likely that intraobserver reliability will be at least as high, and typically higher (14).

Reliability studies can be confounded by *practice effect*, which is improvement of measurement skill by observers as they progress from novice to expert. This is colloquially termed the “learning curve.” The reliability of a new instrument should only be assessed when observers are no longer improving from measurement to measurement.

Several groups have developed guidelines to standardize concepts related to QIBs and to set reliability study quality criteria. Kessler et al. provided terminology consistent with QIB definitions (4), and via the Radiological Society of North America’s QIBA, Sullivan et al. introduced metrology standards applicable to QIB (3), and Raunig et al. reviewed statistical methods for technical performance assessment in detail (13). Kottner et al. developed the *Guidelines for Reporting Reliability and Agreement Studies* to assist investigators in designing and reporting reliability studies. These guidelines are widely used and are applicable in the radiological sciences (15). Lucas et al. developed the *Quality Appraisal of Reliability Studies* checklist as a methodological tool for systematic review of diagnostic reliability studies including diagnostic imaging. This checklist includes 11 items such as spectrum of observers and patients, the observers’ and patients’ sample representativeness, blinding of observers (to other observers, to their own prior measurements, and to the clinical findings), order effects of measurements, appropriateness of time interval between repeated measurements, assessing appropriateness of statistical analysis, and correctness of measurement application and interpretation (8). These papers contain additional detail that would be useful to investigators contemplating reliability study design.

STATISTICAL MEASURES OF RELIABILITY FOR CONTINUOUS VARIABLES

Statistical methods for analyzing reliability were first suggested in the late 1930s, and the ICC was introduced in the 1950s (16). The most commonly used reliability measures for

continuous variables are the ICC and Pearson correlation coefficient. Reliability of a diagnostic test refers to how well patients can be distinguished from each other despite observer measurement error (10). In addition to measurement error, study sample heterogeneity affects reliability; the more heterogeneous the sample, the higher the ICC (12,17).

The observed variability or total variability of a continuous variable in a sample is due to true variance of the continuous variable (σ_t^2) and variance secondary to measurement error (σ_e^2). Mathematically, test reliability is defined as the true variance (σ_t^2), divided by true variability (σ_t^2) plus measurement error variability (σ_e^2).

$$\text{reliability} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \quad (\text{Eq. 1})$$

Measurement error variability can be further divided into systematic error variability (σ_{se}^2) and random error variability (σ_{re}^2).

$$\text{reliability} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{se}^2 + \sigma_{re}^2} \quad (\text{Eq. 2})$$

Intraclass correlation coefficient (ICC)

ICC is the best-known reliability parameter for repeated measurements of continuous variables. It can estimate intra- and interobserver measurement reliability for continuous variables. The ICC formula is defined:

$$\text{ICC} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_{\text{residual}}^2} \quad (\text{Eq. 3})$$

where (σ_p^2) is the true QIB variance, σ_r^2 is the variance between observers, and $\sigma_{\text{residual}}^2$ is the residual variance comprising interaction between observers and patients, in addition to random error.

Acceptable ICC

ICC is reported as a number between 0 and +1 with a 95% confidence interval (CI). When used to discriminate between individuals in a group, an ICC of 1 represents perfect reliability, in which discrimination is unaffected by measurement error. An ICC of 0 represents lack of reliability, in which measurement error precludes distinction of individual patients. If measurement error is small relative to variability between patients, the ICC approaches 1. The acceptability of a given ICC is not a statistical issue but rather a clinical decision based on the consequences of test measurement error. In the literature, an ICC greater than 0.70 is often considered acceptable if a measurement instrument is used in

groups of patients. However, for applications in *individual patients*, an ICC greater than 0.90 or 0.95 is typically required (15,18).

Example 1—Shear wave elastography (SWE) is a modern imaging technique that estimates tissue stiffness numerically as the Young modulus. The estimated Young modulus is a quantitative imaging biomarker measured in kilopascals. Table 1 shows thyroid nodule stiffness data for two observers (observer 1 and observer 2) in 35 patients with benign or malignant lesions. The ICC for these two observers, computed with IBM SPSS for Mac (version 21.0), was 0.87 (95% CI, 0.71–0.92) and 0.86 (95% CI, 0.74–0.92), respectively, and interobserver ICC was 0.87 (95% CI, 0.75–0.93) (19) (Fig 1).

Pearson correlation coefficient

Pearson correlation coefficient (ρ) was used historically as a reliability measure, but has limitations. It is more appropriate for inferring *linear association* between two independent continuous variables (eg, effect of age on height in children), and is not suited to analysis of *agreement* of repeated measurements, which are, by definition, not independent. Statistically significant correlation does not automatically imply good reliability, because it does not account for systematic error (bias), which can result in reliability overestimation (20). In addition, the correlation coefficient depends on the range of the true quantity values or the sample heterogeneity: if the range is wide, the correlation will be greater than if the range is narrow. Studies where investigators compare two QIBs over a broad range of values may show a high correlation, but the two QIBs can be in poor agreement (4).

STATISTICAL MEASURES OF AGREEMENT FOR CONTINUOUS VARIABLES

Agreement refers to how well observer(s) produce similar values on repeated measurements in the same subjects. This parameter, therefore, mostly estimates measurement error inherent to the measurement process. Agreement parameters are particularly applicable when estimating changes over time or between different measurement systems (9,10,21). Statistical measures of agreement include Bland-Altman graphs (which determine bias and limits of agreement), SEM, and CV.

Agreement measures typically have the same unit as the property being measured, which is advantageous in understanding their clinical meaning. By contrast, reliability parameters are dimensionless, with value between 0 and 1.

Bland-Altman Graph and Analysis

Bland and Altman introduced their graphical method comparison system in 1986. Method comparison studies assess measurement reliability between two methods (eg, ultrasound elastography vs magnetic resonance elastography) that measure the same variable (eg, liver stiffness) on the same patients. Method comparison studies are usually performed prior to adoption of a new measurement method in clinical practice to ensure the new measurement method is “substantially equivalent” to the current method (12,22). Bland-Altman graphs and analysis may also be used to evaluate agreement between *two* observers (interobserver agreement) (23). To perform Bland-Altman analysis, two conditions should be satisfied: (1)

data should be continuous with the same unit of measurement used by observers or methods and (2) no more than two observers or methods are compared. For each pair of two continuous measurements (x_1 and x_2), a point is placed on the Bland-Altman graph, with y-axis coordinate of the point the difference between x_1 and x_2 , and x-axis coordinate the mean of x_1 and x_2 . The Bland-Altman graph provides a useful visualization of measurement spread as a function of measurement value, and also shows bias and limits of agreement (Fig 2). *Bias* is calculated as the mean of paired measurement differences of all patients by two observers or methods and represents systematic measurement error, which is the tendency for one of the two observers or methods to underestimate or overestimate the measurement relative to the other observer. This is represented on the Bland-Altman graph as a continuous line, parallel to the x-axis; the closer it is to 0, the less bias. The *limits of agreement* give a range wherein 95% of the measurement differences between the two observers are captured (Eq. 4). This can be interpreted as a measure of the expected random error between two observers. Limits of agreement are depicted as dotted lines on Bland-Altman graphs. Narrower width between dotted lines implies greater agreement between measurements. The acceptability of specific limits of agreement is determined by clinical goals, not statistical factors (24):

$$\text{limit of agreement} = \frac{\sum(x_1 - x_2)}{n} \pm 1.96\sigma_{\text{diff}} \quad (\text{Eq. 4})$$

where x_1 is observer 1, x_2 is the other observer, n is the number of patients, and σ_{diff} is the SD of the difference of measurements between observers. For **Example 2**, using the Example 1 data set, we have drawn a Bland-Altman graph using MedCalc for Windows, version 12.4.0.0. This illustrates agreement between two observers and depicts bias of 0.5 and limits of agreement of (-9.9 to 11.0) (19) (Fig 2).

Standard Error of Measurement

As there is no perfect measurement instrument, the true value of a measurement is almost always unknown. SEM is an agreement measure, which estimates how repeated measurements on the same patients with the same measurement instrument by different observers tend to be distributed around the patients' "true" value. The relationship between SEM (a measure of agreement) and ICC (a measure of reliability) is shown by Equation 5, where σ is the SD of repeated measurements; the larger the SEM, the lower the test reliability (10).

$$\text{SEM} = \sigma \sqrt{1 - \text{ICC}} \quad (\text{Eq. 5})$$

Coefficient of Variation

CV is a measure of the agreement of a measurement method when numerous repeated measurements are performed on the same patients by multiple different observers (Eq. 6), where σ is the SD of repeated measurements and μ is the mean of repeated measurement.

$$CV = \frac{\sigma \text{ of repeated measurements}}{\mu \text{ of repeated measurements}} \quad (\text{Eq. 6})$$

CV is generally used for calibration of a measurement instrument or for determining the reference range of a continuous variable measurement and expressed as a percentage (%) (24). The higher the CV, the greater the measurement dispersion. The advantage of CV is that the SD of measurements generally increases or decreases proportionally as the mean of the measurement increases or decreases, so division by the mean renders CV dimensionless (25). CV can be used with continuous variables that have a real zero and only positive values (4,24).

CONCLUSION

A measurement must be both reliable and valid to be clinically applicable. Reliability and agreement are different but related concepts analyzed by different statistical tests. The most commonly used parameter for the assessment of QIB intraand interobserver reliability is the ICC. Agreement parameters for continuous data between two observers include limits of agreement defined through Bland-Altman analysis. In some situations, SEM and CV are also appropriate measures of agreement.

Acknowledgments

Funded by: NIH – National Institutes of Health (grant numbers: K23 EB020710; HHSN268201300071 C).

References

1. Buckler AJ, Bresolin L, Dunnick NR, et al. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology*. 2011; 258:906–914. [PubMed: 21339352]
2. Buckler, AJ., Ouellette, M., Danagouliau, J., et al. *J Digit Imaging*. Vol. 26. Springer; US: 2013. Quantitative Imaging Biomarker Ontology (QIBO) for knowledge representation of biomedical imaging biomarkers; p. 630-641.
3. Sullivan DC, Bresolin L, Seto B, et al. Introduction to metrology series. *Stat Methods Med Res*. 2015; 24:3–8. [PubMed: 25804285]
4. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat Methods Med Res*. 2014; 24:9–26. 0962280214537333. [PubMed: 24919826]
5. Horn GL Jr, Hahn PF, Tabatabaei S, et al. A practical primer on PI-RADS version 2: a pictorial essay. *Abdom Radiol (NY)*. 2016; 41:899–906. [PubMed: 26984751]
6. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976; 38:388–394. [PubMed: 947531]
7. Kottner, J., Gajewski, BJ., Streiner, DL. *Int J Nurs Stud*. Vol. 48. Elsevier Ltd: 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS); p. 659-660.
8. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010; 10:22. [PubMed: 20298572]
9. Lucas, NP., Macaskill, P., Irwig, L., et al. *J Clin Epidemiol*. Vol. 63. Elsevier Inc; 2010. The development of a Quality Appraisal Tool for Studies of Diagnostic Reliability (QAREL); p. 854-861.

10. de Vet HCW, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006; 59:1033–1039. [PubMed: 16980142]
11. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*. 1987; 40:171–178. [PubMed: 3818871]
12. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol*. 2008; 31:466–475. [PubMed: 18306169]
13. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res*. 2015; 24:27–67. [PubMed: 24919831]
14. de Vet, HCW., Terwee, CB., Mokkink, LB., et al. *Measurement in Medicine*. United Kingdom: Cambridge University Press; 2011.
15. Kottner, J., Audigé, L., Brorson, S., et al. *J Clin Epidemiol*. Vol. 64. Elsevier Inc; 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed; p. 96-106.
16. Tinsley, HEA., Brown, SD. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. San Diego: Academic Press; 2000. p. 1
17. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res*. 2004; 13
18. Fleiss, JL. *Design and analysis of clinical experiments*. John Wiley & Sons; 2011. p. 1
19. Anvari A, Dhyani M, Stephen AE, et al. Reliability of shear-wave elastography estimates of the Young modulus of tissue in follicular thyroid neoplasms. *AJR Am J Roentgenol*. 2016; 206:609–616. [PubMed: 26901019]
20. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol*. 2003; 22:85–93. [PubMed: 12858311]
21. Anvari A, Halpern EF, Samir AE. Statistics 101 for radiologists. *Radiographics*. 2015; 35:1789–1801. [PubMed: 26466186]
22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1:307–310. [PubMed: 2868172]
23. Zaki, R., Bulgiba, A., Ismail, R., Ismail, NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. In: Rapallo, F., editor. *PLoS ONE*. Vol. 7. 2012. p. e37908
24. (a) Euser AM, Dekker FW le, Cessie S. A practical approach to Bland-Altman plots and variation coefficients for log transformed variables. *J Clin Epidemiol*. 2008; 61:978–982. [PubMed: 18468854] (b) Giavarina D. Understanding Bland Altman analysis. *Biochemia Medica*. 2015; 25(2):141–151. [PubMed: 26110027]
25. Reed GF, Lynn F, Meade BD. Use of coefficient of variation in assessing variability of quantitative assays. *Clin Vaccine Immunol*. 2002; 9:1235–1239.

SUMMARY STATEMENTS

1. Measurement is never perfect and variability is intrinsic to all measurement processes.
2. In imaging science, measurement variability may arise from use of differing imaging modalities (eg, ultrasound vs CT), different imaging acquisition algorithms within a single modality, protocol, or equipment variation, differences in observer expertise, and biologic variation.
3. Reliability and agreement should NOT be used interchangeably; *reliability* refers to how well a measurement instrument can distinguish patients in a group from each other, and *agreement* is how well a measure produces the exact same value on repeated measurements in the same patient by one or more observers.
4. ICCs are the best-known reliability parameters for repeated measurements of continuous values.
5. Bland-Altman graph and analysis is the most common statistical test for expressing agreement for continuous values between two observers (interobserver agreement) or two measurement methods.

Intraclass Correlation Coefficient, intra-rater 1

A

	Intraclass Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Single Measures	.851	.726	.921
Average Measures	.919	.841	.959

Intraclass Correlation Coefficient, intra-rater 2

B

	Intraclass Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Single Measures	.862	.745	.928
Average Measures	.926	.854	.962

Intraclass Correlation Coefficient, inter-rater

C

	Intraclass Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Single Measures	.872	.763	.933
Average Measures	.932	.865	.965

Figure 1.

Intraclass correlation coefficients of elasticity values in 35 patients with thyroid nodules in transverse plane for intra-rater 1 (a), intra-rater 2 (b), and inter-rater (c) (data of Table 1).

Figure 2.

Bland-Altman graph stiffness values in 35 patients with thyroid nodules in transverse plane between rater 1 and rater 2. Bias is shown as continuous dark blue line, limits of agreement are shown as the dotted red lines on the graph, and each hollow circle represents the subject that is measured by raters (data of Table 1). (Color version of figure is available online.)

TABLE 1

Quantitative measurements of tissue stiffness values (kPa) in 35 patients with benign and malignant thyroid nodules

	Elasticity Value			
	<i>Rater 1</i>		<i>Rater 2</i>	
	<i>Occasion 1</i>	<i>Occasion 2</i>	<i>Occasion 1</i>	<i>Occasion 2</i>
Patient 1	13.56	19.58	18.24	17.9
Patient 2	13.27	13.45	13.55	13.98
Patient 3	30.22	27.79	33.88	29.79
Patient 4	10.86	11.06	11.75	11.35
Patient 5	46.48	50.46	43.45	44.48
Patient 6	12.42	12.04	12.81	13.59
Patient 7	31.73	17.92	28.51	32.15
Patient 8	10.48	7.16	11.19	8.03
Patient 9	26.15	16.88	26.24	25.38
Patient 10	13.06	17.46	16.81	16.45
Patient 11	9.45	9.39	9.76	9.62
Patient 12	21.64	21.24	23.63	16.76
Patient 13	18.33	17.64	18.66	18.5
Patient 14	8.7	17.34	12.52	9.37
Patient 15	21	21.61	19.05	27.24
Patient 16	33.42	35.16	26.46	32.45
Patient 17	30.28	28.92	32.69	27.74
Patient 18	38.07	43.06	30.12	45.96
Patient 19	11.64	26.26	9.03	25.57
Patient 20	45.12	47.66	46.92	43.31
Patient 21	15.35	15.02	18.12	15.84
Patient 22	15.29	14.52	13.98	14.41
Patient 23	17.53	18.26	20.83	20.68
Patient 24	25.43	25.25	23.42	27.96
Patient 25	16.48	14.83	15.09	18.09
Patient 26	17.38	15.14	16.86	16.37
Patient 27	22.97	11.04	14.19	18.39
Patient 28	24.6	22.4	20	22.1
Patient 29	46.1	32.5	24.8	20.9
Patient 30	17	16.9	17.8	18
Patient 31	7.5	7.7	8.9	8.5
Patient 32	28.1	30	27.3	30.4
Patient 33	31.4	22.7	44.3	38
Patient 34	15.2	22.7	14.6	23.7
Patient 35	14.2	18.9	15.8	17