

Quality Assurance: Accuracy, Precision, Controls and Phantoms¹

Contents

3.1	Introduction	33
	Quality Assurance Concepts • Quantitative Quality Assurance • Multicentre Studies	
3.2	Uncertainty, Error and Accuracy.....	36
	Concepts • Sources of Error • Modelling Error • Uncertainty in Measurement: Type A and Type B Errors • Accuracy	
3.3	Precision.....	39
	Precision Concepts • Within-Subject Standard Deviation • Intraclass Correlation Coefficient or Reliability • Analysis of Variance Components • Other Methods	
3.4	Healthy Controls for QA	43
3.5	Phantoms (Test Objects).....	44
	Phantom Concepts • Single Component Liquids • Multiple Component Mixtures for T_1 and T_2 • Other Materials • Temperature Dependence and Control • Phantom Design	
	References.....	50

Paul S.Tofts
Brighton and Sussex
Medical School

3.1 Introduction

3.1.1 Quality Assurance Concepts

When an instrument such as an MRI scanner is installed and handed over by the vendor (manufacturer) to the user, a series of acceptance tests is often carried out by the customer (de Wilde *et al.*, 2002) (McRobbie and Quest 2002). The vendor's installation engineer will also have carried out extensive testing, according to their own protocols, using phantoms (test objects) to ensure the instrument is operating within the specification of the vendor. For qualitative MRI these may include signal-to-noise ratio, spatial resolution and uniformity tests, gradient calibration and ensuring image artefacts are below certain levels.

Quality assurance (QA, sometimes called *quality control*) is used here to denote an *ongoing process* of ensuring the instrument continues to operate satisfactorily (Barker and Tofts 1992; Firbank *et al.*, 2000).

The QA falls into two groups. Firstly, the vendor's ongoing service contract will include some tests, largely to ensure the machine stays within specification. There may be some periodic recalibrations, for example of transmitter output, as components age. The user will not normally be involved in this process.

The second group of QA measurements will be focussed on monitoring the quantification performance of the scanner. The quantification methods will often have been implemented

in-house, without the explicit support of the vendor, and if they are unreliable the vendor will not be responsible, provided he can ensure the machine is still within the manufacturer's specifications. Thus the user must design, implement and analyse *quantitative quality assurance* (QQA) using appropriate measurements on phantoms and normal subjects (Tofts 1998).

Professional organisations of medical physics sometimes publish material on QA in MRI. The UK Institute of Physics and Engineering in Medicine has published *Report 112: Quality Control and Artefacts in Magnetic Resonance Imaging in 2017*.² This gives a comprehensive description of how to use the Eurospin test objects, and much more. The American Association of Physicists in Medicine has published some guidance on QA (Price *et al.*, 1990; Och *et al.*, 1992).³ The American College of Radiology (ACR)⁴ has an MRI accreditation scheme and an **MRI quality control manual**.⁵

¹ Reviewed by Mara Cercignani.

² <http://www.ipem.ac.uk>

³ *Acceptance Testing and Quality Assurance Procedures for MRI Facilities* free of charge from www.aapm.org.

⁴ <http://www.acr.org/>

⁵ Downloadable from the AAPM website.

AQ: Please note that heading 3.1 was missing. Added in with "Introduction" as title placeholder; please adjust as necessary.

AQ: Please check that footnote 5 is correct as set: it references the AAPM website, but text states "The ACR has... an MRI quality control manual".

3.1.2 Quantitative Quality Assurance

QQA will consume valuable scanner time, yet without it the measurements on research subjects may become valueless. Appropriate QQA provides reassurance that patient data are valid, gives warning if the measurement technique has failed because of a change in equipment or procedure, and may provide some help in rescuing data affected by such a failure. QQA measurements can be carried out in healthy ('normal') controls and in phantoms.

Measurements in *control human subjects* (Section 3.4) are usually completely realistic, provided the parameter is present in normal subjects. Thus brain volume or normal-appearing brain tissue T_1 value could be monitored in this way but lesion volume could not. Increased atrophy or movement in patients might sometimes increase the variability compared to normal control subjects. A few parameters (most notoriously blood perfusion) have large biological intrasubject variation and require special designs for QQA. In addition to long-term monitoring by QQA, short-term reproducibility can be measured in any subject, although there may be ethical issues if Gd contrast agent is to be injected (e.g. for DCE-MRI – see Chapter 14) (Table 3.1).

Phantom measurements (Section 3.5) have the advantages of potentially providing a completely accurate value for the parameter under measurement (e.g. volume or T_1), of potentially being completely stable and of always being available. Often a *loading ring* is inserted into the head coil to provide similar loading to that given by the head. However realism is generally poor, with many potential sources of *in vivo* variation absent (e.g. subject movement, positioning error, partial volume error, variable loading, B_1 variation). Temperature dependence may be a problem (see Section 3.5.4). If a drift is seen in measurements from phantoms, the interpretation is often unclear (was it the scanner or the phantom that was unstable?) (Figure 3.1).

A *short-term test object* may be useful when developing a new measurement technique; this can be made quickly and need not be stable or have good independence of temperature. Later on, as the technique matures and goes into clinical use, full QQA would be needed, using healthy controls or a stable phantom.

A *post-mortem* brain phantom seeks to combine realism with stability and the ability to travel in a multicentre study (Droby *et al.*, 2015).

Frequency: To carry out QQA, controls or phantoms are measured at regular intervals (typically every week or month).

AQ: Please confirm whether the inserted citations for Figure 3.1 to 3.13 are okay.

AQ: Please spell out DCE-MRI at first mention.

AQ: Please check and confirm whether the inserted citations for Table 3.1 to Table 3.8 are okay.

AQ: Please cite Tables 3.1, 3.3, 3.6, 3.7, 3.9 in text in order.

TABLE 3.1 Relative Advantages of Phantoms and Healthy Controls for Quantitative Quality Assurance

	Simple Phantom (Test Object)	Healthy Control Subjects
Availability ^b	Good	Reasonable
Accuracy	Potentially good (e.g. volume)	True value unknown ^a
Uniformity	Poor in gels, good in liquids	Good in white matter
Temperature dependence	D, T_1, T_2 change 2%–3%/°C	Homeostatic temperature control
Stability	Potentially good (e.g. volume) but can be unstable (e.g. gels)	Usually stable
Realism	Generally poor; <i>in vivo</i> changes cannot be realistically modelled; B_1 distribution different	Good but no pathology
Standard design for multicentre studies?	Can be made	Use normal range, or travelling subject(s)

^a Although normal values have a narrow range – see Table 3.5.

^b Though see institutional constraints (Section 3.5.1).

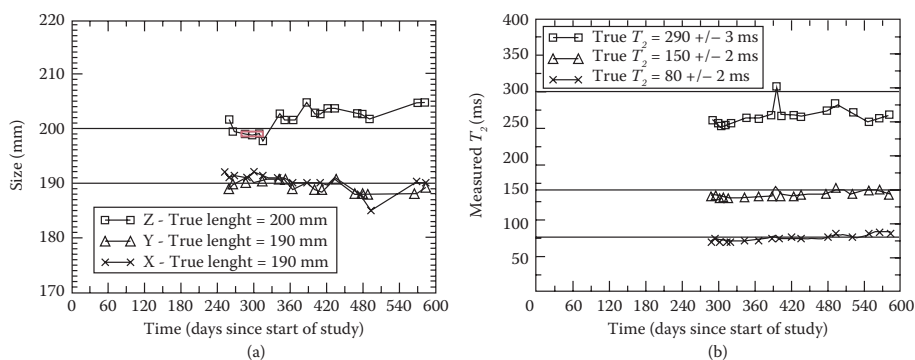


FIGURE 3.1 An early example of quality assurance (QA) measurements of object size (a) and T_2 (b). The apparent size drifts with time, probably because of a fault with gradient calibration. The true size is known accurately and unambiguously. T_2 estimates are inaccurate, particularly for the long- T_2 phantom, and drift with time, suggesting a progressive instrumental error. However inaccuracy and instability in the gel phantoms cannot be ruled out, unless a separate measurement of T_2 is carried out with a procedure known to be reliable. A drift in their temperature is a third possible explanation. (From Barker, G.J. and Tofts, P.S., *Magn. Reson. Imaging*, 10(4), 585–595, 1992.)

TABLE 3.2 Statistical Tests Used for Shewhart Charting

Test Number	Name of Test	Description of Test	Action Required
1	Warning	Measure exceeds control limits of mean \pm 2 SD of previous measures.	Inspect with Tests 2–6
2	3 SD	Measure exceeds control limits of mean \pm 3 SD of previous.	Instrument evaluation
3	2 SD	Two consecutive measures exceed mean \pm 2 SD.	Instrument evaluation
4	Range of 4 SD	Difference between two consecutive measures exceeds 4 SD.	Instrument evaluation
5	Four \pm 1 SD	Four consecutive measures exceed the same limit (+ 1 SD or – 1 SD).	Instrument evaluation
6	Mean \times 10	Ten consecutive measures fall on the same side of the mean.	Instrument evaluation

Source: Adapted from Simmons, A., et al., *Magn. Reson. Med.*, 41(6), 1274–1278, 1999.

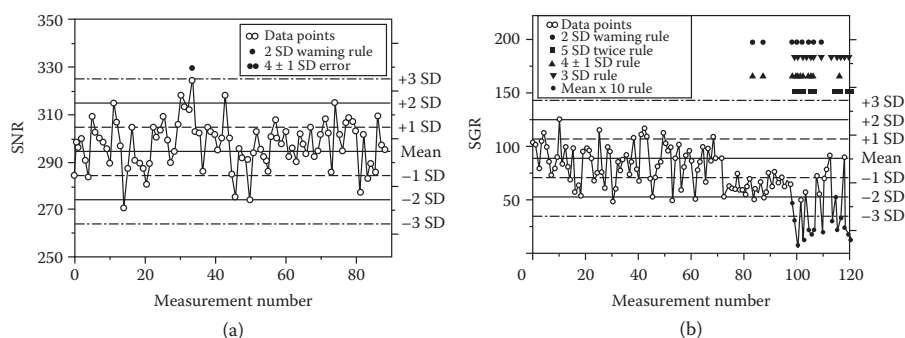


FIGURE 3.2 Shewhart charting of QA parameters. Data points are open symbols; triggering of rules (see Table 3.2) is shown by solid symbols. SNR is signal-to-noise ratio; SGR is signal-to-ghost ratio (used in echoplanar imaging) (From Simmons, A., et al., *Magn. Reson. Med.*, 41(6), 1274–1278, 1999).

The frequency has to be a compromise between rapid detection of a change in the instrument and the limited amount of machine time that is available. If an upgrade is planned, *bunched measurements* should be carried out before and after the change. Analysis should be automated as much as possible, both to save human time and to encourage rapid analysis of scan data (Sun et al., 2015). Shewhart charting (Hajek et al., 1999; Simmons et al., 1999) is a set of statistical rules for automatically deciding when a measurement is abnormal enough to warrant human intervention (Table 3.2 and Figure 3.2).

Calibration was sometimes claimed to be a benefit of scanning phantoms with known magnetic resonance (MR) properties. Calibration is measuring the response of the instrument to a stimulus of known value, with the purpose of then being able to apply that knowledge to *in vivo* measurements. For example, it was hoped that by measuring T_1 estimates for phantoms of known T_1 value, the calibration curve between true and estimated T_1 values could be applied to *in vivo* measurements. This concept has limited validity in the context of *in vivo* measurements, because there are many sources of error that are present *in vivo* but not in the phantom or else have different magnitudes in the two cases. Thus T_1 errors arising from incorrect flip angle settings are unlikely to be the same in a phantom and *in vivo*, and in general any systematic errors present in the phantom do not provide a realistic representation of those present *in vivo*. This is true for both a ‘same-place’ phantom, scanned in the head coil at a different time from the head, and for a ‘same-time’ phantom, attached to the head but in a different place from the brain.

3.1.3 Multicentre Studies

Multicentre studies, where an attempt is made to reproduce the same measurement technique across different centres, or hospitals, often with different kinds of scanners, in different countries, are a challenging test (Podo 1988; Soher et al., 1996; Keevil et al., 1998; Podo et al., 1998; Bauer et al., 2010) (Jerome et al., 2016). The European group MAGNiMS⁶ has conducted multicentre studies for over 20 years (Filippi et al., 1998; Sormani et al., 2016). The Human Connectome Project seeks to map macroscopic human brain circuits in a large population of healthy adults using DTI and other techniques (Van Essen et al., 2013).

AQ: Please spell out DTI.

Multi-centre magnetic resonance imaging (MRI) studies of the human brain enable a more advanced and comprehensive investigation of the disease course of rare and heterogeneous neurological and neuropsychiatric disorders due to increased sample sizes achieved by pooling data from the participating centres. While multi-centre MRI studies allow the acquisition of large amounts of data during a relatively short time period, they are based on the assumption that site-specific differences in MRI equipment do not impose any bias on the data, as this would severely reduce the statistical power of any analysis aimed at detecting differences between groups (Droby et al., 2015).

⁶ Magnetic Resonance Imaging in Multiple Sclerosis, <https://www.magnims.eu/>

Variability among scanners may cancel the benefit of using multiple centers to assess new treatments (Zhou *et al.*, 2017).

Thus the key issue is to minimise contamination of the whole dataset by centres with poor technique.

One approach to minimising between-centre variation is to make the data collection and analysis procedures as identical as possible, so that any systematic errors are replicated across the whole sample of centres. ‘Protocol matching’ for data collection involves attempting to match scanner type, field strength, sequence timing parameters (TR,TE) and also slice profile and RF non-uniformity (which is often not possible). In pharmaceutical trials there is often a travelling quality control officer, who ensures conformity to the agreed scanning protocol. MT measurements from two centres with different scanners were matched by careful attention to sequences, analysis technique, and by using body coil excitation to reduce B_1 differences (Tofts *et al.*, 2006). A second approach is to aim for good accuracy at each centre, measuring the underlying MR or biology parameters independent of the particular measurement procedure, since accurate measurements must necessarily agree with each other. Analysis matching may involve reaching agreement on standardised models, terminology and symbols; this was achieved for the DCE-MRI consensus (Tofts *et al.*, 1999)

Validation can be by measuring healthy controls (which have a narrow spread of values – see Section 3.4 below), measuring travelling controls (which are scanned at each site), measuring a travelling phantom or acquiring a standard phantom at each site.

Thus multicentre studies, although time-consuming and frustrating, are the ultimate test of how good our measurement techniques are. Full discussions of all the issues are available (Padhani *et al.*, 2009; Tofts and Collins 2011; Drobny *et al.*, 2015; Jerome *et al.*, 2016). Early identification of outliers may enable problems at particular contaminating centres to be identified (Walker *et al.*, 2013).

Biomarkers: A major driver for developing quantitative MRI is to produce reliable biomarkers, to be used in multicentre treatment trials. Biomarker concepts come from a drug development paradigm; these are well developed and not always aligned with MRI concepts (Padhani *et al.*, 2009; O’Connor *et al.*, 2017).

3.2 Uncertainty, Error and Accuracy

3.2.1 Concepts

The conventional way to characterise measurement techniques in the physical sciences has been to estimate accuracy and precision (i.e. systematic and random errors). Separating systematic and random error is often helpful, since they occur on different timescales and have different effects on the viability of the measurement. A systematic error, in its ideal form, is one that is constant over the lifetime of the study, whilst a random error is one present in short-term repeated measurements.

A measurement result is complete only when accompanied by a quantitative statement of its uncertainty. The uncertainty is required in order to decide if the result is adequate for its intended purpose and to ascertain if it is consistent with other similar results.⁷

In modern use, *measurement error* is used to mean the difference between the measurement and the true value, whilst *measurement uncertainty* refers to the spread of possible true values that can be inferred from the measurement. Thus, a particular (single) measurement could have zero error but large uncertainty. In psychology and in medicine, the concept of reliability is often used to evaluate the performance of a metric (see Chapter 1, Section 1.3.3).

Accuracy refers to systematic error, the way in which measurements may be consistently different from the truth, or biased. *Precision* refers to random errors, which occur over short time intervals, if the measurement is repeated often. Thus in a determination of T_1 , systematic errors could be caused by a consistently wrong B_1 value, whilst random errors could be caused by image noise (which is different in each image). However the systematic error could vary over a long period of time (for example if the method for setting B_1 was improved or a different head coil was installed). Similarly, the precision could be worse if measured from repeat scans over a long period of time, compared with short-term repeats, as additional sources of variation became relevant (for example a change of data acquisition technologist) (Figure 3.3).

Thus the differences between *long-term precision* and accuracy become blurred, and the difference is merely one of time scale. Some studies of chronic disease can last for long periods (over a decade in the case of MS, epilepsy, dementia and aging) and considerations of accuracy and its variation over time become increasingly important (see Figure 3.3). Precision can be seen as setting the limits of agreement in a short study on the same machine; accuracy sets the limits of agreement in a long-term or multicentre study, where several machines are to be used, possibly extending over different generations of technology.

3.2.2 Sources of Error

Contributions to both inaccuracy and imprecision can arise in both the data collection and the image analysis procedures (see Chapter 2), and both need to be carefully controlled in order to achieve good long-term performance. The major contributors to systematic data collection errors are probably B_1 non-uniformity and partial volume errors. Artefacts arising from imperfect slice selection and k -space sampling (particularly in fast spin echo and echoplanar imaging) can also give systematic error. Patient positioning and movement contribute to random errors; positioning can be improved with technologist training and liberal use

⁷ From the US National Institute of Standards and Technology (NIST) website, <http://physics.nist.gov/cuu/Uncertainty/index.html>. This is a mine of information on constants, units and uncertainty.

AQ: Please check that "see Figure 3.3" is correct as edited; "acc_prec" was deleted.

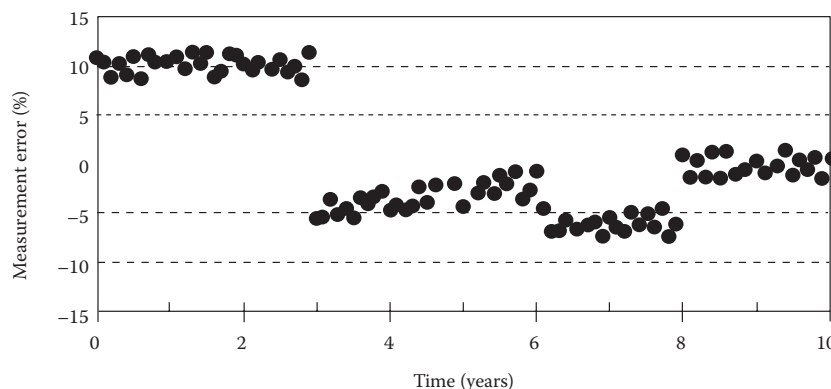


FIGURE 3.3 Long-term precision is dominated by instability in the systematic error. Simulation of fictional change in measurement error over time, during a longitudinal study. Short-term precision is good, and a study completed in the first 3 years is unaffected by the large systematic error (i.e. poor accuracy). A major upgrade at Year 3 dramatically changes the systematic error. A subtle drift in values takes place, followed by two more step changes, at the times of operator change and a minor upgrade. At Year 8 the sources of systematic error are finally identified and removed, giving a system that should provide good accuracy and hence long-term precision for many years.

AQ: Please cite Figures 3.3 through 3.8 in text, in order (or renumber).

AQ: Please check that "Long-term precision ... error. Simulation of..." is correct as set; original had "imulation"

TABLE 3.3 Potential Sources of Error in the MRI Measurement Process^a

	Random Error	Systematic Error
Biology	Normal variation in physiology	
Data collection	Position of subject in head coil Coil loading (corrected by prescan?) Prescan procedure setting B ₁ Position of slices in head	B ₁ error Slice profile K-space sampling (in FSE, EPI) Partial volume Operator training Software upgrade
	Gd injection procedure Patient movement (cardiac pulsation) Patient movement (macroscopic) Image noise Temperature (phantoms only)	Hardware upgrade
Image analysis	ROI creation and placement	Operator training Software upgrade

Note: In their simplest forms, random error is associated with short-term unpredictable variation, whilst systematic error is fixed. However some random processes (e.g. positioning) might only show up over a longer time scale (caused e.g. by change of radiographer [technician]), whilst some sources of systematic error might vary with time (e.g. operator training). ROI = region of interest.

^a See also.

AQ: Please define FSE, EPI beneath Table 3.3

AQ: Please provide the cross-references.

of localiser scans, whilst movement can be reduced by attention to patient comfort, feedback devices to assist the subject in keeping still (Tofts *et al.*, 1990) and spatial registration of images (see Chapter 17). Analysis performance can be characterised by repeat analyses, both by the same observer and by different observers. A change of technologist, either for data collection or for analysis, can introduce subtle changes in procedure and hence results. Early work that measured the reproducibility of an analysis procedure has little value without re-scanning the subject, since patient positioning can be a major source of variation (Tofts 1998). In the case of automatic image analysis this is particularly true, since an automatic procedure, being free of a subjective operator, is intrinsically perfectly reproducible (Table 3.3).

The *analysis software* has to be kept stable, and modern software engineering practice⁸ defines how to do this. The analysis method should be documented in detail, intra- and interrater differences measured, software upgrades should be controlled and documented through version control procedures. In long-term studies, some old data should be kept for re-analysis at a later stage, when operators and software may have changed (Tofts and Collins 2011). Alternatively, all the analysis can be carried out at the end of the study, over a relatively short time. However there is often a value in carrying out a preliminary

⁸ See for example ISO 9001.

analysis, and in any cases studies are often extended beyond their initially planned duration.

3.2.3 Modelling Error

3.2.3.1 Error Propagation Ratio

The error propagation ratio (EPR) is a convenient way of investigating the sensitivity of a parameter estimate to the various assumptions that have gone into the calculation. The EPR is the percent change in a derived parameter arising from a 1% change in one of the model parameters. For example, in a study to measure capillary transfer constant K^{trans} in the breast (Tofts *et al.*, 1995), the estimate is very sensitive to the T_{10} value used (EPR = 1.2) and the relaxivity r_1 (EPR = 1.0) but very insensitive to an error in the echo time (EPR = 0.02). In arterial spin labelling, the sensitivity of the perfusion estimate can similarly be investigated (Parkes and Tofts 2002). Studying error sources in this way immediately brings to light that some errors are truly random, whilst others could be systematic for the same subject in repeated measurements (e.g. a wrongly assumed AIF in T_1w -DCE) but random across other subjects. Uncertainty budgets and type A and B errors are concepts related to EPR (see Section 3.3.6).

AQ: Please spell out AIF

3.2.3.2 Image Noise

The contribution of image noise to imprecision in the final parameter can be calculated. If a simple ratio of images is used (for example T_1 calculated from images at two different flip angles), then propagation of errors (Taylor 1997) allows the effect of noise in each source image to be calculated. An analytic expression can be derived for the total noise, and this can be minimised as a function of imaging parameters such as TR and the number of averages, keeping the total imaging time fixed (see e.g. Tofts 1996, and Chapter 2, Fig. 16).

AQ: Please check "Chapter 2, fig 16" for clarity; if in this book, this should read "Figure 2.16" but there is no Figure 2.16

3.2.3.3 Cramer-Rao Analysis

If least squares curve fitting is used to estimate a parameter from more than two images, simple noise propagation will not work, as the fitted parameter is not a simple function of the source images. However the Cramer-Rao *minimum variance bound* (Cavassila *et al.*, 2001; Brihuega-Moreno *et al.*, 2003) is an analytical method making use of partial derivatives that does calculate the effect of image noise on the fitted parameters. The LC model for estimating spectral areas by fitting uses this method to estimate the minimum uncertainty in the metabolite concentration (Provencher 2001). Only uncertainty arising from data noise is included; other factors (both random and systematic) can make the uncertainty higher than this minimum variance bound.

AQ: Please spell out LC

3.2.3.4 Monte Carlo

Numerical simulation can simulate the effect of image noise. Noise is added to the source data many times and the effect on the fitted parameter measured.

3.2.4 Uncertainty in Measurement: Type A and Type B Errors

The scientific measurement community has moved to refine the traditional concepts of random and systematic error and instead uses a different (though closely related) method of **specifying errors**.⁹ Initiatives have been published from Europe,¹⁰ the USA¹¹ and UK.¹² Type A errors are those estimated by repeated measurements, whilst type B errors are all others. They are combined into a 'standard uncertainty'. This approach was designed by physical metrologists, primarily for reporting uncertainty in physical measurements. An *uncertainty budget* is drawn up, where error components that are considered important are separately identified, quantified (using propagation of errors), then combined to obtain an overall uncertainty. Thus systematic errors are no longer looked on as being benevolent and unchanging. A simple example of an uncertainty budget is that of measuring diffusion coefficient in a test liquid, where the effects of noise, uncertain temperature and uncertain gradient values were analysed and combined (Tofts *et al.*, 2000)

AQ: Please check whether title "Guidance on 'The Expression of Uncertainty in Measurement'" cited in Footnote 9 is an actual publication; if paraphrased, please remove quotation marks and capital letters.

3.2.5 Accuracy

Accuracy is a measure of systematic error, or bias. It estimates how close to the truth the measurements are, on average. It is intrinsically a long-term measure. Often the truth is unknown in MRI, since the brain tissue is not accessible for detailed exhaustive measurements. Thus the true grey matter volume, or total MS lesion volume, would be extremely hard to measure. A physical model (i.e. a phantom) could never be made realistic enough to simulate all the sources of error present in the actual head.

Yet if accuracy is desired, some basic tests can be applied using simple objects. For the example of measuring lesion volume in MS, simple plastic cylinders immersed in a water bath proved too easy, since the major sources of variation (partial volume and low contrast) were missing. However, by tilting the cylinders (to give realistic partial volume effect), inverting the image contrast (to give bright lesions) and adding noise (to give realistically low contrast-to-noise values for the artificial lesions), images were obtained that gave realistic errors in the reported values of volume (Tofts *et al.*, 1997b). Accuracy (and precision)

⁹ The standard work is the *Guide to the Expression of Uncertainty in Measurement (GUM)*, published by the International Standards Organisation (ISO) in 1995. Available from BIPM (Bureau International des Poids et Mesures; www.bipm.org). There is much commercial activity in this field, as organisations selling measurement services seek ISO accreditation. Many national organisations produce guidance on 'The Expression of Uncertainty in Measurement', and publish user-friendly versions of GUM. Books are also available.

¹⁰ The European Accreditation group has produced in 2013 *Evaluation of the Uncertainty of Measurement in Calibration*, document EA-4/02. This gives much detail and good examples of uncertainty budgets. See www.european-accreditation.org

¹¹ The NIST has guidelines from 2000 at <http://physics.nist.gov/cuu/Uncertainty/index.html>. More recent information is at <https://www.nist.gov/>

¹² The United Kingdom Accreditation Service has several useful documents; *M3003* and *LAB 12* are concise expositions of the concepts.

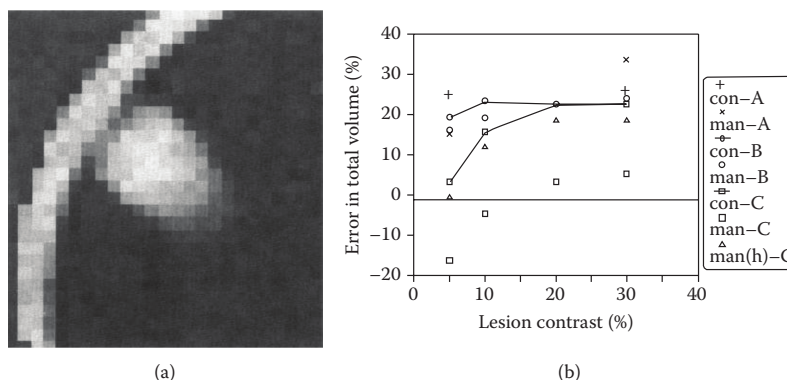


FIGURE 3.4 Lesion volume accuracy measured using an oblique cylinder contrast-adjusted phantom. (a) One small lesion (with a known volume of 0.6 ml), represented as an acrylic cylinder, is mounted on the inside of an acrylic annulus, at an angle to the image slice, giving a realistic partial volume effect. (b) error in total lesion volume (for nine lesions with volumes 0.3–6.2 ml) showing large variation with lesion contrast, observer (A, B or C) and outlining method (con: semi-automatic contouring; man: manual). (From Tofts, P.S. *et al.*, *Magn. Reson. Imaging*, 15, 183–192, 1997b.)

measured on this phantom represent lower limits to what might be achieved with *in vivo* measurements, since additional sources of error would be present with the latter. Nonetheless, this type of study represents a reasonable test to apply to a measurement technique, since it will identify any major problems (Figure 3.4).

Importance of accuracy: It has been argued that accuracy is irrelevant in clinical MR measurements, since the systematic error is always present and does not mask group differences. In principle this is true; however actual systematic errors often do not last forever and can change with time (thus forming a contribution to long-term instability or imprecision). An example from spinal cord atrophy measurements shows this (Tofts 1998). The technique (see Figure 3.9) was estimated to have a 6% systematic error, based on scanning a plastic rod immersed in water. The short-term reproducibility was good (0.8% coefficient of variation, CV), and progressive atrophy in MS patients could be seen after about 12 months. After a scanner *software upgrade*, there was an implausible step increase in the normal control values of about 2%. The step change caused by the upgrade prevented atrophy progression through the time of the upgrade from being measured. If the accuracy had been better, and if the sources of systematic error had been understood and controlled, the upgrade would not have been disastrous for this study.

Machine upgrades cannot be avoided; they can only be planned for, and in this context *accuracy provides long-term stability*. As an additional safeguard, if groups of subjects are being compared, subjects from both groups should be collected during the same period, i.e. ‘interleaved controls’. There is a temptation to leave the controls until the end; if there is a step change in the measurement process characteristics after the patients have been measured, but before the controls have been measured, then a group difference cannot be interpreted as caused by disease, since it may have been caused by the change in procedure.

Subtle left–right *asymmetry* or anterior–posterior differences may be seen in a group of subjects. This could be caused by genuine biological difference between the sides or front and back,

or by a subtle asymmetry in the head coil. This can be resolved by scanning some subjects relocated with respect to the head coil, e.g. prone instead of supine.

3.3 Precision

3.3.1 Precision Concepts

Precision, *reproducibility*¹³ or *repeatability* is concerned with whether a measurement agrees with a second measurement of the same quantity, carried out within a short enough time interval that the underlying quantity is considered to have remained constant. Sometimes this is called the *test–retest* performance in psychology. Good within-subject reproducibility is probably the best indicator of good measurement technique (see Figure 3.8); this is why so much attention is paid here to precision. There is also an ISO definition (Padhani *et al.*, 2009) (Figure 3.5a).¹⁴

Measuring precision: Many studies have been published, for many MRI parameters. Its value at a particular site depends on the method used to measure the parameter and is often very sensitive to the precise details of the data collection procedure (such as patient positioning and prescan procedure) and data analysis (particularly region of interest placement). The results of a study may not be generalisable – a poor value of reproducibility may be a reflection of poor local technique at a particular site.

¹³ A measurement is said to be *reproducible* when it can be repeated (*reproduce*: ‘to bring back into existence again, re-create’). However this term is not used by statisticians, who prefer the more precise term *measurement error*. Reproducibility can include factors such as normal short-term biological variation that are not part of measurement error.

¹⁴ According to ISO 5725, *repeatability* refers to test conditions that are as constant as possible, where the same operator using the same equipment within a ‘short time interval’ obtains independent test results with the same method on identical items in the same laboratory. *Reproducibility* refers to test conditions under which results are obtained with the same method on identical test items but in different laboratories with different operators using equipment.

AQ: Please check that “see Figure 3.8” is correct as edited: “[becky]” was deleted

AQ: Please check “Many studies ... for many MRI parameters. Its value at a particular site” for clarity

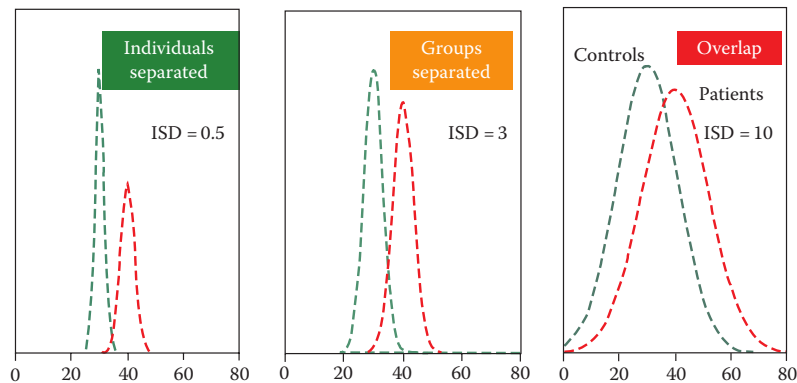


FIGURE 3.5A Simulation showing how magnitude of ISD affects ability to use an MR parameter to separate groups and individuals. Group separation is 10 units. With ISD = 10 (right-hand image), the groups overlap, and considerable statistical power would be needed to separate them (see Chapter 1, Figure 1.3). A reduced ISD = 3 (centre) gives a good group separation c) a further reduction to ISD = 0.5 (left-hand image) enables individuals to be accurately classified into their group.

AQ: Please check that "c" is correct in "gives a good group separation c) a further reduction"

Why measure within-subject reproducibility?

1. It tells you *confidence limits on a single measurement*.

For example, in measuring the concentration of a compound by MRS, the reproducibility (1 sd) is typically 10%. The 95% confidence limit on a single measurement is then 20% (1.96 sd). This means that there is a 95% chance that the true value lies between these limits, and only a 5% chance that it lies outside this range.

2. It tells you the *repeatability* or *minimum detectable difference* that can be measured.

In the above MRS example, the concentration might be estimated on two consecutive occasions, perhaps to look for biochemical effects of progressive disease. The sd in difference measurements is 14% (1.4 times the sd in a single measurement), and the 95%CL on a difference measurement is 28% (1.96 times the sd in difference measurements). Thus unless a measured difference is more than 28%, it cannot be ascribed to a biological cause with a confidence exceeding 95%. If the measured difference is less than 28%, it could have arisen by chance.

FIGURE 3.5B Why measure within-subject reproducibility?

However a good value gives inspiration to other workers to refine their technique. Detailed studies of the various components in a measuring process can identify the major sources of variation; for example rescanning without moving the subject will measure effects such as image noise and patient movement, whilst removing and replacing the subject will also include the effect of positioning the subject in the scanner. This knowledge in turn opens the possibility of reducing the magnitude of the variation by various improvements in technique, ranging from more care, training to reduce interobserver effects (Filippi *et al.*, 1998) to formal mathematical optimisation of the free parameters that define the process (Tofts 1996) (see Chapter 2, Figure 2.12). Measuring the reproducibility of various scanner parameters that are thought to have a large effect on the final MR parameter (such as those set during the prescan procedure) may also be of value.

The methods used to report reproducibility are not always standardised – it is hoped that studies will use instrumental

standard deviation (ISD) and intraclass correlation coefficient (ICC), as described below. Reproducibility may be worse in patients than in normal controls (patients may find it harder to keep still). The reproducibility may depend on the mean value of the parameter (which may be significantly different in patients, for example if there is gross atrophy); see also Figure 3.5. Precision may also have a biological component (see Section 3.3.2.3).

3.3.2 Within-Subject Standard Deviation

3.3.2.1 Bland–Altman and ISD

The simplest and most useful approach to characterising *measurement error* is that of Bland and Altman, which uses pairs of repeated measurements in a range of subjects; the within-subject standard deviation (SD) s of a single measurement, arising primarily from instrumental factors, is estimated (Bland and Altman 1986) (Bland and Altman 1996b;

AQ: Please check that "see also Figure 3.5" is correct as edited; original: "Figure x [BA_PLOT]"

AQ: Please check that "Chapter 2, Figure 2.12" is correct as edited; ["optimisation"] was deleted.

TABLE 3.4 Example of estimating instrumental standard deviation (ISD) using Bland–Altman method

Measurement Set Number	Replicate 1	Replicate 2	Δ Signed Difference		
1	107.14	108.12	0.98	SD of differences $sd\Delta$	6.4
2	103.50	98.60	-4.91	mean_difference	1.6
3	104.65	104.73	0.08		
4	100.97	106.26	5.29	ISD s	4.5
5	96.87	105.76	8.89		
6	90.30	98.76	8.46	σ_s	1.1
7	108.97	98.79	-10.19		
8	104.55	110.24	5.70	95% CL lower	2.4
9	99.55	105.13	5.58	95% CL upper	6.6
10	103.94	99.60	-4.33		

Note: Ten sets of replicate measurements were simulated, drawn from a random normal distribution with mean = 100, SD = 5 (same data set as Figure 3.6). Signed differences were calculated (left-hand table). From these were calculated (right-hand table) their SD ($sd\Delta = 6.4$), ISD $s = 4.5$, the SD of this estimate ($\sigma_s = 1.1$) and 95% confidence limits (CL) for s : 2.4–6.6.

Galbraith *et al.*, 2002; Wei *et al.*, 2002) (Padhani *et al.*, 2002). The 95% confidence limit on a single measurement is 1.96s (Figure 3.1b).

For repeated measurements on the same subject (who is assumed to be unchanging during this process), the measurement values are samples from a normal distribution with SD s . The signed difference Δ between the repeats in pairs of measurements is also normally distributed, with an SD value of $sd\Delta$:

$$sd\Delta = \sqrt{2} s = 1.414s \tag{3.1}$$

Because of the difficulty in making many measurements on the same subject, and because subjects may in any case vary, pairs of measurements (replicates) are usually made on a number of subjects and the difference calculated for each pair. The SD of this set of differences is then calculated ($sd\Delta$), and from this the SD of the measurements on a single subject(s).

Mean absolute difference in pairs of replicates: Instead of taking the signed differences (as in Bland and Altman’s procedure above), the absolute (unsigned) difference is sometimes taken. Its mean value is $0.80 s$ and from this the SD can be found (Table 3.4).¹⁵

The CV in the measurements is the SD divided by the mean value (i.e. $CV = s/\bar{x}$, where \bar{x} is the mean value) and is usually expressed as a percentage.

When using this technique, consideration should be given to what aspect of the measurement process is to be characterised. To assess the whole process, the subject should be taken out of the scanner between replicates, and it may be desirable to carry out the repeat scan a week later, with a different radiographer (technologist). A separate observer, blinded to the first result, could be used for analysis of the replicate. A Bland–Altman plot should be made to check for dependence on the mean value (Figure 3.6).

¹⁵ See 1st edition of this book, page 66.

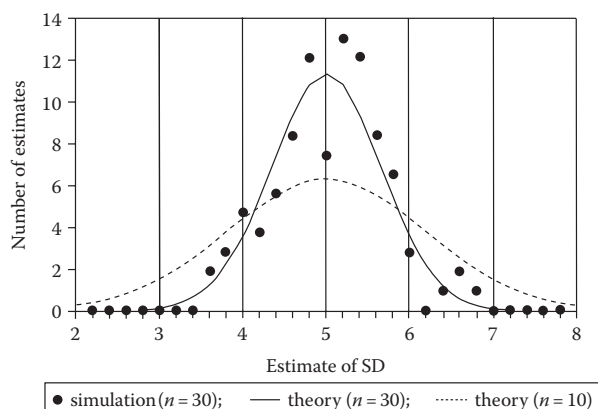


FIGURE 3.6 Simulation of estimation of reproducibility from repeated measurements. Over 8000 samples from a population of random numbers with mean = 100 and SD = 5 were generated. From these, 30 pairs of samples (replicates) were taken, and the differences Δ calculated, retaining the sign of the difference (Δ could be + or -). The SD of the Δ values was found ($sd\Delta$), and from this the SD of the population was estimated (Equation 3.1). Further sets of 30 pairs were taken, to a total of 100 sets, and in each set the population SD estimated. The figure shows the distribution of estimates obtained, showing a mean of 5 (as expected), and clustered mostly between 4 and 6. The theoretical normal distributions are also shown, for 30 and 10 pairs of difference measurements. The theoretical curve for 30 pairs is in agreement with the data. For 30 pairs, an ISD of 0.66 was estimated, which gives a 95% CL of ± 1.3 (Equation 3.2) in estimating s (i.e. 95% of the estimates will lie in the range 3.7–6.3). On the other hand, with only 10 pairs, this range increases to 2.7–7.3; reducing the number of pairs has reduced the precision with which the ISD can be estimated. See also Table 3.4.

Estimation of s , also called the *within-subject variability*, in the underlying distribution of measurements (all with the same mean) characterises the measurement process. From this,

the coefficient of *repeatability* $\sqrt{2} \times 1.96s = 2.77s$ can be found (assuming there is no bias between the first and the second measurement). The difference between two measurements, for the same subject, is expected to be less than the repeatability for 95% of the pairs of observations. Thus for a biological change to be detected in a single subject with 95% confidence, it must exceed the repeatability. These lower and upper limits to differences that can arise from measurement error are sometimes **called the limits of agreement** (Bland & Altman 1986).

Agreement between two instruments has two components: bias (systematic difference) and variability (random differences). Under normal conditions the mean difference between the first and second measurements is expected to be zero, if they come from a set of repeats made under identical conditions. However if two separate occasions, two observers or two scanners are being compared, then a test for bias should be made, using a two-tailed t-test. If the differences are not normally distributed, a Wilcoxon signed rank test is needed.

3.3.2.2 Dependence of SD on Mean Value

The approach above supposes that the mean value in each pair is similar, so that the differences from paired measurements can be pooled. This assumption can be tested in a *Bland–Altman plot*, where the sd is plotted against mean value (Bland and Altman 1986) (Krummenauer and Doll 2000). Any important relationship should be fairly obvious, but an analytic check can be made using a rank correlation coefficient (Kendall's tau) (Bland and Altman 1996a). If SD increases with mean value, (which is often the case) it may need to be transformed in some way to give a quantity that varies less with mean value. For the situation where SD is proportional to the mean, a log transformation is appropriate (Bland & Altman 1996c), although the interpretation of the transformed variable is not so straightforward. An alternative is to use the CV, which is constant under the condition of SD proportional to the mean. For measurements of total lesion volume in MS, the CV is relatively constant over a wide range of volumes (or at least there is no clear evidence of it changing in a systematic way) (see Figure 3.7). In this case the estimates of

CV at different volumes can then legitimately be pooled to give a single, more precise value.

In the Bland–Altman approach, the *uncertainty of the estimate of SD* can be found. The uncertainty (one standard deviation) in estimating an SD (s) from n samples is as follows (Taylor 1997; page 298).

$$\sigma_s = \frac{s}{\sqrt{2(n-1)}} \quad (3.2)$$

See Table 3.4 for an example.

3.3.2.3 Biological Variation

Precision may have a significant biological component, in that intrasubject variation may be significant and limit the usefulness of having good machine precision. Thus blood flow varies by about 10% within a day (Parkes *et al.*, 2004), so if a single number is required to characterise the individual, high precision is not required. However if these biological changes are to be studied in detail, for example to find their origin, then a much better instrumental precision would be needed.

Biological variation at time scales longer than a few minutes can be measured using repeated measurements, provided the machine variation (ISD) is known (e.g. from phantoms or fast repeats when the biology is known to be static). Short-term variation might be accessible by the device of *data fractionation*. The data collection procedure is altered, if necessary, to acquire two independent datasets as simultaneously as possible. The easiest way to do this is to use two signal averages for each phase encode and preserve them without addition. Typically the averages are separated by a second or less of time. Two image datasets are then constructed, and differences measured from these, to estimate instrumental precision. These image datasets are statistically completely independent, yet form samples of the biology separated by a second or less.

Estimation of biological variation is important in the context of creating a 'perfect MRI machine', which contributes no extra variance (see Chapter 1, Section 4.2).

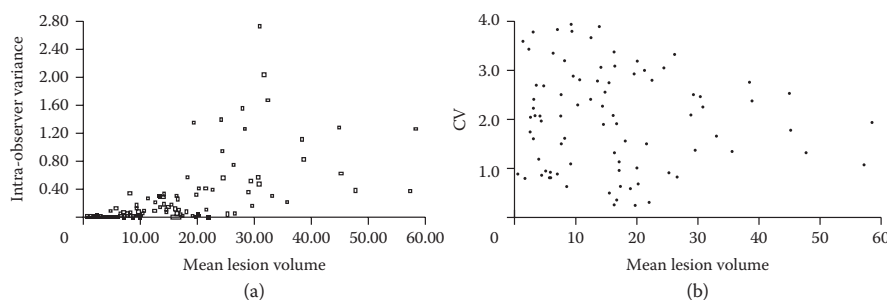


FIGURE 3.7 Bland–Altman plots for estimates of total lesion volume in multiple sclerosis. (a) The variance (var) increases with mean lesion volume (MLV); therefore variance values cannot be pooled. (b) The coefficient of variation (CV) is independent of MLV (i.e. there is no sign of a systematic dependence of CV on MLV); therefore the CV values can be pooled to form a single average value. (From Rovaris, M., *et al.*, *Magn. Reson. Imaging*, 16, 1185–1189, 1998.)

AQ: Please note that text "(see figure box why measure...)" was deleted here.

3.3.3 Intraclass Correlation Coefficient or Reliability

This measure considers both the within-subject (intrasubject) variance arising from measurement error (which we have considered in the previous section) and variance arising from the difference between subjects (Armitage *et al.*, 2001) (Cohen *et al.*, 2000). If there is a large variance between the subjects (intersubject), measurement variance may be less important, particularly if groups are being compared. The ICC is

$$ICC = \frac{\text{variance from subjects}}{\text{variance from subjects} + \text{variance from measurement error}} \quad (3.3)$$

The ICC can be thought of being the fraction of the total variance that is attributed to the subjects (rather than measurement error). Thus if measurement error is small compared to the subject variance, ICC approaches 1. Typical values in good studies would be at least 0.9. ICC as a measure has the benefit of placing measurement error in the context of the subjects, and potentially it can stop us being overly concerned about measurement error when subject variance is large.

However ICC has at least two problems. ICC depends on the group of subjects being studied (Bland and Altman 1996c), and a determination in one group does not tell us the value in another group. For example, in normal subjects (who often form a homogeneous group), ICC may be unacceptably low, whilst in patients (who are naturally more heterogeneous) the ICC may be adequate. Secondly, when studying individual patients, and their subtle MR response to treatment, the crucial parameter is the repeatability (or the within-subject standard deviation, from which it is derived), as this is the smallest biological change that can reliably be detected, and ICC has little value.

The ICC is often called the *reliability* (Cohen *et al.*, 2000; Armitage *et al.*, 2001). Reliability is discussed with insight by Streiner and Norman (1995). Although the ICC is not an absolute characteristic of the instrument, it is favoured by many researchers (Chard *et al.*, 2002); see Chapter 1, Section 1.3.3 on **psychometric measures**. It is probably best to measure both ICC and ISD.

3.3.4 Analysis of Variance Components

This quite complex analysis is carried out by repeating various parts of the measurement procedure, as well as the whole procedure (see e.g. Chard *et al.*, 2002). The variance arising from different parts of the measurement procedure can be estimated, as well as intersubject and interscanner effects. A model of the variance is first prescribed, with possible interactions, such as allowing some of the variance components to depend on subject or scanner. The measurement can be repeated without removing the subject from the scanner ('within-session variance'), then removed and re-scanned ('intersession variance').

Within-session variance has noise and patient movement (including pulsation); intersession variance also has repositioning (and possibly longer-term biological variation).

3.3.5 Other Methods

3.3.5.1 Correlation

In a set of repeated measures, the first result can be correlated with the second one, and high correlation coefficients are usually produced when this is done. However this approach has little value and does not give an indication of agreement between pairs of measurement (Bland and Altman 1986). In a trivial example, the measures could differ by large amounts, e.g. one might be twice the other, and a good correlation could still be produced. A large intersubject variation will also increase its value (Bland and Altman 1996c). Good correlation does not imply good agreement.

3.3.5.2 Kappa Coefficient

This is used for categorical or ordinal data (Armitage *et al.*, 2001), where there are few possible outcomes and is not appropriate for continuous quantitative data.

3.4 Healthy Controls for QA

The range of values measured in healthy controls ('normals') can be quite small for some parameters, notably T_1 , ADC and MTR (Table 3.5). Within-centre CVs of 3%–5% have been achieved for T_1 and ADC, and under 2% for MTR. Between-centre differences are larger (see Section 3.1.3). Values usually depend on location in the brain and age (Silver *et al.*, 1997).

The measured normal range at a centre is influenced by the centre's ISD (measured from repeats – see Section 3.3.2.1). Broadly speaking, the measured spread of values is a convolution of the actual biological spread and that introduced by the instrument. A reduction in ISD can make a dramatic reduction in measured normal range (see Figure 3.8).

Thus healthy controls can be used for QA both within-centre and between-centre. Within-centre stability can be monitored

TABLE 3.5A Normal Range of T_1 Values at 1.5T in White Matter

Study ^a	CV ^d (%)	n^a	Mean (ms)	SD (ms)
Stevenson 2000	5	40	666	36 ^b
Rutgers 2002	6	15	681	40
Ethofer 2003	4	8	770 ^c	30

Source: Adapted from Tofts, P.S., and Collins, D.J., *Br. J. Radiol.*, 84 Spec No 2, S213–S226, 2011.

Note: See also Chapter 5, Table 5.1, for a fuller list of values; coefficients of variation are about 3%.

^a Sample size.

^b Estimated from boxplot in figure.

^c Used spectroscopic technique; probably some cerebrospinal fluid or grey matter contamination.

^d Coefficient of variation = SD/mean.

^e References for all studies are given in the original tables (Tofts and Collins 2011).

AQ: Please check "Chapter 1, Section 1.3.3, on psychometric measures" is correct as edited; original: "Section 3.2.3"

AQ: Please spell out ACD and MTR at first mention.

AQ: Please suggest whether Table 3.5 a,b,c be changed to Table 3.5, 3.6 and 3.7, and the rest of the tables be renumbered accordingly to follow sequential order.

AQ: Please check that "see Figure 3.8" is correct as edited; "becky" deleted

AQ: Please check footnote ^a Estimated from boxplot in figure" for clarity; please provide figure number if applicable.

TABLE 3.5B Normal Range of Mean Diffusivity Values in White Matter

Study	CV (%)	n	Mean (10 ⁻⁹ m ² s ⁻¹)	SD (10 ⁻⁹ m ² s ⁻¹)	
Cercignani	2001	5	20	0.93 ^a	0.04
Emmer	2006	4	12	0.84	0.03
Zhang	2007	5	29	0.69	0.04
Welsh	2007	3	21	0.73	0.02

Source: Adapted from Tofts, P.S., and Collins, D.J., *Br. J. Radiol.*, 84 Spec No 2, S213–S226, 2011.

Note: See also Chapter 8.

^a Some cerebrospinal fluid contamination.

AQ: Please spell out SEM and MTR beneath Table 3.5c

TABLE 3.5C Normal Range of MTR Values in White Matter

Study	CV (%)	n	Mean (pu) ^c	SD (pu)	
Silver	1997	1.9	41	39.5	0.76 ^a
Davies	2005	1.0	19	38.4	0.4
Tofts	2006	1.6	10	37.3 ^b	0.6

Source: Adapted from Tofts, P.S., and Collins, D.J., *Br. J. Radiol.*, 84 Spec No 2, S213–S226, 2011.

Note: See also Figure 3.8, which shows SD values of 0.5–1.0 pu.

^a SEM = 0.17 pu; 4 samples each n = 20 or 21; estimated SD = 0.76 pu.

^b Peak location values in white matter histograms.

^c MTR values not comparable between studies (different sequences).

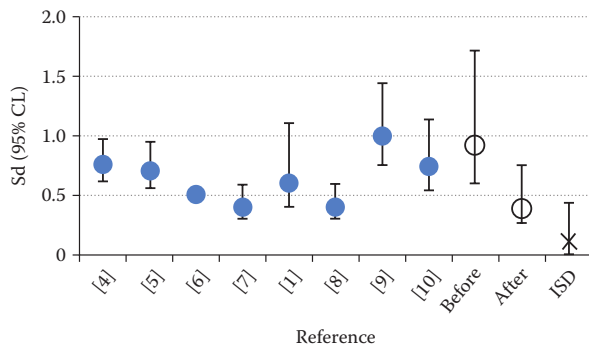


FIGURE 3.8 Normal variation for white matter MTR, and influence of ISD. Blue circles are published values of SD (units for MTR are pu; mean was 38–40 pu) from eight centres; error bars show uncertainty in sd estimate (Equation 3.2). *Before* is authors' first value, almost the highest value of nine centres. After solving a scanner instability problem (Figure [stability] in Chapter 2), ISD was low (≈ 0.2 pu) and the re-measured normal range (*after*) dropped to the lowest value of nine centres. (Adapted from Haynes, B.I., *et al.*, *Measuring scan-rescan reliability in quantitative brain imaging reveals instability in an apparently healthy imager and improves statistical power in a clinical study*, p. 2999, 2010.)

AQ: Please provide correct figure number for "Figure [stability] in Chapter 2".

AQ: Haynes et al. (2010). Please provide publisher location.

AQ: Please check that "see Figure 3.9" is correct as edited; 'leary' deleted.

using a few easily available controls who are likely to remain accessible for a long time (see Figure 3.9). Between-centre differences can be studied and minimised using controls at each centre (Tofts *et al.*, 2006). Although T_1 and ADC are the most explored parameters for QA using healthy controls, other parameters may reach this level of standardisation (e.g. magnetic resonance spectroscopy [MRS] metabolite concentrations).

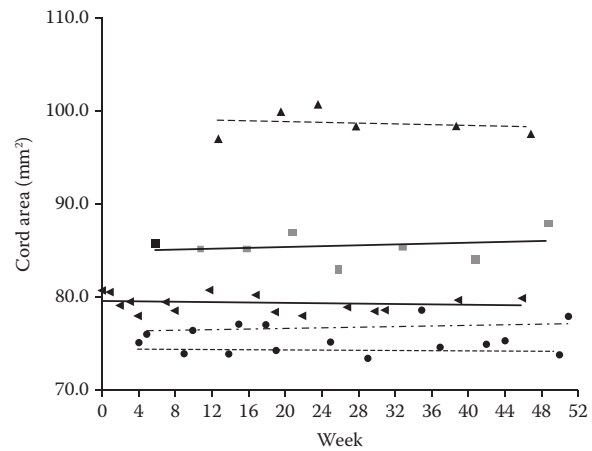


FIGURE 3.9 Early example of quantitative QA in the spinal cord. Data on spinal cord cross-sectional area for five normal controls, which has a short-term precision of 0.8% (CV). The lines are linear regressions. (From Leary, S.M., *et al.*, *Magn. Reson. Imaging*, 17, 773–776, 1999.)

3.5 Phantoms (Test Objects)

3.5.1 Phantom Concepts

Phantom designs for T_1 , T_2 , ADC and PD are the most developed; these can be made from a single component, or mixtures. Geometric objects, used for size or volume standards, are often made of acrylic.¹⁶ These are immersed in water (doped to reduce its T_1 and T_2 values). Objects with a specified T_1 , T_2 or diffusion value can be made from a container filled with liquid or gel, often with various salts added to reduce the relaxation times. Chemical compounds are available from suppliers such as Sigma-Aldrich. Phantoms should ideally be stable with known properties. If a design is to be made into identical phantoms at several centres, as part of a multicentre study, then care is needed on selecting and measuring out the components used in the construction.

AQ: Please spell out PD at first mention.

Institutional constraints: Those wishing to provide quantitative techniques for clinical studies should be warned that some institutional representatives, operating in a paradigm of Health and Safety, or ethics, can object to the use of phantoms and volunteers, and slow down the progress of clinical studies. Phantoms might leak or be damaged, toxic substances might be ingested; ready-made phantoms overcome this objection, though often at considerable cost. Volunteers from the scientist's institution might feel pressurised to volunteer; those from outside might not be covered by insurance. Sometimes a qualitative risk assessment is sufficient to allow progress. Objections might be countered by quoting ethics norms from the paradigm of a chemistry laboratory, or considering the Health and Safety of the patient group whom the clinical study seeks to aid.

AQ: Please check whether "Health and Safety" may be lowercased (if it is not referring to e.g. a department).

¹⁶ Major manufacturers are Perspex in the UK and Plexiglas in North America.

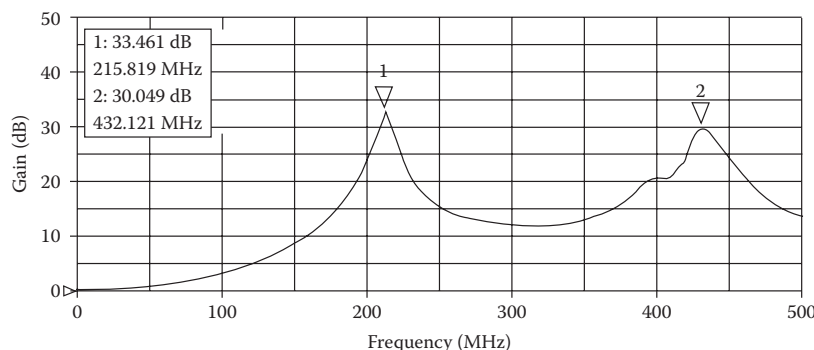


FIGURE 3.10 Dielectric resonance in a spherical flask of water. One small radio frequency coil was placed inside the 2 litre flask (diameter 156 mm), and one outside. The graph shows the transmission between one coil and the other. The plot is the same regardless of whether the inner coil or the outer coil transmits (an example of the principle of reciprocity). The resonances correspond to wavelengths in water of one diameter and half a diameter. Without water the plot is flat. Adding salt to the water damps the resonances. The lower resonance, at 216 MHz, corresponds to 5.1T for protons. (From Hoult, D.I., *Concepts Magn. Reson.*, 12, 173–187, 2000.)

3.5.2 Single Component Liquids

These may be water, oils or organic liquids such as alkanes. They all have the advantage of being readily available either in the laboratory, from laboratory suppliers or from the supermarket, at reasonable prices. No mixing, preparation, weighing or cookery is required. The only equipment needed is a supply of suitable containers. Handling the alkanes should be carried out in accordance with national health and safety regulations.¹⁷

Water has the advantage of being easily available and of a standard composition. Its intrinsic $T_1 \approx 3.3s$, $T_2 \approx 2.5s$ at room temperature (see Table 3.8), and in its pure form these long relaxation times usually cause problems. The long T_1 can lead to incomplete relaxation with sequences that may allow full relaxation with normal brain tissue ($T_1 \approx 600\text{--}800$ ms for normal white matter at 1.5T and 3T – see Chapter 5, Table 5.1). The long T_2 can cause transverse magnetization coherences that would be absent in normal brain tissue ($T \approx 90\text{--}100$ ms). Doped water overcomes these problems (see Section 3.5.3). The low viscosity can also cause problems, with internal movement continuing for some time after a phantom has been moved, giving an artificial and variable loss of transverse magnetization in spin echo sequences used for T_2 or diffusion.

Water has another particular disadvantage when used in large volumes. Its high dielectric constant ($\epsilon = 80$) leads to the presence of radio frequency standing waves (dielectric resonance), where B_1 is enhanced, giving an artificially high flip angle and signal (see Figure 3.10). The high dielectric constant reduces the wavelength of electromagnetic radiation, compared to its value in free space, by a factor $\sqrt{\epsilon}$; at 3T the wavelength is 260 mm, comparable with the dimensions of a head phantom (Glover *et al.*, 1985; Tofts 1994; Hoult 2000). Standing waves are also present

¹⁷ In the UK this involves registering the project with a safety representative, using basic protective clothing and carrying out the pouring operation in a fume cupboard.

TABLE 3.6 Radio Frequency Non-uniformity in Uniform Phantoms

Field B_0	Water ($\epsilon = 80$)	Oil ($\epsilon = 5$)
0.5T	138 mm	551 mm
1.5T	46 mm	184 mm
4.7T	15 mm	59 mm

Note: The maximum diameter of a long cylinder phantom for assessing coil uniformity is given, under the condition that the signal is not to increase by more than 2% as a result of dielectric resonance in the cylinder. A circularly polarised radio frequency coil is assumed. Filling with a low dielectric constant oil ($\epsilon = 5$) allows larger phantoms to be used. (Adapted from Tofts 1994.)

in the head, particularly at high field (see Chapter 2, Figure [RFNU_in_head]), but to a much less extent, because electrical conductivity in the brain tissues damps the resonance. Even at 1.5T this effect is significant, and early attempts to measure head coil non-uniformity using large aqueous phantoms are now seen as fatally flawed.

Iced water has been used as a diffusion standard (Malyarenko *et al.*, 2013) (Table 3.6).

Oil has a low dielectric constant ($\epsilon = 2\text{--}3$) and has been used for non-uniformity phantoms (Tofts *et al.*, 1997a). Several kinds are available, from various sources, with differing properties. It is stable and cheap; cooking oil is a convenient source. Some are too flammable to use in large quantities. Sources with good long-term reproducibility between samples may be hard to find. T_1 and T_2 values may be closer to *in vivo* values (T_2 values are convenient, at 33–110 ms, whilst T_1 values are generally too low, at 100–190 ms, although some flammable oils have higher values).

Silicone oils of different molecular sizes have been used to obtain a range of T_1 and T_2 values (Leach *et al.*, 1995); pure 66.9 Pa s viscosity polydimethylsiloxane gave $T_1 \approx 800$ ms, $T_2 \approx 100$ ms at 1.5T.

Organic liquids such as alkanes have been used for diffusion standards (Holz *et al.*, 2000; Tofts *et al.*, 2000). Cyclic alkanes

AQ: Please check that "see Section 3.5.3" is correct as edited; original: "see next section".

AQ: Please check that "see Figure 3.10" is correct as edited; "[Hoult]" was deleted.

AQ: Please check that RF was correctly spelled out as "radio frequency" in Table 3.6 heading.

AQ: Please check that RF was correctly spelled out as "radio frequency" in heading.

AQ: Please insert correct figure number for "Figure [RFNU_in_head]"

C_nH_{2n} ($n = 6-8$) are the simplest possible set of organic liquids, with a single proton spectroscopic line. There are only three easily available, and they are toxic. Linear alkanes C_nH_{2n+2} ($n = 6-16$) are the next simplest set; 11 are readily available, ranging from hexane (which is very volatile, and inflammable) through octane (a major constituent of petrol [gasoline]), to hexadecane (which freezes at 15°C). Their T_1 values are realistic (670–1900 ms), but the T_2 values are rather long (140–200 ms), and currently it is not possible to dope them to reduce the relaxation times. Their diffusion values are ideal, covering the range found in human tissue. Dodecane ($n = 12$) has a diffusion coefficient of $0.8 \cdot 10^{-9} \text{ m}^2\text{s}^{-1}$, close to the mean diffusivity of normal white matter. Their viscosity is higher than that of water, forcing bulk liquid motion to be rapidly damped. The liquids are anhydrous, so they either should be sealed well or be replaced regularly.

3.5.3 Multiple Component Mixtures for T_1 and T_2

Doped water has reduced T_1 and T_2 , giving a material with more realistic values of relaxation times. Doping compounds are characterised by their relaxivities r_1 and r_2 , which describe how much the relaxation rate $R_{1,2}$ ($R_{1,2} = 1/T_{1,2}$) is increased by adding a particular amount of the compound. In aqueous solution:

$$\frac{1}{T_1} = R_1 = R_{10} + r_1c; \quad \frac{1}{T_2} = R_2 = R_{20} + r_2c \quad (3.4)$$

R_{10} and R_{20} are the relaxation rates of pure water; c is the concentration of the doping compound, and the increase in relaxation rate is proportional to the concentration (Figure 3.11).

The classic compounds used for doping have been copper sulphate CuSO_4 and manganese chloride MnCl_2 ; nickel Ni^{++} has the advantage of a low T_1 temperature coefficient (see Section 3.5.4). **Gd-DPTA** is widely available. Agarose is good for reducing T_2 whilst hardly affecting T_1 . MnCl_2 is a convenient way of reducing T_2 without the complexity of gel manufacture (Table 3.7).

T_1 of water: The value of this is needed to make up mixtures. (T_2 is less important, because tissue-like phantoms have a much lower T_2 than T_1 , and therefore water has less effect on the final T_2 value.) Water T_1 depends on the amount of dissolved oxygen. It is independent of frequency (Krynicky 1966) (Table 3.8).

Dissolved oxygen: Water used for making phantoms is likely to have some oxygen in it (depending on whether it was recently boiled). The relaxivity for oxygen is approximately $1.8 \pm 0.3 \cdot 10^{-4} \text{ s}^{-1} (\text{mmHg})^{-1}$ (measured in plasma at 4.7T by Meyer *et al.*, 1995). Assuming this value still holds at 3T, fully oxygenated water at 23°C ($\text{pO}_2 = 150 \text{ mmHg}$) would then have its T_1 reduced from 3.40s to 3.11s, a reduction of 8%. A modern systematic measurement of water T_1 values, under varying conditions of temperature and pO_2 , would be valuable, particularly if

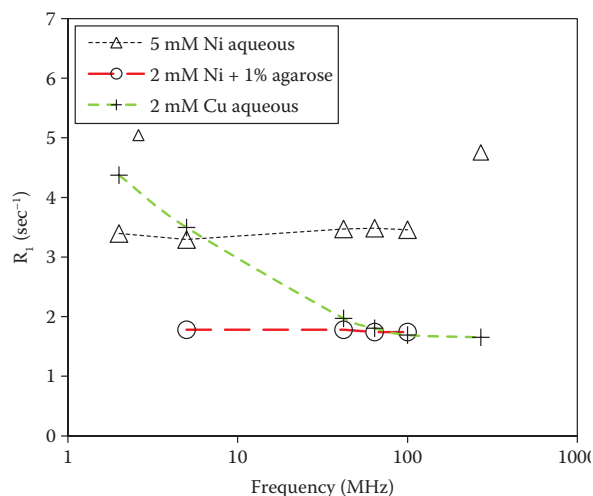


FIGURE 3.11 Field dependence of proton relaxation rate $R_1 = 1/T_1$ for Ni^{++} in aqueous solution and agarose gel and for Cu^{++} in aqueous solution. Cu^{++} has a large frequency dependence. Ni^{++} is independent of frequency up to at least 100 MHz (one point at 270 MHz clearly has a higher R_1 value). Frequency values include 42 MHz (1.0T), 64 MHz (1.5T), 100 MHz (2.4T) and 270 MHz (6.3T). (Re-drawn from Kraft, K.A., *et al.*, *Magn. Reson. Med.*, 5, 555–562, 1987.)

accompanied by T_2 values (high quality measurements of water T_2 seem to be completely lacking).

Doped agarose gels can be made up in a similar way to doped water (Mitchell *et al.*, 1986; Walker *et al.*, 1988) (Walker *et al.*, 1989; Christoffersson *et al.*, 1991; Tofts *et al.*, 1993). There is more control over the values of T_1 and T_2 that can be obtained, since agarose has a high r_2 and low r_1 (see Table 3.10). Agarose flakes are dissolved in hot water, up to concentrations of about 6%, in a similar way to making fruit jelly. A hotplate (Mitchell *et al.*, 1986) or a microwave oven (Tofts *et al.*, 1993) can be used. Stirring is necessary, and care must be taken not to overheat the gel. Fungicide can be added to improve stability. Agarose is relatively expensive if large volumes are to be made up; cooking it is a relatively complex process, and obtaining a uniform gel on cooling also requires skill. Commercially available doped gels with a wide range of T_1 and T_2 values are obtainable (see Section 3.5.6.4); however for many applications single liquids or aqueous solutions will suffice.

By using a *mixture of two compounds*, a range of T_1 and T_2 values can be obtained, intermediate between those that would be obtained with only one of the compounds. (Mitchell *et al.*, 1986; Schneiders 1988; Tofts *et al.*, 1993). It is important to establish that the two components do not interact; this can be done by plotting relaxation rates vs. concentration for the individual components (to establish their relaxivities) and then for mixtures (to show that the individual relaxivities are unaffected). The most useful combinations are pairs where one has high r_2 (much greater than r_1 , i.e. MnCl_2 or agarose) and the other has low r_2 (about the same as r_1). Thus suitable mixtures are Ni^{++} and

AQ: Please spell out DPTA at first mention

AQ: Table 3.10 is not found in this chapter. Please check.

TABLE 3.7 Values of Relaxivity at 1.5T^b and Room Temperature

Relaxation Agent ^t	Source	r_1 (s ⁻¹ mM ⁻¹)	r_2 (s ⁻¹ mM ⁻¹)
T_1			
Ni ⁺⁺	Morgan and Nolle 1959 ^d	0.70 ± 0.06	0.70 ± 0.06
	Kraft <i>et al.</i> , 1987 ^{a,c}	0.64	–
	Jones 1997 ^f	0.644 ± 0.002	0.698 ± 0.005
Gd-DTPA	Tofts <i>et al.</i> , 1993 ^b	4.50 ± 0.04	5.49 ± 0.06
T_2			
Mn ⁺⁺	Morgan and Nolle 1959 ^d	7.0 ± 0.4	70 ± 4
	Bloembergen and Morgan 1961 ^g	8.0 ± 0.4	80 ± 7
Agarose	Mitchell <i>et al.</i> , 1986 ^c	0.05	10
	Tofts <i>et al.</i> , 1993	0.01 ± 0.01	9.7 ± 0.2
	Jones 1997 ^f	0.04 ± 0.01	8.80 ± 0.04

^a See Figure 3.11.^b Gd-DTPA r_1 is independent of field up to 4.7T (data at 37°C; Rohrer *et al.*, 2005).^c Data at 5 and 60 MHz.^d At 60MHz, 27°C, calculated by the author from data points on the published figures; 95% confidence limits estimated from scatter in the plots.^e Estimated from published T_1 value.^f Estimated from data at 1.5T in the MSc thesis of Craig K Jones (University of British Columbia 1997) (for more details see 1st edition); 2mM Ni⁺⁺ in 1% agarose gives $T_1 = 573$ ms, $T_2 = 95$ ms.^g At 60 MHz, 23°C, calculated by the author from data points on the published figures; 95% confidence limits estimated from scatter in the plots.^h There are very few published data at 3T and above; relaxivities for these four agents are similar to values at 1.5T.ⁱ More data are shown in 1st edition, Table 3.5.

AQ: Please provide "Table 3.8" caption is missing.

TABLE 3.8

Temperature (°C)	T_1 (s)
0	1.73
5	2.07
10	2.39
15	2.76
20	3.15
21*	3.23
22*	3.32
23*	3.40
24*	3.49
25	3.57
37*	4.70

T_1 values for pure water. Measurements were made at 28 MHz using a continuous-wave saturation-recovery technique; estimated 95% confidence limits were ±3%. Values marked (*) are linearly interpolated (from data in Krynicky 1966; see also Tofts *et al.*, 2008). Note values are expected to be independent of field strength.

Mn⁺⁺ in aqueous solution (Schneiders 1988), Gd-DTPA¹⁸ and agarose (Walker *et al.*, 1989), Ni⁺⁺ and agarose (Kraft *et al.*, 1987), and Ni-DTPA and agarose (Tofts *et al.*, 1993). Linear equations can be produced giving the concentrations of each compound required, for a target T_1 and T_2 value, given the relaxivities of each component, and the T_1 and T_2 of pure water (Tofts *et al.*, 1993) (Figure 3.12).

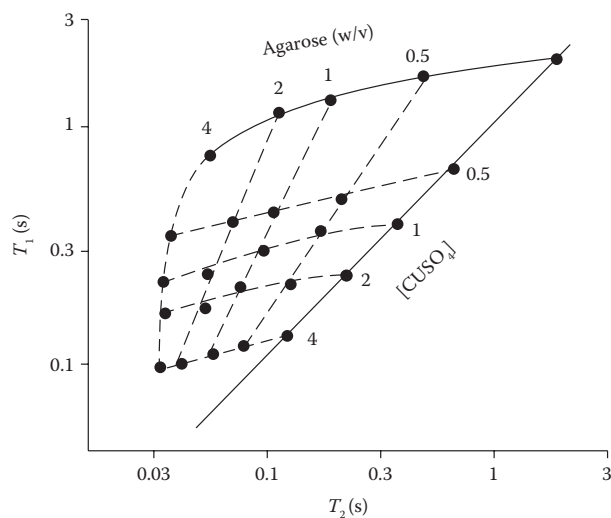
¹⁸ This is preferred to GdCl₃, which interacts with the agarose.

FIGURE 3.12 T_1 and T_2 values for aqueous solutions of agarose and CuSO₄. Agarose concentrations are from 0% to 4% weight/volume, Cu is from 0 to 4 mM. Note agarose decreases T_2 but hardly affects T_1 , whilst Cu decreases T_1 and T_2 about equally. The agarose line (curved, where [Cu] = 0) and the Cu line (straight, where [agarose] = 0) bound the possible values that the mixture can achieve. Dotted lines connect points of equal agarose or Cu concentration. (Data at 5 MHz from Mitchell, M.D., *et al.*, *Magn. Reson. Imaging*, 4, 263–266, 1986).

AQ: Please spell out DTPA beneath Table 3.7

AQ: Note that Jones (1997) is cited here and elsewhere in the text but not in the list. Please provide complete reference details.

A mixture of Ni^{++} in agarose provides reduced temperature dependence for T_1 . For example a phantom with $T_1 = 600$ ms, $T_2 = 100$ ms at 1.5T is produced by mixing 1.77mM Ni^{++} in 0.96% agarose.¹⁹ Relaxation in Ni^{++} is dominated by fast electron interactions, which are independent of temperature; this also increases the frequency up to which relaxation is almost independent of frequency, although above 4T other relaxation mechanisms come into play (Kraft *et al.*, 1987).

The process of making up the mixtures can be simplified by making up concentrated stock solutions of the components. The required T_1 and T_2 values can be entered into a spreadsheet, along with the relaxivities and stock solution concentrations, to give a simple list of how much stock solution must be added to a particular volume of water to give the required relaxation times.

3.5.4 Other Materials

Aqueous *sucrose solutions* have been used for diffusion standards; these are easily made up, and T_1 and T_2 can be controlled by doping (Laubach *et al.*, 1998; Delakis *et al.*, 2004; Lavdas *et al.*, 2013).

PVP (Polyvinylpyrrolidone) is biologically benign (used in postage stamp adhesive and shampoo) and stable. A 25% solution in water at 0°C has $T_1 = 533$ ms, $T_2 = 519$ ms at 1.5T and $T_1 = 610$ ms, $T_2 = 500$ ms at 3T. ADC is $0.49 \cdot 10^{-9} \text{ m}^2\text{s}^{-1}$ (Jerome *et al.*, 2016; Pullens *et al.*, 2017). PVP was first suggested by Pierpaoli *et al.* in 2009.

Various gels have been used as *MRI radiation dosimeters* (Lepage *et al.*, 2001); the T_2 decreases with dose and is read out after irradiation. In this context, there has been much attention devoted to designing stable gels (De Deene *et al.*, 2000), and this work may result in new designs for MRI QA materials.

3.5.5 Temperature Dependence and Control

3.5.5.1 Temperature Dependence

Temperature dependence of phantom parameter values can be a major problem. The scanner room environment may vary by 1° or 2°C, unless special precautions are taken. The magnet bore may have a colder supply of air blown in to assist the breathing of the MRI subject. A refrigerated phantom could be as cold as 5°C. A phantom positioned next to the subject could warm above room temperature.

T_1 , T_2 and D all vary by about 1%–3%/°C, corresponding to errors of about 5% in the parameter value in an uncontrolled environment. The Eurospin gels (Lerski and de Certaines 1993) have a T_1 temperature coefficient of +2.6%/°C.²⁰ In alkane phantoms, the diffusion coefficient changes by 2%–3%/°C

(Tofts *et al.*, 2000). Agarose has a T_2 coefficient of about –1.25%/°C (Tofts *et al.*, 1993).²¹

PD and MRS concentration measurements are also vulnerable to temperature change; both the density and the magnetic susceptibility vary considerably between refrigerator, room and body temperature. An accurate correction is possible (Section 3.5.5.5).

The effects of temperature dependence can be mitigated by controlling or correcting for the environmental temperature or by using compounds (principally *Ni*) with reduced temperature dependence.

3.5.5.2 Controlling Environmental Temperature

Phantoms should be stored in the scanner room at room temperature (not refrigerated). The phantoms should be thermally insulating during their time in the magnet bore (which may have a different and varying temperature), and their temperature should be measured (ideally whilst in the bore, using a thermocouple²² or a liquid-in-glass thermometer; Tofts and Collins 2011). Temperatures should be known to within better than 1°C, and ideally 0.2°C, in order to allow MR measurements to within 1%. Temperature gradients within the phantoms can be minimised by avoiding both rapid changes in temperature and the presence of any electrical conductors, which might attract induced RF currents and consequent heating. Thermal insulation foam is effective and widely available for house insulation.

Iced water can also be used as a bath for temperature control (Malyarenko *et al.*, 2013; Jerome *et al.*, 2016).

3.5.5.3 Correcting for Phantom Temperature

If the phantom temperature varies with each measurement occasion, a correction procedure may still be possible. The temperature coefficient has to be known, the phantom must have a well-defined single temperature (without any temperature gradients), and the temperature must be measured on each occasion. The measured parameter value (e.g. T_1) can then be converted to its estimated value at a standard temperature (e.g. 20°C) (Vassiliou *et al.*, 2016).

3.5.5.4 Compounds with Reduced Temperature Dependence for T_1

Ni has a minimum in its T_1 relaxation rate, fortuitously at room temperature (Kraft *et al.*, 1987; Tofts *et al.*, 1993), allowing a brain-equivalent *Ni*-DTPA agarose gel to have a flat temperature response (Figure 3.12). At 1.5T and 37°C, Ni^{++} agarose phantoms had temperature coefficients of +0.05% K^{-1} (for 530 ms) and +0.7 K^{-1} (for 900 ms) (Vassiliou *et al.*, 2016) (Figure 3.13).

¹⁹ See 1st edition, page 73.

²⁰ This coefficient, calculated from values given in the manual, is approximately independent of T_1 and T_2 , since the T_1 behaviour is almost completely determined by the Gd-DTPA.

²¹ Walker *et al.* (1988) give a theoretical temperature coefficient in 2% agarose ($T_2 = 60$ ms) at 20 MHz of 1.7%/°C. The Eurospin gels closest to brain tissue have a T_2 coefficient of about 1.5%/°C; most of this originates in the agarose, but it may be attenuated slightly by the positive coefficient of the Gd-DTPA.

²² A thin T-type thermocouple has signal dropout limited to within about 3 mm of the tip.

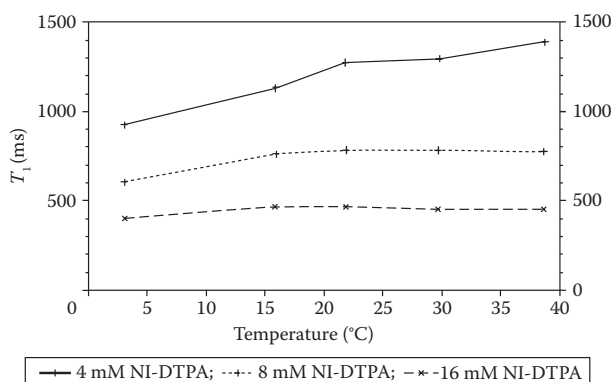


FIGURE 3.13 Temperature dependence of T_1 in a Ni-doped gel at 1.5T. In the tissue-equivalent material (Ni-DTPA in 2% agarose), T_1 is dominated by the relaxation from Ni-DTPA, particularly at the lower T_1 values, and therefore has little dependence on temperature. At room temperature the 8 mM and 16 mM data are flat. Materials with these concentrations of Ni-DTPA, in 1% agarose, have ($T_1 = 909$ ms, $T_2 = 99$ ms) and ($T_1 = 510$ ms, $T_2 = 89$ ms), covering the range of normal brain tissue. (From Tofts, P.S., *et al.*, *Magn. Reson. Imaging*, 11, 125–133, 1993.)

A more general solution is to use a Gd polymer pair (Kellar and Briley-Saebo 1998).²³ One component (NC663868) has a zero temperature coefficient for T_1 . The second component (NC22181) has a negative coefficient (about $-1.2\%/^{\circ}\text{C}$) and can be used to neutralise the small effect arising from the positive coefficient of the host material (water and/or agarose). Thus the pair, used in agarose solution, can give zero temperature coefficients for a range of T_1 values.

3.5.5.5 PD and MRS Concentrations: Correction for Temperature

When measuring proton density (Chapter 4) or metabolite concentration by spectroscopy (Chapters 12 and 13), the signal from a concentration standard is often used to measure the absolute gain of the MR system.

The signal from a test object or standard at *room temperature* differs from the same object at *body temperature* for two reasons: first the magnetization M_0 of a given number of protons is inversely proportional to absolute temperature (Chapter 2, Equation 2.2). This corresponds to a reduction of $0.34\%/K$ at 20°C . Thus the signal at room temperature (about 20°C) will be approximately 5.5% higher than at body temperature (37°C). This was confirmed by phantom measurements of effective spin density vs temperature, over the range $17\text{--}36^{\circ}\text{C}$, which did indeed show a decrease of $0.32\%/K$ (Venkatesan *et al.*, 2000).

Second, the density of water at room temperature is about 0.5% higher than at body temperature; this small increase in the

number of protons will increase the signal at room temperature by this amount. These two factors reinforce.

Thus the signal from a standard, S_{T_s} , measured at room temperature $T_s^{\circ}\text{C}$, should be converted to the equivalent (lower) value (S_{37}) that it would have at body temperature ($37.0^{\circ}\text{C} = 310.2$ K):

$$S_{37} = S_{T_s} \frac{\rho_{37}}{\rho_{T_s}} \frac{273.2 + T_s}{310.2} \quad (3.5)$$

For a room temperature of 20°C , the correction factor is 0.9406.²⁴ If the phantom has been kept refrigerated before being imaged, the correction factor will be even larger (up to 11%). Thus for high accuracy the phantom temperature should be recorded, and the signal from the standard corrected to obtain the body temperature value.

3.5.6 Phantom Design

3.5.6.1 Phantoms for All Quantitative MR Parameters

Phantoms for T_1 , T_2 and ADC are well developed and have been described above. Phantoms for other parameters are less developed; these may use doped agarose as a host matrix, to obtain realistic T_1 and T_2 values; for example an R_2^* phantom uses USPIO particles in agarose doped with Gd-DTPA (Brown *et al.*, 2017). The chapters on each MR parameter will give information on any available phantoms. Desirable qualities are summarised in Table 3.1.

3.5.6.2 Phantom Containers

Aqueous and gel-based materials can be conveniently contained in cylindrical polythene containers, about 20–25 mm in diameter. These have plastic screw tops. Foil inserts should probably be avoided. Organic liquids need to be in glass, and polypropylene snap tops are available, although in the author's experience they do allow significant evaporation. For some applications (particularly spectroscopy) spherical containers may be advised, to eliminate the internal susceptibility field gradient. A long cylinder can also give a uniform internal field. Glass spheres with a neck attached for filling are available. Larger objects can be machined from acrylic, although this can be time-consuming and expensive. Convenient airtight polythene containers are often available sold as food containers (lunch boxes).

A matrix of small cylindrical battles can conveniently be supported in a block of expanded polystyrene, with holes drilled in it. Slabs of polystyrene, about 50 mm thick, are available from builders' merchants for use as wall cavity insulation material. Drilling is a messy operation and should be carried out with a bit that has a tangential blade that rotates around the circumference of the hole. The polystyrene slab can be cut to a circular shape that is a tight fit inside the head coil. Tool blades should be new, with no history of cutting ferrous materials.

AQ: Note that Venkatesan *et al.* (2000) is cited in the text but not in the list. Please provide complete reference details.

AQ: Please spell out USPIO at first mention

²³ These compounds have to be made up; they are not, to the author's knowledge, available commercially.

²⁴ Correction factors for signal measured from a standard at a range of temperatures are given in 1st edition, p 98.

AQ: Please spell out EPI

EPI sequences probably need the phantoms to be in a water bath (to reduce the susceptibility effects). This can be done by placing bottles in as large a food container as can be fitted into the head coil. Alternatively, a close-fitting water bath, with holes for bottles to be slid in, can be made from acrylic.

3.5.6.3 Stability of Phantom Materials

The stability of agarose gels is still under discussion.²⁵ Although there is evidence of stability (Mitchell *et al.*, 1986; Walker *et al.*, 1988; Christoffersson *et al.*, 1991), other workers have reported changes over time, possibly related to how well the containers are sealed or to contamination of the gel. The temperatures involved in melting the gel should sterilise the container; alternatively, a fungicide can be added. Care should be taken to avoid the entrance of air into the container during the gel cooling process (Vassiliou *et al.*, 2016). A glass container with a narrow neck that can be melted to provide a permanent seal is ideal; if the air is pumped out, then as the neck melts, air pressure forces it to narrow and seal.²⁶ An alternative is to use a cylindrical glass bottle, pour melted wax over the solid gel, then glue the lid on.²⁷ Evaporation of water (from an aqueous solution or a gel), or water entering and mixing with anhydrous liquids, can be detected by regular weighing of the test objects. However instability of the gel could not be detected by a weight change.

Nonetheless, a slow change in parameter value may be measured, and it is almost impossible to be sure whether this is caused by scanner or phantom change over time (see Figure 3.1) (Vassiliou *et al.*, 2016). The only reliable way to use liquid- or gel-based test objects is to regularly calibrate them (i.e. measure their true parameter value) or else, in the case of single-component liquids, to replace them regularly.

3.5.6.4 Ready-made Phantoms and Designs

The Eurospin set of test objects (Lerski 1993; Lerski and de Certaines 1993) from Diagnostic Sonar Ltd²⁸ is comprehensive. The ACR phantom (aqueous $NiCl_2 + NaCl_2$) is widely used for evaluation of geometric parameters, weighted images and even diffusion (Ihalainen *et al.*, 2011; Panych *et al.*, 2016; Wang *et al.*, 2016). The Alzheimer's Disease Neuroimaging Initiative phantom is for multicentre geometric measurements (Gunter *et al.*, 2009). The ISMRM/NIST²⁹ phantom contains multiple compartments with standardised PD, T_1 and T_2 values; aqueous solutions of $NiCl_2$ and $MnCl_2$ are used for T_1 and T_2 , respectively (Jiang *et al.*, 2016).

²⁵ Commercial fruit jams are stable for many years, and these may inspire suitable gel design.

²⁶ A chemistry glassmaker can often make such a container.

²⁷ This approach appears to have been used with the Eurospin gels.

²⁸ <http://www.diagnosticsonar.com>

²⁹ ISMRM, International Society for Magnetic Resonance in Medicine; NIST, US National Institute of Standards and Technology.

References

- Armitage P, Matthews JNS, Berry G. *Statistical Methods in Medical Research*. Blackwell; 2001.
- Barker GJ, Tofts PS. Semiautomated quality assurance for quantitative magnetic resonance imaging. *Magn Reson Imaging* 1992; 10(4): 585–95.
- Bauer CM, Jara H, Killiany R, Initiative AsDN. Whole brain quantitative T2 MRI across multiple scanners with dual echo FSE: applications to AD, MCI, and normal aging. *NeuroImage* 2010; 52(2): 508–14.
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1(8476): 307–10.
- Bland JM, Altman DG. Measurement error. *BMJ* 1996a; 313(7059): 744.
- Bland JM, Altman DG. Measurement error. *BMJ* 1996b; 312(7047): 1654.
- Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996c; 313(7048): 41–2.
- Bloembergen N, Morgan LO. Proton relaxation times in paramagnetic solutions. Effects of electron spin relaxation. *J Chem Phys* 1961; 34(3): 842–50.
- Brihuega-Moreno O, Heese FP, Hall LD. Optimization of diffusion measurements using Cramer-Rao lower bound theory and its application to articular cartilage. *Magn Reson Med* 2003; 50(5): 1069–76.
- Brown GC, Cowin GJ, Galloway GJ. A USPIO doped gel phantom for $R2^*$ relaxometry. *MAGMA* 2017; 30(1): 15–27.
- Cavassila S, Deval S, Huegen C, van Ormondt D, Graveron-Demilly D. Cramer-Rao bounds: an evaluation tool for quantitation. *NMR Biomed* 2001; 14(4): 278–83.
- Chard DT, McLean MA, Parker GJ, MacManus DG, Miller DH. Reproducibility of in vivo metabolite quantification with proton magnetic resonance spectroscopic imaging. *J Magn Reson Imaging* 2002; 15(2): 219–25.
- Christoffersson JO, Olsson LE, Sjoberg S. Nickel-doped agarose gel phantoms in MR imaging. *Acta Radiol* 1991; 32(5): 426–31.
- Cohen JA, Fischer JS, Bolibrush DM, Jak AJ, Kniker JE, Mertz LA, et al. Intrarater and interrater reliability of the MS functional composite outcome measure. *Neurology* 2000; 54(4): 802–6.
- De Deene Y, Hanselaer P, De Wagter C, Achten E, De Neve W. An investigation of the chemical stability of a monomer/polymer gel dosimeter. *Phys Med Biol* 2000; 45(4): 859–78.
- Delakis I, Moore EM, Leach MO, De Wilde JP. Developing a quality control protocol for diffusion imaging on a clinical MRI system. *Phys Med Biol* 2004; 49(8): 1409–22.
- de Wilde J, Price D, Curran J, Williams J, Kitney R. Standardization of performance evaluation in MRI: 13 Years' experience of intersystem comparison. *Concepts Magn Reson Part A* 2002; 15(1): 111–6.

AQ: Armitage et al. (2001): Please provide publisher location.

- Droby A, Lukas C, Schänzer A, Spiwox-Becker I, Giorgio A, Gold R, et al. A human post-mortem brain model for the standardization of multi-centre MRI studies. *NeuroImage* 2015; 110: 11–21.
- Filippi M, Gawne-Cain ML, Gasperini C, vanWaesberghe JH, Grimaud J, Barkhof F, et al. Effect of training and different measurement strategies on the reproducibility of brain MRI lesion load measurements in multiple sclerosis. *Neurology* 1998; 50(1): 238–44.
- Firbank MJ, Harrison RM, Williams ED, Coulthard A. Quality assurance for MRI: practical experience. *Br J Radiol* 2000; 73(868): 376–83.
- Galbraith SM, Lodge MA, Taylor NJ, Rustin GJ, Bentzen S, Stirling JJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR Biomed* 2002; 15(2): 132–42.
- Glover GH, Hayes CE, Pelc NJ, Edelstein WA, Mueller OM, Hart HR, et al. Comparison of linear and circular polarization for magnetic resonance imaging. *J Magn Reson* 1985; 64: 255–70.
- Gunter JL, Bernstein MA, Borowski BJ, Ward CP, Britson PJ, Felmlee JP, et al. Measurement of MRI scanner performance with the ADNI phantom. *Med Phys* 2009; 36(6): 2193–205.
- Hajek M, Babis M, Herynek V. MR relaxometry on a whole-body imager: quality control. *Magn Reson Imaging* 1999; 17(7): 1087–92.
- Haynes BI, Dowell NG, Tofts PS. *Measuring scan-rescan reliability in quantitative brain imaging reveals instability in an apparently healthy imager and improves statistical power in a clinical study.* ISMRM annual scientific meeting; 2010; Stockholm; 2010. p. 2999.
- Holz M, Heil SR, Sacco A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate H-1 NMR PFG measurements. *Phys Chem Chem Phys* 2000; 2(20): 4740–2.
- Hoult DI. The principle of reciprocity in signal strength calculations - a mathematical guide. *Concepts Magn Reson* 2000; 12: 173–87.
- Ihalainen TM, Lönnroth NT, Peltonen JI, Uusi-Simola JK, Timonen MH, Kuusela LJ, et al. MRI quality assurance using the ACR phantom in a multi-unit imaging center. *Acta Oncol* 2011; 50(6): 966–72.
- Jerome NP, Papoutsaki MV, Orton MR, Parkes HG, Winfield JM, Boss MA, et al. Development of a temperature-controlled phantom for magnetic resonance quality assurance of diffusion, dynamic, and relaxometry measurements. *Med Phys* 2016; 43(6): 2998.
- Jiang Y, Ma D, Keenan KE, Stupic KF, Gulani V, Griswold MA. Repeatability of magnetic resonance fingerprinting T1 and T2 estimates assessed using the ISMRM/NIST MRI system phantom. *Magn Reson Med* 2016.
- Keevil SF, Barbiroli B, Brooks JC, Cady EB, Canese R, Carlier P, et al. Absolute metabolite quantification by in vivo NMR spectroscopy: II. A multicentre trial of protocols for in vivo localised proton studies of human brain. *Magn Reson Imaging* 1998; 16(9): 1093–106.
- Kellar KE, Briley-Saebo K. Phantom standards with tempera. *Invest Radiol* 1998; 33(8): 472–9.
- Kraft KA, Fatouros PP, Clarke GD, Kishore PR. An MRI phantom material for quantitative relaxometry. *Magn Reson Med* 1987; 5(6): 555–62.
- Krummenauer F, Doll G. Statistical methods for the comparison of measurements derived from orthodontic imaging. *Eur J Orthod* 2000; 22(3): 257–69.
- Krynicky K. Proton spin lattice relaxation in pure water between 0°C and 100°C. *Physica* 1966; 32: 167–78.
- Laubach HJ, Jakob PM, Loevblad KO, Baird AE, Bovo MP, Edelman RR, et al. A phantom for diffusion-weighted imaging of acute stroke. *J Magn Reson Imaging* 1998; 8(6): 1349–54.
- Lavdas I, Behan KC, Papadaki A, McRobbie DW, Aboagye EO. A phantom for diffusion-weighted MRI (DW-MRI). *J Magn Reson Imaging* 2013; 38(1): 173–9.
- Leach MO, Collins DJ, Keevil S, Rowland I, Smith MA, Henriksen O, et al. Quality assessment in in vivo NMR spectroscopy: III. Clinical test objects: design, construction, and solutions. *Magn Reson Imaging* 1995; 13(1): 131–7.
- Leary SM, Parker GJ, Stevenson VL, Barker GJ, Miller DH, Thompson AJ. Reproducibility of magnetic resonance imaging measurements of spinal cord atrophy: the role of quality assurance. *Magn Reson Imaging* 1999; 17(5): 773–6.
- Lepage M, Whittaker AK, Rintoul L, Back SA, Baldock C. The relationship between radiation-induced chemical processes and transverse relaxation times in polymer gel dosimeters. *Phys Med Biol* 2001; 46(4): 1061–74.
- Lerski RA. Trial of modifications to Eurospin MRI test objects. *Magn Reson Imaging* 1993; 11(6): 835–9.
- Lerski RA, de Certaines JD. Performance assessment and quality control in MRI by Eurospin test objects and protocols. *Magn Reson Imaging* 1993; 11(6): 817–33.
- Malyarenko D, Galbán CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, et al. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. *J Magn Reson Imaging* 2013; 37(5): 1238–46.
- McRobbie DW, Quest RA. Effectiveness and relevance of MR acceptance testing: results of an 8 year audit. *Br J Radiol* 2002; 75(894): 523–31.
- Meyer ME, Yu O, Eclancher B, Grucker D, Chambron J. NMR relaxation rates and blood oxygenation level. *Magn Reson Med* 1995; 34(2): 234–41.
- Mitchell MD, Kundel HL, Axel L, Joseph PM. Agarose as a tissue equivalent phantom material for NMR imaging. *Magn Reson Imaging* 1986; 4(3): 263–6.

AQ: Haynes et al. (2010): Please provide publisher location.

AQ: Jiang et al. (2016): Please provide volume number and page range.

- Morgan LO, Nolle AW. Proton spin relaxation in aqueous solutions of paramagnetic ions II Cr $^{+++}$, Mn $^{++}$, Ni $^{++}$, Cu $^{++}$ and Gd $^{+++}$. *J Chem Phys* 1959; 31: 365.
- O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* 2017; 14(3): 169–86.
- Och JG, Clarke GD, Sobol WT, Rosen CW, Mun SK. Acceptance testing of magnetic resonance imaging systems: report of AAPM Nuclear Magnetic Resonance Task Group No. 6. *Med Phys* 1992; 19(1): 217–29.
- Padhani AR, Hayes C, Landau S, Leach MO. Reproducibility of quantitative dynamic MRI of normal human tissues. *NMR Biomed* 2002; 15(2): 143–53.
- Padhani AR, Liu G, Koh DM, Chenevert TL, Thoeny HC, Takahara T, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia* 2009; 11(2): 102–25.
- Panych LP, Chiou JY, Qin L, Kimbrell VL, Bussolari L, Mulkern RV. On replacing the manual measurement of ACR phantom images performed by MRI technologists with an automated measurement approach. *J Magn Reson Imaging* 2016; 43(4): 843–52.
- Parkes LM, Rashid W, Chard DT, Tofts PS. Normal cerebral perfusion measurements using arterial spin labeling: reproducibility, stability, and age and gender effects. *Magn Reson Med* 2004; 51(4): 736–43.
- Parkes LM, Tofts PS. Improved accuracy of human cerebral blood perfusion measurements using arterial spin labeling: accounting for capillary water permeability. *Magn Reson Med* 2002; 48(1): 27–41.
- Podo F. Tissue characterization by MRI: a multidisciplinary and multi-centre challenge today. *Magn Reson Imaging* 1988; 6(2): 173–4.
- Podo F, Henriksen O, Bovee WM, Leach MO, Leibfritz D, de Certaines JD. Absolute metabolite quantification by in vivo NMR spectroscopy: I. Introduction, objectives and activities of a concerted action in biomedical research. *Magn Reson Imaging* 1998; 16(9): 1085–92.
- Price RR, Axel L, Morgan T, Newman R, Perman W, Schneiders N, et al. Quality assurance methods and phantoms for magnetic resonance imaging: report of AAPM nuclear magnetic resonance Task Group No. 1. *Med Phys* 1990; 17(2): 287–95.
- Provencher SW. Automatic quantitation of localized in vivo ^1H spectra with LCModel. *NMR Biomed* 2001; 14(4): 260–4.
- Pullens P, Bladt P, Sijbers J, Maas AI, Parizel PM. Technical note: a safe, cheap, and easy-to-use isotropic diffusion MRI phantom for clinical and multicenter studies. *Med Phys* 2017; 44(3): 1063–70.
- Rohrer M, Bauer H, Mintorovitch J, Requardt M, Weinmann HJ. Comparison of magnetic properties of MRI contrast media solutions at different magnetic field strengths. *Invest Radiol* 2005; 40(11): 715–24.
- Rovaris M, Mastronardo G, Sormani MP, Iannucci G, Rodegher M, Comi G, et al. Brain MRI lesion volume measurement reproducibility is not dependent on the disease burden in patients with multiple sclerosis. *Magn Reson Imaging* 1998; 16(10): 1185–9.
- Schneiders NJ. Solutions of two paramagnetic ions for use in nuclear magnetic resonance phantoms. *Med Phys* 1988; 15(1): 12–16.
- Silver NC, Barker GJ, MacManus DG, Tofts PS, Miller DH. Magnetisation transfer ratio of normal brain white matter: a normative database spanning four decades of life. *J Neurol Neurosurg Psychiatry* 1997; 62(3): 223–8.
- Simmons A, Moore E, Williams SC. Quality control for functional magnetic resonance imaging using automated data analysis and Shewhart charting. *Magn Reson Med* 1999; 41(6): 1274–8.
- Soher BJ, Hurd RE, Sailasuta N, Barker PB. Quantitation of automated single-voxel proton MRS using cerebral water as an internal reference. *Magn Reson Med* 1996; 36(3): 335–9.
- Sormani MP, Gasperini C, Romeo M, Rio J, Calabrese M, Cocco E, et al. Assessing response to interferon- β in a multicenter dataset of patients with MS. *Neurology* 2016; 87(2): 134–40.
- Streiner DL, Norman GR. *Health Measurement Scales: a practical guide to their development and use*. Oxford University Press; 1995.
- Sun J, Barnes M, Dowling J, Menk F, Stanwell P, Greer PB. An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom. *Australas Phys Eng Sci Med* 2015; 38(1): 39–46.
- Taylor JR. *An introduction to error analysis: the study of uncertainties in physical measurements*. Sausalito, CA, USA: University Science Books; 1997.
- Tofts PS. Standing waves in uniform water phantoms. *J Magn Reson series B* 1994; 104: 143–7.
- Tofts PS. Optimal detection of blood-brain barrier defects with Gd-DTPA MRI—the influences of delayed imaging and optimised repetition time. *Magn Reson Imaging* 1996; 14(4): 373–80.
- Tofts PS. Standardisation and optimisation of magnetic resonance techniques for multicentre studies. *J Neurol Neurosurg Psychiatry* 1998; 64 Suppl 1: S37–43.
- Tofts PS, Barker GJ, Dean TL, Gallagher H, Gregory AP, Clarke RN. A low dielectric constant customized phantom design to measure RF coil nonuniformity. *Magn Reson Imaging* 1997a; 15(1): 69–75.
- Tofts PS, Barker GJ, Filippi M, Gawne-Cain M, Lai M. An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis. *Magn Reson Imaging* 1997b; 15(2): 183–92.
- Tofts PS, Berkowitz B, Schnall MD. Quantitative analysis of dynamic Gd-DTPA enhancement in breast tumors using a permeability model. *Magn Reson Med* 1995; 33(4): 564–8.

AC: Streiner and Norman (1995): Please provide publisher location.

- Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, 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(3): 223–32.
- Tofts PS, Collins DJ. Multicentre imaging measurements for oncology and in the brain. *Br J Radiol* 2011; 84 Spec No 2: S213–26.
- Tofts PS, Jackson JS, Tozer DJ, Cercignani M, Keir G, MacManus DG, et al. Imaging cadavers: cold FLAIR and noninvasive brain thermometry using CSF diffusion. *Magn Reson Med* 2008; 59(1): 190–5.
- Tofts PS, Kermod AG, MacManus DG, Robinson WH. Nasal orientation device to control head movement during CT and MR studies. *J Comput Assist Tomogr* 1990; 14(1): 163–4.
- Tofts PS, Lloyd D, Clark CA, Barker GJ, Parker GJ, McConville P, et al. Test liquids for quantitative MRI measurements of self-diffusion coefficient in vivo. *Magn Reson Med* 2000; 43(3): 368–74.
- Tofts PS, Shuter B, Pope JM. Ni-DTPA doped agarose gel - a phantom material for Gd-DTPA enhancement measurements. *Magn Reson Imaging* 1993; 11(1): 125–33.
- Tofts PS, Steens SC, Cercignani M, Admiraal-Behloul F, Hofman PA, van Osch MJ, et al. Sources of variation in multi-centre brain MTR histogram studies: body-coil transmission eliminates inter-centre differences. *Magma* 2006; 19(4): 209–22.
- Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, et al. The WU-Minn human connectome project: an overview. *NeuroImage* 2013; 80: 62–79.
- Vassiliou VS, Heng EL, Gatehouse PD, Donovan J, Raphael CE, Giri S, et al. Magnetic resonance imaging phantoms for quality-control of myocardial T1 and ECV mapping: specific formulation, long-term stability and variation with heart rate and temperature. *J Cardiovasc Magn Reson* 2016; 18(1): 62.
- Walker L, Curry M, Nayak A, Lange N, Pierpaoli C, Group BDC. A framework for the analysis of phantom data in multi-center diffusion tensor imaging studies. *Hum Brain Mapp* 2013; 34(10): 2439–54.
- Walker P, Lerski RA, Mathur-De Vre R, Binet J, Yane F. Preparation of agarose gels as reference substances for NMR relaxation time measurement. EEC Concerted Action Program. *Magn Reson Imaging* 1988; 6(2): 215–22.
- Walker PM, Balmer C, Ablett S, Lerski RA. A test material for tissue characterisation and system calibration in MRI. *Phys Med Biol* 1989; 34(1): 5–22.
- Wang ZJ, Seo Y, Babcock E, Huang H, Bluml S, Wisnowski J, et al. Assessment of diffusion tensor image quality across sites and vendors using the American College of Radiology head phantom. *J Appl Clin Med Phys* 2016; 17(3): 442–51.
- Wei X, Warfield SK, Zou KH, Wu Y, Li X, Guimond A, et al. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. *J Magn Reson Imaging* 2002; 15(2): 203–9.
- Zhou X, Sakaie KE, Debbins JP, Kirsch JE, Tatsuoka C, Fox RJ, et al. Quantitative quality assurance in a multicenter HARDI clinical trial at 3T. *Magn Reson Imaging* 2017; 35: 81–90.

