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).

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 inrasubject 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).

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

<table>
<thead>
<tr>
<th>Simple Phantom (Test Object)</th>
<th>Healthy Control Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability(^a)</td>
<td>Good</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Potentially good (e.g. volume)</td>
</tr>
<tr>
<td>Uniformity</td>
<td>Poor in gels, good in liquids</td>
</tr>
<tr>
<td>Temperature dependence</td>
<td>$D, T_1, T_2$ change 2%–3%/°C</td>
</tr>
<tr>
<td>Stability</td>
<td>Potentially good (e.g. volume) but can be unstable (e.g. gels)</td>
</tr>
<tr>
<td>Realism</td>
<td>Generally poor; in vivo changes cannot be realistically modelled; $B_1$ distribution different</td>
</tr>
</tbody>
</table>

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

\(^b\) Though see institutional constraints (Section 3.5.1).

\[^\text{Use normal range, or travelling subject(s)}\]

**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.)
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).

**Quality Assurance: Accuracy, Precision, Controls and Phantoms**

**Table 3.2** Statistical Tests Used for Shewhart Charting

<table>
<thead>
<tr>
<th>Test Number</th>
<th>Name of Test</th>
<th>Description of Test</th>
<th>Action Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Warning</td>
<td>Measure exceeds control limits of mean ± 2 SD of previous measures</td>
<td>Inspect with Tests 2–6</td>
</tr>
<tr>
<td>2</td>
<td>3 SD</td>
<td>Measure exceeds control limits of mean ± 3 SD of previous</td>
<td>Instrument evaluation</td>
</tr>
<tr>
<td>3</td>
<td>2 SD</td>
<td>Two consecutive measures exceed mean ± 2 SD</td>
<td>Instrument evaluation</td>
</tr>
<tr>
<td>4</td>
<td>Range of 4 SD</td>
<td>Difference between two consecutive measures exceeds 4 SD</td>
<td>Instrument evaluation</td>
</tr>
<tr>
<td>5</td>
<td>Four ± 1 SD</td>
<td>Four consecutive measures exceed the same limit (+1 SD or –1 SD)</td>
<td>Instrument evaluation</td>
</tr>
<tr>
<td>6</td>
<td>Mean × 10</td>
<td>Ten consecutive measures fall on the same side of the mean</td>
<td>Instrument evaluation</td>
</tr>
</tbody>
</table>


**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).

### 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, and with different kinds of scanners, in different centres, for over 20 years (Filippi et al., 1998; Podo et al., 1998; Bauer et al., 2010) (Jerome et al., 2016). The European group MAGNIMS® 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).

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).

---

*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 B1 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; Droby 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.

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 T1, systematic errors could be caused by a consistently wrong B1 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 B1 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 become 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 B1 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/acc/uncertainty/index.html. This is a mine of information on constants, units and uncertainty.
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). 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. 

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).

* See also.

### TABLE 3.3 Potential Sources of Error in the MRI Measurement Process

<table>
<thead>
<tr>
<th>Random Error</th>
<th>Systematic Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
</tr>
</tbody>
</table>

*See also.*
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 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 \( K^{\text{app}} \) in the breast (Tofts et al., 1995), the estimate is very sensitive to the \( T_1 \) 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 \( T_1 \)-w-DCE) but random across other subjects. Uncertainty budgets and type A and B errors are concepts related to EPR (see Section 3.3.6).

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).

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; Brihugue-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.

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. Initiatives have been published from Europe, the USA, and UK. 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).

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 volumes of value (Tofts et al., 1997b). Accuracy (and precision)

---

Footnote 9


2. 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.


4. The United Kingdom Accreditation Service has several useful documents; M3003 and LAB 12 are concise expositions of the concepts.
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\(^3\) 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 3.8).\(^4\)

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.

\(^3\) 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.

\(^4\) 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.
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;
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 unchanged 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 sdδ:

\[ sd\delta = \sqrt{2} s = 1.414 s \]  

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 (sdδ), 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).\(^5\)

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).


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

<table>
<thead>
<tr>
<th>Measurement Set Number</th>
<th>Replicate 1</th>
<th>Replicate 2</th>
<th>ΔSigned Difference</th>
<th>SD of differences sdΔ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>107.14</td>
<td>108.12</td>
<td>0.98</td>
<td>6.4</td>
</tr>
<tr>
<td>2</td>
<td>103.50</td>
<td>98.60</td>
<td>–4.91</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>104.65</td>
<td>104.73</td>
<td>0.08</td>
<td>4.5</td>
</tr>
<tr>
<td>4</td>
<td>100.97</td>
<td>106.26</td>
<td>5.29</td>
<td>ISD s</td>
</tr>
<tr>
<td>5</td>
<td>96.87</td>
<td>105.76</td>
<td>8.89</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>90.30</td>
<td>98.76</td>
<td>8.46</td>
<td>σδ</td>
</tr>
<tr>
<td>7</td>
<td>108.97</td>
<td>98.79</td>
<td>–10.19</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>104.55</td>
<td>110.24</td>
<td>5.70</td>
<td>95% CL lower</td>
</tr>
<tr>
<td>9</td>
<td>99.55</td>
<td>105.13</td>
<td>5.58</td>
<td>95% CL upper</td>
</tr>
<tr>
<td>10</td>
<td>103.94</td>
<td>99.60</td>
<td>–4.33</td>
<td></td>
</tr>
</tbody>
</table>

**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 (sdδ), the SD of this set of differences (sdΔ), and from this the SD of the population (SD). 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).
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).

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

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 I, Section 4.2).
3.3.3 Intraclass Correlation Coefficient or Reliability

This measure considers both the within-subject (intra-subject) 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

\[
 ICC = \frac{\text{variance from subjects}}{\text{variance from subjects} + \text{variance from measurement error}} \tag{3.3}
\]

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 like this 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 interscaner 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 3.2.1). 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

<table>
<thead>
<tr>
<th>Study ( ^* )</th>
<th>CV( ^* ) (%)</th>
<th>( n^* )</th>
<th>Mean (ms)</th>
<th>SD (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stevenson 2000</td>
<td>5</td>
<td>40</td>
<td>666</td>
<td>36 ( ^* )</td>
</tr>
<tr>
<td>Rutgers 2002</td>
<td>6</td>
<td>15</td>
<td>681</td>
<td>40</td>
</tr>
<tr>
<td>Ethofer 2003</td>
<td>8</td>
<td>4</td>
<td>770 ( ^* )</td>
<td>30</td>
</tr>
</tbody>
</table>


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

\( ^* \) Sample size.

\( ^* \) Estimated from boxplot in figure.

\( ^* \) Used spectroscopic technique; probably some cerebrospinal fluid or grey matter contamination.

\( ^* \) Coefficient of variation = SD/mean.

\( ^* \) References for all studies are given in the original tables (Tofts and Collins 2011).
TABLE 3.5B  Normal Range of Mean Diffusivity Values in White Matter

<table>
<thead>
<tr>
<th>Study</th>
<th>CV (%)</th>
<th>n</th>
<th>Mean (10⁻⁶ m²/s)</th>
<th>SD (10⁻⁶ m²/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cercignani</td>
<td>5</td>
<td>20</td>
<td>0.93</td>
<td>0.04</td>
</tr>
<tr>
<td>Emmer</td>
<td>2006</td>
<td>4</td>
<td>0.84</td>
<td>0.03</td>
</tr>
<tr>
<td>Zhang</td>
<td>2007</td>
<td>5</td>
<td>0.69</td>
<td>0.04</td>
</tr>
<tr>
<td>Velosh</td>
<td>2007</td>
<td>3</td>
<td>0.73</td>
<td>0.02</td>
</tr>
</tbody>
</table>


Note: See also Chapter 2.

* Some cerebrospinal fluid contamination.

TABLE 3.5C  Normal Range of MTR Values in White Matter

<table>
<thead>
<tr>
<th>Study</th>
<th>CV (%)</th>
<th>n</th>
<th>Mean (pu)</th>
<th>SD (pu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver</td>
<td>1.9</td>
<td>41</td>
<td>39.5</td>
<td>0.76</td>
</tr>
<tr>
<td>Davies</td>
<td>1.0</td>
<td>19</td>
<td>38.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Tofts</td>
<td>1.6</td>
<td>10</td>
<td>37.3³</td>
<td>0.6</td>
</tr>
</tbody>
</table>


Note: See also Figure 3.8, which shows SD values of 0.5–1.0 pu. Object was a polyurethane phantom (Equation 3.11). SEM = 0.17 pu; 4 samples each n = 20 or 21; estimated SD = 0.76 pu.

³ Peak location values in white matter histograms.

³ 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 estimates (Equation 3.2). Before is authors first value, almost the highest value of nine centres. After solving a scanner instability problem (Figure 3.9), ISD was dropped to the lowest value of nine centres. (Adapted from Haynes, B.L., 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, 2003.)

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 T₁ 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 T₁, T₂, 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. These are immersed in water (doped to reduce their T₁ and T₂ values). Objects with a specified T₁, T₂ 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.

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.

Σ Major manufacturers are Perspex in the UK and Plexiglas in North America.
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 ($T_1 \approx 600–800$ 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 ($T_2 \approx 90–100$ 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 ($\varepsilon \approx 80$) leads to the presence of radio frequency standing waves (dielectric resonance), where $B_0$ is enhanced, giving an artificially high flip angle and signal (see Figure 3.30). The high dielectric constant reduces the wavelength of electromagnetic radiation, compared to its value in free space, by a factor $\varepsilon$; 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 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 ($\varepsilon = 2-3$) 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 ($T_1$ values are convenient, at 33–110 ms, whilst $T_2$ 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

---

**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.)
3.5.3 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 $R_2(R_1 = 1/T_1)$ is increased by adding a particular amount of the compound. In aqueous solution:

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

$R_m$ and $R_m$ 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 $\text{CuSO}_4$ and manganese chloride $\text{MnCl}_2$; nickel $\text{Ni}^{++}$ 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 $T_1$. $\text{MnCl}_2$ 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 $T_1$ value.) Water $T_1$ depends on the amount of dissolved oxygen. It is independent of frequency (Krynicki 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 relativity for oxygen is approximately $1.8 \pm 0.3 \times 10^{-4} \text{ s}^{-1} \text{ mmHg}^{-1}$ (measured in plasma at 4.7T by Meyer et al., 1995). Assuming this value still holds at 3T, fully oxygenated water at 23°C ($\text{pO}_2 = 150$ 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 $\text{pO}_2$, would be valuable, particularly if accompanied by $T_2$ values (high quality measurements of water $T_1$ 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 $r_1$ and low $r_2$ (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 $r_1$ (much greater than $r_2$, i.e. $\text{MnCl}_2$ or agarose) and the other has low $r_2$ (about the same as $r_1$). Thus suitable mixtures are $\text{Ni}^{++}$ and

![Figure 3.11](image_url)
Mn$^{++}$ in aqueous solution (Schneiders 1988), Gd-DTPA$^{a}$ 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).

$^{a}$ This is preferred to GdCl$_3$, which interacts with the agarose.

### TABLE 3.7 Values of Relaxivity at 1.5T$^a$ and Room Temperature

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

$^a$ 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).

$^c$ 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.

---

### TABLE 3.8

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>$T_1$ (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.73</td>
</tr>
<tr>
<td>5</td>
<td>2.07</td>
</tr>
<tr>
<td>10</td>
<td>2.39</td>
</tr>
<tr>
<td>15</td>
<td>2.76</td>
</tr>
<tr>
<td>20</td>
<td>3.15</td>
</tr>
<tr>
<td>21</td>
<td>3.23</td>
</tr>
<tr>
<td>22</td>
<td>3.32</td>
</tr>
<tr>
<td>23</td>
<td>3.40</td>
</tr>
<tr>
<td>24</td>
<td>3.49</td>
</tr>
<tr>
<td>25</td>
<td>3.57</td>
</tr>
<tr>
<td>26</td>
<td>3.67</td>
</tr>
<tr>
<td>27</td>
<td>4.70</td>
</tr>
</tbody>
</table>

$T_1$ 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 Krynicki 1966; see also Tofts et al., 2008). Note values are expected to be independent of field strength.

---

**FIGURE 3.12** $T_1$ and $T_2$ values for aqueous solutions of agarose and CuSO$_4$. 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).
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. 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 $0.49 \times 10^{-3}$ m$^2$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%$/^\circ$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. 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).

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 thermocouple21 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 3T, 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).

---

21 Walker et al. (1988) give a theoretical temperature coefficient in 2% agarose ($T_1 = 60$ ms) at 20 MHz of 1.7%/°C. The Eurospin gels closest to brain tissue have a $T_1$ 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.
A more general solution is to use a Gd polymer pair (Kellar et al. 2000). One component (NC663868) has a zero temperature coefficient for \( T_1 \). The second component (NC22181) has a negative coefficient (about –1.2%/°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 concentrations 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 magnetisation \( M_0 \) 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 17–36°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, \( S_{90} \), measured at room temperature \( T_{90} \)°C, should be converted to the equivalent (lower) value \( S_{5} \) that it would have at body temperature \( (37.0°C = 310.2 K) \):

\[
S_{5} = S_{90} \frac{\rho_{90}}{\rho_{5}} \left( \frac{273.2 + T_{90}}{310.2} \right)
\]

(3.5)

For a room temperature of 20°C, the correction factor is 0.9406. 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 \( R_1 \) 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 polystyrene 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 polystyrene 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.
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. 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., 2011). 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. An alternative is to use a cylindrical glass bottle, pour melted wax over the solid gel, then glue the lid on. 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., 2011). 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 Ltd is comprehensive. The ACR phantom (aqueous NiCl₂ + NaCl) 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/NIST phantom contains multiple compartments with standardised PD, T₁ and T₂ values; aqueous solutions of NiCl₂ and MnCl₂ are used for T₁ and T₂, respectively (Jiang et al., 2016).

---

References


Quantitative MRI of the Brain


Quality Assurance: Accuracy, Precision, Controls and Phantoms


