OPEN

Variability and Standardization of Quantitative Imaging
Monoparametric to Multiparametric Quantification, Radiomics, and Artificial Intelligence

Akifumi Hagiwara, MD, PhD, Shohei Fujita, MD,*† Yoshiharu Ohno, MD, PhD,‡ and Shigeki Aoki, MD, PhD**

Abstract: Radiological images have been assessed qualitatively in most clinical settings by the expert eyes of radiologists and other clinicians. On the other hand, quantification of radiological images has the potential to detect early disease that may be difficult to detect with human eyes, complement or replace biopsy, and provide clear differentiation of disease stage. Further, objective assessment by quantification is a prerequisite of personalized/precision medicine. This review article aims to summarize and discuss how the variability of quantitative values derived from radiological images are induced by a number of factors and how these variabilities are mitigated and standardization of the quantitative values are achieved. We discuss the variabilities of specific biomarkers derived from magnetic resonance imaging and computed tomography, and focus on diffusion-weighted imaging, relaxometry, lung density evaluation, and computer-aided computed tomography volumetry. We also review the sources of variability and current efforts of standardization of the rapidly evolving techniques, which include radiomics and artificial intelligence.

Key Words: quantitative imaging, quantitative imaging biomarker alliance, standardization, diffusion-weighted imaging, synthetic MRI, magnetic resonance fingerprinting, chest CT, radiomics, artificial intelligence, deep learning

(*Invest Radiol* 2020;55: 601–616)

Quantitative imaging, defined as the extraction of quantifiable features from radiological images,[1] has been increasingly performed for the measurement of normal biological and pathological processes, patient risk stratification, evaluation of treatment response and outcome, and drug development.[2] Such features of quantitative imaging in clinical settings are called biomarkers (quantitative imaging biomarkers [QIBs]), which is a characteristic that is objectively measured and evaluated.[3] Although the term "biomarker" is often meant to imply a measurand (the true value of the quantity intended to be measured) of laboratory assays, such as blood sugar tests, it can also denote clinical measurands such as blood pressure and metrics obtained with quantitative imaging. Quantitative imaging biomarkers are continuous variables, whereas ordinal variables, such as the PI-RADS (Prostate Imaging Reporting and Data System) with 5 numbered categories for assessment of prostate carcinoma,[4] are not considered to be QIBs.[5] Biomarkers are important in healthcare for a physician to determine the most appropriate management for a patient's unique state of disease at the molecular level. This concept is called personalized or precision medicine. A biopsied specimen is only a small fraction of the entire tissue that is sampled at a certain time point, and spatial/temporal sampling biases are not negligible.[1] On the other hand, QIB covers a wide segment or the whole of a subject and can provide more comprehensive spatial information concerning the tissues. Repetitive sampling is also much easier for imaging than a biopsy, and imaging data can be dynamically obtained in some cases in the order of seconds to milliseconds.[6,7] In addition, QIBs may enable the detection of a subclinical presentation of disease that is too subtle to be detected by human eyes[8]; this leads to a better outcome for patients than when disease is detected after the clinical presentation is recognized. Reliable QIBs can also help foster the development of medical products in regulatory settings.[9] For example, if a QIB is qualified by the Food and Drug Administration for drug development, it could help deliver a new therapy to the public through either a traditional or accelerated approval pathway.[10]

In addition to the clinical relevance and sensitivity to the disease process, good reproducibility is the key element of a qualified biomarker.[11] Although QIBs can be used similarly to laboratory assays, its clinical application has been hindered by its generally lower reproducibility. This is partly because the extraction of most QIBs from radiological images is not yet fully automated, and it requires a radiologist or other experienced practitioner to engage in the analysis process, which introduces an inevitable variability arising from human perception.[12] Further, the variability of a QIB is also derived from acquisition hardware, software, procedures, operators, and the measurement methods. The Quantitative Imaging Biomarkers Alliance (QIBA) was established by the Radiological Society of North America (RSNA) in 2007 to proceed quantitative imaging and introduce the use of QIBs in clinical trials and practice by engaging researchers, healthcare professionals, and the industry (https://www.rsna.org/en/research/quantitative-imaging-biomarkers-alliance). The mission of QIBA is to improve the value and practicability of QIB by reducing variability across devices/sites, patients, and time. The QIBA has been developing QIBA Profiles that standardize methods for each selected QIB to achieve a useful level of performance.[13] Claims written in the Profiles describe the performance of the QIB and focus on a quantitative interpretation of the measurements for the individual subject. Conformance to the specifications of a Profile is required not only for hardware, software, and analysis methods, but also for operators and analysts. In collaboration with QIBA, the Japan Radiological Society and European Society of Radiology have also established Japan QIBA (J-QIBA) (http://www.radiology.jp/j-qiba/english/index.html) and European Imaging Biomarker Alliance, respectively, both of which have the same goal.

This review article aims to summarize and discuss how the variability of quantitative values derived from radiological images are induced by a number of factors and how variabilities are mitigated and standardization of the quantitative values are achieved. For the interpretation of studies related to evaluating the performance of QIBs, terminology and key statistics will be explained. We also discuss the variabilities of specific biomarkers derived from magnetic resonance imaging (MRI) and computed tomography (CT). Further, we review the sources of variability and current standardization efforts for rapidly evolving techniques,

Received for publication January 20, 2020; and accepted for publication, after revision, January 28, 2020.

From the *Department of Radiology, Juntendo University School of Medicine, Tokyo; †Department of Radiology, Graduate School of Medicine, The University of Tokyo, Tokyo; and ‡Department of Radiology, Fujita Health University School of Medicine, Toyoake, Aichi, Japan.

Conflicts of interest and sources of funding: We have no conflict of interest to declare. This work was supported by AMED under grant number JP19lk1010025h9902; JSPS KAKENHI grant number 19K17150, 19K17177, 18H02772, and JP16H06280; Health, Labour and Welfare Policy Research Grants for Research on Region Medical; and a Grant-in-Aid for Special Research in Subsidies for ordinary expenses of private schools from The Promotion and Mutual Aid Corporation for Private Schools of Japan.

Correspondence to: Akifumi Hagiwara, MD, PhD, Department of Radiology, Juntendo University School of Medicine, 1-2-1, Hongo, Bunkyo-ku, Tokyo, Japan, 113-8421. E-mail: a-hagiwara@juntendo.ac.jp.

ISSN: 0020-9996/20/5509–0601
DOI: 10.1097/RLI.0000000000000666

including radiomics and artificial intelligence (AI). Overall, the terminologies related to variability used in this article conform to those suggested by the QIBA Terminology Working Group in 2015.[9]

STATISTICS

In this section, we explain the statistics and related terminology to understand the literature describing the performance of QIBs and to help readers conduct an appropriate evaluation of a QIB by themselves.

Terminology

For a QIB to be clinically useful, it is desired to be reliably comparable to known reference measurements or true value, and it must be comparable to one another in the same subjects for repetitive measurements.[14] These properties of a QIB can be characterized by *accuracy* (systematic measurement error or bias) and *precision* (random measurement error), which are together called *uncertainty*.[9] Measurement bias can be estimated only when the true value is known, and it can be mitigated by improving the calibration of a measurement system. The term *measurand* refers to the true value of the quantity intended to be measured.[15]

Accuracy commonly describes a range of characteristics including how a measured value relates to a known reference. *Accuracy* in terms of quantitative imaging usually consists of *linearity* and *bias*. *Linearity* is the ability of a measurement to provide a directly proportional value to the measurand or a known reference. *Bias* is an estimate of systemic measurement error; it describes the difference between the average of a measurement made on a subject and its true value or known reference.

Although *reliability, agreement, precision, repeatability,* and *reproducibility* are often used interchangeably, these terms are distinctive.[9,15] *Reliability* is defined as the ratio of variance based on the between-subject measurement to total variance based on the observed measurement. In other words, reliability represents how well different subjects can be distinguished from each other despite a QIB's uncertainty or measurement errors. *Reliability* is typically assessed by an intraclass correlation coefficient (ICC). *Agreement* has a broader meaning than reliability and indicates the degree of closeness between measurements made on the same subject by different observers or measurement methods. *Precision* or *repeatability* represents the closeness between measured values obtained by replicate measurements of the same subject with the same measurement method under the identical or near-identical conditions—including subject, measurement procedure, environment, and scanner—over a short period. Repeatability studies are often referred to as test-retest or scan-rescan.

Reproducibility describes the closeness of measurements under a set of conditions that includes different locations, operators, measuring systems, or replicate measurements on the same or similar subjects. These conditions are analogous to real clinical practice where various external factors cannot be tightly regulated.

Linearity

Linearity can be evaluated by regressing the measurements (Y values) on the true values (X values). A linear model can be fit by least squares as:

$$Y = \beta_0 + \beta_1 X$$

where β_0 is the intercept and β_1 is the slope. If the relationship between Y and X is well explained by a line (ie, $R^2 > 0.90$), then the assumption of linearity is met.[16] Although linearity is the ideal condition, monotonic relation (ie, the relationship of a QIB and the measurand can be described as a strictly increasing or decreasing function) is necessary and generally sufficient for a QIB to be clinically useful; it does so by discriminating every distinct value of measurands.[9] However, the slope of a function is related to sensitivity. If the relationship between Y and X is nonlinear, the ability of a measurement to detect change in the measurand is inconstant.[15] If there is lack of a standard reference, another imaging measurement method for which proportionality is established can be used as the reference standard to evaluate the new imaging measurement method.

Bias

Bias is the difference between the sampled mean and true value or known reference. %Bias is calculated by dividing the bias by the true value or known reference. If the true value or known reference is unavailable, bias cannot be evaluated. Hence, bias is typically calculated using validated phantoms with a well-defined reference. At least 5 to 7 similarly spaced values over the relevant range of true values should be chosen.[15] If the data are from various cohorts and the bias is inconsistent, a bias profile should be reported rather than a single bias value.[16] For example, bias for tumors with different sizes, shapes, and densities can be reported as a bias profile for CT volumetry.[17] Inconstant biases should be specified cautiously, especially when assessment of change in the QIB is the focus; this is because the different biases do not cancel out in calculating the change.[15] In this case, transformations, such as log-transformation, may render an inconstant bias constant. The assessments of linearity and bias are directly linked to each other; both should be presented when assessing either for the technical performance of a QIB.

Precision (Repeatability)

Precision, or repeatability, is concerned with whether a measurement agrees with a second measurement of the same quantity; high precision is a good indicator of the ability of a QIB to reveal an effect of treatment, identify disease, or discriminate between groups using the same scanner, sequence, software, and analysis method. Precision can be assessed by repeated imaging of a phantom, although it does not perfectly reflect the real clinical situation. When precision is assessed by repeated imaging of human subjects, the variance in measurements can be contaminated by subject-related variability due to a variety of reasons including behavioral, physiological, and psychological factors that may have changed between scans, even if the actual process of imaging acquisition remains unchanged. Further, it may be ethically inappropriate to scan subjects with repeated doses of radiation or with the use of contrast agents or tracers. A washout period also needs to be considered before a rescan if a contrast agent or tracer is used.

Precision can be expressed numerically by measures of variability such as within-subject standard deviation (wSD), within-subject coefficient of variation (wCV), or 95% precision limit.[9] The wSD represents the standard deviation of measurements from the same or similar subjects under specified conditions. The wCV for repeated measurements of a subject is the wSD divided by the mean. The wCV as a group is typically acquired by taking the square root of the mean of wCV^2 per subject. Only precision, not biological variation, is recommended to be included when reporting the performance of a QIB. Within-subject variance may also arise from patient repositioning and scanner calibrations. If the precision varies over a range of relevant magnitudes in the measurands, a precision profile should be considered; it should be reported as a table or plot showing estimates of precision—possibly stratified by one or more variables affecting the precision. The 95% precision limit is calculated as the repeatability coefficient (RC) or % repeatability coefficient (%RC). The standard deviation of the difference between 2 repeated measurements is $\sqrt{2}$wSD. Repeatability coefficient is the least significant difference between 2 repeated measurements at a 2-sided significance of $\alpha = 0.05$, and it is calculated as[15]:

$$\mathrm{RC} = 1.96\sqrt{2}\ \mathrm{wSD} = 2.77\ \mathrm{wSD}$$

Likewise, %RC is calculated as:

$$\%\mathrm{RC} = 1.96\sqrt{2}\ \mathrm{wCV} = 2.77\ \mathrm{wCV}$$

The limit of agreement (LOA), the interval containing 95% differences between repeated measurements on the same subjects, is −RC

to +RC. It represents the minimum detectable difference in 2 measurements with 95% confidence. A meta-analysis of the literature can summarize an RC by taking a weighted average of the reported values.[16]

Bland-Altman Graph Analysis

The Bland-Altman plot provides a graphic representation of agreement in addition to the 95% LOA.[18] The 95% LOA is the interval that is expected to contain 95% of differences between the measurement and true value or the other measurement, and it is calculated using the standard deviation of the difference. The Bland-Altman plot illustrates the differences between a measurement method and another one, or the true value, plotted against their mean. If the true value is used, one may plot the differences against the true value instead of their mean. The differences can also be expressed as percentages, which is useful when the variability of the difference increases as the magnitude of the measurement increases. The Bland-Altman plot also helps to demonstrate the relationship between bias and variance.

Intraclass Correlation Coefficient

Instead of reporting the components of uncertainty (eg, bias and precision) in a separate manner, ICC can also be used to summarize the uncertainty.[9] Intraclass correlation coefficient considers both the within-subject variance originating from measurement error and variance originating from the difference between subjects.[19] The ICC is the fraction of the total variance that is attributed to the subjects and is calculated as:

$$\mathrm{ICC} = \frac{\text{Between}-\text{Subject variance}}{\text{Between}-\text{Subject variance} + \text{Variance from measurement error}}$$

If the measurement error is small compared with the true variance between subjects, ICC approaches 1. Although subjective, adjectives to describe ranges of ICC values include the following: poor (0 to 0.5), moderate (0.5 to 0.75), substantial (0.75 to 0.9), and excellent (0.9 to 1).[20] A moderate ICC can be considered sufficient when a measurement is used for group-level comparisons for research purposes. However, if a measurement is used in individual patients for important clinical decisions, an excellent ICC is required.[21] Intraclass correlation coefficient can help us stop being excessively concerned about measurement error when between-subject variance is large. However, ICC depends on the subject population being studied,[22,23] and ICC calculated for a group of subjects may not be applicable to another population. For example, when ICC is calculated for a group of healthy subjects, it may become unacceptably low because a group of healthy subjects tends to be homogenous and biological variance is low. However, ICC may be acceptable when calculated for a group of patients, which may typically be more heterogenous than a group of healthy subjects. Further, when we assess the subtle differences within a subject (eg, evaluating treatment response), ICC is often impractical. In this case, precision reported by RC would be more suitable, as the RC shows the smallest within-subject change that can be reliably detected.

Pearson Correlation Coefficient

The Pearson correlation coefficient has been frequently used to compare repeated measurements or a new measurement technique with the old one. However, this approach only evaluates the linear association between 2 measurements without consideration of bias, and it does not give an indication of repeatability or agreement.[18] Further, a large between-subject variation makes the correlation coefficient higher. High correlation coefficients may be achieved for 2 QIBs with wide ranges, even when they are in poor agreement, such as when one is twice the size of the other.

Reproducibility Coefficient and Multicenter Study

The reproducibility coefficient (RDC) is a measure of precision that is used when scanners, imaging procedures, location, operator, analysts, and/or algorithms differ at 2 time points. It shows the minimum detectable difference between 2 repeated measurements performed under different conditions with a 95% confidence, and it can be measured directly from clinical studies.[16] Just like RC and %RC, RDC and %RDC are calculated as 2.77 wSD and 2.77 wCV, respectively, under different rather than unchanged imaging acquisitions. An example of a reproducibility study may compare the volume of an organ measured by CT with that measured by MRI for the same organ of the same subject.

Reproducibility is especially important in multicenter studies where reproduction of the same measurement is required across different centers and often with different kinds of scanners.[24] Multicenter studies of human subjects enable a comprehensive investigation of the disease. This is an advantage, especially when the disease is rare. However, if reproducibility across scanners is low, variability across scanners may reduce the statistical power of detecting differences between groups and annul the benefit of using data from multiple centers.[25] Mitigation of variation across centers can be achieved by (1) setting scanning and analysis procedures to be as identical as possible so that any systemic errors are replicated across participating centers[26] or (2) aiming for high accuracy at each center.[27] Although standardizing the scanning protocol is the simplest method for reducing measurement variabilities, differences in the scanners produced by different vendors may prevent identical protocols from being used at every site. In a cross-sectional study that compares groups, all groups should be included at each center, and the effect of the center should be added as a covariate in the statistical analysis.[11]

Meta-analysis of Technical Performance Studies

Before a QIB is accepted for clinical use, performance metrics, such as repeatability and reproducibility, should be evaluated. Ideally, this evaluation should involve summaries from multiple studies to overcome any limitations arising from a small sample size (typically 10 to 20 subjects) of a study concerning technical performance and include a wider range of relevant clinical settings and patient populations.[28] Although a meta-analysis of any technical performance metric is theoretically feasible, a meta-analysis of reproducibility and agreement is more complicated than that of repeatability because the studies that assess reproducibility and agreement are more heterogenous than those of repeatability. For example, a reproducibility study can be performed using scanners of the same type across different sites, scanners of different types from the same vendor, or different scanners from multiple vendors. Generally, reproducibility of the measurement decreases in this order.

VARIABILITY SOURCES, STANDARDIZATION, AND HARMONIZATION

This section focuses on the variability sources common to QIBs (Fig. 1) and how these variabilities could be mitigated. Variability sources specific to the modality or each biomarker will be discussed later in the corresponding section. The degree of measurement imperfections in comparison to the pathophysiological changes due to disease determines the significance of measurement imperfections for each QIB and hence the amount of effort required to be taken to reduce such variabilities. This effort may include building and keeping quality assurance (QA) at each center and improving the acquisition/analytical method. For example, the MAGNIMS (magnetic resonance in multiple sclerosis) research group has led a number of multicenter studies on MS, which occasionally included MRI physicists traveling to different centers in Europe, sometimes with a phantom, to decrease the measurement variability.[11]

Patient Positioning and Movement

The operator should be trained for adequate and consistent positioning of the phantoms/human subjects. Movement of the subject during and between scan sequences can cause artifacts and degrade the image quality. This can be mitigated by paying careful attention to the

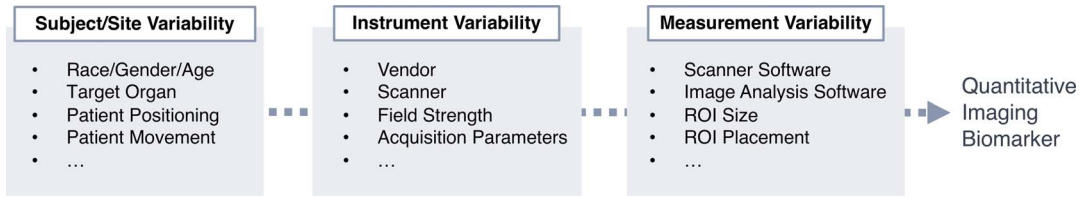

FIGURE 1. Variability sources of quantitative imaging biomarkers.

comfort level of the subject. Involuntary movement, such as respiratory, cardiac, and gut motion, can also cause degradation of the image.

Region of Interest

The size of the region of interest (ROI) has been known to affect the repeatability and reproducibility of QIBs.[29] A QIB is often measured as the mean value of a map within an ROI, where an increase in the size of the ROI leads to a smaller variance (ie, higher repeatability/reproducibility) of the measurement. This is important when the ROI size can be variable due to treatment response or disease progression, when monitoring the effect of treatment on lesions such as tumors,[30] or when the ROI size is small such as in the case of measuring multiple sclerosis focal plaques.[31–33] The appropriate selection of an ROI size and estimation of size effect would help adjust the decision threshold by a QIB in monitoring a treatment effect.[29] The ROI placement procedure can also be variable among radiologists. Before starting a clinical trial involving a number of radiologists, especially when they are from different sites, the variability across them should be assessed and desirably standardized. Region of interest placement using automated techniques, such as deep learning, is a possible approach[34] that reduces the burden of clinicians and may increase both repeatability and reproducibility.

Observer

If an analysis (eg, ROI placement) involves observers, observers should be trained to a set of well-defined rules. Agreement across observers should also be assessed using ICC. There is a possibility of a practice effect, so observers should perform the actual analysis after reaching the plateau of their learning curve.[5] Software for semiautomated or fully automated analyses will increase the repeatability. Automated techniques using deep learning trained with a large dataset is expected to reduce the variability in tasks such as tumor segmentation.[35]

Hardware and Software Upgrades

Hardware and scanner software upgrades may introduce more bias and/or less precision in the derived QIB.[36] For example, Keenan et al[36] showed that the variable flip angle (VFA) T1 measurements on upgraded systems (hardware and software) had an overestimation of approximately 18% compared with the measurements of the original system (Fig. 2). Lee et al[37] also showed that a consistent bias of up to 3% was observed between VFA T1 measurements before and after a scanner software upgrade. Even when performing a study that uses only a single scanner, consistent versions should be used for both hardware and software.

Standardization

The performance of QIBs can be assessed with the true value (eg, phantoms, digital reference objects [DROs], simulation, and test-retest datasets—assuming no change), a reference standard, or without a reference standard (eg, agreement studies between algorithms and studies of algorithm precision). Phantoms and simulation data are cost-effective and reliable, and can be in large amount. Phantoms can be scanned repeatedly without any ethical constraints and are relatively easy to transport between centers. However, one must bear in mind that optimization of a QIB to a phantom or simulation data may not work well on in vivo data, due to the lack of realism. For example, pulmonary nodules in a phantom have several characteristics that may differ from human pulmonary nodules, including sharp margins, smooth surfaces, elemental shapes (spheroids and conics), homogeneous density, no vascular interaction, and no motion artifact. An algorithm that is optimized for any of these properties may appear to have an overly optimistic performance and may not show high performance for real in vivo data.

In human studies assessing QIB performance, the true value is often unavailable. Although histology or pathology tests are usually

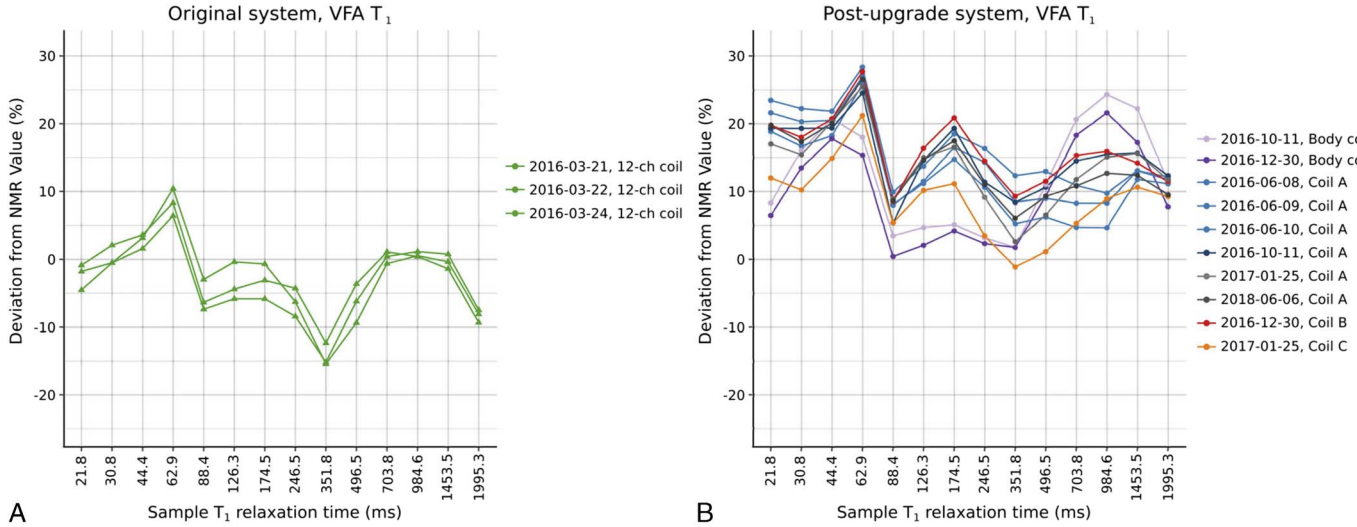

FIGURE 2. Variable flip angle measurement of T1 relaxation time on the original and postupgrade systems. Following upgrade, variable flip angle overestimates the T1 compared with the original measurements. Postupgrade measurements were completed using 3 different head coils (coils A–C) and the body coil (reproduced with permission from Keenan et al[36]).

considered as the true value, these are more appropriately referred to as reference standards; they are described as well-accepted or commonly used methods for measuring a biomarker but also have associated bias and/or random error.[38] For example, histology and pathology are known to be affected by fixation and staining, spatial and temporal sampling errors due to heterogeneity in tissue and the difference in time between imaging and sampling, the nonquantitative nature of the histopathology examinations, and subjective interpretations by humans. Specialist markings (eg, setting a boundary of a region for volumetry) are also sometimes considered as the true value, but they often have a variable degree of interreader variation and should be considered as a reference standard.[39]

Data Harmonization

Although a multicenter study has the potential to increase statistical power, the inclusion of different scanner vendors, acquisition protocols, image reconstruction algorithms, and field strengths results in unwanted systematic variation. Data harmonization aims to remove these variations retrospectively after acquisition while preserving the biological variability. Harmonization can be performed using traveling human data acquired at each site by determining a scanner-specific correction factor.[40–42] If only postprocessed data (eg, fractional anisotropy map and cortical thickness) are available, regression analysis or more sophisticated statistical approaches can be performed for harmonization.[43,44] Harmonization of raw data is particularly important for diffusion-weighted imaging (DWI) data to be analyzed by multicompartment models or tractography, and model-free methods for harmonization of raw DWI signals have also been suggested.[45,46] Harmonization by deep learning has also been proposed, but the algorithm should be trained on the data of same subjects acquired on different scanners that are intended to be harmonized.[47]

MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging can extract a variety of quantitative tissue properties that include not only length and volume but also relaxation properties (T1, T2, and T2*), diffusion, perfusion, phase, fat fraction, temperature, tissue chemical properties (eg, spectroscopy and chemical exchange saturation transfer), and physical properties (eg, elastography).[48] However, a large number of variabilities in image acquisition methods and postprocessing algorithms hinder the extraction of accurate and reproducible quantitative information from MRI. In this section, we discuss sources of variability in QIBs that are specific to MRI and the importance of periodic QA to maintain sufficient accuracy and precision. We also discuss the current body of knowledge regarding the standardization of quantitative MRI metrics that are fundamental in MRI.

Temperature

Temperature control is required for phantom scanning. T1 and T2 of Ni-DTPA were reported to change 0.2% to 1% and approximately 1.3% to 1.5%, respectively, per °C at a temperature approximately 21°C.[49] The apparent diffusion coefficient (ADC) of the pure water changes approximately 3% per °C at room temperature.[50] The phantom should be stored in the MRI room and reach a temperature close to that in the magnet bore so that the temperature of the phantom does not fluctuate during the scan. The temperature of the phantom should be measured and recorded after the scan is complete. Conversion of the acquired values to those at a standard temperature might be possible.[11] In case of a human scan, temperature control is assumed to be unnecessary because homeostasis provides intrinsic temperature control. However, core temperature can increase more than 1°C by MRI scan, especially at 3 T for obese subjects, and this thermal effect on quantitative MRI remains to be investigated.

B_1 Field Nonuniformity

Nonuniformity in the radiofrequency transmit field (B_1^+) is the major cause of error in quantitative MRI, especially when using high magnetic fields and surface coils for transmission.[50] Body coil excitation is preferable for uniform transfmission.[26] Calibration of the transmitter output can be carried out periodically as part of routine maintenance. The accuracy of flip angle depends on the B_1^+ inhomogeneity at a given spatial location, which can be measured by B_1^+ mapping.[51] Notably, every vendor uses their own radiofrequency pulse shapes that lead to variability in flip angle, complicating comparison across scanners from different vendors. The acquired B_1^+ map can be used for the measurement of tissue parameters such as T1 and magnetization transfer to correct the achieved flip angle for the intended one.[52,53] The B_1^+ field is smooth compared with anatomical structures, even at high field strengths, so B_1^+ maps are often acquired at low resolution to spare acquisition time.[54]

With the increasing use of a large number of coils and parallel imaging, the receive sensitivity field (B_1^-) nonuniformity should be addressed. B_1^- nonuniformity used to be measured from a B_1^+ map based on the reciprocity principle ($B_1^+ = B_1^-$) if the excitation and receiving are done by the same coil. The reciprocity principle can still be used when different coils are used by performing an additional acquisition in which the transmit coil is used for receiving.[55] However, the reciprocity principle becomes less accurate at a field strength of 3 T or higher.[56,57] B_1^- nonuniformity affects the spatial distribution of image intensity and thereby any quantitative MRI, especially the measurement of proton density and absolute metabolite concentration. The receiver gain can be automatically set or changed during the prescan procedure, but it is desired to be fixed during the acquisition of the image series.

B_0 Field Nonuniformity

When an object is placed in the magnet, the magnetic susceptibility of the object alters the static magnetic field B_0 in the object slightly. The shim coil usually adjusts to obtain a spatially uniform B_0 distribution. However, for extended fields of view, observable deviations from uniformity and image degradation can occur in the periphery.[58] Spatially varying tissue susceptibility, especially at the air-tissue interface, can also induce B_0 field nonuniformity. This is one cause of a generally higher repeatability and reproducibility for in vitro phantom studies than for in vivo human studies. In human subjects, the ROIs have to be put on spatially variable places, and signal variability becomes higher as the pixel departs from isocenter of the magnet. Proton density fat fraction is vulnerable to B_0 field nonuniformity because differentiation between the phase shifts, due to B_0 nonuniformity and those due to chemical shift utilized for extracting fat signal, is difficult.[59]

Field Strength

Some tissue parameters, including ADC, diffusion tensor imaging, proton density, volume, and perfusion, are independent of field strength. However, a higher field strength may contribute to an increased signal-to-noise ratio. Other parameters, including T1, T2, and magnetization transfer, are dependent on field strength.[60]

Quality Assurance in Quantitative MRI

Quality assurance is an ongoing process of ensuring that the instrument continues to operate adequately.[61–63] To use a QIB in a clinical routine, regular QA on a weekly basis (possibly on a daily basis as an initial assessment) is required. Quality assurance for quantitative MRI can be performed in healthy controls and/or in phantoms. Phantoms have the advantage of providing accurate values and being stable and always available. Phantoms and analysis software are ideally developed specifically for each QIB to address some of the variabilities. Anthropomorphic phantoms have been developed for certain body parts including the breast,[64] prostate,[65] and brain,[66] considering the fact that spatial relationship between scan objects and the coil affects the patterns of field inhomogeneity. For example, the breast phantoms were developed partly because the previous phantoms were not physically compatible with a breast coil.[64] The properties of the phantom may vary over time

due to the instability of the material due to fungal invasion, chemical decay, evaporation, or contamination by water vapor. Temperature dependence is also a problem, whereas human temperature is homeostatically controlled.[11] Further, the realism of a phantom may not be sufficient because many potential sources of variability in vivo (eg, movement, positioning variability, and B_1 variation due to subject shape) are absent. Normal white matter can be a standard for some MRI parameters (eg, ADC or magnetization transfer) because the normal biological range is narrow. In a multicenter study, standardized QA procedures should be followed by all institutions to keep the acquired data as uniform as possible.

Computer-simulated phantoms, or DROs, can also be used to evaluate the performance of the propagation of error in quantitative MRI regarding error from both the measurement and bias of parameter constraints or assumptions, as well as that from noise. However, simulations often do not match measurements in vivo due to the negligence of biological effects different than those being simulated.[67]

Diffusion-Weighted Imaging

One of the most widely investigated MRI QIBs in clinical trials is the ADC derived from DWI, which is sensitive to the random motion of water molecules.[68] Although DWI is used clinically as a qualitative indicator of disease presence, ADC has been investigated in clinical trials for diagnosis, staging tumors, assessing treatment response, and predicting tumor aggressiveness. However, confidence in its use has not been fully established due to differences across scanners and populations, which hinders the use of ADC in the clinical workflow. The complexity of the tissue structures makes ADC dependent on a number of factors including pulse sequence construction,[69] acquisition parameters, modeling techniques, anatomic regions being evaluated, and the subject orientation with respect to the diffusion directions.[68,70–72] An example of systematic ADC variations arises from a scanner upgrade to a high-end machine that allows shorter echo time for improving DWI quality, which would shorten the diffusion time, lead to a possible decrease in the visibility of acute brain infarction, and increase in the measured ADC value (Fig. 3).[71,73] It is accepted that ADC is independent of field strengths,[74,75] although higher field strengths may be beneficial due to improvements in signal-to-noise ratio. Huo et al[76] reported lower variance in ADC measurement at 3 T compared with 1.5 T. Control of these variabilities may enable ADC to replace biopsy, such as for the differential diagnosis between tumor recurrence and necrosis.[77]

The presumption for using ADC in clinical practice for managing tumors is that treatment-associated change in the microenvironment precedes changes in the lesion size, thereby encouraging the use of ADC as a biomarker of treatment response.[78] Conformance to the specifications of the QIBA DWI Profile[13] by all relevant staff, scanner, and software involved in ADC acquisition/measurement supports the following claim: a measured change in the ADC of lesions in the brain,[79–81] liver,[82–85] prostate,[86–90] and breast[91,92] of 11%, 26%, 47%, and 13% (each denotes %RC), respectively, or larger indicates that a true change has occurred with 95% confidence. Due to the intrinsic dependence of the measured ADC on biophysical tissue properties, these claims are organ specific. Notably, the Profile requires usage of the same scanner and image acquisition parameters for baseline and subsequent measurements with periodic QA (Fig. 4). Estimation of reproducibility based on previous studies is more complicated than repeatability because the reproducibility condition is heterogenous among studies. When interscanner CV is evaluated, one should be careful if the scanners are from the same vendor or from different vendors as significant intervendor bias in ADC measurement of the brain has been reported with a %bias up to 7%.[93]

Before a multicenter trial, qualification of each site should be assessed according to the specific protocol for the site's ability to adopt a standardized acquisition protocol and image analysis. The performance of ADC in each site should be assessed by the ice-water DWI phantom.[50,94,95] Although substrates, such as sucrose,[96] alkane,[97] and copper sulfate,[98] have been used to achieve a wide range of ADC values, sensitivity of ADC values to temperature variation has been problematic; ADC of pure water changes approximately 3% per °C.[50] The ice-water phantom was designed to eliminate thermal variability and keep

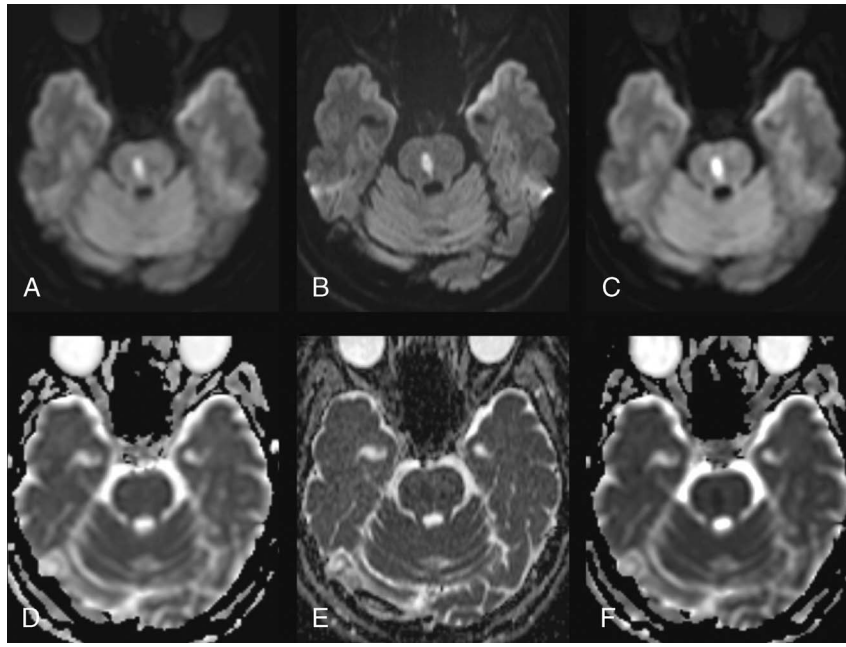

FIGURE 3. Apparent diffusion coefficient dependency on diffusion time. Diffusion-weighted imaging with short (A), intermediate (B), and long diffusion times (Δeff) (C) shows acute infarction at the right paramedian aspect of the pons, responsible for medial longitudinal fasciculus syndrome. Diffusion-weighted imaging with short Δeff (A) demonstrated decreased contrast of the lesion with the surrounding tissue compared with diffusion-weighted imaging with longer Δeff (B and C). D–F, Images show ADC maps of corresponding diffusion-weighted imaging. The apparent diffusion coefficient values of the lesion were increased with short Δeff (D) compared with long Δeff (F) (reproduced with permission from Boonrod et al[71]).

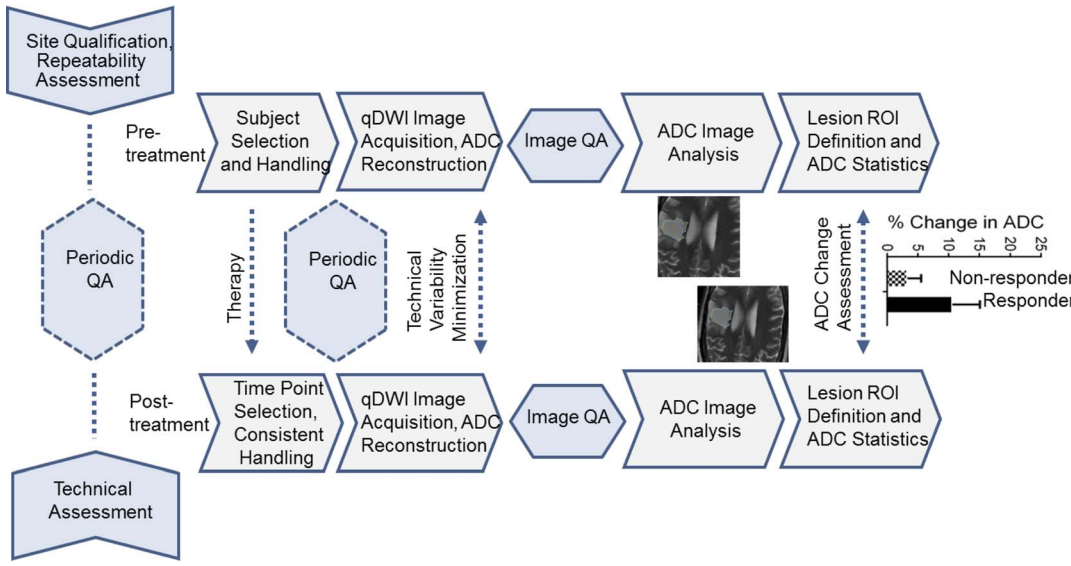

FIGURE 4. Typical quantitative diffusion-weighted magnetic resonance imaging trial workflow for treatment response assessment with key QIBA (Quantitative Imaging Biomarkers Alliance) Profile activities (reprinted with permission from QIBA RSNA diffusion-weighted imaging profile v:12.20.2019: https://qibawiki.rsna.org/index.php/Profiles).

the phantom at 0°C by filling the phantom with an ice-water bath. The inner tube is typically filled with distilled water,[94,95] giving an ADC of 1.1×10^{-3} mm^2/s, or alternatively, polyvinylpyrrolidone solutions with a range of ADC values.[99]

Magnetic Resonance Relaxometry

The signal intensity of conventional MR images, such as T1- and T2-weighted images, depends on many acquisition parameters and MR scanner variations. Thus, absolute signal intensity has no direct meaning, and the evaluation of MRI scans mainly involves comparison with surrounding tissues in the same slice. Absolute quantification of longitudinal relaxation time (T1), transverse relaxation time (T2) or their inverse relaxation rates (R1 and R2), and proton density (PD) provides an absolute scale; hence, it enables a more objective evaluation of development,[100] aging,[101] and diseases.[102]

Proton density indicates the amount of detectable protons by MRI and is proportional to the MRI signal intensity. Calculation of PD is based on the estimation of the magnetization at equilibrium (M_0), which represents the signal intensity in the absence of any relaxation.[103]

T1 relaxation time characterizes the approach of the polarized spins to equilibrium in the direction of the external magnetic field, and it is affected by a number of tissue properties including free water content, macromolecules, iron, and gadolinium chelate. T1 values significantly increase with the field strength.[60] The criterion standard method for measuring T1 relaxation time is the inversion recovery technique, in which only one echo is acquired at a time and full spin relaxation is awaited (approximately 5 T1 periods) before the next spin inversion. This technique is time-consuming and infeasible in clinical settings. To aim for time efficiency, 2 other techniques, namely, the Look-Locker (LL)[104] and VFA[105] techniques, were introduced. Stikov et al[52] compared inversion recovery, LL, and VFA techniques using a phantom and the brains of healthy volunteers. Although these techniques agreed well on the phantom, LA and VFA respectively showed consistently underestimated and overestimated T1 values measured by the inversion recovery technique. The deviations reached over 30% in WM, from 750 milliseconds (LL) to 1070 milliseconds (VFA). They found that major sources of differences were inaccurate B_1^+ mapping and incomplete spoiling of transverse magnetizations. Thus, they concluded that quality assessment of T1 mapping techniques should be performed both for a phantom and in vivo.

T2 relaxation time indicates the rate at which the transverse component of magnetization decays to zero, and it is primarily driven by nearby nuclei. By changing the echo time, T2 relaxation time can be measured by spin echo technique with as few as 2 measurements—assuming monoexponential decay; however, the measurement suffers from partial volume and is susceptible to noise.[106] Further, the acquisition time is very long because full T1 relaxation is required during the acquisitions. Multiecho T2 (MET_2) accelerates the spin echo method by using multiple refocusing pulses at increasing TE. A version of MET_2, termed the Carr-Purcell-Meiboom-Gill sequence, accelerates MET_2 by incorporating a 180-degree phase increment to refocusing pulses and is now considered to be the criterion standard for T2 measurement.[107,108] Another approach for T2 measurement is driven equilibrium single-pulse observation of T2 (DESPOT2), which uses a balanced steady-state free precession pulse sequence.[109] An alternative approach to T2 measurement is to separate the imaging section of the sequence from T2-weighting using the T2 preparation pulse (T2-prep), which enables acquisition of multiple echo with fast imaging technique.[110] As with T1 measurement, all these methods are affected by B_1^+ and B_1^- inhomogeneities. T2 measurements are also affected by magnetization transfer effects between the water and macromolecular protons, resulting in diminished signal in the free water and inaccurate T2 measurement.[111,112] Similar to the discrepancy among T1 measurement methods, T2 measurement methods are known to show disagreement with each other.[113] Jutras et al[114] reported that the WM T2 of 70 and 50 milliseconds was measured by MET2 and DESPOT2, respectively, in the same subject, partly due to the different weighting of each tissue component by these methods.

Instead of measuring T1, T2, and PD separately, these values can also be measured simultaneously. Simultaneous measurement has attracted attention for its merit in inherent alignment of the acquired maps and potential reduction in scan time. Two major approaches are quantitative synthetic MRI[102] and MR fingerprinting (MRF).[115] Quantitative synthetic MRI is commonly performed by a 2D multidynamic multiecho (MDME) sequence, which is a turbo spin-echo sequence typically performed with 4 delay times and 2 echo times in the brain so that the scan time is clinically feasible—approximately 6 minutes for the whole head coverage (Fig. 5). B_1^+ field measurements can be simultaneously performed based on the same acquisition data.[116] The performance of the MDME sequence was examined on 3 scanners from 3 different vendors.[117] The highest intrascanner wCVs for T1, T2, and

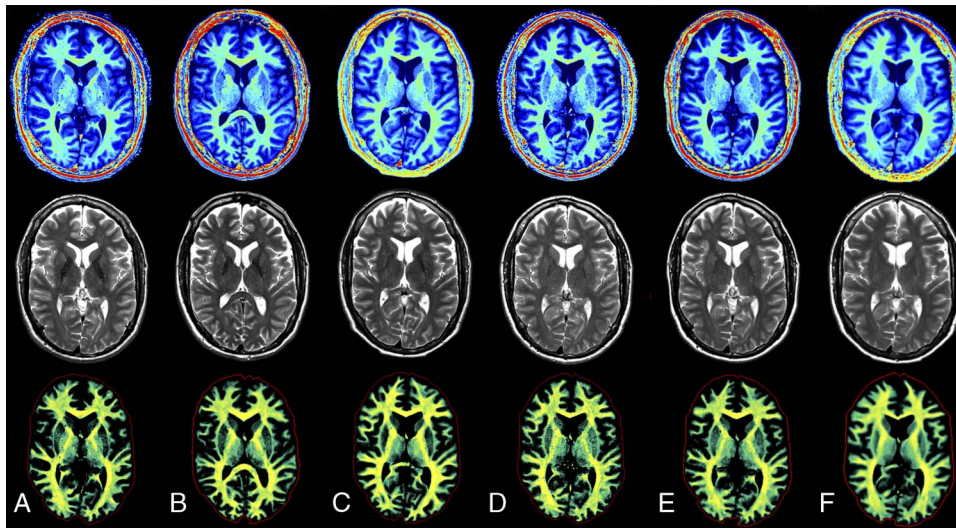

FIGURE 5. Quantification using quantitative synthetic magnetic resonance imaging. The QRAPMASTER acquisition was applied to retrieve the R1 map (top row), R2 map, and proton density map. Based on these maps, conventional (eg, T2-weighted) images can be synthesized (middle row). Furthermore, the R1, R2, and proton density maps provide an absolute scale and hence a robust input to brain segmentation. An example of one of these segmentations (of myelin) is shown in the bottom row. The quantitative synthetic magnetic resonance imaging method provides maps that are independent of the magnetic resonance scanner and hence provide the same result on all major platforms. For this example, the subject was scanned at 3.0 T on a GE 750 (A), Siemens Skyra (B), Philips Ingenia (C), and at 1.5 T on a GE 450 W (D), Siemens Aera (E), and Philips Ingenia (F) (adapted and reproduced with permission from Hagiwara et al[116]).

PD were 2.07%, 7.60%, and 12.86%, respectively, for the ISMRM/NIST standardized phantom, and 1.33%, 0.89%, and 0.77%, respectively, for healthy volunteer brains. The highest interscanner wCVs of T1, T2, and PD were 10.86%, 15.27%, and 9.95%, respectively, for the phantom, and 3.15%, 5.76%, and 3.21%, respectively, for the volunteer brains. Estimating the myelin volume fraction in each voxel by using a 4-compartment model based on the MDME data has also been suggested[118] and applied to diseases such as multiple sclerosis[33,119] and Sturge-Weber syndrome.[120] For applications in other organs, the sequence may have to be adjusted to each target tissue.[121,122] Because radiologists are not accustomed to reading parameter maps, synthetic MRI techniques have also been applied to the MDME data. Synthetic MRI enables the creation of clinically used contrast-weighted images including T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) images based on the T1, T2, and PD maps.[123] Although the image quality of synthetic FLAIR is generally perceived to be inferior to FLAIR acquired by conventional methods, improvement of synthetic FLAIR quality by deep learning has also been suggested.[124] Hence, relaxometry data derived from MDME may become an adjunct to contrast-weighted images in clinical settings. The effect of variability in in-plane resolution on the volumetry based on the MDME data was found to be little in healthy volunteers and MS patients, presumably because the segmentation algorithm considers tissue partial volumes in the interval of 0% to 100% rather than assigning a single tissue type to each voxel.[125,126] The 3-dimensional (3D) version of the MDME, namely, 3D-QALAS (3D-quantification using an interleaved LL acquisition sequence with T2-prep pulse), was recently developed for the heart[127] and has also been applied to the brain.[128,129] Synthetic MR angiography constructed by deep learning is also feasible based on the 3D-QALAS data of high resolution.[130]

Another promising approach of simultaneous relaxometry is MRF. In contrast to quantitative synthetic MRI, MRF adopts a novel approach that does not rely on a traditional curve fitting approach. In MRF, radiofrequency pulses and repetition times are simultaneously varied in a pseudorandom fashion to create signal evolutions that characterize the various relaxation processes unique for each type of tissue (so-called fingerprint).[131] The acquired signal evolutions are pattern-matched against a separately simulated dictionary data, allowing the extraction of multiple tissue properties, including but not limited to T1, T2, PD, and B_0. Proton density is estimated as a scaling factor between the acquired and simulated signal evolutions. Magnetic resonance fingerprinting can measure any property that can be simulated by the Bloch equation, for example, and recent works have also incorporated the measurements of B_1^+,[132,133] T2*,[134] magnetization transfer,[135] amide,[136] spectroscopy, perfusion,[137] and microvascular characteristics into the MRF.[138] The pattern matching can be performed even in the presence of undersampling artifacts; hence, the scan can be highly accelerated to reduce the scan time.[139] The effect of motion on the resulting image is also small as long as the errors are incoherent in such a way that pattern-matching is still possible; however, MRF is known to be more vulnerable to through-plane motion than to in-plane motion.[140] The dictionary of MRF should cover the signal evolutions of a physiologically possible range of tissue properties. The dictionary size presents a trade-off between accuracy and the speed of pattern-matching. The pattern-matching process may benefit from deep learning in terms of both accuracy and speed.[141] The pattern-matching process is a distinctive factor of MRF in view of standardization because resulting maps are dependent on the structure of the dictionary (Fig. 6). The dictionary should be carefully prepared based on the intended purpose, computational resource, and acceptable matching time.

Sequence design is flexible in MRF, and several sequence designs have been used, including balanced steady-state free precession,[131] fast imaging with steady-state precession (FISP),[142] RF-spoiled gradient echo, and quick echo splitting nuclear magnetic resonance.[143] The MRF has been primarily investigated for brain and phantom imaging, but methods for adjusting MRF acquisition to the heart,[144] abdomen,[145] and prostate[146] have also been proposed. High-resolution 3D MRF with the resolution of 1 mm isovoxel was also proposed with a scan time less than 8 minutes for full brain coverage.[147]

Kato et al[148] investigated FISP-based MRF with B_1^+ correction and T1 and T2 measurements on the ISMRM/NIST phantom scanned for 100 days and showed high repeatability with a CV of T1 less than 1% and that of T2 less than 3%, which were better than the values reported by Jiang et al[149] without B_1^+ correction. Korzdorfer et al[150]

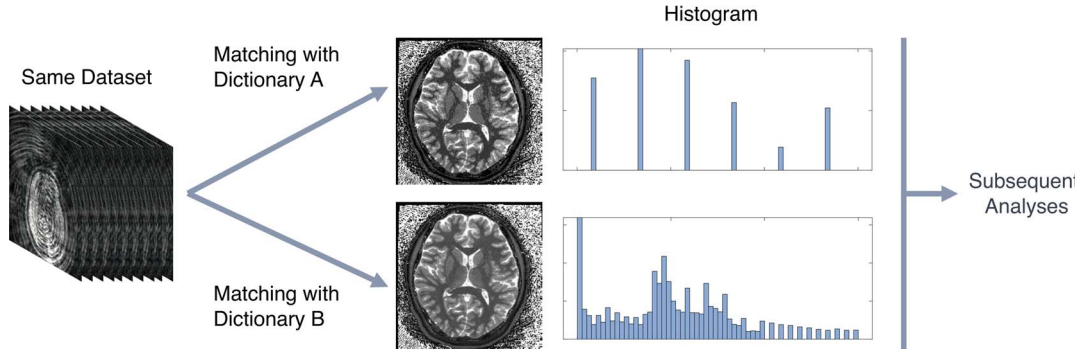

FIGURE 6. Magnetic resonance fingerprinting relies on pattern-matching of through-time signals of highly undersampled images to separately simulated dictionary data. The resulting maps depend not only on the pulse sequence, like conventional magnetic resonance imaging, but also on the dictionary with which the raw data are processed. Although the images processed with 2 different dictionaries in this figure look similar, their histograms are quite different from each other.

investigated the repeatability and reproducibility of T1 and T2 measurements by FISP-based MRF with B_1^+ correction on the brains of 10 healthy volunteers, each of which were scanned 4 times using 4 or more of the 10 MRI scanners; these include 3 different models at 3 T from the same vendor at 4 sites. Repeatability, defined by 95% confidence intervals on relative difference, was 2.0% to 3.1% for T1 and 3.1% to 7.9% for T2 in GM and WM, respectively. Interscanner reproducibility was 3.4% for T1 and 8% for T2 in GM and WM. Larger variations of T2 were likely attributed to scanner imperfections related to certain system characteristics, such as different eddy current behaviors and the diffusion effect.[132]

Multiparametric Quantitative MRI

There is a growing amount of evidence showing that multiparametric MRI can offer better diagnostic ability[151] or more specific biological information[152] than each quantitative measurement. However, this approach is limited by time constraint; hence, a balance between benefit and time should be considered before clinical implementation. One possible approach for implementing multiparametric MRI into clinical practice is setting a cutoff value for each parameter.[146,153] Pinker et al[154] evaluated the diagnostic accuracy of contrast-enhanced MRI, DWI, and MR spectroscopy in combinations of 2 or 3 and used a single measure by setting a threshold for each parameter. As a result, 3 parameters achieved higher accuracy for differentiating between benign and malignant breast lesions than did 1 or 2 parameters. Although the current practice of cancer assessment by multiparametric MRI relies largely on qualitative analysis, if interscanner differences can be overcome or quantified, the current practice may be replaced by more objective and precise quantitative analyses. However, approaches taken toward a single QIB may not be appropriate for multiparametric imaging. For example, the use of multiple QIBs at the same time leads to increases in false-positives for declaring change in at least one QIB, especially when the correlations between the QIBs are low.[155] This could be solved by introducing Mahalanobis distance (ie, distance between a point and zero in a multivariate space that is corrected by variance), resulting in an appropriate type I error rate. Currently, the QIBA Multiparametric Metrology Group is working on developing guidelines for treating multiparametric imaging data.[155]

Another possible application of multiparametric MRI in clinical practice is feeding the data into machine learning for diagnosis,[156] tumor grading,[157] or the prediction of treatment response.[158] Multiparametric quantitative MRI can also be used for extracting new measures that reflect subvoxel microstructural information such as myelin and axon density, axon diameter, and membrane permeability.[152,159] Geometric distortion of images, image misregistration, and different interpolation techniques will introduce errors in created maps; hence, these issues should be cautiously handled.

COMPUTED TOMOGRAPHY

Since the beginning of the clinical application of multidetector row CT (MDCT) in the late 1990s, CT has played a critical role in routine clinical practice. Further, in the last decade, some societies have considered the application of quantitative CT-based indexes as QIBs for patient management, including therapeutic planning and treatment response assessment, and have worked on standardizing the CT protocols for quantitative assessment of CT-based indexes.[160,161] In addition, several investigators have proposed CT-based QIBs for the management of chronic obstructive pulmonary disease (COPD), pulmonary nodules, interstitial lung disease, pulmonary thromboembolism, and pulmonary hypertension.[162–169] In line with this, RSNA QIBA has been working on standardizing the CT protocols and has published profiles through the following committees: (1) CT angiography, (2) CT volumetry, (3) lung density, and (4) small lung nodule.[170]

However, academic and social interests in radiation dose reduction for CT examinations without any accompanying reduction in diagnostic capability have been steadily on the rise. In addition, newly developed iterative reconstruction (IR) methods have been introduced and applied in routine clinical practice.[171,172] In fact, dose reduction strategies have been realized by employing a variety of techniques for data acquisition, such as tube current reduction, tube voltage reduction, increased helical pitch, scan length optimization, scan protocol individualization, and utilization of automatic exposure control (AEC).[172–176] In contrast to RSNA QIBA, J-QIBA primarily aims to determine state-of-the-art CT protocols while keeping the suggested accuracy of CT numbers and bronchial wall thickness, as well as volumetry, within QIBA CT profiles.[170]

Lung Density Evaluation for Quantitative Assessment of Chronic Obstructive Pulmonary Disease

Computed tomography is currently the most widely used modality to evaluate morphologic and pulmonary functional changes for the assessment of COPD.[168,177–180] For both clinical and academic purposes, several commercially available and proprietary software and visual scoring systems have been adopted for the CT-based assessment of pulmonary emphysema.[168,177] Two major approaches have been reported for the quantitative assessment of COPD.[168,177,181–183] One approach determines the percentage of low attenuation area in the lung, which reflects the destruction of the lung parenchyma,[168,177,181–183] and the other determines the percentage of wall area in the bronchi, which reflects bronchial narrowing and wall thickening.[183] In addition, 3D airway luminal volumetry has been introduced as another quantitative airway evaluation method for COPD patients.[184–186] Taking these quantitative CT assessments of COPD and the current situation regarding radiation dose reduction strategies into consideration,[168,177,181–186] the application of IR can be viewed as an important issue not only

related to radiation dose reduction, but also the accuracy of quantitative CT evaluation of COPD.

In the meanwhile, Chen-Mayer et al[187] members of RSNA QIBA published an article regarding the standardization of CT protocols for 64-detector row CT using a variety of scanner models. They provided a quantitative assessment of the variations observed in CT lung density measurements attributed to nonbiological sources, including scanner calibration, the x-ray spectrum, and filtration. However, this study did not address the differences in scan protocols, reconstruction methods, or tube current, and so on. Hence, Ohno et al,[188] as part of J-QIBA activity, compared the effect of different acquisition and reconstruction algorithms on the radiation dose and accuracy of CT number measurements using a 320-detector row CT and the same phantom used by Chen-Mayer et al.[187] They found that the use of a forward projected model-based iterative reconstruction (FIRST, model-based IR method) and adaptive iterative dose reduction using 3D processing (AIDR 3D, hybrid-type IR method) for the 80-detector row helical and wide-volume acquisitions can reduce the radiation dose to a level of 10 mA while keeping the CT number accuracy smaller than the RSNA QIBA Profile request. Therefore, a collaboration between RSNA QIBA and J-QIBA will provide not only standard CT protocols, but also state-of-the art CT protocols for lung density measurement and the application of CT number as one of the QIBs for pulmonary diseases.

Computer-Aided Volumetry for Quantitative Assessment of a Small Pulmonary Nodule

Several large cohort trials, including the National Lung Screening Trial for reducing lung cancer mortality,[189] showed that lung cancer screening with low-dose CT could reduce lung cancer-specific mortality.[163,190–194] Many studies have reported the importance of volume measurements and/or doubling time assessment by computer-aided volumetry (CADv) software in nodule management.[163,190–195] In line with this, the RSNA-QIBA has evaluated the measurement accuracy of various CADv software programs provided by many vendors in a QIBA recommended phantom study[196] and given feedback to suppliers. The J-QIBA contributed to this study by providing scan data. However, this study did not address the effect of differences in scan methods, tube currents, or reconstruction methods. Hence, Ohno et al[197] performed a phantom study in accordance with QIBA recommendations to evaluate the effects of tube current and reconstruction methods on the nodule volume measured with 3D CADv software (CT Lung Nodule Analysis; Vital Images Inc, Minnetonka, MN).[197] In this study, an anthropomorphic thoracic phantom with 30 simulated nodules with various densities and diameters were scanned with an area-detector CT at several tube currents. The mean absolute measurement errors of AIDR 3D and FIRST methods were significantly lower than those of the FBP algorithm in ultra-low-dose CT. For all nodule types, absolute measurement errors of the FBP method in ultra-low-dose CT were significantly higher than those of standard-dose CT. Both IR algorithms were thus shown to be more effective than the FBP algorithm for radiation dose reduction. Ohno et al are now considering to perform a study with the RSNA-QIBA investigating clinical application of the 3D CADv software with deep learning technique in routine clinical practice.

RADIOMICS

Radiomics is based on the high-throughput computer extraction of potentially innumerable numbers of quantitative imaging metrics, or "radiomic features," which will be collectively used for the prediction of diagnosis and prognosis and gene expression profiling.[198] These radiomic features can be combined with other patient characteristics to increase the accuracy of prediction. Because radiomics analyses can be conducted with conventionally used clinical images such as T1-weighted images, FLAIR images, and ADC maps, it is conceivable that conversion of radiological images to mineable data will become routine practice for improving decision-making in precision medicine. Radiomic features are often categorized into shape and first- and higher-order features. First-order features are based on histogram-based analyses and include mean, maximum, minimum, and entropy. Higher-order features are described as texture features related to spatial patterns of voxel intensities. Due to the complexity of radiomic features, there is the danger of overfitting, and hence, dimensionality should be reduced by prioritizing the features. This can be performed by detecting redundant features that are highly correlated with each other. Determining the repeatability and reproducibility of each feature and extracting stable ones can also help the prioritization process in reducing redundant dimensions.[199] Radiomics-specific phantoms with known features are useful in evaluating the effect of scanner and vendor variance on radiomic features, optimizing protocols and image processing in obtaining radiomic features.[200–202]

In general, a lack of reproducibility in radiomic features is a limitation for radiomics to be widely used in clinical practice.[199,203,204] The stability of radiomic features is sensitive, at various degrees, to a number of processing factors, including image acquisition parameters, reconstruction algorithms, digital image preprocessing, and feature extracting methods (Fig. 7). For example, Zhao et al[205] reported that different reconstruction formulas (sharp or smooth) for lung CT introduce variability in radiomic

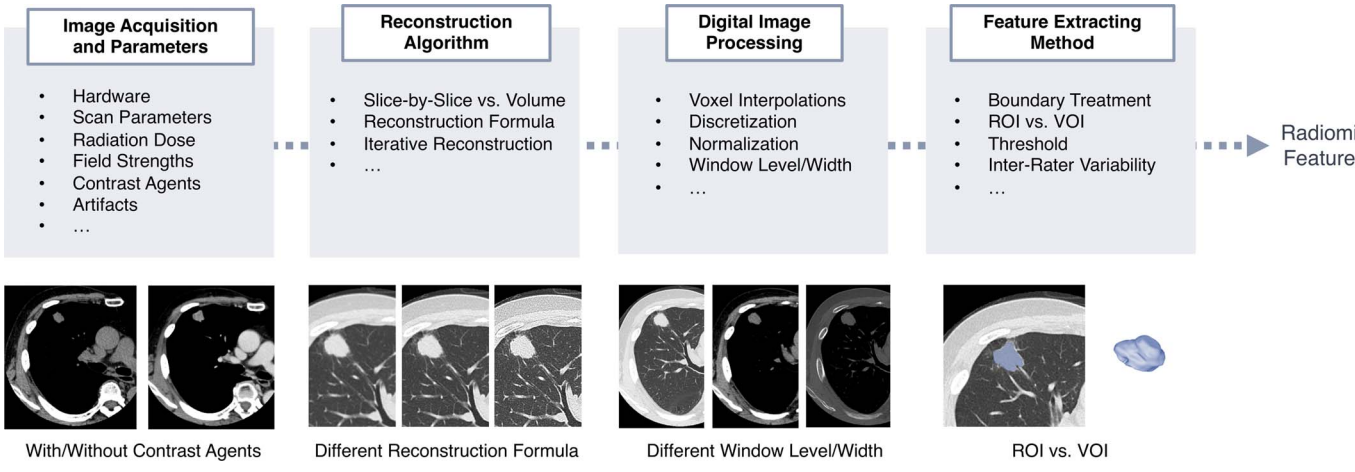

FIGURE 7. Typical variability sources in radiomics analysis. Each feature is the result of multiple processes performed on radiological images. Example factors that introduce variabilities in the resulting features are shown for each process in the bottom row.

features of tumors. Voxel size resampling is often performed for CT datasets acquired with variable voxel sizes in order to obtain more reproducible CT features.[206–208] However, the selection of interpolation methods (eg, nearest neighbor, trilinear and tricubic interpolation) has been revealed to affect the reproducibility of radiomic features.[209] Also, how the software treats the boundary of the volume of interest also affects radiomic features; when the voxels outside the ROI are treated as zero, these boundary regions can have an extremely high gradient and may affect the resulting feature values. Eroding the ROI to include only the core of the target tissue[210] and generating 3 different regions (the tumor, boundary, and peritumor)[211] are possible approaches to achieve reproducible features.

In 2018, Traverso et al[206] reported the results of a systematic review of previous articles investigating repeatability and reproducibility of radiomic features. Overall, first-order features had higher reproducibility than shape and higher-order features, with entropy being consistently reported as one of the most stable features. Among higher-order features, coarseness and contrast were generally poorly reproducible. Interobserver differences in segmentation affected radiomic features, and variabilities were higher for higher-order features. Semiautomated or fully automated methods improved feature reproducibility.[212] However, the articles included in this systematic review were mostly about CT and PET. A more recent study investigated the stability of radiomic features extracted from ADC maps by a multicenter trial and concluded that 122 of 1322 features were stable with a concordance correlation coefficient of more than 0.85 for all tumor entities investigated (ie, ovarian cancer, lung cancer, and colorectal liver metastasis).[213] For magnetic field strength and vendor differences, 245 and 209 features, respectively, were stable. In a phantom study, Baeßler et al[214] showed that FLAIR provided the highest number of stable radiomic features among T1-weighted, T2-weighted, and FLAIR images. It is largely unknown how the repeatability and reproducibility of each feature are propagated to the final result of radiomics, and therefore, validation of a radiomics algorithm against another independent dataset is considered to be crucial.[215]

There are several multi-institutional efforts to standardize and increase the reproducibility of the radiomics approach; this includes providing guidelines, standardized framework, and DROs. The Imaging Biomarker Standardization Initiative provides consensus-based recommendations, nomenclature, and guidelines to improve the reproducibility of radiomic studies.[208] Radiomics Ontology, which is publicly accessible via the NCBO BioPortal (https://bioportal.bioontology.org/ontologies/RO), provides a semantic framework for radiomic features that is in line with the nomenclature addressed by Imaging Biomarker Standardization Initiative. The NRG Oncology investigators have provided recommendations and a guideline specifically for use in the National Clinical Trials Network.[216] They suggest that the radiomics quality score[217,218] may serve to evaluate the quality of radiomics studies.

ARTIFICIAL INTELLIGENCE

Artificial intelligence, including machine learning and deep learning, has been increasingly applied to medical imaging.[219] Promising results have been shown in various tasks related to radiological images, such as the detection of lesions,[220] segmentation (eg, labeling organs),[221] classification (eg, pneumonia vs cancer),[222] reconstruction (eg, MRI k-space to clinical image),[223] and noise reduction.[224] In relation to standardization of QIB, an AI algorithm that automates the process of QIB extraction has the capability to decrease variability, such as through an automated pipeline that can reduce ambiguity and variability in lesion segmentation.[225] Extracting a QIB using AI in a fully automated manner is also feasible. For example, it can predict the functional flow reserve from cardiac CT data by point estimatioin.[226] The major advantages of AI approaches over manual approaches in terms of decreasing variability in QIB are as follows: (1) no variance is caused by fatigue as in human analysts, and (2) AI returns consistent results from the same input. There are several recommendations and guidelines for the development and evaluation of an AI algorithm in the medical field.[219,227,228] In brief, desired steps to develop a reliable AI algorithm include the following: (1) using reliable reference standards, (2) using a training dataset that matches the intended use, (3) tuning hyperparameters on a dataset independent of the training dataset, and (4) using external datasets to evaluate the model performance. To develop an AI algorithm that is robust to variability in acquisition parameters, machine settings, and clinical conditions, the algorithm should be trained with a heterogeneous dataset.[229] The standardization/harmonization of the input images could be one approach to making an algorithm that is generalizable to multiple scanners—although this is unrealistic for multiple vendors, especially when the inputs are multimodal. Although a QIB extracted using AI can be assessed through conventional approaches, there comes a possible issue specific to AI; AI models can be further fine-tuned at each institution using its own data, and the repeatability and reproducibility may change through each update. Quality assessment methods of AI algorithms that are easy to be implemented at each institution still remain to be established.

CONCLUSIONS

Quantification of radiological images has the potential to enable earlier detection of disease, complement or replace biopsy, provide clear differentiation of disease stage, and play an important role in precision medicine. Various sources of variabilities in QIBs have been identified, and extensive efforts have been made to achieve accurate and precise results. Artificial intelligence, especially deep learning techniques, may also further mitigate the variabilities of QIB. In recent years, there has been a surge of interest in multiparametric imaging, including radiomics, but evaluation methods of accuracy and precision of the end results for such techniques still remain to be investigated.

REFERENCES

1. Sullivan DC, Bresolin L, Seto B, et al. Introduction to metrology series. *Stat Methods Med Res*. 2015;24:3–8.
2. 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.
3. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001; 69:89–95.
4. 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.
5. Anvari A, Halpern EF, Samir AE. Essentials of statistical methods for assessing reliability and agreement in quantitative imaging. *Acad Radiol*. 2018;25:391–396.
6. Souchon R, Gennisson JL, Tanter M, et al. Measurement of pulsatile motion with millisecond resolution by MRI. *Magn Reson Med*. 2012;67:1787–1793.
7. Sagawa H, Kataoka M, Kanao S, et al. Impact of the number of iterations in compressed sensing reconstruction on ultrafast dynamic contrast-enhanced breast MR imaging. *Magn Reson Med Sci*. 2019;18:200–207.
8. Jeong D, Malalis C, Arrington JA, et al. Mean apparent diffusion coefficient values in defining radiotherapy planning target volumes in glioblastoma. *Quant Imaging Med Surg*. 2015;5:835–845.
9. 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*. 2015;24:9–26.
10. Buckler AJ, Bresolin L, Dunnick NR, et al. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. *Radiology*. 2011;259:875–884.
11. Tofts PS, Collins DJ. Multicentre imaging measurements for oncology and in the brain. *Br J Radiol*. 2011;84 Spec No :S213–S226.
12. Kundel HL. History of research in medical image perception. *J Am Coll Radiol*. 2006;3:402–408.
13. Radiological Society of North America, Diffusion-Weighted Imaging Task Force subgroup of the Perfusion Diffusion and Flow Biomarker Committee. QIBA Profile: Diffusion-Weighted Magnetic Resonance Imaging (DWI). 2019. Available at: http://qibawiki.rsna.org/images/6/63/QIBA_DWIProfile_Consensus_Dec2019_Final.pdf. Accessed January 15, 2020.
14. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976;38:388–394.

15. 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.
16. Obuchowski NA, Buckler A, Kinahan P, et al. Statistical issues in testing conformance with the quantitative imaging biomarker Alliance (QIBA) profile claims. *Acad Radiol*. 2016;23:496–506.
17. Petrick N, Kim HJ, Clunie D, et al. Comparison of 1D, 2D, and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images. *Acad Radiol*. 2014;21:30–40.
18. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
19. Cohen JA, Fischer JS, Bolibrush DM, et al. Intrarater and interrater reliability of the MS functional composite outcome measure. *Neurology*. 2000;54:802–806.
20. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
21. Kottner J, Audige L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64:96–106.
22. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ*. 1996;313:41–42.
23. Nevill AM, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med*. 1997;31:314–318.
24. Prohl AK, Scherrer B, Tomas-Fernandez X, et al. Reproducibility of structural and diffusion tensor imaging in the TACERN multi-center study. *Front Integr Neurosci*. 2019;13:24.
25. Zhou X, Sakaie KE, Debbins JP, et al. Quantitative quality assurance in a multicenter HARDI clinical trial at 3T. *Magn Reson Imaging*. 2017;35:81–90.
26. Tofts PS, Steens SC, Cercignani M, et al. Sources of variation in multi-Centre brain MTR histogram studies: body-coil transmission eliminates inter-Centre differences. *MAGMA*. 2006;19:209–222.
27. Tofts PS, Brix G, Buckley DL, et al. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. *J Magn Reson Imaging*. 1999;10:223–232.
28. Huang EP, Wang XF, Choudhury KR, et al. Meta-analysis of the technical performance of an imaging procedure: guidelines and statistical methodology. *Stat Methods Med Res*. 2015;24:141–174.
29. Jafari-Khouzani K, Paynabar K, Hajighasemi F, et al. Effect of region of interest size on the repeatability of quantitative brain imaging biomarkers. *IEEE Trans Biomed Eng*. 2019;66:864–872.
30. Voglein J, Tuttenberg J, Weimer M, et al. Treatment monitoring in gliomas: comparison of dynamic susceptibility-weighted contrast-enhanced and spectroscopic MRI techniques for identifying treatment failure. *Invest Radiol*. 2011;46:390–400.
31. Hagiwara A, Hori M, Yokoyama K, et al. Utility of a multiparametric quantitative MRI model that assesses myelin and edema for evaluating plaques, periplaque white matter, and normal-appearing white matter in patients with multiple sclerosis: a feasibility study. *AJNR Am J Neuroradiol*. 2017;38:237–242.
32. Hagiwara A, Hori M, Yokoyama K, et al. Analysis of white matter damage in patients with multiple sclerosis via a novel in vivo MR method for measuring myelin, axons, and G-ratio. *AJNR Am J Neuroradiol*. 2017;38:1934–1940.
33. Hagiwara A, Kamagata K, Shimoji K, et al. White matter abnormalities in multiple sclerosis evaluated by quantitative synthetic MRI, diffusion tensor imaging, and neurite orientation dispersion and density imaging. *AJNR Am J Neuroradiol*. 2019;40:1642–1648.
34. Le Berre A, Kamagata K, Otsuka Y, et al. Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin MRI. *Neuroradiology*. 2019;61:1387–1395.
35. Perkuhn M, Stavrinou P, Thiele F, et al. Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine. *Invest Radiol*. 2018;53:647–654.
36. Keenan KE, Gimbutas Z, Dienstfrey A, et al. Assessing effects of scanner upgrades for clinical studies. *J Magn Reson Imaging*. 2019;50:1948–1954.
37. Lee Y, Callaghan MF, Acosta-Cabronero J, et al. Establishing intra- and intervendor reproducibility of T1 relaxation time measurements with 3T MRI. *Magn Reson Med*. 2019;81:454–465.
38. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. *Stat Methods Med Res*. 2015;24:68–106.
39. Reeves AP, Biancardi AM, Apanasovich TV, et al. The lung image database consortium (LIDC): a comparison of different size metrics for pulmonary nodule measurements. *Acad Radiol*. 2007;14:1475–1485.
40. Pohl KM, Sullivan EV, Rohlfing T, et al. Harmonizing DTI measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the NCANDA study. *Neuroimage*. 2016;130:194–213.
41. Yamashita A, Yahata N, Itahashi T, et al. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biol*. 2019;17:e3000042.
42. Wrobel J, Martin M, Bakshi R, et al. Intensity warping for multisite MRI harmonization. *bioRxiv 679357*. 2019.
43. Fortin JP, Parker D, Tunc B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*. 2017;161:149–170.
44. Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*. 2018;167:104–120.
45. Huynh KM, Chen G, Wu Y, et al. Multi-site harmonization of diffusion MRI data via method of moments. *IEEE Trans Med Imaging*. 2019;38:1599–1609.
46. Cetin Karayumak S, Bouix S, Ning L, et al. Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *Neuroimage*. 2019;184:180–200.
47. Dewey BE, Zhao C, Reinhold JC, et al. DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. *Magn Reson Imaging*. 2019;64:160–170.
48. Keenan KE, Biller JR, Delfino JG, et al. Recommendations towards standards for quantitative MRI (qMRI) and outstanding needs. *J Magn Reson Imaging*. 2019;49:e26–e39.
49. Tofts PS, Shuter B, Pope JM. Ni-DTPA doped agarose gel–a phantom material for Gd-DTPA enhancement measurements. *Magn Reson Imaging*. 1993;11:125–133.
50. Keenan KE, Ainslie M, Barker AJ, et al. Quantitative magnetic resonance imaging phantoms: a review and the need for a system phantom. *Magn Reson Med*. 2018;79:48–61.
51. Cheng HL, Wright GA. Rapid high-resolution T(1) mapping by variable flip angles: accurate and precise measurements in the presence of radiofrequency field inhomogeneity. *Magn Reson Med*. 2006;55:566–574.
52. Stikov N, Boudreau M, Levesque IR, et al. On the accuracy of T1 mapping: searching for common ground. *Magn Reson Med*. 2015;73:514–522.
53. Samson RS, Wheeler-Kingshott CA, Symms MR, et al. A simple correction for B1 field errors in magnetization transfer ratio measurements. *Magn Reson Imaging*. 2006;24:255–263.
54. Brink WM, Bornert P, Nehrke K, et al. Ventricular B1 (+) perturbation at 7 T - real effect or measurement artifact? *NMR Biomed*. 2014;27:617–620.
55. Hoult DI. The principle of reciprocity in signal strength calculations—a mathematical guide. *Concepts Magn Reson*. 2000;12:173–187.
56. Wang J, Qiu M, Yang QX, et al. Measurement and correction of transmitter and receiver induced nonuniformities in vivo. *Magn Reson Med*. 2005;53:408–417.
57. Volz S, Noth U, Deichmann R. Correction of systematic errors in quantitative proton density mapping. *Magn Reson Med*. 2012;68:74–85.
58. Huang SY, Seethamraju RT, Patel P, et al. Body MR imaging: artifacts, k-space, and solutions. *Radiographics*. 2015;35:1439–1460.
59. Bray TJ, Chouhan MD, Punwani S, et al. Fat fraction mapping using magnetic resonance imaging: insight into pathophysiology. *Br J Radiol*. 2018;91:20170344.
60. Rooney WD, Johnson G, Li X, et al. Magnetic field and tissue dependencies of human brain longitudinal 1H2O relaxation in vivo. *Magn Reson Med*. 2007;57:308–318.
61. Barker GJ, Tofts PS. Semiautomated quality assurance for quantitative magnetic resonance imaging. *Magn Reson Imaging*. 1992;10:585–595.
62. Firbank MJ, Harrison RM, Williams ED, et al. Quality assurance for MRI: practical experience. *Br J Radiol*. 2000;73:376–383.
63. Belli G, Busoni S, Ciccarone A, et al. Quality assurance multicenter comparison of different MR scanners for quantitative diffusion-weighted imaging. *J Magn Reson Imaging*. 2016;43:213–219.
64. Keenan KE, Wilmes LJ, Aliu SO, et al. Design of a breast phantom for quantitative MRI. *J Magn Reson Imaging*. 2016;44:610–619.
65. Neumann W, Bichert A, Fleischhauer J, et al. A novel 3D printed mechanical actuator using centrifugal force for magnetic resonance elastography: initial results in an anthropomorphic prostate phantom. *PLoS One*. 2018;13:e0205442.
66. Chen SJ, Hellier P, Marchal M, et al. An anthropomorphic polyvinyl alcohol brain phantom based on Colin27 for use in multimodal imaging. *Med Phys*. 2012;39:554–561.
67. Fieremans E, Lee HH. Physical and numerical phantoms for the validation of brain microstructural MRI: a cookbook. *Neuroimage*. 2018;182:39–61.
68. Padhani AR, Liu G, Koh DM, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia*. 2009;11:102–125.

69. Stocker D, Manoliu A, Becker AS, et al. Image quality and geometric distortion of modern diffusion-weighted imaging sequences in magnetic resonance imaging of the prostate. *Invest Radiol*. 2018;53:200–206.
70. Maekawa T, Hori M, Murata K, et al. Choroid plexus cysts analyzed using diffusion-weighted imaging with short diffusion-time. *Magn Reson Imaging*. 2019;57:323–327.
71. Boonrod A, Hagiwara A, Hori M, et al. Reduced visualization of cerebral infarction on diffusion-weighted images with short diffusion times. *Neuroradiology*. 2018;60:979–982.
72. Andica C, Hori M, Kamiya K, et al. Spatial restriction within intracranial epidermoid cysts observed using short diffusion-time diffusion-weighted imaging. *Magn Reson Med Sci*. 2018;17:269–272.
73. Hori M, Irie R, Suzuki M, et al. Teaching Neuroimages: obscured cerebral infarction on MRI. *Clin Neuroradiol*. 2017;27:519–520.
74. Jafar MM, Parsai A, Miquel ME. Diffusion-weighted magnetic resonance imaging in cancer: reported apparent diffusion coefficients, in-vitro and in-vivo reproducibility. *World J Radiol*. 2016;8:21–49.
75. Grech-Sollars M, Hales PW, Miyazaki K, et al. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. *NMR Biomed*. 2015;28:468–485.
76. Huo J, Alger J, Kim H, et al. Between-scanner and between-visit variation in normal white matter apparent diffusion coefficient values in the setting of a multicenter clinical trial. *Clin Neuroradiol*. 2016;26:423–430.
77. Verma N, Cowperthwaite MC, Burnett MG, et al. Differentiating tumor recurrence from treatment necrosis: a review of neuro-oncologic imaging strategies. *Neuro Oncol*. 2013;15:515–534.
78. Galban S, Brisset JC, Rehemtulla A, et al. Diffusion-weighted MRI for assessment of early cancer treatment response. *Curr Pharm Biotechnol*. 2010; 11:701–708.
79. Bonekamp D, Nagae LM, Degaonkar M, et al. Diffusion tensor imaging in children and adolescents: reproducibility, hemispheric, and age-related differences. *Neuroimage*. 2007;34:733–742.
80. Paldino MJ, Barboriak D, Desjardins A, et al. Repeatability of quantitative parameters derived from diffusion tensor imaging in patients with glioblastoma multiforme. *J Magn Reson Imaging*. 2009;29:1199–1205.
81. Pfefferbaum A, Adalsteinsson E, Sullivan EV. Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. *J Magn Reson Imaging*. 2003;18:427–433.
82. Braithwaite AC, Dale BM, Boll DT, et al. Short- and midterm reproducibility of apparent diffusion coefficient measurements at 3.0-T diffusion-weighted imaging of the abdomen. *Radiology*. 2009;250:459–465.
83. Deckers F, De Foer B, Van Mieghem F, et al. Apparent diffusion coefficient measurements as very early predictive markers of response to chemotherapy in hepatic metastasis: a preliminary investigation of reproducibility and diagnostic value. *J Magn Reson Imaging*. 2014;40:448–456.
84. Heijmen L, Ter Voert EE, Nagtegaal ID, et al. Diffusion-weighted MR imaging in liver metastases of colorectal cancer: reproducibility and biological validation. *Eur Radiol*. 2013;23:748–756.
85. Miquel ME, Scott AD, Macdougall ND, et al. In vitro and in vivo repeatability of abdominal diffusion-weighted MRI. *Br J Radiol*. 2012;85:1507–1512.
86. Gibbs P, Pickles MD, Turnbull LW. Repeatability of echo-planar-based diffusion measurements of the human prostate at 3 T. *Magn Reson Imaging*. 2007; 25:1423–1429.
87. Jambor I, Merisaari H, Aronen HJ, et al. Optimization of b-value distribution for biexponential diffusion-weighted MR imaging of normal prostate. *J Magn Reson Imaging*. 2014;39:1213–1222.
88. Jambor I, Merisaari H, Taimen P, et al. Evaluation of different mathematical models for diffusion-weighted imaging of normal prostate and prostate cancer using high b-values: a repeatability study. *Magn Reson Med*. 2015; 73:1988–1998.
89. Litjens GJ, Hambrock T, Hulsbergen-van de Kaa C, et al. Interpatient variation in normal peripheral zone apparent diffusion coefficient: effect on the prediction of prostate cancer aggressiveness. *Radiology*. 2012;265: 260–266.
90. Fedorov A, Vangel MG, Tempany CM, et al. Multiparametric magnetic resonance imaging of the prostate: repeatability of volume and apparent diffusion coefficient quantification. *Invest Radiol*. 2017;52:538–546.
91. Newitt DC, Zhang Z, Gibbs JE, et al. Test-retest repeatability and reproducibility of ADC measures by breast DWI: results from the ACRIN 6698 trial. *J Magn Reson Imaging*. 2019;49:1617–1628.
92. Sorace AG, Wu C, Barnes SL, et al. Repeatability, reproducibility, and accuracy of quantitative MRI of the breast in the community radiology setting. *J Magn Reson Imaging*. 2018.
93. Sasaki M, Yamada K, Watanabe Y, et al. Variability in absolute apparent diffusion coefficient values across different platforms may be substantial: a multivendor, multi-institutional comparison study. *Radiology*. 2008;249:624–630.
94. Chenevert TL, Galban CJ, Ivancevic MK, et al. Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. *J Magn Reson Imaging*. 2011;34:983–987.
95. Malyarenko DI, Newitt D, J Wilmes L, et al. Demonstration of nonlinearity bias in the measurement of the apparent diffusion coefficient in multicenter trials. *Magn Reson Med*. 2016;75:1312–1323.
96. Laubach HJ, Jakob PM, Loevblad KO, et al. A phantom for diffusion-weighted imaging of acute stroke. *J Magn Reson Imaging*. 1998;8:1349–1354.
97. Maekawa T, Hori M, Murata K, et al. Changes in the ADC of diffusion-weighted MRI with the oscillating gradient spin-echo (OGSE) sequence due to differences in substrate viscosities. *Jpn J Radiol*. 2018;36:415–420.
98. Delakis I, Moore EM, Leach MO, et al. Developing a quality control protocol for diffusion imaging on a clinical MRI system. *Phys Med Biol*. 2004;49:1409–1422.
99. Wagner F, Laun FB, Kuder TA, et al. Temperature and concentration calibration of aqueous polyvinylpyrrolidone (PVP) solutions for isotropic diffusion MRI phantoms. *PLoS One*. 2017;12:e0179276.
100. Lee SM, Choi YH, You SK, et al. Age-related changes in tissue value properties in children: simultaneous quantification of relaxation times and proton density using synthetic magnetic resonance imaging. *Invest Radiol*. 2018;53:236–245.
101. Badve C, Yu A, Rogers M, et al. Simultaneous T_1 and T_2 brain relaxometry in asymptomatic volunteers using magnetic resonance fingerprinting. *Tomography*. 2015;1:136–144.
102. Hagiwara A, Warntjes M, Hori M, et al. SyMRI of the brain: rapid quantification of relaxation rates and proton density, with synthetic MRI, automatic brain segmentation, and myelin measurement. *Invest Radiol*. 2017;52:647–657.
103. Meyers SM, Kolind SH, Laule C, et al. Measuring water content using T2 relaxation at 3T: phantom validations and simulations. *Magn Reson Imaging*. 2016; 34:246–251.
104. Look D, Locker D. Time saving in measurement of NMR and EPR relaxation times. *Rev Sci Instrum*. 1970;41:250–251.
105. Wang HZ, Riederer SJ, Lee JN. Optimizing the precision in T1 relaxation estimation using limited flip angles. *Magn Reson Med*. 1987;5:399–416.
106. Whittall KP, MacKay AL, Li DK. Are mono-exponential fits to a few echoes sufficient to determine T2 relaxation for in vivo human brain? *Magn Reson Med*. 1999;41:1255–1257.
107. Carr HY, Purcell EM. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Phys Rev*. 1954;94:630–638.
108. Meiboom S, Gill D. Modified spin-Echo method for measuring nuclear relaxation times. *Rev Sci Instrum*. 1958;29:688–691.
109. Deoni SC, Peters TM, Rutt BK. High-resolution T1 and T2 mapping of the brain in a clinically acceptable time with DESPOT1 and DESPOT2. *Magn Reson Med*. 2005;53:237–241.
110. Oh J, Han ET, Pelletier D, et al. Measurement of in vivo multi-component T2 relaxation times for brain tissue using multi-slice T2 prep at 1.5 and 3 T. *Magn Reson Imaging*. 2006;24:33–43.
111. Maier CF, Tan SG, Hariharan H, et al. T2 quantitation of articular cartilage at 1.5 T. *J Magn Reson Imaging*. 2003;17:358–364.
112. Liu F, Block WF, Kijowski R, et al. Rapid multicomponent relaxometry in steady state with correction of magnetization transfer effects. *Magn Reson Med*. 2016; 75:1423–1433.
113. Yuan J, Patterson AJ, Ruetten PPR, et al. A comparison of black-blood T_2 mapping sequences for carotid vessel wall imaging at 3T: an assessment of accuracy and repeatability. *Magn Reson Med Sci*. 2019;18:29–35.
114. Jutras JD, Wachowicz K, De Zanche N. Analytical corrections of banding artifacts in driven equilibrium single pulse observation of T2 (DESPOT2). *Magn Reson Med*. 2016;76:1790–1804.
115. Bipin Mehta B, Coppo S, Frances McGivney D, et al. Magnetic resonance fingerprinting: a technical review. *Magn Reson Med*. 2019;81:25–46.
116. Warntjes JB, Leinhard OD, West J, et al. Rapid magnetic resonance quantification on the brain: optimization for clinical usage. *Magn Reson Med*. 2008; 60:320–329.
117. Hagiwara A, Hori M, Cohen-Adad J, et al. Linearity, bias, intrascanner repeatability, and interscanner reproducibility of quantitative multidynamic multiecho sequence for rapid simultaneous Relaxometry at 3 T: a validation study with a standardized phantom and healthy controls. *Invest Radiol*. 2019;54:39–47.
118. Warntjes M, Engström M, Tisell A, et al. Modeling the presence of myelin and edema in the brain based on multi-parametric quantitative MRI. *Front Neurol*. 2016;7:16.
119. Andica C, Hagiwara A, Kamagata K, et al. Gray matter alterations in early and late relapsing-remitting multiple sclerosis evaluated with synthetic quantitative magnetic resonance imaging. *Sci Rep*. 2019;9:8147.

120. Andica C, Hagiwara A, Hori M, et al. Aberrant myelination in patients with Sturge-Weber syndrome analyzed using synthetic quantitative magnetic resonance imaging. *Neuroradiology*. 2019;61:1055–1066.
121. Lee SH, Lee YH, Song HT, et al. Quantitative T2 mapping of knee cartilage: comparison between the synthetic MR imaging and the CPMG sequence. *Magn Reson Med Sci*. 2018;17:344–349.
122. Chougar L, Hagiwara A, Andica C, et al. Synthetic MRI of the knee: new perspectives in musculoskeletal imaging and possible applications for the assessment of bone marrow disorders. *Br J Radiol*. 2018;91:20170886.
123. Wallaert L, Hagiwara A, Andica C, et al. The advantage of synthetic MRI for the visualization of anterior temporal pole lesions on double inversion recovery (DIR), phase-sensitive inversion recovery (PSIR), and myelin images in a patient with CADASIL. *Magn Reson Med Sci*. 2018;17:275–276.
124. Hagiwara A, Otsuka Y, Hori M, et al. Improving the quality of synthetic FLAIR images with deep learning using a conditional generative adversarial network for pixel-by-pixel image translation. *AJNR Am J Neuroradiol*. 2019;40:224–230.
125. Andica C, Hagiwara A, Hori M, et al. Automated brain tissue and myelin volumetry based on quantitative MR imaging with various in-plane resolutions. *J Neuroradiol*. 2018;45:164–168.
126. Saccenti L, Andica C, Hagiwara A, et al. Brain tissue and myelin volumetric analysis in multiple sclerosis at 3T MRI with various in-plane resolutions using synthetic MRI. *Neuroradiology*. 2019;61:1219–1227.
127. Kvernby S, Warntjes MJ, Haraldsson H, et al. Simultaneous three-dimensional myocardial T1 and T2 mapping in one breath hold with 3D-QALAS. *J Cardiovasc Magn Reson*. 2014;16:102.
128. Fujita S, Hagiwara A, Hori M, et al. 3D quantitative synthetic MRI-derived cortical thickness and subcortical brain volumes: scan-rescan repeatability and comparison with conventional T1 -weighted images. *J Magn Reson Imaging*. 2019; 50:1834–1842.
129. Fujita S, Hagiwara A, Hori M, et al. Three-dimensional high-resolution simultaneous quantitative mapping of the whole brain with 3D-QALAS: an accuracy and repeatability study. *Magn Reson Imaging*. 2019;63:235–243.
130. Fujita S, Hagiwara A, Otsuka Y, et al. Deep learning approach for generating MRA images from 3D quantitative synthetic MRI without additional scans. *Invest Radiol*. 2020.
131. Ma D, Gulani V, Seiberlich N, et al. Magnetic resonance fingerprinting. *Nature*. 2013;495:187–192.
132. Kobayashi Y, Terada Y. Diffusion-weighting caused by spoiler gradients in the fast imaging with steady-state precession sequence may lead to inaccurate T2 measurements in MR fingerprinting. *Magn Reson Med Sci*. 2019;18:96–104.
133. Buonincontri G, Sawiak SJ. MR fingerprinting with simultaneous B1 estimation. *Magn Reson Med*. 2016;76:1127–1135.
134. Wyatt CR, Smith TB, Sammi MK, et al. Multi-parametric T_2 * magnetic resonance fingerprinting using variable echo times. *NMR Biomed*. 2018;31:e3951.
135. Hilbert T, Xia D, Block KT, et al. Magnetization transfer in magnetic resonance fingerprinting. *Magn Reson Med*. 2019.
136. Cohen O, Huang S, McMahon MT, et al. Rapid and quantitative chemical exchange saturation transfer (CEST) imaging with magnetic resonance fingerprinting (MRF). *Magn Reson Med*. 2018;80:2449–2463.
137. Su P, Mao D, Liu P, et al. Multiparametric estimation of brain hemodynamics with MR fingerprinting ASL. *Magn Reson Med*. 2017;78:1812–1823.
138. Christen T, Pannetier NA, Ni WW, et al. MR vascular fingerprinting: a new approach to compute cerebral blood volume, mean vessel radius, and oxygenation maps in the human brain. *Neuroimage*. 2014;89:262–270.
139. Korzdorfer G, Pfeuffer J, Kluge T, et al. Effect of spiral undersampling patterns on FISP MRF parameter maps. *Magn Reson Imaging*. 2019;62:174–180.
140. Yu Z, Zhao T, Asslander J, et al. Exploring the sensitivity of magnetic resonance fingerprinting to motion. *Magn Reson Imaging*. 2018;54:241–248.
141. Hoppe E, Thamm F, Korzdorfer G, et al. Magnetic resonance fingerprinting reconstruction using recurrent neural networks. *Stud Health Technol Inform*. 2019; 267:126–133.
142. Jiang Y, Ma D, Seiberlich N, et al. MR fingerprinting using fast imaging with steady state precession (FISP) with spiral readout. *Magn Reson Med*. 2015; 74:1621–1631.
143. Jiang Y, Ma D, Jerecic R, et al. MR fingerprinting using the quick echo splitting NMR imaging technique. *Magn Reson Med*. 2017;77:979–988.
144. Hamilton JI, Jiang Y, Chen Y, et al. MR fingerprinting for rapid quantification of myocardial T1, T2, and proton spin density. *Magn Reson Med*. 2017;77:1446–1458.
145. Chen Y, Jiang Y, Pahwa S, et al. MR fingerprinting for rapid quantitative abdominal imaging. *Radiology*. 2016;279:278–286.
146. Panda A, O'Connor G, Lo WC, et al. Targeted biopsy validation of peripheral zone prostate cancer characterization with magnetic resonance fingerprinting and diffusion mapping. *Invest Radiol*. 2019;54:485–493.
147. Liao C, Bilgic B, Manhard MK, et al. 3D MR fingerprinting with accelerated stack-of-spirals and hybrid sliding-window and GRAPPA reconstruction. *Neuroimage*. 2017;162:13–22.
148. Kato Y, Ichikawa K, Okudaira K, et al. Comprehensive evaluation of B_1^+-corrected FISP-based magnetic resonance fingerprinting: accuracy, repeatability and reproducibility of T_1 and T_2 relaxation times for ISMRM/NIST system phantom and volunteers. *Magn Reson Med Sci*. 2019. [Epub ahead of print].
149. Jiang Y, Ma D, Keenan KE, et al. Repeatability of magnetic resonance fingerprinting T_1 and T_2 estimates assessed using the ISMRM/NIST MRI system phantom. *Magn Reson Med*. 2017;78:1452–1457.
150. Korzdorfer G, Kirsch R, Liu K, et al. Reproducibility and repeatability of MR fingerprinting Relaxometry in the human brain. *Radiology*. 2019;292:429–437.
151. Scheenen TW, Rosenkrantz AB, Haider MA, et al. Multiparametric magnetic resonance imaging in prostate cancer management: current status and future perspectives. *Invest Radiol*. 2015;50:594–600.
152. Cercignani M, Bouyagoub S. Brain microstructure by multi-modal MRI: is the whole greater than the sum of its parts? *Neuroimage*. 2018;182:117–127.
153. Pinker K, Moy L, Sutton EJ, et al. Diffusion-weighted imaging with apparent diffusion coefficient mapping for breast cancer detection as a stand-alone parameter: comparison with dynamic contrast-enhanced and multiparametric magnetic resonance imaging. *Invest Radiol*. 2018;53:587–595.
154. Pinker K, Bogner W, Baltzer P, et al. Improved diagnostic accuracy with multiparametric magnetic resonance imaging of the breast using dynamic contrast-enhanced magnetic resonance imaging, diffusion-weighted imaging, and 3-dimensional proton magnetic resonance spectroscopic imaging. *Invest Radiol*. 2014;49:421–430.
155. Quantitative Imaging Biomarkers Alliance (QIBA) Multiparametric Metrology Group. Multiparametric Quantitative Imaging Biomarkers: A Framework for Estimating and Testing Technical Performance. *Presented at: 105th Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA)*. 2019; Available at: https://qibawiki.rsna.org/index.php/QIBA_posters_from_RSNA_2019_Annual_Meeting. Accessed January 12, 2020.
156. Eshaghi A, Riyahi-Alam S, Saeedi R, et al. Classification algorithms with multimodal data fusion could accurately distinguish neuromyelitis optica from multiple sclerosis. *Neuroimage Clin*. 2015;7:306–314.
157. Vamvakas A, Williams SC, Theodorou K, et al. Imaging biomarker analysis of advanced multiparametric MRI for glioma grading. *Phys Med*. 2019;60:188–198.
158. Tahmassebi A, Wengert GJ, Helbich TH, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Invest Radiol*. 2019;54:110–117.
159. Hori M, Hagiwara A, Fukunaga I, et al. Application of quantitative microstructural MR imaging with atlas-based analysis for the spinal cord in cervical spondylotic myelopathy. *Sci Rep*. 2018;8:5213.
160. Hsia CC, Hyde DM, Ochs M, et al. An official research policy statement of the American Thoracic Society/European Respiratory Society: standards for quantitative assessment of lung structure. *Am J Respir Crit Care Med*. 2010; 181:394–418.
161. Mulshine JL, Gierada DS, Armato SG 3rd, et al. Role of the quantitative imaging biomarker alliance in optimizing CT for the evaluation of lung cancer screen-detected nodules. *J Am Coll Radiol*. 2015;12:390–395.
162. Matsuoka S, Washko GR, Yamashiro T, et al. Pulmonary hypertension and computed tomography measurement of small pulmonary vessels in severe emphysema. *Am J Respir Crit Care Med*. 2010;181:218–225.
163. Henschke CI, Yankelevitz DF, Yip R, et al. Lung cancers diagnosed at annual CT screening: volume doubling times. *Radiology*. 2012;263:578–583.
164. Wells JM, Washko GR, Han MK, et al. Pulmonary arterial enlargement and acute exacerbations of COPD. *N Engl J Med*. 2012;367:913–921.
165. Goldin JG. Computed tomography as a biomarker in clinical trials imaging. *J Thorac Imaging*. 2013;28:291–297.
166. Shaw JG, Vaughan A, Dent AG, et al. Biomarkers of progression of chronic obstructive pulmonary disease (COPD). *J Thorac Dis*. 2014;6:1532–1547.
167. Hansell DM, Goldin JG, King TE Jr., et al. CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the Fleischner society. *Lancet Respir Med*. 2015;3:483–496.
168. Hoffman EA, Lynch DA, Barr RG, et al. Pulmonary CT and MRI phenotypes that help explain chronic pulmonary obstruction disease pathophysiology and outcomes. *J Magn Reson Imaging*. 2016;43:544–557.
169. Matsuoka S, Kotoku A, Yamashiro T, et al. Quantitative CT evaluation of small pulmonary vessels in patients with acute pulmonary embolism. *Acad Radiol*. 2018;25:653–658.
170. RSNA QIBA Profiles. Available at: http://qibawiki.rsna.org/index.php/Profiles. Accessed January 10, 2010.

171. Stiller W. Basics of iterative reconstruction methods in computed tomography: a vendor-independent overview. *Eur J Radiol*. 2018;109:147–154.
172. Ohno Y, Koyama H, Seki S, et al. Radiation dose reduction techniques for chest CT: principles and clinical results. *Eur J Radiol*. 2019;111:93–103.
173. Kubo T. Vendor free basics of radiation dose reduction techniques for CT. *Eur J Radiol*. 2019;110:14–21.
174. Kubo T, Lin PJ, Stiller W, et al. Radiation dose reduction in chest CT: a review. *AJR Am J Roentgenol*. 2008;190:335–343.
175. Kubo T, Ohno Y, Kauczor HU, et al. Radiation dose reduction in chest CT—review of available options. *Eur J Radiol*. 2014;83:1953–1961.
176. Kubo T, Ohno Y, Seo JB, et al. Securing safe and informative thoracic CT examinations—progress of radiation dose reduction techniques. *Eur J Radiol*. 2017;86:313–319.
177. Coxson HO, Mayo J, Lam S, et al. New and current clinical imaging techniques to study chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2009;180:588–597.
178. Park EA, Goo JM, Park SJ, et al. Chronic obstructive pulmonary disease: quantitative and visual ventilation pattern analysis at xenon ventilation CT performed by using a dual-energy technique. *Radiology*. 2010;256:985–997.
179. Remy-Jardin M, Faivre JB, Pontana F, et al. Thoracic applications of dual energy. *Semin Respir Crit Care Med*. 2014;35:64–73.
180. Ohno Y, Yoshikawa T, Takenaka D, et al. Xenon-enhanced CT using subtraction CT: basic and preliminary clinical studies for comparison of its efficacy with that of dual-energy CT and ventilation SPECT/CT to assess regional ventilation and pulmonary functional loss in smokers. *Eur J Radiol*. 2017;86:41–51.
181. Mishima M, Oku Y, Kawakami K, et al. Quantitative assessment of the spatial distribution of low attenuation areas on x-ray CT using texture analysis in patients with chronic pulmonary emphysema. *Front Med Biol Eng*. 1997;8:19–34.
182. Nakano Y, Muller NL, King GG, et al. Quantitative assessment of airway remodeling using high-resolution CT. *Chest*. 2002;122:271S–275S.
183. Hasegawa M, Nasuhara Y, Onodera Y, et al. Airflow limitation and airway dimensions in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2006;173:1309–1315.
184. Koyama H, Ohno Y, Yamazaki Y, et al. Quantitative bronchial luminal volumetric assessment of pulmonary function loss by thin-section MDCT in pulmonary emphysema patients. *Eur J Radiol*. 2012;81:384–388.
185. Koyama H, Ohno Y, Nishio M, et al. Three-dimensional airway lumen volumetry: comparison with bronchial wall area and parenchymal densitometry in assessment of airway obstruction in pulmonary emphysema. *Br J Radiol*. 2012;85:1525–1532.
186. Koyama H, Ohno Y, Nishio M, et al. Iterative reconstruction technique vs filter back projection: utility for quantitative bronchial assessment on low-dose thin-section MDCT in patients with/without chronic obstructive pulmonary disease. *Eur Radiol*. 2014;24:1860–1867.
187. Chen-Mayer HH, Fuld MK, Hoppel B, et al. Standardizing CT lung density measure across scanner manufacturers. *Med Phys*. 2017;44:974–985.
188. Ohno Y, Fujisawa Y, Fujii K, et al. Effects of acquisition method and reconstruction algorithm for CT number measurement on standard-dose CT and reduced-dose CT: a QIBA phantom study. *Jpn J Radiol*. 2019;37:399–411.
189. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365:395–409.
190. Saghir Z, Dirksen A, Ashraf H, et al. CT screening for lung cancer brings forward early disease. The randomised Danish lung cancer screening trial: status after five annual screening rounds with low-dose CT. *Thorax*. 2012;67:296–301.
191. Nawa T, Nakagawa T, Mizoue T, et al. A decrease in lung cancer mortality following the introduction of low-dose chest CT screening in Hitachi, Japan. *Lung Cancer*. 2012;78:225–228.
192. Horeweg N, Scholten ET, de Jong PA, et al. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. *Lancet Oncol*. 2014;15:1342–1350.
193. Infante M, Cavuto S, Lutman FR, et al. Long-term follow-up results of the DANTE trial, a randomized study of lung cancer screening with spiral computed tomography. *Am J Respir Crit Care Med*. 2015;191:1166–1175.
194. Xu DM, Gietema H, de Koning H, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer*. 2006;54:177–184.
195. Yanagawa M, Tanaka Y, Leung AN, et al. Prognostic importance of volumetric measurements in stage I lung adenocarcinoma. *Radiology*. 2014;272:557–567.
196. Li Q, Gavrielides MA, Sahiner B, et al. Statistical analysis of lung nodule volume measurements with CT in a large-scale phantom study. *Med Phys*. 2015;42:3932–3947.
197. Ohno Y, Yaguchi A, Okazaki T, et al. Comparative evaluation of newly developed model-based and commercially available hybrid-type iterative reconstruction methods and filter back projection method in terms of accuracy of computer-aided volumetry (CADv) for low-dose CT protocols in phantom study. *Eur J Radiol*. 2016;85:1375–1382.
198. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563–577.
199. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology*. 2018;288:407–415.
200. Nordstrom RJ. The quantitative imaging network in precision medicine. *Tomography*. 2016;2:239–241.
201. Mackin D, Fave X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*. 2015;50:757–765.
202. Shafiq-Ul-Hassan M, Latifi K, Zhang G, et al. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep*. 2018;8:10545.
203. Kalpathy-Cramer J, Mamomov A, Zhao B, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. *Tomography*. 2016;2:430–437.
204. Nyflot MJ, Yang F, Byrd D, et al. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J Med Imaging (Bellingham)*. 2015;2:041002.
205. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:23428.
206. Traverso A, Wee L, Dekker A, et al. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102:1143–1158.
207. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44:1050–1062.
208. Zwanenburg A, Leger S, Vallières M, et al. Image biomarker standardisation initiative. *arXiv:1612.07003v11*. 2019.
209. Larue RTHM, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol*. 2017;56:1544–1553.
210. Echegaray S, Gevaert O, Shah R, et al. Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. *J Med Imaging (Bellingham)*. 2015;2:041011.
211. Tunali I, Stringfield O, Guvenis A, et al. Radial gradient and radial deviation radiomic features from pre-surgical CT scans are associated with survival among lung adenocarcinoma patients. *Oncotarget*. 2017;8:96013–96026.
212. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One*. 2014;9:e102107.
213. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep*. 2019;9:4800.
214. Baeßler B, Weiss K, Pinto dos Santos D. Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. *Invest Radiol*. 2019;54:221–228.
215. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One*. 2015;10:e0124165.
216. Nie K, Al-Hallaq H, Li XA, et al. NCTN assessment on current applications of Radiomics in oncology. *Int J Radiat Oncol Biol Phys*. 2019;104:302–315.
217. Sanduleanu S, Woodruff HC, de Jong EEC, et al. Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score. *Radiother Oncol*. 2018;127:349–360.
218. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14:749–762.
219. Weikert T, Cyriac J, Yang S, et al. A practical guide to artificial intelligence-based image analysis in radiology. *Invest Radiol*. 2020;55:1–7.
220. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol*. 2017;52:434–440.
221. Weston AD, Korfiatis P, Kline TL, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology*. 2019;290:669–679.
222. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284:574–582.
223. Hyun CM, Kim HP, Lee SM, et al. Deep learning for undersampled MRI reconstruction. *Phys Med Biol*. 2018;63:135007.

224. Kidoh M, Shinoda K, Kitajima M, et al. Deep learning based noise reduction for brain MR imaging: tests on phantoms and healthy volunteers. *Magn Reson Med Sci*. 2019. [Epub ahead of print].

225. Kuzina A, Egorov E, Burnaev E. Bayesian generative models for knowledge transfer in MRI semantic segmentation problems. *Front Neurosci*. 2019;13:844.

226. Kumamaru KK, Fujimoto S, Otsuka Y, et al. Diagnostic accuracy of 3D deep-learning-based fully automated estimation of patient-level minimum fractional flow reserve from coronary computed tomography angiography. *Eur Heart J Cardiovasc Imaging*. 2019. [Epub ahead of print].

227. Liu Y, Chen PC, Krause J, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322: 1806–1816.

228. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286:800–809.

229. Mårtensson G, Ferreira D, Granberg T, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *arXiv: 1911.00515v1*. 2019.