Check for updates

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–14
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280217754231
journals.sagepub.com/home/smm
SAGE

Reference-free error estimation for multiple measurement methods

Hennadii Madan, Franjo Pernuš and Žiga Špiclin

Abstract

We present a computational framework to select the most accurate and precise method of measurement of a certain quantity, when there is no access to the true value of the measurand. A typical use case is when several image analysis methods are applied to measure the value of a particular quantitative imaging biomarker from the same images. The accuracy of each measurement method is characterized by systematic error (bias), which is modeled as a polynomial in true values of measurand, and the precision as random error modeled with a Gaussian random variable. In contrast to previous works, the random errors are modeled jointly across all methods, thereby enabling the framework to analyze measurement methods based on similar principles, which may have correlated random errors. Furthermore, the posterior distribution of the error model parameters is estimated from samples obtained by Markov chain Monte-Carlo and analyzed to estimate the parameter values and the unknown true values of the measurand. The framework was validated on six synthetic and one clinical dataset containing measurements of total lesion load, a biomarker of neurodegenerative diseases, which was obtained with four automatic methods by analyzing brain magnetic resonance images. The estimates of bias and random error were in a good agreement with the corresponding least squares regression estimates against a reference.

Keywords

Quantitative imaging biomarker, Bayesian inference, linear regression, Markov chain Monte-Carlo, gold standard

1 Introduction

Objective, accurate, and reliable assessment of patient status is the foundation of medical research and is nowadays possible through several clinical and paraclinical tests that aim to quantify certain physical or functional characteristics of a patient. For certain medical conditions, there may be several tests available to measure the same characteristic, and the choice of the best test is often a complex compromise between tests' performance (i.e. sensitivity and specificity), availability, and associated costs. A compelling class of widely available and relatively inexpensive paraclinical tests emanates from the field of medical imaging, where computational image analysis methods enable *in vivo* extraction of quantitative biological characteristics of tissue structure or function.

Examples of quantitative measurements derived from medical images include lesion size and count, chemical tumor marker concentration, relative position of surgical targets, and nearby vulnerable anatomical structures and rate of change of the above quantities over time. Such data are already being used in diagnosis of a vast multitude of conditions including trauma, kidney stones and cysts, cancer, multiple sclerosis (MS), Parkinsons disease progression etc.[1–4] as well as in cancer staging[5,6] and treatment decision-making.[7] Quantitative measurements are also used for preoperative planning, intraoperative guidance, and postoperative assessment in image-guided procedures such as surgery, endoscopy, radiation therapy, and biopsy.[8,9] Recently a class of scalar measurements called quantitative imaging biomarker (QIB) defined as "an objective characteristic derived from an in vivo image measured on a ratio or interval scale as an indicator of normal biological processes, pathogenic processes or

Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

Corresponding author:
Hennadii Madan, Faculty for Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia.
Email: hennadii.madan@fe.uni-lj.si

a response to a therapeutic intervention''[10] has gained attention for its potential for application as surrogate endpoints in clinical trials.[11–14] However, evaluating the performance and choosing the best image analysis method to extract a certain QIB is a difficult task.[15]

Characteristics of image analysis methods generally contribute to the overall cost of measurement. There is a cost associated with *acquisition* of images. For example, tumor volume can be measured from different modalities: computed tomography is generally cheaper than magnetic resonance (MR), which is in turn cheaper than positron emission tomography (PET). There are costs associated with the *act of measurement*, i.e. with conversion from image intensities to numerical values that represent the measured physical quantity. For instance, employing an experienced radiologist to manually outline a tumor in order to measure its size is more expensive than using automatic software.

In medical practice, the *errors* inherent in the measurement method induce further costs for the involved parties—patients, clinical centers, and society at large. For example, a wrong diagnosis and/or treatment decision based on an erroneous measurement may cost the patient a great deal of time and money and may lead to irreversible health deterioration or even death. For a hospital, such a situation is at least a waste of staffs time and medical supplies that are erroneously prescribed. For the economy, the consequence is a certain loss of productivity and resources depending on frequency and severity of errors.

As a rule of thumb, methods with smaller errors tend to have higher acquisition and measurement costs. Therefore, the choice of a measurement method has to be based on the balance between the aforementioned costs in a given application. Accurate prediction and estimation of the measurement errors is, therefore, of utmost importance.

1.1 Reference-based error estimation

In medical imaging, the usual way to estimate errors of a given measurement method or methods is based on the concept of a reference method, often called gold standard. Out of many methods to measure the same quantity, one that is considered to produce reasonably small errors is chosen as a substitute for the true values of the measurand. The errors are then estimated from the differences in measurements obtained with a test and the reference method, both applied to the same sample of patient population. With these data, a comparison of several measurement methods is enabled through statistical tests for superiority, equivalence, or noninferiority.[16,17] To be useful, reference-based error estimation has to be performed with large samples, representative of the underlying population. For in vivo images, however, measurements with a reference method are generally prohibitively costly for large-scale application.

Another, often overlooked, fact is that reference-based error estimation fails inconspicuously when the reference is not accurate. As was demonstrated by simulation studies for a number of measurement error estimators, the reported estimates are generally biased and overconfident at the same time when the reference has low accuracy.[16] This means that decisions based on error estimation with low accuracy reference are likely to be erroneous. Therefore, special care has to be taken to ensure accuracy of the reference and in interpretation of the results of reference-based error estimation.

In medical imaging, the reference is often not accurate indeed. For example, for measurements based on segmentation, on which the majority of QIBs are based, the reference is produced from the outlines of structures of interest made by human experts. It is, however, a well-established fact that these outlines will not be identical when made by two different experts (inter-rater variability) nor even when the same expert creates two different outlines (intra-rater variability).[18,19] The extent of this variability depends on the properties of the imaging process, the object imaged, and the experience of the experts. Even excluding the subjective effects, for objects with high surface to volume ratio (like focal brain lesions), the combination of partial volume effects and low resolution of clinical scanning protocols may lead to high relative errors in physical size (e.g. volume and linear dimensions) estimation.[20]

To deal with the problem of unreliable reference, error-in-variable models have been proposed.[21–24] They work by explicitly modeling the variance of the values of the reference method around the unobserved true value of the measurand. Error-in-variable models have been shown to produce more accurate uncertainty estimates in synthetic tests with imperfect reference measurements.[16] This result comes at a price—reduced amount of information compared to the known-truth reference has to be compensated by increased sample size in order to achieve a given level of confidence. More importantly, error-in-variable models deal only with costs of errors, but do not address the measurement and/or acquisition costs of the reference method.

1.2 Reference-free error estimation

The emerging field of reference-free error estimation aims to reduce the costs of acquisition, measurement, and the errors associated with the reference method. Instead of relying on some dedicated measurement method as representative of truth, the statistical properties of an ensemble of several different measurement methods are exploited.[25,26]

Regression without truth (RWT) proposed in 2002 by Kupinski and Hoppin[27] represents a framework for reference-free error estimation for several measurement methods of a bounded continuous physical quantity. In RWT, the measurement error of each method is usually modeled as a sum of a linear systematic error and Gaussian random error. By assuming a prior distribution on the true value of the measurand, it is possible to marginalize out the unknowns and calculate the point of maximum marginal likelihood using quasi-Newton optimization. The RWT framework was applied for error estimation in methods for measuring cardiac ejection fraction,[28,29] volume biological activity in SPECT images,[30,31] and apparent diffusion coefficient in diffusion-weighted MR images.[32]

A number of issues can be identified when applying RWT in practice. First, random errors of a measurement method are assumed to be independent of those of other methods. However, it is very often of interest to compare measurement methods that are based on similar principles, but differ in details. Random errors of such variant methods cannot be considered statistically independent. Second, it is important to initialize the iterative optimizer close to the unknown true values of the error model parameters or risk convergence to a non-global maximum of marginal likelihood. Third, as a consequence of the use of an iterative optimizer, only point estimates are returned without uncertainty quantification. To some extent, this can be remedied by bootstrap application.[33] Besides, to the best of our knowledge, there are no reports of RWT validation against traditional least squares (LS) regression with a reference method on a clinical dataset.

A drawback expected of any reference-free error estimation, compared to the reference-based, is that methods have to be applied to larger samples to achieve a given degree of certainty. Nevertheless, the attractiveness of reference-free estimation is based on the premise that eliminating the costs associated with the reference will outweigh the costs of acquiring a larger sample.

1.3 Contributions

In this work, we continue the line of thought underlying RWT and aim to rectify the aforementioned deficiencies. First, possible lack of statistical independence between the random error of different measurement methods is modeled explicitly. Second, instead of seeking point estimates based on quasi-Newton optimization of marginal likelihood we employ sampling of the full posterior distribution using Markov chain Monte-Carlo (MCMC). This enables detailed characterization of modes of the distribution, uncertainty estimates, computation of various sample statistics, and statistical tests. Third, a set of validation experiment is performed on total lesion load (TLL) data calculated by four automated methods from clinical magnetic resonance images (MRIs) of 22 MS patients, for which a gold standard reference was available, as well as on six synthetic datasets, each exhibiting a varying degree of random error correlation between the methods in ensemble.

2 Framework description

Consider a dataset of images of N patients and M different measurement methods for a certain QIB. Note that QIB is only a practical example of a measurand to which the framework can be applied. In the rest of the paper, we use terms QIB and measurand interchangeably. Let x_{pm} denote the value measured with method m for patient p and let x_{pt} denote the corresponding true value, which is unknown. Given a table of all measurements $X = [x_{pm}] \in \mathbb{R}^{N \times M}$, the question we want to answer is: "Which method is the most accurate or precise?" The answer is obtained by estimating systematic and random errors of each method.

2.1 Error model

The error model relates the measured x_{pm} to the unknown true value of the measurand x_{pt}. We consider error models of the form

$$x_{pm} = \sum_{k=0}^{K} b_{km} x_{pt}^{k} + \epsilon_{pm} \tag{1}$$

where the polynomial represents the systematic error (bias) and ϵ_{pm} the random error (noise). Measurement methods, albeit different, are often based on similar principles, thus the corresponding random errors ϵ_{pm} are generally not statistically independent. We model this explicitly by a multivariate Gaussian (MVG) distribution

$$\boldsymbol{\epsilon_p} \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{2}$$

where $\boldsymbol{\epsilon_p} = [\epsilon_{p1}, \epsilon_{p2}, \ldots, \epsilon_{pm}, \ldots, \epsilon_{pM}]$ and Σ is an $M \times M$ covariance matrix.

2.2 Posterior probability

Let L_p denote the likelihood of observing the measurements for a patient p given the true value of the measurand and error model parameters. By expressing ϵ_{pm} from equation (1) and using equation (2) we obtain

$$L_p \triangleq P\left(\boldsymbol{x_p} | B, \Sigma, x_{pt}\right) = \mathcal{N}\left(\boldsymbol{\epsilon_p}, \Sigma\right) \tag{3}$$

where $\boldsymbol{x_p} = [x_{p1}, x_{p2}, \ldots, x_{pm}, \ldots, x_{pM}]$ and $B = [b_{km}] \in \mathbb{R}^{K \times M}$. Since the true values x_{pt} across different patients p can be considered statistically independent, the likelihood of observing the entire table of measurements X is given by

$$L \triangleq P(X | \theta) = \prod_{p=1}^{N} L_p \tag{4}$$

where $\theta = (B, \Sigma, \boldsymbol{x_t})$ is the set of all parameters, including the vector of true values $\boldsymbol{x_t} = \left[x_{1t}, x_{2t}, \ldots, x_{pt}, \ldots, x_{Nt}\right]$ of the measurand.

By Bayes's theorem, the posterior probability of θ and $\boldsymbol{x_t}$ given the measurements X is

$$P(\theta \mid X) = \frac{L \cdot P(\theta)}{P(X)} \tag{5}$$

where $P(\theta)$ is prior probability of parameters, while $P(X)$ is evidence probability, which is a fixed normalization constant for any observed dataset.

We use MCMC to sample from unnormalized posterior distribution $P(\theta \mid X) \propto L \cdot P(\theta)$. The sample is then analyzed to arrive at the estimates of the error model parameters and their uncertainties. Since MCMC is a well-established method, we omit the theory behind its workings and refer an interested reader to a simple and short introduction.[34]

2.3 Prior specification

To use MCMC, it is necessary to specify the prior distribution $P(\theta)$. The dependence between components of θ is defined by the model equation (1) and therefore is encoded in the likelihood function L. When sufficient amount of data is available, the priors on individual components of θ may be specified separately

$$P(\theta) = P(B) \cdot P(\Sigma) \cdot P(\boldsymbol{x_t}) \tag{6}$$

Regarding the systematic error coefficients B, it is reasonable to assume for all m that b_{0m} and b_{1m} are likely close to zero and one, respectively, while all $b_{km}, k > 1$ are close to zero. Note that although correlations between $b_{0m}, b_{1m}, \cdots, b_{Km}$ are expected regardless of the observed data, specifying these in the prior is superfluous: this information is ingrained in the model and, therefore, is already encoded in the likelihood. We have found experimentally that $N = K + 1$, i.e. the absolute minimum of patients to consider K-th degree polynomial for bias, is enough to observe these correlations in the posterior. Therefore, $P(B)$ can be specified as a product of univariate distributions $P(B) = \prod_m \prod_k P(b_{km})$, where each $P(b_{km})$ attains a maximum at values $0, 1, 0, \ldots$ for $k = 0, 1, 2, \ldots$

Given the following decomposition:

$$\Sigma = SRS \tag{7}$$

where $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_M)$ is a diagonal matrix of standard deviations and R is a symmetric correlation matrix, R can be assigned Lewandowski–Kurowicka–Joe prior[35] with $\eta = 1$, providing uniform distribution of R_{ij}, while standard deviations can be assigned truncated Jeffreys priors

$$\sigma_m \sim \begin{cases} \frac{1}{\sigma_m}, & \sigma_{min(m)} < \sigma_m < \sigma_{max(m)} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Truncation guarantees that the posterior is proper, and boundaries can be assigned from physical considerations, e.g $\sigma_{min(m)}$ may be set to measurement resolution, while $\sigma_{max(m)}$ is limited by the span of measurements.

Upper and lower bounds dictated by the nature of the measurement or by physiological constraints can be established for any QIB. Given only this knowledge, for each patient, true QIB values can be assigned a uniform prior $x_{pt} \sim P(x_t) \stackrel{\Delta}{=} \mathcal{U}(x_{min}, x_{max})$. Then, since QIB values of one patient do not depend on those of other patients, we write

$$P(\boldsymbol{x_t}) = \prod_{p=1}^{N} P(x_t) = P(x_t)^N \tag{9}$$

2.4 Parameter estimation

The expected values of error model parameters can be estimated from the posterior distribution sample obtained by MCMC. If the posterior is unimodal or has a dominant mode, the expected values of the parameters are approximated by the expected value of the sample. If the posterior has several well-separated modes with comparable probability, it means that several distinct mechanisms, i.e. several distinct sets of parameters, explain the data. In this case, the sample will consist of several clusters—one per mode. The expected values of parameters for each mechanism are approximated by the expected value of the corresponding cluster (and not the expected value of the entire sample). In Bayesian model selection the ratio of probabilities of each mechanism is equal to the ratio of mode masses. The latter can be approximated by the ratio of the number of sample points belonging to each cluster.

With the error model parameter estimates at hand, the original question can be answered: the measurement methods can be ranked according to their precision, i.e. σ_m. Alternatively, methods can be ranked according to accuracy, e.g. using Chebyshev norm of the estimated bias as a metric

$$C_{\Omega m} = \max_{x \in \Omega} \left| \sum_k b_{km} x^k - x \right| \tag{10}$$

where Ω is the interval of measurand values that is of practical interest.

3 Validation

The proposed framework was validated on six synthetic and one clinical dataset of TLL measurements, obtained from brain MRI by four different automated methods. For all datasets reference TLL values were given and we evaluated the framework's capability to estimate the error model parameters in comparison to LS regression with respect to the reference.

3.1 Datasets

The clinical dataset was based on the analysis of MR images of 22 patients diagnosed with MS (41.3 ± 10.5 years old, 13 females) obtained from the University Medical Centre Ljubljana (UMCL). All patients signed a written

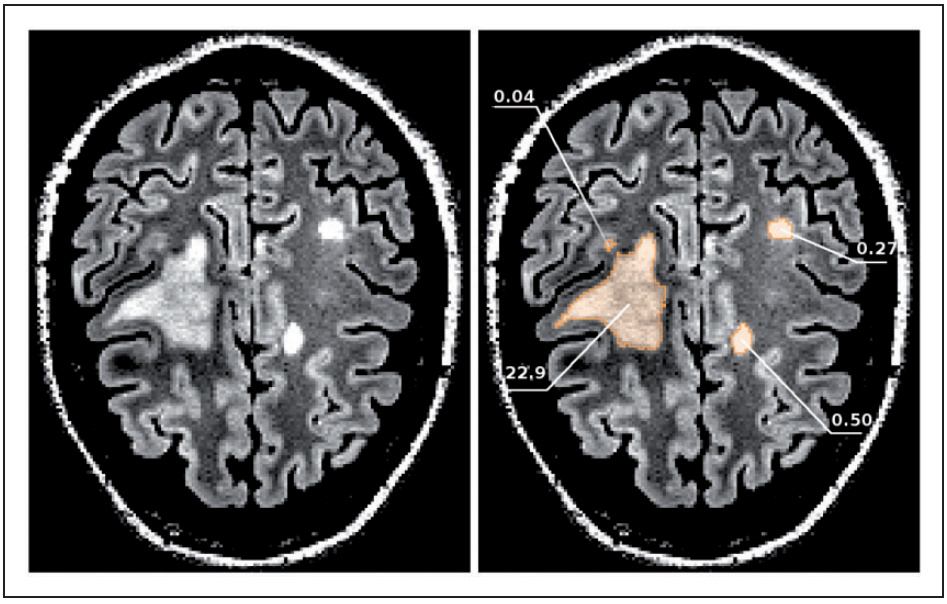

Figure 1. A slice through FLAIR MRI image of a patient with multiple sclerosis presenting characteristic hyperintense lesions (left). Reference lesion segmentation of the given slice with volumes of corresponding lesions indicated in cm^3 (right).

informed consent at the time of enrollment for imaging, and for this study, the UMCL approved the use of these data, which were analyzed anonymously.

Each patient's images were acquired on a 3T Siemens Magnetom Trio MR system at the UMCL using conventional sequences such as 2D T1- and T2-weighted and 3D FLAIR, from which white matter lesions were segmented by four automatic algorithms. Three of the algorithms were unsupervised and detected lesions as abnormal T1-weighted and FLAIR intensities[36–38] as compared to major brain tissues, while the fourth algorithm was a supervised random forest classifier.[39] Two of the unsupervised algorithms were similar, one[38] being an incremental upgrade of the other.[37] Reference lesion segmentations (Figure 1) were created by three neuroradiologists using semi-automated image analysis tools.[40] The three segmentations were merged and jointly revised by the neuroradiologists to obtain a consensus segmentation, which was then used as a reference.[20] To specify a TLL value from a lesion segmentation, the count of voxels labeled as lesions was multiplied by the volume of a voxel in an image.

In order to obtain the estimates of the "true" parameter values of the error model in equation (1), an LS regression of automatic versus manual reference TLL measurements was performed to determine the coefficients $b_{km} \in B$ of quadratic polynomial ($K=2$) and random error terms $\sigma_m \in S$ for each method ($M=4$). Correlations of the residuals estimated the elements of the correlation matrix R. The fit curves are shown in Figure 2, while the estimated parameter values are summarized in the first section of Table 1.

Synthetic TLL datasets were created such that they resembled the clinical dataset; however, a controlled amount of random error correlation between one pair of measurement methods was simulated. For each synthetic dataset, $N=22$ points were drawn from $\mathcal{U}(0, 55)$, where upper bound of $55\,cm^3$ was chosen to include the maximal reference TLL value with a certain margin ($2.3\ cm^3$). Measurements with $M=4$ methods were synthesized by applying equation (1) with bias parameters b_{km} obtained by LS regression from clinical data (Table 1). Random errors were drawn from an MVG with covariance matrix structure as defined in equation (7), where the diagonal elements in matrix S were set to σ_m as reported in Table 1 and correlation matrix R had the following form

$$R = \begin{pmatrix} 1 & R_{12} & 0 & 0 \\ R_{12} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{11}$$

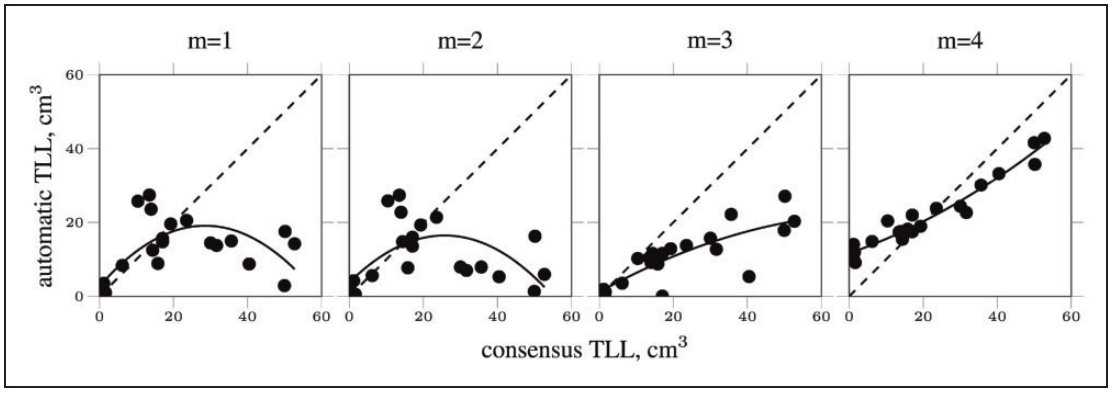

Figure 2. Lines of least-squares fit of quadratic polynomial model to TLL measurements extracted from MRI images by four automatic methods. Corresponding TLL measurements computed from expert consensus segmentations were used as a reference. The obtained estimates of model parameter values are given in Table 1.

Table 1. Error model estimates and Chebyshev norm of the bias obtained with LS regression using a gold standard, with MCMC without a gold standard, in the correct mode of the "proposed" experiment, and in the "control" experiment, and in the incorrect mode of the "proposed" experiment.

m	b_{0m} cm^3	b_{1m} 1	b_{2m} $10^{-3}\,cm^{-3}$	σ^m cm^3	$C_{[0,55]m}$ cm^3	b_{0m} cm^3	b_{1m} 1	b_{2m} $10^{-3}\,cm^{-3}$	σ^m cm^3	$C_{[0,55]m}$ cm^3
	LS (reference)					MCMC correct mode				
1	3.7	0.99	−19.2	7.0	54.9	−0.3	1.07	−18.8	6.5	53.0
2	2.7	1.15	−20.1	6.2	49.9	−1.6	1.22	−19.5	5.7	48.3
3	1.2	0.55	−3.6	4.3	34.4	−2.1	0.69	−5.1	4.9	35.1
4	11.6	0.35	4.0	2.2	12.1	9.3	0.38	4.4	2.8	11.2
	Control experiment					MCMC incorrect mode				
1	−0.3	0.25	5.8	2.4	23.6	−0.9	0.28	5.4	2.2	24.1
2	−1.8	0.56	−0.1	1.3	26.4	−2.5	0.63	−1.4	2.0	27.1
3	−1.7	0.87	−12.5	6.7	46.7	−2.2	0.86	−12.5	6.4	45.7
4	10.6	1.01	−16.9	9.0	40.3	9.8	0.99	−16.2	8.9	39.5

MCMC: Markov chain Monte-Carlo; LS: least squares.

Six different datasets were created, each with a different value of $R_{12} \in R$ set as $\{0.0, 0.5, 0.6, 0.7, 0.8, 0.9\}$ in order to simulate situations from zero to progressively higher degrees of random error correlation.

3.2 Experiments

Two sets of experiments were performed on all datasets, so that two sets of estimates were obtained. One experiment involved the proposed joint modeling of random errors, while the other assumed independence of random errors, equivalent to constraining R to the identity matrix (cf. equation (7)). In the following, we will refer to the respective experiments as "proposed" and "control." The assumed prior distributions of parameters were as follows: $b_{0m} \sim \mathcal{N}(0, 55/3)$ (cm^3); $b_{1m} \sim \mathcal{N}(1, 0.5)$; $b_{2m} \sim \mathcal{N}(0, 1/55)$ (cm^{-3}); $\sigma_m \sim 1/\sigma_m, 0.001 < \sigma_m < 55$ (cm^3); $R_{ij} \sim \mathcal{U}(-1, 1)$ (only in "proposed" experiment); $x_{pt} \sim \mathcal{U}(0, 55)$ (cm^3). For evaluation purposes, the reference-free estimates obtained with MCMC were compared to the LS estimates based on a reference.

An implementation of No U-Turn Sampler(NUTS)[41] from Python package pymc3 was used in the experiments. A draw from an approximation provided by automatic differential variational inference algorithm[42] was used to initialize six parallel chains. For each chain, 19,000 samples were collected, first 9000 samples were discarded as

burn-in, while last 10,000 samples were used for further analysis. This resulted in Gelman–Rubin potential scale reduction factor $\hat{R} < 1.04$ for all variables.

To quantify the data prediction ability of obtained error model estimates, we use the following procedure. For each method, the estimates and the reference values of measurand are plugged into the model equation (1) assuming $\epsilon_{pm} = 0$ to obtain predicted measurements $\tilde{x}_{pm}$. These are used to calculate the coefficient of determination R^2 for each method

$$R_m^2 = 1 - \frac{\sum_{p=1}^{N} \left(\tilde{x}_{pm} - x_{pm}\right)^2}{\sum_{p=1}^{N} \left(x_{pm} - \bar{x}_{pm}\right)^2} \tag{12}$$

These are normalized by the coefficient of determination obtained with the reference bias polynomial coefficients (LS estimates for the clinical and values used for data generation for the synthetic experiments) and averaged over all methods to arrive at a single scalar measure of performance

$$q = \frac{1}{M} \sum_{m=1}^{M} \frac{R_m^2}{R_m^{2ref}} \tag{13}$$

The larger the value the better the prediction. This summary statistic is only dependent on bias coefficient estimates. To quantify the ability to predict random error dispersion, we use the coefficient of determination of σ_m estimates with respect to reference standard deviation

$$R_\sigma^2 = 1 - \frac{\sum \left(\sigma_m - \sigma_m^{ref}\right)^2}{\sum \left(\sigma_m^{ref} - \overline{\sigma_m^{ref}}\right)^2} \tag{14}$$

3.3 Results

In Figure 3, both the ''control'' and ''proposed'' estimates of b_{km} and σ_m along with 90% credible intervals from the experiments on the six synthetic datasets are plotted against the values used to generate the data. It is evident that the true values are always within the 90% credible region for ''proposed'' experiments, while the ''control'' estimates become incorrect for $R_{12} > 0.7$.

For the clinical dataset, the posterior distribution obtained in the ''proposed'' experiment contained two well separated modes (cf. example b_{21} histogram from method 1 in Figure 4), indicating two possible mechanisms that could have produced the data. Based on visual assessment of b_{21} histogram, the modes were separated at $b_{21} = 5 \times 10^{-3}$ cm^{-3} and, since the mass was slightly higher for the left mode (i.e. 51% versus 49% of the sample), it was designated as *correct* while the right mode was designated as *incorrect*. In the ''control'' experiment, the posterior always contained only one mode.

The predictive curves based on a sample from the posterior distribution in the ''proposed'' experiment are shown in Figure 5. Apparently, the estimates based on the correct mode adequately describe the data and the width of the posterior distribution seems representative of the actual uncertainty. This is not the case for the incorrect mode.

Figure 6 and Table 1 give comparisons of the error model estimates obtained by the proposed framework to those obtained by the LS regression based on the reference. Parameters b_{1m}, b_{2m}, and σ_m are in good agreement, while b_{0m} is slightly offset. This is expected as a result of multicollinearity of the polynomial bias model: small error in b_{1m} and b_{2m} estimates has a large impact on b_{0m}, but the model's overall data prediction ability is not affected.

As mentioned earlier, the posterior sample can be used to estimate the true values of the TLL, as shown in Figure 7 for the clinical dataset. Here, the nature of the incorrect mode is especially apparent—the estimates of the true value of TLL are virtually identical those from the ''control'' experiment, both being close to TLL values measured by the related methods $m = 1$ and $m = 2$. On the other hand, the correct mode yields true TLL estimates remarkably close to the reference.

The methods may be ranked according to the estimated precision σ_m or according to estimated accuracy $C_{[0,55]m}$ shown in Table 1. According to the MCMC estimates, the best precision and accuracy were achieved by method $m = 4$, which was the supervised method based on random forest classification.[39] This result is in agreement with the LS estimates based on reference. High precision and accuracy of method $m = 4$ is also apparent in Figure 2.

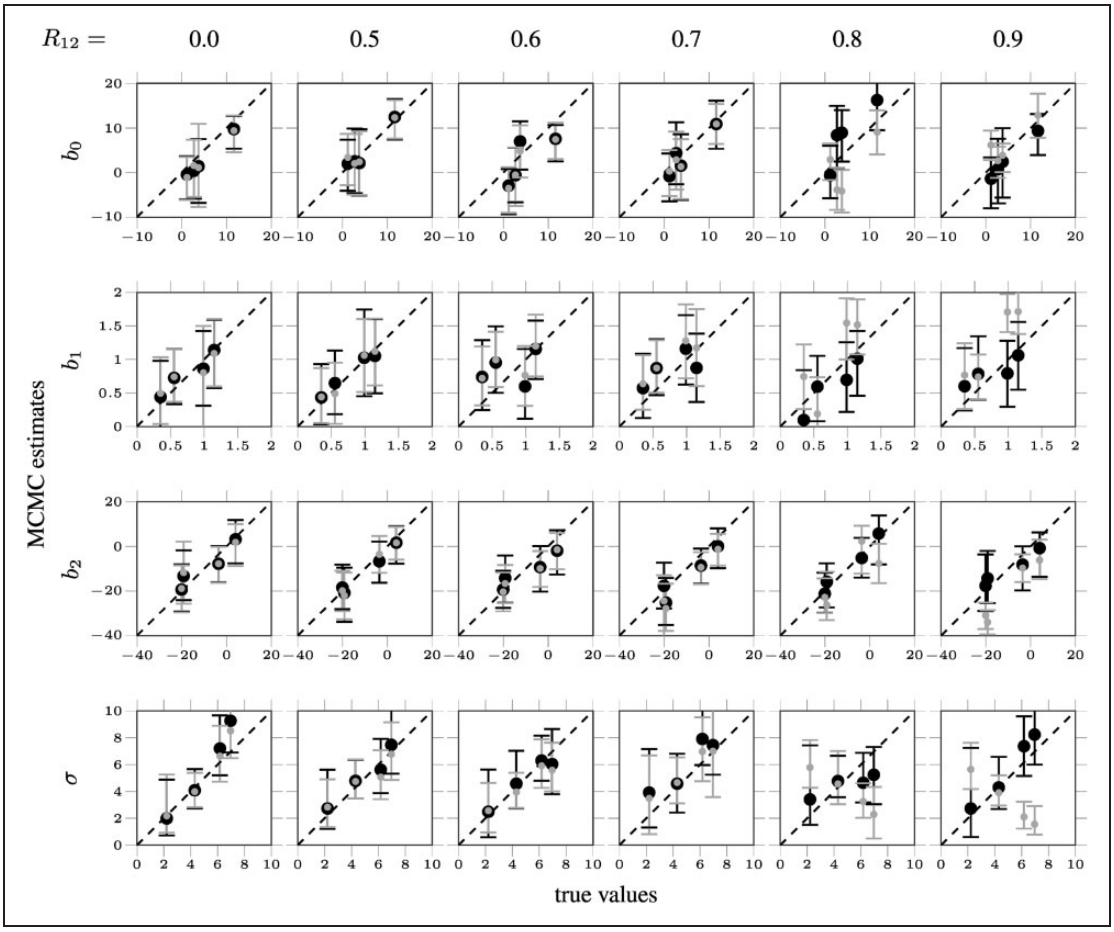

Figure 3. Experiments on synthetic datasets: parameter estimates with "proposed" (*large black circles*) and "control" (*small gray circles*) models versus the values used to generate the data. The error bars stretch between 5th and 95th sample percentiles. R_{12} indicates the correlation coefficient between random errors of methods $m = 1$ and $m = 2$ that was used as a parameter when generating the datasets.

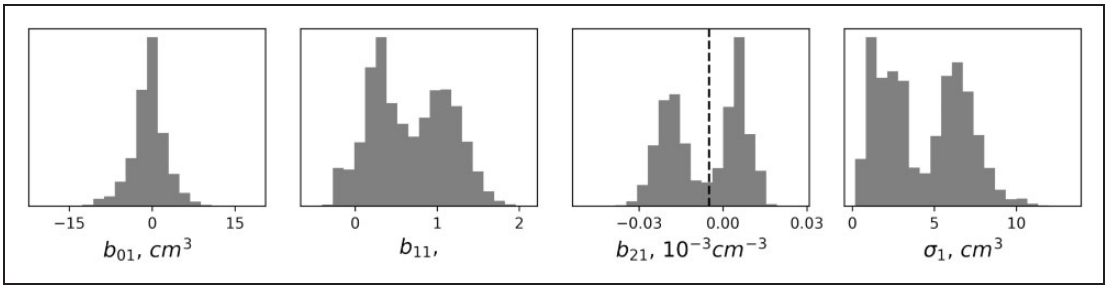

Figure 4. Experiments on the clinical dataset: posterior sample histograms for error model parameters of the measurement method $m = 1$. *Dashed line* indicates the value used to split the sample.

Note that a reference is required to obtain both the LS estimates and Figure 2, whereas the MCMC estimates were obtained without the reference.

From Figures 6 and 7, it is evident that the "proposed" experiment yields two modes, where one corresponds to the correct estimates, while the other is close to the "control" experiment estimates. The correct mode had a higher maximal value of the posterior probability and higher mass hinting that it represents the mechanism underlying the data. Using additional knowledge that two of the methods are related, it is possible to select the mode that

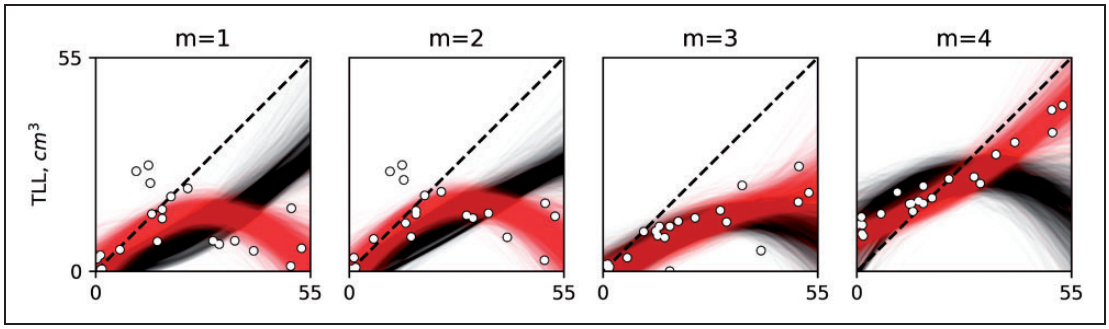

Figure 5. Experiments on the clinical dataset: model predictions of individual points from the posterior sample in the "proposed" experiment for each measurement method. *Red and black curves* correspond to the correct and incorrect mode, respectively.

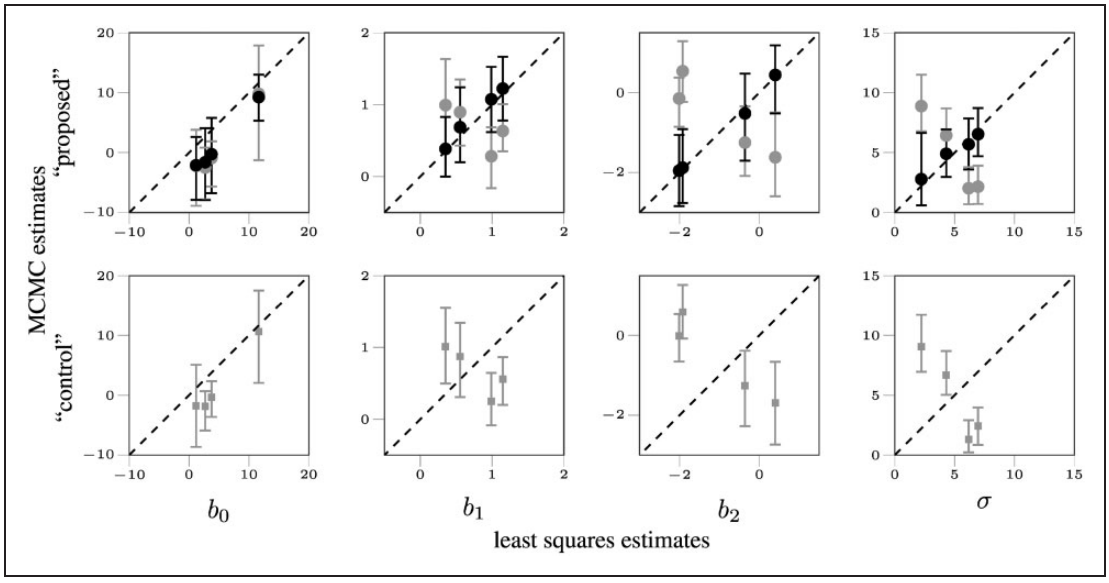

Figure 6. Experiments on the clinical dataset: parameter estimates with "proposed" (*top row*) and "control" (*bottom row*) model versus their estimates with least-squares regression against manual reference. *Black and gray markers* correspond to the correct and incorrect modes, respectively. The error bars stretch between 5th and 95th sample percentiles.

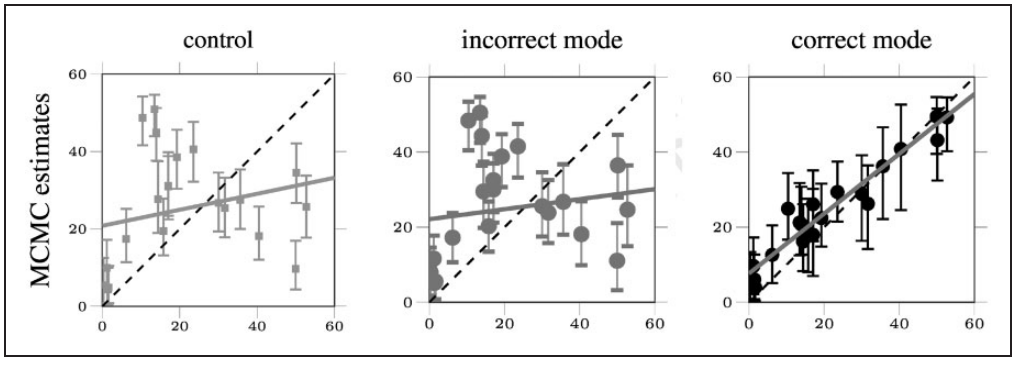

Figure 7. Experiments on the clinical dataset: estimated true values x_{pt} of the measurand, where the *gray line* indicates corresponding linear trendline and *dashed line* the reference TLL.
MCMC: Markov chain Monte-Carlo.

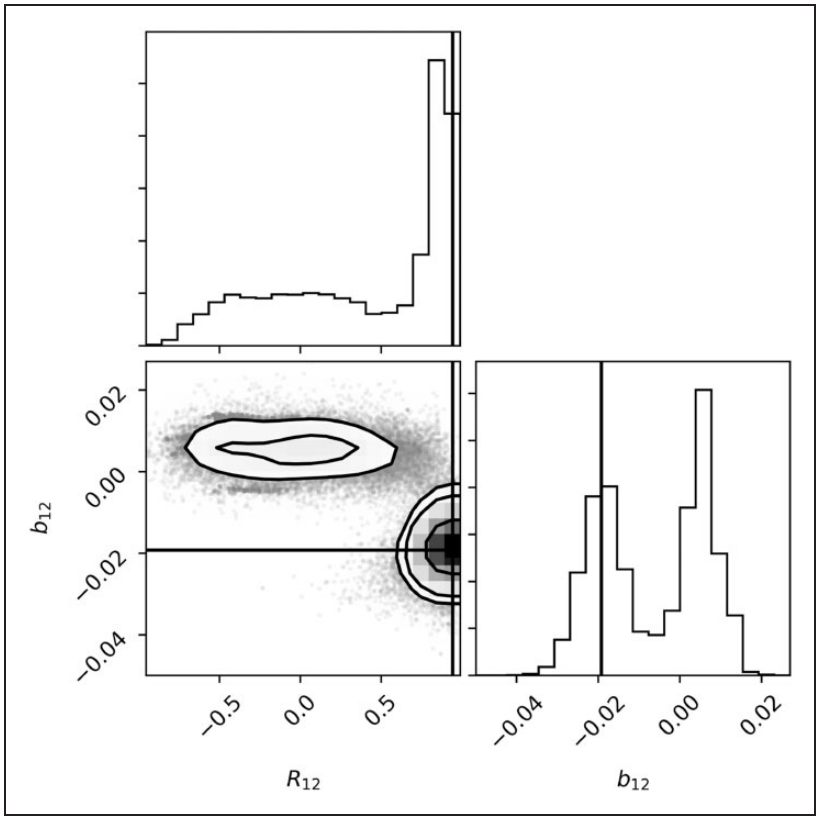

Figure 8. Experiments on the clinical dataset: posterior sample corner plot for correlation coefficient R_{12} between methods $m = 1$ and $m = 2$ and quadratic bias coefficient b_{12} of method $m = 1$. The "correct" mode corresponds to high correlation between random errors of methods $m = 1$ and $m = 2$, while the "incorrect" mode corresponds to insignificant amount of correlation. *Straight black lines* indicate the reference least squares estimates based on reference.

Table 2. Experiments summary.

	q		R^2_σ	
R_{12}	"Proposed"	"Control"	"Proposed"	"Control"
0.00	1.00	0.99	0.51	0.48
0.50	1.10	1.11	0.92	0.87
0.60	1.04	1.04	0.92	0.84
0.70	1.06	0.98	0.54	0.82
0.80	1.23	0.26	0.48	−2.21
0.90	0.85	0.29	0.75	−3.29
Clinical	0.82	−1.31	0.89	−5.77

Note: For synthetic datasets the reference is the values used for data generation for the clinical dataset the reference is the least squares estimates. q: Mean normalized coefficient of determination; R^2_σ: coefficient of determination of σ_m.

corresponds to the mechanism behind the data by taking into account the posterior distribution of correlation coefficient R_{12} between random errors of methods $m = 1$ and $m = 2$ (cf. Figure 8).

The values of q and R^2_σ obtained on the synthetic and clinical datasets are summarized in Table 2. Apparently for correlation coefficients up to 0.7, the ability to predict the data and estimate the variance in both "proposed" and "control" experiments is comparable. When significant correlation between random errors of two or more methods is present, the "control" model starts to yield incorrect estimates.

4 Discussion

A framework for performance comparison and ranking of multiple measurement methods in the absence of a reference was presented. The framework estimates error model parameters along with corresponding uncertainty of each parameter. A unique feature of the framework is that it is applicable even in situations when using both related and unrelated measurement methods, which was achieved by modeling correlations between random errors of the methods. The framework was validated on six synthetic and one clinical MS datasets and produced error model parameter estimates in good agreement with the truth and the reference-based estimates respectively.

The framework is based on full Bayesian inference and estimates the posterior probability density of model parameters. This density represents all the knowledge about the model parameters that can be extracted from the data, the model and the prior densities. Depending on the data, the inference might be ambiguous—multiple mechanisms, i.e. multiple distinct sets of parameter values, would explain the data. This manifests itself in posterior being multimodal, as was the case for our clinical dataset. If we want to resolve this ambiguity, further information (beyond the data, the model, and the priors) is necessary. In our case, this information was the fact that two of the measurement methods are related and therefore are likely to have correlated random errors. Generally speaking, possible reasons for this kind of ambiguity include small datasets, a mismatch between the model and the data-generating process, distinct subpopulations within the data, etc. Investigation into specific reasons for inference ambiguity in our clinical dataset goes beyond the scope of this article.

In our framework, MVG distribution of random errors is assumed. This must be justified by the physics of the measurement—it should either grant normal distribution of errors or a distribution for which normal approximation holds. Our TLL data are an example of the latter: in an approximation of constant per-voxel true positive and true negative rates of each segmentation algorithm,[18] the measurements are distributed binomially, however, due to high voxel counts normal approximation holds. Note that with LS regression against known reference deviations from this assumption can be diagnosed by residual analysis. Without a reference, we see little possibility for such post hoc analysis, so one has to check the assumptions prior to application.

Another assumption is encoded in the choice of priors for bias coefficients. The priors on b_{km} express the belief that the measurements do not deviate too much from the true value of the measurand at least around zero. Generally this should be true for a genuine measurement method. If ''measurement'' methods significantly deviate from these assumptions, the estimates will be biased.

With the error model parameter estimates at hand the methods can be ranked according to some figure of merit. Early works on RWT[27] used a linear bias model and $\frac{\sigma_m}{b_{1m}}$ to rank the methods. Such figure of merit can only be meaningfully interpreted in the context of a linear bias model, but not for bias models based on higher degree polynomials. Later works[33] used

$$F_m = E\left(\left(x_t - \sum_k^K b_{km} x_t^k - \epsilon_m\right)^2\right) \tag{15}$$

which was calculated analytically. This is again a special situation, since $P(x_t)$ assumed to be beta-distributed and bias model linear. Both of the two figures of merit take into account both systematic and random errors. We prefer to treat these errors separately, since it is possible to have a highly precise measurement method with a large bias, which is easy to compensate. For example, in context of TLL a segmentation algorithm may consistently, but incorrectly label the flow-induced artifacts within the ventricles as lesions, thereby generating a large constant bias. Such consistent incorrect labeling is generally easy to detect and to remove using simple segmentation post-processing techniques. Hence, we suggest that the choice of the best measurement method should be based primarily on σ_m, whereas one should verify that the bias is monotonous within the range of expected measurand values and the resolution of the method (i.e. $d[\sum_k^K b_{km} x_t^k]/dx_t$) is sufficiently high compared to σ_m.

To the best of our knowledge, this work is the first to successfully validate a reference-free error estimation against LS regression estimates on a clinical dataset with a gold standard reference. Although the proposed framework was inspired by the RWT[27] technique, important novel methodology was introduced, which seems to have contributed to the success of validation. Compared to the RWT, our improved error model captures correlated random errors, uses joint posterior probability criterion that, besides the error model parameters, enables the estimation of measurand values, and employs MCMC that technically enables discovery and characterization of the multiple modes of the posterior that arise when random errors of some of the methods are sufficiently correlated. Lack of modeling of these correlations is likely to lead to results not consistent with the

estimates based on a reference. This might be a possible reason that there are no previous reports on RWT validation on clinical in vivo datasets.

Besides the ability to rank the measurement methods, the analysis of the joint posterior provided by the MCMC allows to estimate the unknown true values (Figure 7). This opens an avenue for a clinical application in which several methods are employed to extract a certain QIB value measurements that are further processed with the proposed framework to compute the estimates of true QIB value. Such estimates may be more representative of the true value than the measurement values obtained by any individual method and can be better interpreted since the framework also estimates their uncertainty in the form of a credible region.

Application of the technique to larger clinical datasets is an intriguing opportunity to obtain the estimates with narrow credible regions and apply more complex error models.

TLL measurements with all measurement methods and with the gold standard, all synthetic datasets and Python code are available on Github https://github.com/madanh/smmr_code.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Slovenian Research Agency under grants J7-6781, J2-7211, J2-7118, and P2-0232.

References

1. Giorgio A and De Stefano N. Clinical use of brain volumetry. *J Magn Reson Imaging* 2012; **37**: 1–14.
2. Popescu V, Agosta F, Hulst HE, et al. Brain atrophy and lesion load predict long term disability in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2013; **84**: 1082–1091.
3. Marek K, Innis R, van Dyck C, et al. [^{123}I]β-CIT SPECT imaging assessment of the rate of Parkinson's disease progression. *Neurology* 2001; **57**: 2089–2094.
4. Bigler ED. Quantitative magnetic resonance imaging in traumatic brain injury. *J Head Trauma Rehabil* 2001; **16**: 117–134.
5. Huang EP, Wang XF, Choudhury KR, et al. Meta-analysis of the technical performance of an imaging procedure: guidelines and statistical methodology. *Stat Methods Med Res* 2015; **24**: 141–174.
6. Freudenberg LS, Antoch G, Schütt P, et al. FDG-PET/CT in re-staging of patients with lymphoma. *Eur J Nucl Med Mol Imaging* 2004; **31**: 325–329.
7. Rischin D. Prognostic significance of [^{18}F]-misonidazole positron emission tomography-detected tumor hypoxia in patients with advanced head and neck cancer randomly assigned to chemoradiation with or without tirapazamine: a substudy of Trans-Tasman radiation oncology group study 98.02. *J Clin Oncol* 2006; **24**: 2098–2104.
8. Jolesz FA. 1996 RSNA Eugene P. Pendergrass new horizons lecture. Image-guided procedures and the operating room of the future. *Radiology* 1997; **204**: 601–612.
9. Vannier MW and Marsh JL. Three-dimensional imaging, surgical planning, and image-guided therapy. *Radiol Clin North Am* 1996; **34**: 545–563.
10. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat Methods Med Res* 2015; **24**: 9–26.
11. Weber WA. Positron emission tomography as an imaging biomarker. *J Clin Oncol* 2006; **24**: 3282–3292.
12. Tabrizi SJ, Reilmann R, Roos RAC, et al. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol* 2012; **11**: 42–53.
13. Richter WS. Imaging biomarkers as surrogate endpoints for drug development. *Eur J Nucl Med Mol Imaging* 2006; **33**: 6–10.
14. Abramson RG and Yankeelov TE. Imaging biomarkers and surrogate endpoints in oncology clinical trials. In: Luna A, Vilanova JC, da Cruz LCH, Jr, et al. (eds) *Functional imaging in oncology*. Heidelberg, Germany: Springer, 2013, pp. 29–42.
15. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res* 2015; **24**: 27–67.
16. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. *Stat Methods Med Res* 2015; **24**: 68–106.
17. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm* 1987; **15**: 657–680.

18. Warfield SK, Zou KH and Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; **23**: 903–921.
19. Armato SG, 3rd, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 2011; **38**: 915–931.
20. Lesjak Ž, Galimzianova A, Koren A, et al. A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*. Epub ahead of print 4 November 2017. DOI: 10.1007/s12021-017-9348-7.
21. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement error in nonlinear models: a modern perspective*. Boca Raton, FL: CRC press, 2006.
22. Gustafson P. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Boca Raton, FL: CRC Press, 2003.
23. Dunn G and Roberts C. Modelling method comparison data. *Stat Methods Med Res* 1999; **8**: 161–179.
24. Dunn G. Regression models for method comparison data. *J Biopharm Stat* 2007; **17**: 739–756.
25. Christensen R and Blackwood LG. Tests for precision and accuracy of multiple measuring devices. *Technometrics* 1993; **35**: 411–420.
26. Pennello G. Comparing monitoring devices when a gold standard in unavailable: application to pulse oximeters. In: *ASA proceedings of the joint statistical meetings*. American Statistical Association, 2003, pp. 3256–3263.
27. Kupinski MA, Hoppin JW, Clarkson E, et al. Estimation in medical imaging without a gold standard. *Acad Radiol* 2002; **9**: 290–297.
28. Kupinski MA, Hoppin JW, Krasnow J, et al. Comparing cardiac ejection fraction estimation algorithms without a gold standard. *Acad Radiol* 2006; **13**: 329–337.
29. Lebenberg J, Lalande A, Clarysse P, et al. Improved estimation of cardiac function parameters using a combination of independent automated segmentation results in cardiovascular magnetic resonance imaging. *Plos One* 2015; **10**: e0135715.
30. Jha AK, Song N, Caffo B, et al. Objective evaluation of reconstruction methods for quantitative SPECT imaging in the absence of ground truth. *Proc SPIE Int Soc Opt Eng* 2015; **9416**: 94161K.
31. Jha AK, Caffo B and Frey EC. A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods. *Phys Med Biol* 2016; **61**: 2780–2800.
32. Jha AK, Kupinski MA, Rodríguez JJ, et al. Task-based evaluation of segmentation algorithms for diffusion-weighted MRI without using a gold standard. *Phys Med Biol* 2012; **57**: 4425–4446.
33. Lebenberg J, Buvat I, Lalande A, et al. Nonsupervised ranking of different segmentation approaches: application to the estimation of the left ventricular ejection fraction from cardiac cine mri sequences. *IEEE Trans Med Imag* 2012; **31**: 1651–1660.
34. van Ravenzwaaij D, Cassey P and Brown SD. A simple introduction to Markov Chain Monte Carlo sampling. *Psychonomic Bull Rev* 2016. Epub ahead of print 11 March 2016. DOI: 10.3758/s13423-016-1015-8.
35. Lewandowski D, Kurowicka D and Joe H. Generating random correlation matrices based on vines and extended onion method. *J Multivariate Anal* 2009; **100**: 1989–2001.
36. Jain S, Sima DM, Ribbens A, et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage Clin* 2015; **8**: 367–375.
37. Galimzianova A, Lesjak Z, Likar B, et al. Locally adaptive MR intensity models and MRF-based segmentation of multiple sclerosis lesions. *Proc SPIE Int Soc Opt Eng* 2015; **9413**: 94133G.
38. Galimzianova A, Pernus F, Likar B, et al. Stratified mixture modeling for segmentation of white-matter lesions in brain MR images. *NeuroImage* 2016; **124**: 1031–1043.
39. Jerman T, Galimzianova A, Pernus F, et al. Combining unsupervised and supervised methods for lesion segmentation. In: Crimi A, Menze B, Maier O, et al. (eds) *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*. Number 9556 in Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2016, pp. 45–56.
40. Lesjak Z, Galimzianova A, Likar B, et al. Increased accuracy and reproducibility of MS lesion volume quantification by using publicly available BrainSeg3d image analysis software. *Multiple Scleros J* 2015; **21**: 500–501.
41. Hoffman MD and Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 2014; **15**: 1593–1623.
42. Kucukelbir A, Tran D, Ranganath R, et al. Automatic differentiation variational inference. *J Mach Learn Res* 2017; **18**: 1–45.