3

Quality Assurance, Accuracy, Precision and Phantoms

Paul Tofts

Department of Medical Physics, NMR Research Unit, Institute of Neurology, University College London, Queen Square, London WCIN 3BG, UK

3.1 Quality Assurance Concepts 55
3.2 Accuracy and Systematic Errors 59
3.3 Precision 63
3.4 Phantoms (Test Objects) 68

3.1 QUALITY ASSURANCE CONCEPTS

3.1.1 Acceptance Testing

When an instrument such as an MRI scanner is installed and handed over to the user, a series of acceptance tests is often carried out by the customer (Och *et al.*, 1992; McRobbie and Quest, 2002; Och *et al.*, 1992). The vendor's installation engineer will also have carried out extensive testing, according to the manufacturer's protocols, using phantoms (test objects), to ensure the instrument is operating within the specification of the manufacturer. For MRI these may include signal-to-noise ratio, spatial resolution and uniformity tests, gradient calibration, 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

Quantitative MRI of the Brain. Edited by Paul Tofts
© 2003 John Wiley & Sons, Ltd ISBN: 0-470-84721-2

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 will be some periodic re-calibrations, 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 they 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 *et al.*, 1991a,b; Tofts, 1998). This will consume valuable scanner time, yet without these tests the measurements on patients 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 phantoms and in subjects. Phantom measurements 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). Temperature dependence is a major problem with many parameters, since the scanner room environment may vary by 1 or 2 °C, unless special precautions are taken, such as thermal insulation of the phantoms. T_1, T_2 and diffusion (D) all vary by about 2–3 % per °C. Phantoms based on liquids may be unstable, for example through evaporation. If a drift is seen in measurements from phantoms, the interpretation is often unclear (Figure 3.1).

Measurements in control human subjects are usually completely realistic, provided the parameter is present in normal subjects (Table 3.1). Thus brain volume or normal-appearing brain tissue T_1 value could be monitored in this way, but lesion

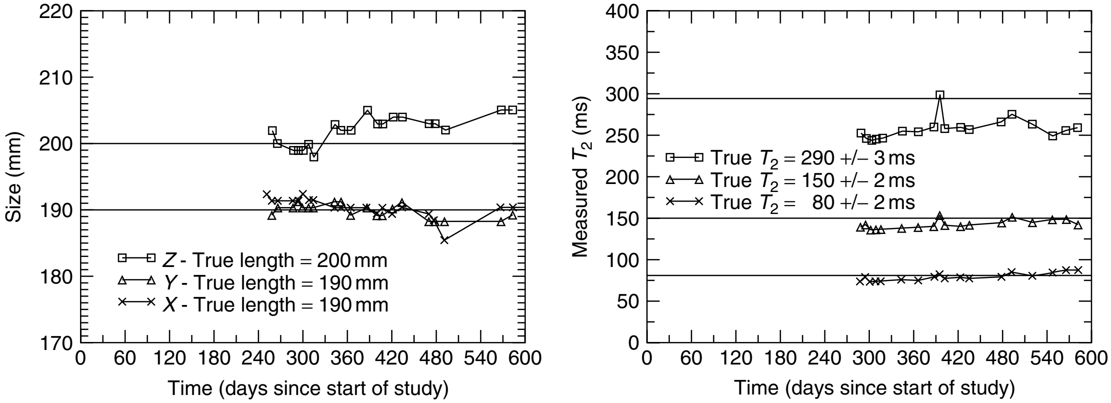

Figure 3.1. QA measurements of object size (left) and T_2 (right). The apparent size drifts with time, probably because of a fault with gradient calibration. The true size is known accurately and unambiguously. T_2 estimates are inaccurate, particularly for the long-T_2 phantom, and drift with time, 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 third possible explanation. Reproduced from Barker, G. J. and Tofts, P. S. 1992, Semiautomated quality assurance for quantitative magnetic resonance imaging, in *Magn. Reson. Imag.* **10**, 585–595, Copyright 1992 with permission from Elsevier Science Ltd

Table 3.1. Relative advantages of phantoms and normal controls for QQA

	Phantom (test object)	Normal control subject
Availability	Good	Reasonable
Accuracy	Potentially good (e.g. volume)	True value unknown
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; *in vivo* changes cannot be realistically modelled	Good; but no pathology

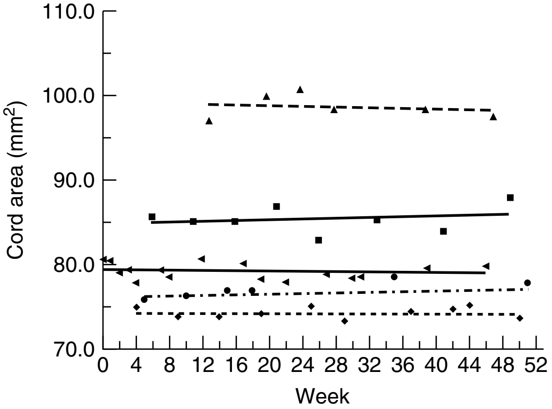

Figure 3.2. QQA in the spinal cord. Data on spinal cord cross-sectional area for five normal controls, measured by the method of Losseff *et al.* (1996), which has a short-term precision of 0.8 % (CV). The lines are linear regressions. Reproduced from Leary, S. M., Parker, G. J., Stevenson, V. L., Barker, G. J., Miller, D. H. and Thompson, A. J. 1999, Reproducibility of magnetic resonance imaging measurements of spinal cord atrophy: the role of quality assurance, in *Magn. Reson. Imag.* **17**, 773–776, Copyright 1999, with permission from Elsevier Science Ltd

volume could not. Increased atrophy or movement in patients might sometimes increase the variability compared with normal control subjects (Figure 3.2). A few parameters (most notoriously blood perfusion) have large biological intra-subject 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. In general, when introducing a new technique, phantom measurements are likely to be needed at first, to characterize instrumental sources of variation such as noise (and possibly reduce them using techniques such as sequence optimization), before going on to *in vivo* QQA with normal controls. The subject of QQA is discussed further in Chapter 16, Section 16.6.

Professional organizations of medical physics sometimes publish material on QA in MRI. The UK Institute of Physics and Engineering in Medicine (IPEM)[1] has published *Report 80: Quality Control in Magnetic Resonance Imaging*. This gives a comprehensive description of how to use the Eurospin test objects. The American Association of Physicists in Medicine (AAPM) has published some guidance on QA (Price *et al.*, 1990; Och *et al.*, 1992). The American College of Radiology (ACR)[2] has an MRI accreditation scheme and an *MRI Quality Control Manual* (published in 2001). A phantom test is used during the accreditation process. In the UK, MagNET[3] provides help in evaluating MR machines. The Eurospin set of test objects (Lerski, 1993; Lerski and de Certaines, 1993) from Diagnostic Sonar Ltd[4] is comprehensive.

To carry out QQA, phantoms and normal subjects are measured at regular intervals (typically every week or month). 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

[1] www.ipem.org.uk
[2] www.acr.org/
[3] www.magnet-mri.org/
[4] www.bigwig.net/diagnosticsonar/

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 ± 2 SD of previous measures	Inspect with tests 2–6
2	3 SD	Measure exceeds control limits of mean ± 3 SD of previous	Instrument evaluation
3	2 SD	Two consecutive measures exceed mean ± 2 SD	Instrument evaluation
4	Range of 4 SD	Difference between two consecutive measures exceeds 4 SD	Instrument evaluation
5	Four ± 1 SD	Four consecutive measures exceed the same limit (+1 SD or −1 SD)	Instrument evaluation
6	Mean × 10	Ten consecutive measures fall on the same side of the mean	Instrument evaluation

Adapted from Simmons *et al.* (1999).

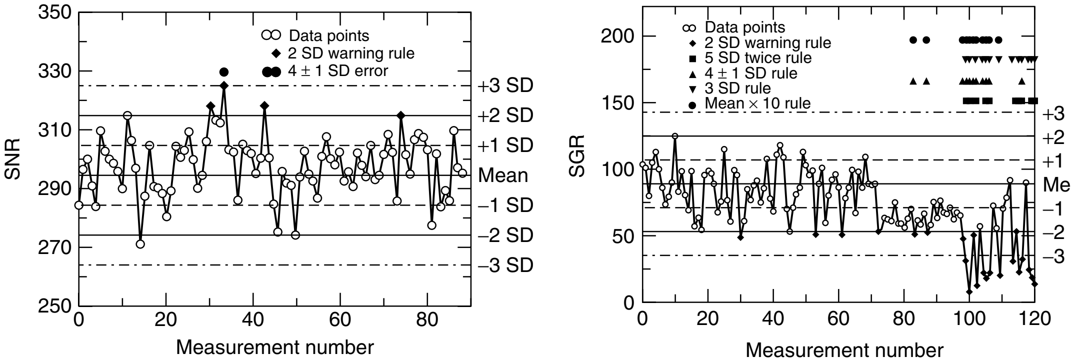

Figure 3.3. Shewhart charting of QA parameters. Data points are open symbols; triggering of rules (see Table 3.2) is shown by solid symbols. SNR, signal-to-noise ratio; SGR, signal-to-ghost ratio (used in echoplanar imaging). Reproduced with permission from Simmons, A., Moore, E., and Williams, S. C. 1999, Quality control for functional magnetic resonance imaging using automated data analysis and Shewhart charting, in *Magn. Reson. Med.*, Copyright 1999 John Wiley & Sons Inc

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. Shewhart charting (Simmons *et al.*, 1999; Hajek *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; Figure 3.3).

3.1.2 Calibration

Calibration is sometimes claimed to be a benefit of scanning phantoms with known 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 'same-place' phantoms, scanned in the head coil at a different time from the head, and for a 'same-time' phantoms, attached to the head.

3.2 ACCURACY AND SYSTEMATIC ERRORS

3.2.1 Specifying Uncertainty

According to the tradition of measurement:

> *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.*[5]

The conventional way to characterize measurement techniques in the physical sciences has been to estimate accuracy and precision (i.e. systematic and random errors), although there are at least two other paradigms which should be considered. First, the concepts of type A and type B uncertainty are superseding random and systematic error in the physical sciences (see Section 3.3.6 below). 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 values. Thus, a particular (single) measurement could have zero error but large uncertainty. Second, in psychology and in medicine, the concepts of validity, sensitivity and reliability are often used to evaluate the performance of a metric (see Section 3.3.7 below).

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). Thus the differences between long-term precision and accuracy become blurred, and as the demand grows for studies to carry on for longer periods, for over a decade in the case of MS, epilepsy, dementia and aging, considerations of accuracy become increasingly important. 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. Precision may have a biological component, in that intra-subject variation may be significant and limits the usefulness of having good machine precision. Blood flow varies by about 10 % within a day, so if a single number is required to characterize 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.

Sources of error (contributing to both inaccuracy and imprecision) can be in the data collection procedure (as described in Section 2.1), and in the image analysis procedure (Section 2.2), and both procedures 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 nonuniformity and partial volume errors. Artefacts arising from imperfect slice selection and k-space sampling (particularly in fast spin echo and echoplanar imaging) give systematic error. Patient positioning and movement contribute to random errors; positioning can be improved with technologist training and liberal use of localizer scans, whilst movement can be

[5] From the USA National Institute of Standards and Technology (NIST) website: http://physics.nist.gov/cuu/index.html. This is a mine of information on constants, units and uncertainty.

Table 3.3. Potential sources of error in the MRI measurement process

	Random error	Systematic error
Biology	Normal variation in physiology	
Data collection	Position of subject in head coil	B_1 error (nonuniformity)
	Coil loading (? corrected by pre-scan)	Slice profile
	Pre-scan procedure setting B_1	k-Space sampling (in fast spin echo, EPI)
	Position of slices in head	Operator training
	Gd injection procedure	Software upgrade
	Patient movement (cardiac pulsation)	Hardware upgrade
	Patient movement (macroscopic)	Partial volume
	Image noise	
	Temperature (phantoms only)	
Image analysis	ROI creation and placement	Operator training software upgrade

Reproduced with permission from Tofts, P. S., Standing waves in uniform water phantoms, in *J. Magn. Reson. Ser. B* **104**, 143–147, Copyright 1994 Elsevier Science Ltd.

reduced by attention to patient comfort, feedback devices to assist the subject in keeping still (Tofts *et al.*, 1990), and spatial registration images (see Chapter 15). Analysis performance can be characterized 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 (Gawne-Cain *et al.*, 1996; 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 practice[6] defines how to do this. The analysis method should be documented in detail, intra- and inter-rater differences measured, and 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. 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 analysis, and in many cases studies are often extended beyond their initially planned duration.

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 which gave realistic errors in the reported values of volume. Accuracy (and precision) 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

[6] See for example ISO 9001.

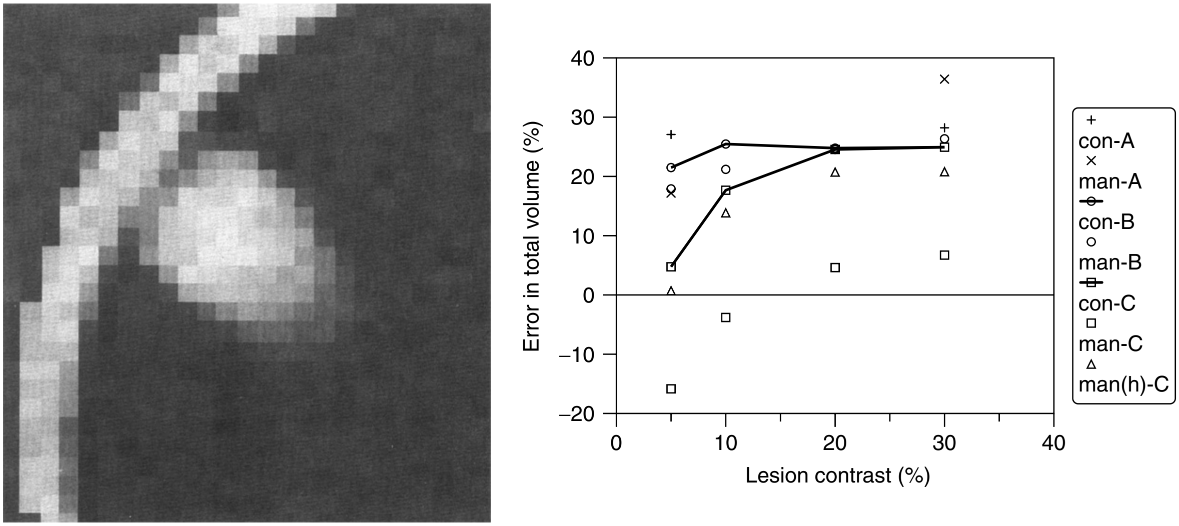

Figure 3.4. Lesion volume accuracy measured using an oblique cylinder contrast-adjusted phantom. Left: one small lesion (0.6 ml in volume), represented as an acrylic cylinder, is mounted on the inside of an acrylic annulus, at an angle to the image slice, giving a realistic partial volume effect. Right: 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, semiautomatic contouring; man manual). Reproduced from Tofts, P. S., Barker, G. J., Filippi, M., Gawne-Cain, M. and Lai, M. 1997b, An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis, in *Magn. Reson. Imag.* **15**, 183–192, Copyright 1997, with permission from Elsevier Science Ltd

to a measurement technique, since it will identify any major problems (Figure 3.4).

3.2.2 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 3.2) 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. (Figure 3.5). As an additional safeguard, if groups of subjects are being compared, subjects from both groups should be collected during the same period. 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 (LR) asymmetry or anterior–posterior (AP) differences may be seen in a group of subjects. This could be caused by genuine

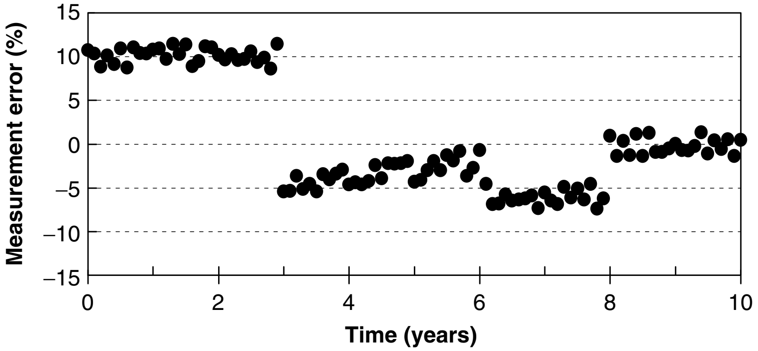

Figure 3.5. Long-term precision is dominated by instability in the systematic error. Simulation of fictional change in measurement error over time, during a longitudinal study. Short-term precision is good, and a study completed in the first 3 years is unaffected by the large systematic error (i.e. poor 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

biological difference between the sides or front and back, or by a subtle asymmetry in the head coil. The only unequivocal way to resolve this uncertainty is to scan some subjects relocated with respect to the head coil, i.e. prone (to resolve a AP issue), or with the left and right interchanged (for an LR issue).

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 percentage 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 in the breast (Tofts *et al.*, 1995), the estimate is very sensitive to the T_1 value used (EPR = 1.2), and the relaxivity (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; and see Figure 13.11 in this book). Studying error sources in this way immediately brings to light that some errors are truly random, whilst others are systematic for the same subject in repeated measurements, but random across other subjects. The distinction between random and systematic errors soon becomes blurred. 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. Uncertainty budgets and type A and B errors are concepts related to EPR (see the next section).

Multicentre studies, where an attempt is made to reproduce the same measurement technique across different centres or hospitals, often with different kinds of scanner, in different countries, are a challenging test (Podo, 1988; Soher *et al.*, 1996; Podo *et al.*, 1998; Keevil *et al.*, 1998; Filippi *et al.*, 1998a; Berry *et al.*, 1999). One approach is to make the data collection and analysis procedures as similar as possible, so that any systematic errors are replicated across the whole sample of centres. This involves attempting to match scanners, sequence timing parameters (*TR*, *TE*), and also slice profile and RF nonuniformity (which is often not possible). A second approach is to aim for good accuracy at each centre, measuring the underlying biology independently of the particular measurement procedure, since accurate measurements must necessarily agree with each other. Thus multicentre studies, although time-consuming and frustrating, are the ultimate test of how good our measurement techniques are.

3.3 PRECISION

Precision, or reproducibility,[7] 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. It has been measured for many parameters, for example spectroscopy (Charles *et al.*, 1996; Marshall *et al.*, 1996; Simmons *et al.*, 1998; Bartha *et al.*, 2000; Chard *et al.*, 2002), dynamic Gd imaging (Buckley, 2002; Galbraith *et al.*, 2002; Padhani *et al.*, 2002), fMRI (Tegeler *et al.*, 1999; Loubinoux *et al.*, 2001), tissue volume (Fox and Freeborough, 1997; Lemieux *et al.*, 2000; Cardenas *et al.*, 2001; Gasperini *et al.*, 2001), lesion volume (Grimaud *et al.*, 1996; Vaidyanathan *et al.*, 1997; Rovaris *et al.*, 1998; Filippi *et al.*, 1998b), lesion counting (Rovaris *et al.*, 1999; Wei *et al.*, 2002), spinal cord cross-sectional area (Leary *et al.*, 1999) and clinical scores (Cohen *et al.*, 2000). Its value 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 pre-scan procedure) and data analysis (particularly ROI placement). The results of a study may not be generalizable – a poor value of reproducibility may be a reflection of poor technique. 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 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 inter-observer effects (Filippi *et al.*, 1998b) to formal mathematical optimization of the free parameters that define the process (Tofts, 1996) (see Figure 2.12). Measuring the reproducibility of various scanner parameters that are thought to have large effect on the final MR parameter (such as those set during the pre-scan procedure) may also be of value.

The methods used to report reproducibility are not always standardized – it is hoped that studies will use CV and intra-class 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 Figure 3.6). Genuine biological variation may be included in the estimate; if so, this should be made clear when reporting the study.

3.3.1 Within-subject Standard Deviation and Repeatability

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 (CL) 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

[7] A measurement is said to be *reproducible* when it can be repeated (reproduce; to bring back into existence again, recreate). 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, which are not part of measurement error.

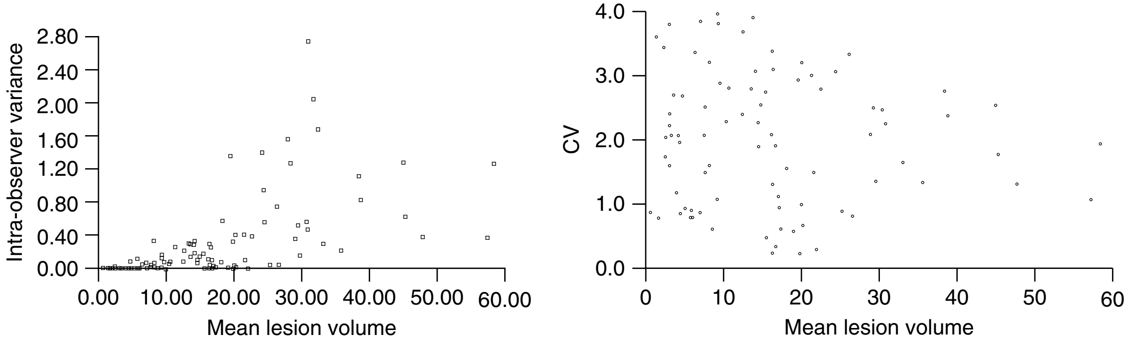

Figure 3.6. Bland–Altman plots for estimates of total lesion volume in MS. Left: the variance (var) increases with mean lesion volume (MLV). Modelling the variance as var $= A$ MLV^B, where A and B are constants, gives $B = 1.84$, with 95 % confidence limits 1.55–2.13. The data are therefore consistent with a quadratic behavior for var ($B = 2$). Right: the CV is independent of MLV, as predicted by the model (since CV $= \sqrt{\text{var}}/\text{MLV}$). Reproduced from Rovaris, M., Mastronardo, G., Sormani, M. P., Iannucci, G., Rodegher, M., Comi, G. and Filippi, M. 1998, Brain MRI lesion volume measurement reproducibility is not dependent on the disease burden in patients with multiple sclerosis, in *Magn. Reson. Imag.* **16**, 1185–1189, Copyright 1998 with permission from Elsevier Science Ltd

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.

The simplest and most useful approach to characterizing measurement error is that of Bland and Altman, which uses pairs of repeated measurements in a range of subjects (Bland and Altman, 1986, 1996a; Wei *et al.*, 2002; Galbraith *et al.*, 2002; Padhani *et al.*, 2002).

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 standard deviation s. The 95 % confidence limit on a single measurement is $1.96s$. The difference between the repeats in pairs of measurements is also normally distributed, with SD $= \sqrt{2}s = 1.414s$. 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, and from this the SD of the measurements on a single subject. When using this technique, consideration should be given to what aspect of the measurement process is to be characterized. 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. 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.

Estimation of s, also called the *within-subject variability*, in the underlying distribution of measurements (all with the same mean) characterizes 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 measurements). 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. Other thresholds than 95 % can be specified (Padhani *et al.*, 2002). These lower and upper limits to differences that can arise from measurement error are sometimes called the *limits of agreement* (Bland and 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.

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

3.3.1.1 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 analytical check can be made using a rank correlation coefficient (Kendall's tau; Bland and Altman, 1996b). 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 the mean value. For the situation where SD is proportional to the mean, a log transformation is appropriate (Bland and 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 (see Figure 3.6). 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 SD of a population can be estimated, without bias, from n samples, and is

$$s = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2 \tag{3.1}$$

where

$$\overline{x} = \frac{1}{n}\sum_{1}^{n} x_i$$

The uncertainty (1 SD) in estimating an SD from n samples is (Taylor, 1997, p. 298):

$$\sigma_{\mathrm{s}} = \frac{s}{\sqrt{[2(n-1)]}} \tag{3.2}$$

There may be a biological component to differences in repeated measurements, even in the short term. Normal biological variation at timescales greater than about a second (e.g. in blood perfusion) can be removed from an estimate of machine precision 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.

3.3.1.2 Mean Absolute Difference in Pairs of Replicates

Instead of taking the signed differences (as in Bland and Altmans' procedure above), the absolute (unsigned) difference Δ may be taken. Its mean value is

$$\begin{aligned} < |\Delta| > &= 2\int_0^{\infty} xP(x) = 2\int \frac{x\mathrm{e}^{-x^2/2\sigma^2}\mathrm{d}x}{\sigma\sqrt{2\pi}} \\ &= \sqrt{2/\pi}\sigma = 0.7979\sigma \end{aligned} \tag{3.3}$$

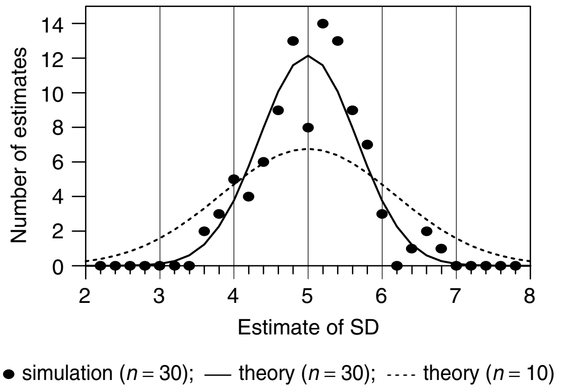

Figure 3.7. Simulation of estimation of reproducibility from repeated measurements. Over 8000 samples from a population of random numbers with mean = 100 and SD = 5 were generated. From these, 30 pairs of samples (replicates) were taken, and the differences Δ calculated, retaining the sign of the difference (it could be $+$ or $-\Delta$. The SD of the Δ) values was found (SD Δ), and from this the SD of the population was estimated ($S = \text{SD}\Delta/1.414$). 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 Q1 distribution [from Equations (3.9) and (3.11)]• is 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 SD of 0.66 was estimated, which gives a 95 % CL of ±1.3 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

where $P(x)$ is the normal distribution:

$$P(x) = \frac{e^{-x^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \tag{3.4}$$

and from this σ, the SD can be found ($\sigma = 1.2533 < |\Delta| >$).

This expression was confirmed numerically using the simulation reported in Figure 3.7.

3.3.2 Intra-class 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 (inter-subject), measurement variance may be less important, particularly if groups are being compared. The ICC is:

$$\text{ICC} = \frac{\text{variance from subjects}}{\begin{array}{c}\text{variance from subjects}\\ +\ \text{variance from measurement error}\end{array}} \tag{3.5}$$

The ICC can be thought of being the fraction of the total variance that is attributed to the subjects (rather than measurement error), i.e. ICC = inter-subject variance/(inter-subject variance + intra-subject variance). Thus if measurement error is small compared with 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 it has at least two problems. ICC depends on the group of subjects being studied (Bland and Altman, 1996d), 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); it is probably best to measure both ICC and CV.

3.3.3 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 for example Chard *et al.*, 2002). The variance arising from different parts of the measurement procedure can be estimated, as well as inter-subject and inter-scanner 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).

3.3.4 Other Methods

3.3.4.1 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). For 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 inter-subject variation will also increase its value (Bland and Altman, 1996d). Good correlation does not imply good agreement.

3.3.4.2 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. Appropriate methods for analysing MR data are still under intense discussion. The clinical metrics are also being scrutinized and redesigned (Hobart *et al.*, 2000). Developments in psychology are probably ahead of those used in this field – see the books on biostatistics and biometrics (Table 2.2). The review by Krummenauer and Doll (2000) is helpful.

3.3.5 Psychometric Measures: Sensitivity, Validity and Reliability

From a clinical point of view, a potential new quantity to characterize brain tissue can be evaluated by considering:

- Sensitivity – does the quantity alter with disease? Is the false negative rate low?
- Validity – is it relevant to the biological changes that are taking place?
- Reliability – is it reproducible? Is the false positive rate low?

Thus the concept of validity (which is absent from a judgement based merely on accuracy and precision) enables the relevance of a metric to be considered. For example, intra-cranial volume could be measured very accurately and precisely, but would be completely irrelevant in most situations. In Chapter 12, Section 12.6, an excellent discussion is given by Dr Nick Ramsey of these three concepts (in the context of fMRI). An alternative viewpoint (closely related) is the set of four psychometric properties often used to assess scores: acceptability, reliability, validity and responsiveness (Hobart *et al.*, 2000).

3.3.6 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.[8] Type A errors are those estimated by repeated measurements, whilst type B errors are all

[8] From *Estimating Uncertainties in Testing*, Measurement Good Practice Guide no. 36, by Keith Birch, of the British Measurement and Testing Association, on behalf of the National Physical Laboratory, UK. NIST technical note TN1297, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, from the USA National Institute of

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 are analysed and combined (Tofts *et al.*, 2000).

3.3.7 Noise Propagation and the Cramér–Rao Minimum Variance Bound

The contribution of image noise to imprecision in the final parameter can be calculated. If a simple ratio of images is used (for example magnetization transfer ratio – see Chapter 8), then propagation of errors (Taylor, 1997) allows the effect of noise in each source image to be calculated. An analytical expression can be derived for the total noise, and this can be minimized 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). 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 Cramér–Rao minimum variance bound (van den Bos, 1982; Cavassila *et al.*, 2001) 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 uses this method to estimate the minimum uncertainty in the metabolite concentration. Only uncertainty arising from data noise is included; other factors (both random and systematic) can make the uncertainty higher than this minimum variance bound. Another way to model noise propagation is to use numerical simulation, such as the Monte-Carlo technique, where known noise is added to the source data and the effect on the fitted parameter measured.

3.4 PHANTOMS (TEST OBJECTS)

Phantoms can be made from a single component or mixtures. Geometric objects, used for size or volume standards, are often made of acrylic. Major manufacturers are Perspex in the UK and Plexiglas in North America. 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.

3.4.1 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.[9]

Water has the advantage of being easily available, and of a standard composition. Its intrinsic $T_1 \approx 3.3$ s, $T_2 \approx 2.5$ s at room temperature (see Table 3.6 below), and in its pure form these long relaxation times usually cause problems. The long T_1 can lead to incomplete relaxation with

Standards and Technology, covers similar material, and is available online. The standard work is the *Guide to the Expression of Uncertainty in Measurement* (*GUM*), published by the International Standards Organization (ISO) in 1995. There is much commercial activity in this field, as organizations selling measurement services seek ISO accreditation. Many national organizations produce guidance on 'the expression of uncertainty in measurement', and publish user-friendly versions of GUM. These can be located via the Internet.

[9] 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.

sequences that may allow full relaxation with normal brain tissue ($T_1 \approx 600$ ms for normal white matter at 1.5 T; see Chapter 5, Figure 5.3). 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 the next section). 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 ($\varepsilon = 80$) leads to the presence of radiofrequency standing waves (dielectric resonance), where B_1 is enhanced, giving an artificially high flip angle and signal (see Figure 3.8). The high dielectric constant reduces the wavelength of electromagnetic radiation, compared with its value in free space, by a factor $\sqrt{\varepsilon}$; at 1.5 T the wavelength is 0.52 m, comparable with the dimensions of the subject (Glover *et al.*, 1985; Tofts, 1994; Hoult, 2000). Standing waves are also present in the head, particularly at high field (see Figure 3.9), but to a much less extent, because electrical conductivity in the brain tissues damps the resonance. Even at 1.5 T this effect is significant, and early attempts to measure head coil nonuniformity using large aqueous phantoms are now seen as fatally flawed (Table 3.4).

Oil has a low dielectric constant ($\varepsilon = 2 - 3$), and has been used for nonuniformity phantoms (Tofts *et al.*, 1997a). Several kinds are available, from various sources, and their properties have been described by Tofts *et al.* (1997a). 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 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_1s are generally too low, at 100–190 ms, although some flammable oils have higher values).

Organic liquids such as alkanes have been used for diffusion standards (Tofts *et al.*, 2000). Cyclic alkanes C_nH_{2n} ($n = 6$–8) are the simplest possible 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; 11 are readily available, ranging from hexane (which is very volatile and inflammable), through octane [a major constituent of petrol (gasoline), used to power vehicles],

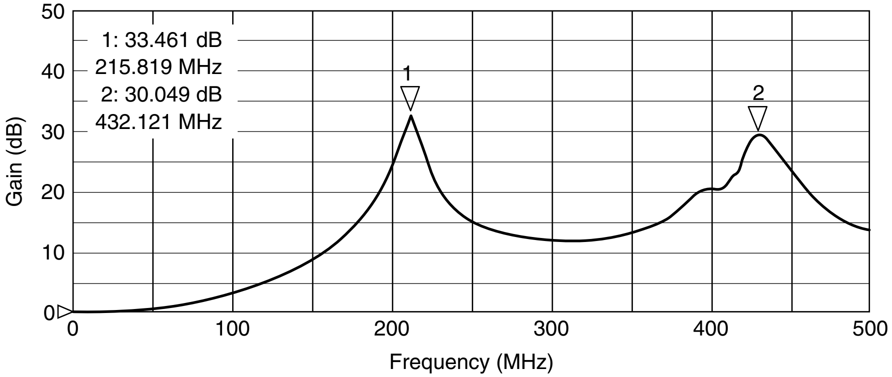

Figure 3.8. Dielectric resonance in a spherical flask of water. One small RF coil was placed inside the 2 l flask (diameter 15.6 cm), and one outside. The graph shows the transmission between one coil and the other. The plot is the same regardless of whether the inner coil or the outer coil transmits (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 5.1 T for protons. Reproduced with permission from Hoult, D. I. 2000, The principle of reciprocity in signal strength calculations – a mathematical guide, in *Conc. Magn. Reson.*, Copyright 2000 John Wiley & Sons Inc

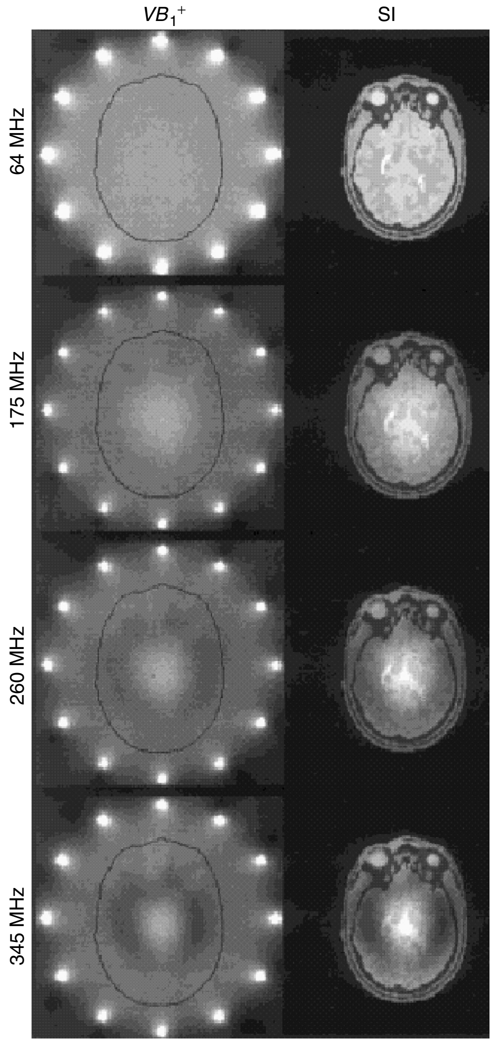

Figure 3.9. RF nonuniformity in the head. An accurate mathematical model of the head inside a birdcage coil, with FA = 90° at the head centre. Frequencies correspond to fields of 1.5, 4.1, 6.1 and 8.1 T. $VB_1{}^+$ is the excitation field (normalized by the factor V), SI is the signal intensity from a gradient echo sequence. The doming effect at the centre of the head becomes increasingly pronounced at higher fields, and an annular region of reduced signal is visible further from the centre. Reproduced with permission from Collins, C. M. and Smith, M. B. 2001, Signal-to-noise ratio and absorbed power as functions of main magnetic field strength, and definition of '90 degrees' RF pulse for the head in the birdcage coil', in *Magn. Reson. Med.*, Copyright 2001 John Wiley & Sons Inc

Table 3.4. RF nonuniformity in uniform phantoms. 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 polarized RF coil is assumed. Filling with a low dielectric constant oil ($\varepsilon = 5$) allows larger phantoms to be used

field B_0	water ($\varepsilon = 80$)	oil ($\varepsilon = 5$)
0.5 T	13.8 cm	55.1 cm
1.5 T	4.6 cm	18.4 cm
4.7 T	1.5 cm	5.9 cm

Adapted from Tofts (1994).

to hexadecane (which freezes at 15 °C). Their T_1 values are realistic (670–1900 ms), but their 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 0.8×10^{-9} m^2 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 replaced regularly.

3.4.2 Multiple Component Mixtures

Doped water has reduced T_1 and T_2, giving a material with more realistic values of relaxation times. The classic compounds used for doping are copper sulfate, $CuSO_4$, and manganese chloride, $MnCl_2$; Gd salts such as $GdCl_3$ can also be used. Ni^{2+} has the advantage of a low T_1 temperature coefficient (see Section 3.4.3 below). Agarose is good for reducing T_2 whilst hardly affecting T_1. They are characterized 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:

$$\frac{1}{T_1} = R_1 = R_{10} + r_1 c$$
$$\frac{1}{T_2} = R_2 = R_{20} + r_2 c \qquad (3.6)$$

Table 3.5. Values of relaxivity at 1.5 T at room temperature

Relaxation agent	Source	$r_1(\mathrm{s}^{-1}\ \mathrm{mM}^{-1})$	$r_2(\mathrm{s}^{-1}\ \mathrm{mM}^{-1})$
Mn^{2+}	Morgan and Nolle (1959)[a]	7.0 ± 0.4	70 ± 4
	Bloembergen and Morgan (1961)[b]	8.0 ± 0.4	80 ± 7
Ni^{2+}	Morgan and Nolle (1959)[a]	0.70 ± 0.06	0.70 ± 0.06
	Hertz and Holz (1985)[c]	0.64	–
	Kraft *et al.* (1987)[d]	0.64	–
	Jones (1997)[e]	0.644 ± 0.002	0.698 ± 0.005
Ni-DTPA	Tofts *et al.* (1993)[f]	0.114 ± 0.005	0.10 ± 0.01
Cu^{2+}	Morgan and Nolle (1959)[a]	0.69 ± 0.04	0.77 ± 0.04
	Mitchell *et al.* (1986)[g]	0.71 ± 0.03	0.72 ± 0.04
Gd^{3+}	Morgan and Nolle (1959)[a]	9.8 ± 0.7	11.6 ± 0.9
Gd-DTPA	Tofts *et al.* (1993)[h]	4.50 ± 0.04	5.49 ± 0.06
Agarose	Mitchell *et al.* (1986)	0.05	10
	Tofts *et al.* (1993)	0.01 ± 0.01	9.7 ± 0.2[j]
	Jones (1997)[i]	0.04 ± 0.01	8.80 ± 0.04

[a] At 60 MHz, 27 °C, calculated by the author from data points on the published figures; 95 % confidence limits estimated from scatter in the plots.

[b] 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.

[c] $Ni(ClO_4)_2$ 98 mM at 60 MHz, 12–30 °C, pH = 3.0. other pH values were also measured.

[d] Estimated by the author from published T_1 value (5 mM aqueous solution: $R_1 = 3.5\mathrm{s}^{-1}$; T_1 of water = 3.5 s).

[e] Estimated by the author, from data at 1.5 T in the MSc thesis of Craig K. Jones (University of British Columbia, 1997) available on-line at www.physics.ubc.ca/~craig/msc/node46.html. r_1 and r_2 of Ni^{2+} in the table were estimated from measurements in 0.2 % agarose of Ni^{2+} concentrations up to 5 mM. At higher concentrations of agarose r_1 increased slightly, and r_2 substantially (1 % agarose: $r_1 = 0.650 \pm 0.010; r_2 = 0.725 \pm 0.010$; 4 % agarose: $r_1 = 0.681 \pm 0.002, r_2 = 0.99 \pm 0.16$). The binding of Ni^{2+} with agarose appears to be minimal for concentrations required to make a mixture like normal white matter. 2 mM Ni^{2+} in 1 % agarose gives $T_1 = 573$ ms, $T_2 = 95$ ms. To simulate pathology, with higher T_1 and T_2 will require lower concentrations with less chance of binding. The Ni^{2+} plots are linear at 1 % agarose, thus T_1 can be reduced by the addition of more Ni^{2+} to simulate the presence of Gd in tissue, if required.

[f] Values are in aqueous solution. In 2 % agarose gel: $r_1 = 0.105 \pm 0.001\mathrm{s}^{-1}\ \mathrm{mM}^{-1}, r_2 = 0.0 \pm 0.1\mathrm{s}^{-1}\ \mathrm{mM}^{-1}$; the zero value for r_2 may be caused by binding of Ni-DTPA with agarose.

[g] Estimated by the author from published data at 60 MHz.

[h] Values are in aqueous solution. In 2 % agarose gel: $r_1 = 4.37 \pm 0.04\ \mathrm{s}^{-1}\ \mathrm{mM}^{-1}, r_2 = 5.44 \pm 0.05\ \mathrm{s}^{-1}\ \mathrm{mM}^{-1}$.

[i] Estimated from data of Jones in 0.5 mM Ni^{2+} (see footnote *e*). In 2 mM Ni^{2+}, $r_1 = 0.06 \pm 0.02$; $r_2 = 8.71 \pm 0.01$.

[j] Agarose r_2 value in water. Values in Ni-DTPA are lower; in 4 mM Ni-DTPA, $r_2 = 8.2 \pm 0.1$.

where 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.

Cu, Gd and Ni have similar r_1 and r_2 relaxivities, however Mn has r_2 about 10 times r_1 and is useful for making solutions with T_2 lower than T_1, as is the case in tissue (Table 3.5).

3.4.2.1 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 (Krynicki, 1966; Table 3.6).

Doped agarose gels can be made up in a similar way to doped water (Mitchell *et al.*, 1986; Walker *et al.*, 1988, 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 3.5). Agarose flakes are dissolved in hot water, up to concentrations of about

Table 3.6. 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 %.[a] Values are linearly interpolated (from data in Krynicki, 1966, reported in Duck, 1990)

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

[a] Water used for making phantoms is likely to have some oxygen in it (depending on whether it was recently boiled). Its T_1 value is crucially dependent on the presence of dissolved oxygen, and in the vitreous humour of the eye the reduction in T_1 is used to monitor the concentration of oxygen diffusing from the retina (Berkowitz *et al.*, 2001). The relaxivity for oxygen is approximately $1.8 \pm 0.3 \times 10^{-4} \text{s}^{-1} (\text{mmHg})^{-1}$ [measured in plasma at 4.7 T by Meyer *et al.* (1995)]. Assuming this value still holds at 1.5 T, fully oxygenated water at 23 °C ($pO_2 = 150$ mmHg) would then have its T_1 reduced from 3.40 to 3.11 s, 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 accompanied by T_2 values (high quality measurements of water T_2 seem to be completely lacking).

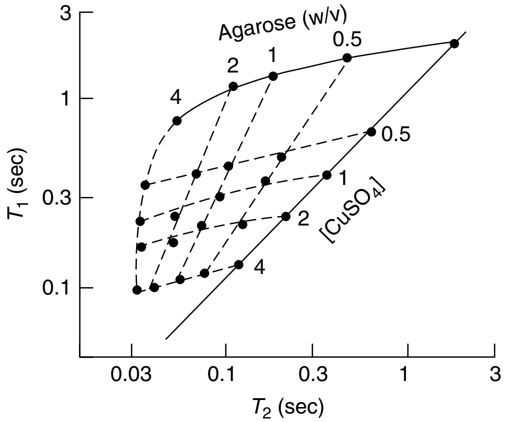

Figure 3.10. T_1 and T_2 values for aqueous solutions of agarose and $CuSO_4$. Agarose concentrations are from 0 to 4 % weight/volume, Cu is from 0 to 4 mM. Note agarose decreases T_2 but hardly affects T_1, whilst Cu decreases T_1 and T_2 about equally. The agarose line (curved, where [Cu] = 0) and the Cu line (straight, where [agarose] = 0) bounds the possible values that the mixture can achieve. Dotted lines connect points of equal agarose or Cu concentration. Reproduced from Mitchell, M. D., Kundel, H. L., Axel, L. and Joseph, P. M. 1986, Agarose as a tissue equivalent phantom material for NMR imaging, in *Magn. Reson. Imag.* **4**, 263–266, Copyright 1986 with permission from Elsevier Science Ltd

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 3.1); 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 made intermediate between those that would be obtained with only one of the compounds. (Figure 3.10; 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^{2+} and Mn^{2+} in aqueous solution (Schneiders, 1988), Gd-DTPA[10] and agarose (Walker *et al.*, 1989), Ni^{2+} and agarose (Kraft *et al.*, 1987), and Ni-DTPA and agarose (Tofts *et al.*, 1993). Linear

[10] This is preferred to $GdCl_3$, described in Walker's earlier work (Walker *et al.*, 1988), which interacted with the agarose.

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. For a mixture of Ni (N) in agarose (a), the concentrations required are (Tofts *et al.*, 1993):

$$C_a = \frac{R_2 - R_{2w} - (r_{2N}/r_{1N})(R_1 - R_{1w})}{r_{2a} - (r_{2N}/r_{1N})r_{1a}} \quad (3.7)$$

$$C_N = \frac{R_1 - R_{1w} - (r_{1a}/r_{2a})(R_2 - R_{2w})}{r_{1N} - (r_{1a}/r_{2a})r_{2N}} \quad (3.8)$$

where R_1 and R_2 are the desired relaxation rates ($R_1 = 1/T_1$; $R_2 = 1/T_2$), and R_{1w}, R_{2w} are the values for water.

A mixture of Ni^{2+} in agarose provides temperature independence for T_1. Using Jones' values of relaxivity[11] (see Table 3.5) and water relaxation,[12] the equations for this mixture then become:

$$C_a = 115.7/T_2 - 129.0/T_1 + 0.019\,\% \quad (3.9)$$

$$C_N = 1550/T_1 - 10.70/T_2 - 0.705 \text{ mM} \quad (3.10)$$

Thus a phantom with $T_1 = 600$ ms, $T_2 = 100$ ms is produced by mixing 1.77 mM Ni^{2+} in 0.96 % agarose.

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.

3.4.2.2 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. Mixtures of 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 $T_2 \approx 100$ ms, $T_1 \approx 800$ ms. Agar (Mathur-De Vre *et al.*, 1985; Walker *et al.*, 1988) and gelatine have also been used, but these are relatively impure and have been superseded by agarose. 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.

3.4.3 Temperature Dependence

The temperature dependence of phantom properties can be a major limitation to their usefulness. Typical changes in T_1, T_2 and D are 1–3 %/°C. The Eurospin gels have a T_1 temperature coefficient of +2.6 %/°C.[13] The best temperature performance T_1 and T_2 phantoms are obtained using Ni or a Gd polymer pair in agarose. Ni has a minimum in its relaxation rate, fortunately at room temperature (Kraft *et al.*, 1987), allowing a brain-equivalent Ni-DTPA agarose gel to have a flat temperature response (Figure 3.11). A more general solution is to use a Gd polymer pair[14] (Kellar and Briley-Saebo, 1998). One component (NC663 868) 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 neutralize 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 coefficient for a range of T_1 values. T_2 behaviour cannot currently be made independent of temperature. Agarose has a T_2 coefficient of about −1.25 %/°C (Tofts *et al.*, 1993). [15] In alkane phantoms, the coefficient changes by

[11] Agarose relaxivity values for 2 mM Ni^{2+} and Ni^{2+} relaxivity values for 1 % agarose were used (given in Table 3.5 footnote).

[12] Analysis of Jones' data (see Table 3.5) gives $R_{1w} = 0.457$ s^{-1}, $R_{2w} = 0.345$ s^{-1}.

[13] 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.

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

[15] Walker *et al.* (1988) gives 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.

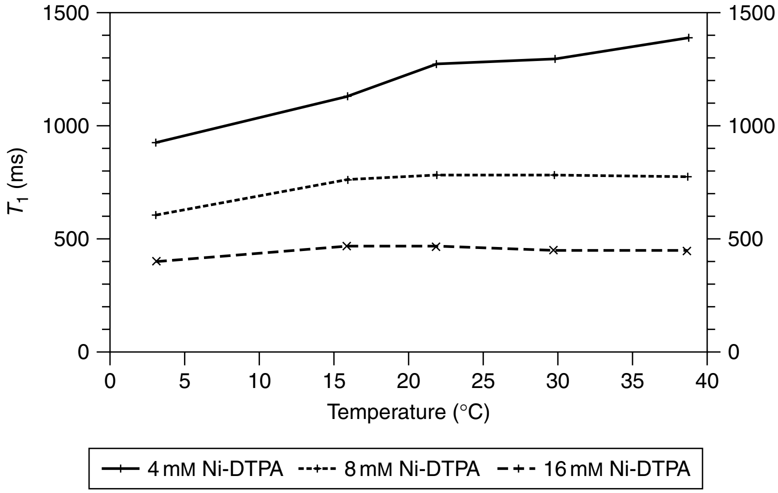

Figure 3.11. Temperature dependance of T_1 in a Ni-doped gel. In the tissue-equivalent material (Ni-DTPA in 2 % agarose), T_1 is dominated by the relaxation from Ni-DTPA, particularly at the lower T_1 values, and therefore has little dependence on temperature. At room temperature the 8 and 16 mM data 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. Reproduced from Tofts, P. S., Shuter, B. and Pope, J. M. 1993, Ni-DTPA doped agarose gel – a phantom material for Gd-DTPA. enhancement measurements, in *Magn. Reson. Imag.* **11**, 125–133, Copyright 1993, with permission from Elsevier Science Ltd

2–3 %/°C (Tofts *et al.*, 2000), and some form of temperature control and monitoring are needed.

General methods for controlling and monitoring temperature have yet to be developed. Storing the phantoms near room temperature, thermally insulating them during their time in the magnet bore (which may have a different and varying temperature) and measuring their temperature (ideally whilst in the bore, using a thermocouple or a liquid-in-glass thermometer) are all essential components of a temperature protocol. 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 minimized by avoiding both rapid changes in temperature and the presence of any electrical conductors which might attract induced RF currents and consequent heating.

Unless special precautions are taken, the measurements made on T_1, T_2 and D phantoms are all subject to uncertainty in temperature of about 2 °C, which corresponds to errors of about 5 % in the parameter value. PD and spectroscopy measurements may have a temperature error of 18 °C, if no correction has been made for the increase in signal at room temperature, which corresponds to a similar error of 6 % in the parameter value.[16]

3.4.4 Phantom Design

Some parameters have obvious phantom designs associated with them; for others there is currently nothing established that is available (see Table 3.7). Measurements in normal white matter are often preferable (see Table 3.1).

3.4.4.1 Proton Density

Water has a well-defined proton density, although the small temperature dependence of the signal (−0.3 %/°C) should be taken into account (see Chapter 4). Variation can be provided by the use of D_2O, or glass beads (see Chapter 4).

[16] The signal increases by 6 % between room temperature (21 °C = 292 K) and body temperature (37 °C = 310 K).

Table 3.7. Phantoms and normal scanning for quantitative QA measurements

MR quantity	Phantom	Normal human study
PD	Water (? + D_2O)	Normal white matter
T_1	Doped water, doped gel or alkanes	
T_2		
D	Alkanes or sucrose solution	
MT	Cooked egg-white or BSA	
MRS	Metabolite solutions	
T_1-weighted DCE	? Complex phantom	? Ethics prevent repeated Gd
T_2-weighted DCE	? Complex phantom	
fMRI	None	? Normal activation
ASL	? Complex phantom	Intra-subject variation ≈ 10 %
Volume	Acrylic objects in water bath	Brain or intra-cranial volume
Shape		? Cortical folds
Texture	? Small beads set in gel	? Normal white matter

3.4.4.2 T_1 *and* T_2

The simplest option is a single component liquid, although a realistic T_2 will probably not be obtainable. An aqueous solution of $CuSO_4$ or $MnCl_2$ will provide a reasonable range of T_1 and T_2 values. If a long T_2 is required, Cu is preferred. If a more realistic T_2 is wanted, Mn can provide values close to those in brain.[17] However longer T_1 values will require correspondingly longer T_2 values, which may not be acceptable.[18] Using agarose gels provides greatly increased flexibility, at the expense of increased preparation time. Doping with $GdCl_3$ (Walker *et al.*, 1988) or Gd-DTPA (Walker *et al.*, 1989) is straightforward. If reduced dependence on temperature (for T_1) and on field is desired, $NiCl_2$ (Kraft *et al.*, 1987; see Figure 3.12) or Ni-DTPA (Tofts *et al.*, 1993) should be used. For ultimate independence of temperature for T_1, the Gd polymer pair can give zero temperature coefficient for a range of T_1 values (Kellar and Briley-Saebo, 1998).

3.4.4.3 Diffusion

Alkanes are ideal for ADC. An accepted anisotropic phantom for tensor imaging does not

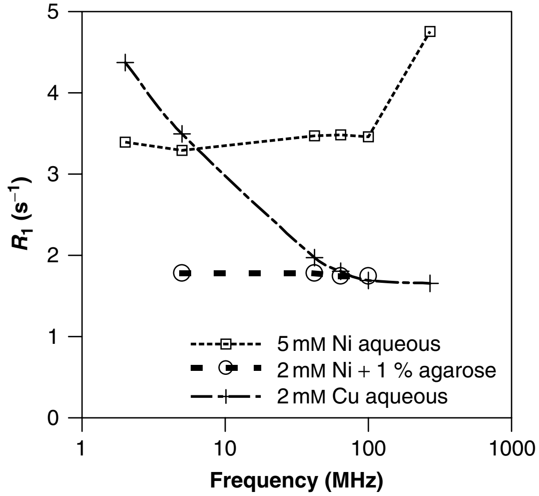

Figure 3.12. Field dependence of proton relaxivity for Ni^{2+} and Cu^{2+} in aqueous solution and agarose gels. Cu^{2+} has a large frequency dependence, whilst Ni^{2+} is independent of frequency up to at least 100 MHz. Frequency values include 42 MHz (1.0 T), 64 MHz (1.5 T), 100 MHz (2.4 T) and 270 MHz (6.3 T). The dotted lines do not imply the values of R_1 between the symbols. [Re-drawn from Kraft *et al.* (1987); data on Cu^{2+} in agarose have been omitted for clarity; these show similar field dependence to Cu^{2+} in aqueous solution.] Reproduced with permission from Kraft, K. A., Fatouros, P. P., Clarke, G. D. and Kishore, P. R. 1987, An MRI phantom material for quantitative relaxometry in *Magn. Reson. Med.*, Copyright 1987 John Wiley & Sons Inc

[17] Using the relaxivity values in Table 3.5, and assuming pure water has $T_1 = 3.5$ s, $T_2 = 2.5$ s, a 0.2 mM aqueous solution of $MnCl_2$ has $T_1 = 615$ ms, $T_2 = 96$ ms.

[18] A 0.1 mM solution of $MnCl_2$ has $T_1 = 1046$ ms, $T_2 = 185$ ms.

yet exist; the difficulty is finding a substance which is stable, and contains long compartments of water that are narrow enough to restrict diffusion. This diameter[19] needs to be less than about 70 μm. A phantom made of sheets of parallel plastic capillaries, of internal diameter 50 μm, has recently been described (Lin *et al.*, 2002).

3.4.4.4 Magnetization Transfer

Bovine serum albumin, cross-linked by heat treatment, gives a good MTR, up to 72 pu, and better than agarose (Mendelson *et al.*, 1991).

3.4.4.5 Spectroscopy

Metabolite solutions are used to measure the scanner sensitivity. It may be desirable to refrigerate these for stability; if so, the temperature should be controlled, and the user should be aware of how this affects signal (see Chapter 9). Test objects for spectroscopy have been described by Leach *et al.* (1995) and Keevil *et al.* (1998).

3.4.4.6 Volume

Standards can be made from acrylic (Tofts *et al.*, 1997b; Lemieux and Barker, 1998).

3.4.5 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, because of the lack of an internal susceptibility field gradient. A long cylinder can also give a uniform internal field (see Figure 12.4). 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 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 close fit inside the head coil.

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 that 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.

3.4.6 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 sterilize the container; alternatively, a fungicide can be added. 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.[20] An alternative is to use a cylindrical glass bottle, pour melted wax over the solid gel, then glue the lid on.[21] 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

[19] A rough calculation for free water ($D \approx 3 \times 10^{-9} \mathrm{m}^2\mathrm{s}^{-1}$ at body temperature, $TE = 80$ ms) shows that the rms displacement is 38 μm. Thus a 70 μm diameter cylinder of water would show appreciable anisotropy.

[20] A chemistry glassmaker can often make such a container.

[21] This approach appears to have been used with the Eurospin gels.

change. 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. The gel dosimeters (see Section 3.4.2 above) may ultimately provide better solutions to the problems of stability and evaporation.

Acknowledgements

Craig Jones and Cornelia Laule advised on sealing glass tubes.

REFERENCES

Armitage, P., Matthews, J. N. S. and Berry, G. 2001, *Statistical Methods in Medical Research*. Blackwell, Oxford.

Barker, G. J. and Tofts, P. S. 1992, Semiautomated quality assurance for quantitative magnetic resonance imaging, *Magn. Reson. Imag.*, **10**, 585–595.

Bartha, R., Drost, D. J., Menon, R. S. and Williamson, P. C. 2000, Comparison of the quantification precision of human short echo time (1)H spectroscopy at 1.5 and 4.0 Tesla, *Magn. Reson. Med.*, **44**, 185–192.

Berkowitz, B. A., McDonald, C., Ito, Y., Tofts, P. S., Latif, Z. and Gross, J. 2001, Measuring the human retinal oxygenation response to a hyperoxic challenge using MRI: eliminating blinking artifacts and demonstrating proof of concept, *Magn. Reson. Med.*, **46**, 412–416.

Berry, I., Barker, G. J., Barkhof, F., Campi, A., Dousset, V., Franconi, J. M., Gass, A., Schreiber, W., Miller, D. H. and Tofts, P. S. 1999, A multicenter measurement of magnetization transfer ratio in normal white matter, *J. Magn. Reson. Imag.*, **9**, 441–446.

Bland, J. M. and Altman, D. G. 1986, Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet*, **1**(8476), 307–310.

Bland, J. M. and Altman, D. G. 1996a, Measurement error, *Br. Med. J.*, **312**(7047), 1654.

Bland, J. M. and Altman, D. G. 1996b, Measurement error, *Br. Med. J.*, **313**(7059), 744.

Bland, J. M. and Altman, D. G. 1996c, Measurement error proportional to the mean, *Br. Med. J.*, **313**(7049), 106.

Bland, J. M. and Altman, D. G. 1996d, Measurement error and correlation coefficients, *Br. Med. J.*, **313**(7048), 41–42.

Bloembergen, N. and Morgan, L. O. 1961, Proton relaxation times in paramagnetic solutions. Effects of electron spin relaxation, *J. Chem. Phys.*, **34**, 842–850.

Buckley, D. L. 2002, Uncertainty in the analysis of tracer kinetics using dynamic contrast-enhanced T_1-weighted MRI, *Magn. Reson. Med.*, **47**, 601–606.

Cardenas, V. A., Ezekiel, F., Di, S., V., Gomberg, B. and Fein, G. 2001, Reliability of tissue volumes and their spatial distribution for segmented magnetic resonance images, *Psychiat. Res.*, **106**, 193–205.

Cavassila, S., Deval, S., Huegen, C., van Ormondt, D. and Graveron-Demilly, D. 2001, Cramer-Rao bounds: an evaluation tool for quantitation, *NMR Biomed.*, **14**, 278–283.

Chard, D. T., McLean, M. A., Parker, G. J., MacManus, D. G. and Miller, D. H. 2002, Reproducibility of in vivo metabolite quantification with proton magnetic resonance spectroscopic imaging, *J. Magn. Reson. Imag.*, **15**, 219–225.

Charles, H. C., Lazeyras, F., Tupler, L. A. and Krishnan, K. R. 1996, Reproducibility of high spatial resolution proton magnetic resonance spectroscopic imaging in the human brain, *Magn. Reson. Med.*, **35**, 606–610.

Christoffersson, J. O., Olsson, L. E. and Sjoberg, S. 1991, Nickel-doped agarose gel phantoms in MR imaging, *Acta Radiol.*, **32**, 426–431.

Cohen, J. A., Fischer, J. S., Bolibrush, D. M., Jak, A. J., Kniker, J. E., Mertz, L. A., Skaramagas, T. T. and Cutter, G. R. 2000, Intrarater and interrater reliability of the MS functional composite outcome measure, *Neurology*, **54**, 802–806.

Collins, C. M. and Smith, M. B. 2001, Signal-to-noise ratio and absorbed power as functions of main magnetic field strength, and definition of '90 degrees' RF pulse for the head in the birdcage coil, *Magn. Reson. Med.*, **45**, 684–691.

De Deene, Y., Hanselaer, P., De Wagter, C., Achten, E. and De Neve, W. 2000, An investigation of the chemical stability of a monomer/polymer gel dosimeter, *Phys. Med. Biol.*, **45**, 859–878.

Duck, F. A. 1990, *Physical Properties of Tissue* Academic Press, London.

Filippi, M., Horsfield, M. A., Ader, H. J., Barkhof, F., Bruzzi, P., Evans, A., Frank, J. A., Grossman, R. I., McFarland, H. F., Molyneux, P., Paty, D. W., Simon, J., Tofts, P. S., Wolinsky, J. S. and Miller, D. H. 1998a, Guidelines for using quantitative measures of brain magnetic resonance imaging abnormalities in monitoring the treatment of multiple sclerosis, *Ann. Neurol.*, **43**, 499–506.

Filippi, M., Gawne-Cain, M. L., Gasperini, C., van-Waesberghe, J. H., Grimaud, J., Barkhof, F., Sormani,

M. P. and Miller, D. H. 1998b, Effect of training and different measurement strategies on the reproducibility of brain MRI lesion load measurements in multiple sclerosis, *Neurology*, **50**, 238–244.

Firbank, M. J., Harrison, R. M., Williams, E. D. and Coulthard, A. 2000, Quality assurance for MRI: practical experience, *Br. J. Radiol.*, **73**(868), 376–383.

Fox, N. C. and Freeborough, P. A. 1997, Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease, *J. Magn. Reson. Imag.*, **7**, 1069–1075.

Galbraith, S. M., Lodge, M. A., Taylor, N. J., Rustin, G. J., Bentzen, S., Stirling, J. J. and Padhani, A. R. 2002, Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis, *NMR Biomed.*, **15**, 132–142.

Gasperini, C., Rovaris, M., Sormani, M. P., Bastianello, S., Pozzilli, C., Comi, G. and Filippi, M. 2001, Intraobserver, inter-observer and inter-scanner variations in brain MRI volume measurements in multiple sclerosis, *Mult. Scler.*, **7**, 27–31.

Gawne-Cain, M. L., Webb, S., Tofts, P. and Miller, D. H. 1996, Lesion volume measurement in multiple sclerosis: how important is accurate repositioning?, *J. Magn. Reson. Imag.*, **6**, 705–713.

Glover, G. H., Hayes, C. E., Pelc, N. J., Edelstein, W. A., Mueller, O. M., Hart, H. R., Hardy, C. J., O'Donnell, M. and Barber, W. D. 1985, Comparison of linear and circular polarization for magnetic resonance imaging, *J. Magn. Reson.*, **64**, 255–270.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G. J., Plummer, D. L., Tofts, P. S., McDonald, W. I. and Miller, D. H. 1996, Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques, *Magn. Reson. Imag.*, **14**, 495–505.

Hajek, M., Babis, M. and Herynek, V. 1999, MR relaxometry on a whole-body imager: quality control, *Magn. Reson. Imag.*, **17**, 1087–1092.

Hertz, H. G. and Holz, M. 1985, Longitudinal proton relaxation rates in aqueous Ni++ solutions as a function of the temperature, frequency, and pH value, *J. Magn. Reson.*, **63**, 64–73.

Hobart, J., Freeman, J. and Thompson, A. 2000, Kurtzke scales revisited: the application of psychometric methods to clinical intuition, *Brain*, **123**(Pt 5), 1027–1040.

Hoult, D. I. 2000, The principle of reciprocity in signal strength calculations – a mathematical guide, *Conc. Magn. Reson.*, **12**, 173–187.

Keevil, S. F., Barbiroli, B., Brooks, J. C., Cady, E. B., Canese, R., Carlier, P., Collins, D. J., Gilligan, P., Gobbi, G., Hennig, J., Kugel, H., Leach, M. O., Metzler, D., Mlynarik, V., Moser, E., Newbold, M. C., Payne, G. S., Ring, P., Roberts, J. N., Rowland, I. J., Thiel, T., Tkac, I., Topp, S., Wittsack, H. J. and Podo, F., 1998, 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. Imag.*, **16**, 1093–1106.

Kellar, K. E. and Briley-Saebo, K. 1998, Phantom standards with tempera, *Invest. Radiol.*, **33**, 472–479.

Kraft, K. A., Fatouros, P. P., Clarke, G. D. and Kishore, P. R. 1987, An MRI phantom material for quantitative relaxometry, *Magn. Reson. Med.*, **5**, 555–562.

Krummenauer, F. and Doll, G. 2000, Statistical methods for the comparison of measurements derived from orthodontic imaging, *Eur. J. Orthod.*, **22**, 257–269.

Krynicki, K. 1966, Proton spin lattice relaxation in pure water between 0 °C and 100 °C, *Physica*, **32**, 167–178.

Leach, M. O., Collins, D. J., Keevil, S., Rowland, I., Smith, M. A., Henriksen, O., Bovee, W. M. and Podo, F. 1995, Quality assessment in in vivo NMR spectroscopy: III. Clinical test objects: design, construction, and solutions, *Magn. Reson. Imag.*, **13**, 131–137.

Leary, S. M., Parker, G. J., Stevenson, V. L., Barker, G. J., Miller, D. H. and Thompson, A. J. 1999, Reproducibility of magnetic resonance imaging measurements of spinal cord atrophy: the role of quality assurance, *Magn. Reson. Imag.*, **17**, 773–776.

Lemieux, L. and Barker, G. J. 1998, Measurement of small inter-scan fluctuations in voxel dimensions in magnetic resonance images using registration, *Med. Phys.*, **25**, 1049–1054.

Lemieux, L., Liu, R. S. and Duncan, J. S. 2000, Hippocampal and cerebellar volumetry in serially acquired MRI volume scans, *Magn. Reson. Imag.*, **18**, 1027–1033.

Lepage, M., Whittaker, A. K., Rintoul, L., Back, S. A. and Baldock, C. 2001, The relationship between radiation-induced chemical processes and transverse relaxation times in polymer gel dosimeters, *Phys. Med. Biol.*, **46**, 1061–1074.

Lerski, R. A. 1993, Trial of modifications to Eurospin MRI test objects, *Magn. Reson. Imag.*, **11**, 835–839.

Lerski, R. A. and de Certaines, J. D. 1993, Performance assessment and quality control in MRI by Eurospin test objects and protocols, *Magn. Reson. Imag.*, **11**, 817–833.

Lin, C. P., Wedeen, V. J., Yao, C., Chen, J. H. and Tseng, W. Y. T. 2002, Validation of diffusion spectrum magnetic resonance imaging with registered manganese-enhanced optic tracts and phantom., *Proc. Int. Soc. Mag. Reson. Med.*, **10**, 442.

Losseff, N. A., Webb, S. L., O'Riordan, J. I., Page, R., Wang, L., Barker, G. J., Tofts, P. S., McDonald, W. I., Miller, D. H. and Thompson, A. J. 1996, Spinal cord atrophy and disability in multiple sclerosis. A new reproducible and sensitive MRI method with potential to monitor disease progression, *Brain*, **119**(Pt 3), 701–708.

Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., Rascol, O., Celsis, P. and Chollet, F. 2001, Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging, *J. Cereb. Blood Flow Metab.*, **21**, 592–607.

Marshall, I., Wardlaw, J., Cannon, J., Slattery, J. and Sellar, R. J. 1996, Reproducibility of metabolite peak areas in 1H MRS of brain, *Magn. Reson. Imag.*, **14**, 281–292.

Mathur-De Vre, R., Grimee, R., Parmentier, F. and Binet, J. 1985, The use of agar gel as a basic reference material for calibrating relaxation times and imaging parameters, *Magn. Reson. Med.*, **2**, 176–179.

McRobbie, D. W. and Quest, R. A. 2002, Effectiveness and relevance of MR acceptance testing: results of an 8 year audit, *Br. J. Radiol.*, **75**(894), 523–531.

Mendelson, D. A., Heinsbergen, J. F., Kennedy, S. D., Szczepaniak, L. S., Lester, C. C. and Bryant, R. G. 1991, Comparison of agarose and cross-linked protein gels as magnetic resonance imaging phantoms, *Magn. Reson. Imag.*, **9**, 975–978.

Meyer, M. E., Yu, O., Eclancher, B., Grucker, D. and Chambron, J. 1995, NMR relaxation rates and blood oxygenation level, *Magn. Reson. Med.*, **34**, 234–241.

Mitchell, M. D., Kundel, H. L., Axel, L. and Joseph, P. M. 1986, Agarose as a tissue equivalent phantom material for NMR imaging, *Magn. Reson. Imag.*, **4**, 263–266.

Morgan, L. O. and Nolle, A. W. 1959, Proton spin relaxation in aqueous solutions of paramagnetic ions II Cr^{+++}, Mn^{++}, Ni^{++}, Cu^{++} and Gd^{+++}, *J. Chem. Phys.*, **31**, 365.

Och, J. G., Clarke, G. D., Sobol, W. T., Rosen, C. W. and Mun, S. K. 1992, Acceptance testing of magnetic resonance imaging systems: report of AAPM Nuclear Magnetic Resonance Task Group No. 6, *Med. Phys.*, **19**, 217–229.

Padhani, A. R., Hayes, C., Landau, S. and Leach, M. O. 2002, Reproducibility of quantitative dynamic MRI of normal human tissues, *NMR Biomed.*, **15**, 143–153.

Parkes, L. M. and Tofts, P. S. 2002, Improved accuracy of human cerebral blood perfusion measurements using arterial spin labeling: accounting for capillary water permeability, *Magn. Reson. Med.*, **48**, 27–41.

Podo, F. 1988, Tissue characterization by MRI: a multidisciplinary and multi-centre challenge today, *Magn. Reson. Imag.*, **6**, 173–174.

Podo, F., Henriksen, O., Bovee, W. M., Leach, M. O., Leibfritz, D. and de Certaines, J. D. 1998, Absolute metabolite quantification by in vivo NMR spectroscopy: I. Introduction, objectives and activities of a concerted action in biomedical research, *Magn. Reson. Imag.*, **16**, 1085–1092.

Price, R. R., Axel, L., Morgan, T., Newman, R., Perman, W., Schneiders, N., Selikson, M., Wood, M. and Thomas, S. R. 1990, Quality assurance methods and phantoms for magnetic resonance imaging: report of AAPM nuclear magnetic resonance Task Group No. 1, *Med. Phys.*, **17**, 287–295.

Rovaris, M., Mastronardo, G., Sormani, M. P., Iannucci, G., Rodegher, M., Comi, G. and Filippi, M. 1998, Brain MRI lesion volume measurement reproducibility is not dependent on the disease burden in patients with multiple sclerosis, *Magn. Reson. Imag.*, **16**, 1185–1189.

Rovaris, M., Barkhof, F., Bastianello, S., Gasperini, C., Tubridy, N., Yousry, T. A., Sormani, M. P., Viti, B., Miller, D. H. and Filippi, M. 1999, Multiple sclerosis: interobserver agreement in reporting active lesions on serial brain MRI using conventional spin echo, fast spin echo, fast fluid-attenuated inversion recovery and post-contrast T_1-weighted images, *J. Neurol.*, **246**, 920–925.

Schneiders, N. J. 1988, Solutions of two paramagnetic ions for use in nuclear magnetic resonance phantoms, *Med. Phys.*, **15**, 12–16.

Simmons, A., Smail, M., Moore, E. and Williams, S. C. 1998, Serial precision of metabolite peak area ratios and water referenced metabolite peak areas in proton MR spectroscopy of the human brain, *Magn. Reson. Imag.*, **16**, 319–330.

Simmons, A., Moore, E. and Williams, S. C. 1999, Quality control for functional magnetic resonance imaging using automated data analysis and Shewhart charting, *Magn. Reson. Med.*, **41**, 1274–1278.

Soher, B. J., Hurd, R. E., Sailasuta, N. and Barker, P. B. 1996, Quantitation of automated single-voxel proton MRS using cerebral water as an internal reference, *Magn. Reson. Med.*, **36**, 335–339.

Streiner, D. L. and Norman, G. R. 1995, *Health Measurement Scales: a Practical Guide to their Development and Use.* Oxford University Press, Oxford.

Taylor, J. R. 1997, *An Introduction to Error Analysis: the Study of Uncertainties in Physical Measurements*, 2nd edn., University Science Books, Sausalito, CA, USA.

Tegeler, C., Strother, S. C., Anderson, J. R. and Kim, S. G. 1999, Reproducibility of BOLD-based functional MRI obtained at 4 T, *Hum. Brain Mapp.*, **7**, 267–283.

Tofts, P. S. 1994, Standing waves in uniform water phantoms, *J. Magn. Reson. Ser. B*, **104**, 143–147.

Tofts, P. S. 1996, Optimal detection of blood-brain barrier defects with Gd-DTPA MRI-the influences of delayed imaging and optimised repetition time, *Magn. Reson. Imag.*, **14**, 373–380.

Tofts, P. S. 1998, Standardisation and optimisation of magnetic resonance techniques for multicentre studies, *J. Neurol. Neurosurg. Psychiat.*, **64** (Suppl. 1), S37–S43.

Tofts, P. S., Kermode, A. G., MacManus, D. G. and Robinson, W. H. 1990, Nasal orientation device to control head movement during CT and MR studies, *J. Comput. Assist. Tomogr.*, **14**, 163–164.

Tofts, P. S., Wicks, D. A. and Barker, G. J. 1991a, The MRI measurement of NMR and physiological parameters in tissue to study disease process, *Prog. Clin. Biol. Res.*, **363**, 313–325.

Tofts, P. S., Wicks, D. A. G. and Barker, G. J. 1991b, The MRI measurement of NMR and physiological parameters in tissue to study disease process, in *Information Processing in Medical Imaging*, Ortendahl, D. A. and Llacer, J. (eds). Wiley-Liss, New York, 313–325.

Tofts, P. S., Shuter, B. and Pope, J. M. 1993, Ni-DTPA doped agarose gel – a phantom material for Gd-DTPA enhancement measurements, *Magn. Reson. Imag.*, **11**, 125–133.

Tofts, P. S., Berkowitz, B. and Schnall, M. D. 1995, Quantitative analysis of dynamic Gd-DTPA enhancement in breast tumors using a permeability model, *Magn. Reson. Med.*, **33**, 564–568.

Tofts, P. S., Barker, G. J., Dean, T. L., Gallagher, H., Gregory, A. P. and Clarke, R. N. 1997a, A low dielectric constant customized phantom design to measure RF coil nonuniformity, *Magn. Reson. Imag.*, **15**, 69–75.

Tofts, P. S., Barker, G. J., Filippi, M., Gawne-Cain, M. and Lai, M. 1997b, An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis, *Magn. Reson. Imag.*, **15**, 183–192.

Tofts, P. S., Lloyd, D., Clark, C. A., Barker, G. J., Parker, G. J., McConville, P., Baldock, C. and Pope, J. M. 2000, Test liquids for quantitative MRI measurements of self-diffusion coefficient *in vivo*, *Magn. Reson. Med.*, **43**, 368–374.

Vaidyanathan, M., Clarke, L. P., Hall, L. O., Heidtman, C., Velthuizen, R., Gosche, K., Phuphanich, S., Wagner, H., Greenberg, H. and Silbiger, M. L. 1997, Monitoring brain tumor response to therapy using MRI segmentation, *Magn. Reson. Imag.*, **15**, 323–334.

van den Bos, A. 1982, Parameter estimation, in *Handbook of Measurement Science*, Vol. 1, Sydenham, P. H. (ed.). Wiley, Chichester, 331–337.

Walker, P., Lerski, R. A., Mathur-De Vre, R., Binet, J. and Yane, F. 1988, Preparation of agarose gels as reference substances for NMR relaxation time measurement. EEC Concerted Action Program, *Magn. Reson. Imag.*, **6**, 215–222.

Walker, P. M., Balmer, C., Ablett, S. and Lerski, R. A. 1989, A test material for tissue characterisation and system calibration in MRI, *Phys. Med. Biol.*, **34**, 5–22.

Wei, X., Warfield, S. K., Zou, K. H., Wu, Y., Li, X., Guimond, A., Mugler, J. P., III, Benson, R. R., Wolfson, L., Weiner, H. L. and Guttmann, C. R. 2002, Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy, *J. Magn. Reson. Imag.*, **15**, 203–209.

Queries in Chapter 3:

Q1. In Figure 3.7 caption, Equations (3.9) and (3.11) – there are only 10 equations?