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

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

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.

Key Words: 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)

Acad Radiol 2017; ■:■■–■■

From the Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114. Received March 3, 2017; revised September 8, 2017; accepted September 9, 2017. Funded by: NIH – National Institutes of Health (grant numbers: K23 EB020710; HHSN268201300071 C). **Address correspondence to:** A.A. e-mail: anvari.arash@mgh.harvard.edu

https://doi.org/10.1016/j.acra.2017.09.010

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