Benchmarking Quantitative Imaging Biomarker Measurement Methods Without a Gold Standard

Hennadii Madan(✉), Franjo Pernuš, and Žiga Špiclin

Faculty of Electrical Engineering, Laboratory of Imaging Technologies, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
{hennadii.madan,franjo.pernus,ziga.spiclin}@fe.uni-lj.si

Abstract. Validation of quantitative imaging biomarker (QIB) measurement methods is generally based on the concept of a reference method, also called a gold standard (GS). Poor quality of the GS, for example due to inter- and intra-rater variabilities in segmentation, may lead to biased error estimates and thus adversely impact the validation. Herein we propose a novel framework for benchmarking multiple measurement methods without a GS. The framework consists of (i) an error model accounting for correlated random error between measurements extracted by the methods, (ii) a novel objective based on a joint posterior probability of the error model parameters (iii) Markov chain Monte Carlo to sample the posterior. Analysis of the posterior enables not only to estimate the error model parameters (systematic and random error) and thereby benchmark the methods, but also to estimate the unknown true values of QIB. Validation of the proposed framework on multiple sclerosis total lesion load measurements by four automated segmentation methods applied to a clinical brain MRI dataset showed a very good agreement of the error model and true value estimates with corresponding least squares estimates based on a known GS.

Keywords: Bayesian inference · Markov Chain Monte Carlo · Validation · Brain lesion segmentation · Clinical dataset

1 Introduction

Computational analysis of medical images is increasingly used in clinical routine to extract quantitative measurements providing an objective insight into patients' disease status and progression. Such measurements or QIBs are usually scalar indices that characterize a certain morphological or functional aspect of the anatomy of interest. Often there are many different methods to measure the same quantity, e.g. various segmentation methods for brain volumetry. The classical way to validate a given measurement method or methods is based on the concept of a reference method, often called a "gold standard".

Gold standard (GS) method is considered to produce reasonably small errors, but usually requires substantial effort to execute (e.g. manual segmentation)

M. Descoteaux et al. (Eds.): MICCAI 2017, Part II, LNCS 10434, pp. 763–771, 2017.
DOI: 10.1007/978-3-319-66185-8_86

and/or is related to high costs. To address this there is a proliferation of grand challenges [1], which distribute image datasets with a GS for the purpose of method validation. However, these GSs themselves are generally not validated and may be of poor quality. For instance, due to inter- and intra-rater variabilities if based on manual segmentation. A likely outcome of validation when an inaccurate GS is used to evaluate another measurement method is that estimated errors will be both biased and overconfident [9]. This is critical, since validation based on a poor GS may lead one to use an inappropriate measurement method to extract a certain QIB, which may be hazardous to patient health if the QIB is a surrogate endpoint in a clinical trial or treatment process.

Without using a GS, a statistical framework called regression without truth (RWT) [8] may be used to benchmark a group of methods, all applied to measure the same quantity on the same datasets. For each method the RWT estimates a *model* consisting of systematic and random measurement error through iterative likelihood maximization, in which the unknown true values of measurand are treated as nuisance parameters. The approach has several deficiencies, namely, the error model assumes statistical independence between the measurement methods, which is often not the case. Second, it is important to initialize the optimizer close to the unknown "true model" in order to avoid convergence to a non-global maximum. Finally, it is important to note that RWT has not yet been validated against a least squares (LS) regression with a known gold standard method.

In this paper, we propose a novel RWT framework for benchmarking image analysis methods aimed at QIBs measurement. First, an improved error model accounts for correlated random error between measurements of different methods. Second, compared to the maximum likelihood estimate (MLE) used in original RWT, a joint posterior probability of the error models across all methods is formulated and sampled using Markov chain Monte-Carlo (MCMC). One of the advantages of the novel formulation is that, besides obtaining the error model parameters to benchmark the methods, one can also estimate the unknown true values. Validation of the proposed framework on total lesion load (TLL) measurements based on four automated lesion segmentation methods applied to clinical brain magnetic resonance image (MRI) datasets showed a very good agreement with the LS estimates based on known GS.

2 Framework Description

Consider a dataset of images of N patients and M different measurement methods for a certain QIB. Let x_{pm} denote the value of the QIB 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 for QIB extraction?" It can be answered by estimating systematic and random errors of each method on the same dataset.

Error Model. For each measurement method m the measured x_{pm} and the unknown true value x_{pt} are related as:

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

where the polynomial represents the systematic error (bias) and ϵ_{pm} the random error (noise). Some of the measurement methods, albeit different, may be based on similar principles thus the corresponding random errors ϵ_{pm} may be correlated. We model this explicitly by a multivariate Gaussian (MVG) distribution:

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

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

Posterior Probability. Let L_p denote the likelihood of observing the measurements for a single given patient p. By expressing ϵ_{pm} from (1) and using (2) we obtain:

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

where $\boldsymbol{x_p} = [x_{p1}, x_{p2}, \ldots, x_{pm}, \ldots, x_{pM}]^\top$, $B = [b_{km}] \in \mathbb{R}^{K \times M}$ and $\boldsymbol{\chi} = [1, x_{pt}, \ldots, x_{pt}^{k}, \ldots, x_{pt}^{K}]^\top$. 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 as:

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

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

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

$$P(\theta|X) = L \cdot P(\theta)/P(X), \tag{5}$$

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 Markov chain Monte-Carlo (MCMC) to draw samples from the $P(\theta|X)$ without specifying $P(X)$, and then estimate error model parameters from the samples.

Prior Specification. Before applying MCMC it is necessary to specify $P(\theta)$, which, when sufficient amount of data is available, can be simplified by assuming statistical independence between B, Σ and $\boldsymbol{x_t}$, i.e.:

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

Regarding the systematic error coefficients B it is reasonable to assume that the measurement methods response is at least approximately linear, i.e. b_{0m}

and b_{1m} are likely close to zero and one, respectively, while all b_{km}, $k > 1$ are close to zero. 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 $k = 0, 1, 2, \ldots$

Following Barnard et al. [2] the covariance matrix is decomposed as:

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

where $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_M)$ is a diagonal matrix of standard deviations and R is a symmetric correlation matrix. The standard deviations are assigned uninformative Jeffreys' priors for scale parameters, i.e. $\sigma_m \sim \frac{1}{\sigma_m}$, while correlation coefficients are assigned uniform priors, i.e. $R_{ij} \sim \mathcal{U}(-1, 1)$, $i \neq j$.

Prior on true values $\boldsymbol{x_t}$ may be defined based on a certain population-based distribution $P(x_t)$ of the QIB in question. Then, x_{pt} are modeled as i.i.d. according to this distribution as $P(\boldsymbol{x_t}) = P(x_t)^N$. Depending on the particular QIB an informed decision about the family or shape of $P(x_t)$ distribution can be made. In a general situation, some physical limits of the QIB values can be established and the prior on x_{pt} is then assigned a uniform distribution according to these limits.

Parameter Estimation. The posterior (5) specified up to a proportionality constant can be sampled using MCMC. The expected values of error model parameters can be estimated from this sample. 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. In Bayesian model selection the ratio of probabilities of each mechanism is equal to the ratio of mode masses (evidences). The latter is approximated by the ratio of the number of sample points belonging to each cluster. The expected values of parameters for each mechanism are approximated by the expected value of the corresponding cluster.

With the error model parameter estimates at hand the original question can be answered: the methods can be ranked according to their precision, i.e. σ_m. Alternatively, methods can be ranked according to accuracy, e.g. using root mean square error (RMSE) obtained by plugging the estimates into (1) and simulating measurements based on a random sample of x_{pt}.

3 Validation

The proposed framework was validated on a set of TLL measurements, extracted from MRI brain images by four different automated lesion segmentation methods. We evaluated the capability of the proposed framework to recover the values of error model parameters and the unknown true TLL in comparison to the reference values obtained by LS regression with respect to a gold standard TLL.

Dataset and Gold Standard. Clinical dataset was based on the analysis of MRI images of 22 patients diagnosed with multiple sclerosis (MS) (41.3 ± 10.5 years old, 13 females). Each patient's images were acquired on a 3T Siemens MRI using conventional sequences. Three unsupervised methods segmented lesions as abnormal T1-weighted and FLAIR intensity [4–6] as compared to major brain tissues, while the fourth method was a supervised random forest classifier learning algorithm [7]. Additional lesion segmentations were created by three neuroradiologists, who used local semi-automated image analysis tools to segment the lesions. Then they merged and revised the segmentations to reach a consensus lesion segmentation, which was used as a GS. The TLL value was obtained from lesion segmentations by counting lesion voxels and multiplying by voxel volume. Quadratic LS regression of automatic versus gold standard TLL was performed to determine reference "true" values of the polynomial coefficients $b_{km} \in B$ ($K = 2$) and the standard deviations $\sigma_m \in S$ of the error model (Table 1).

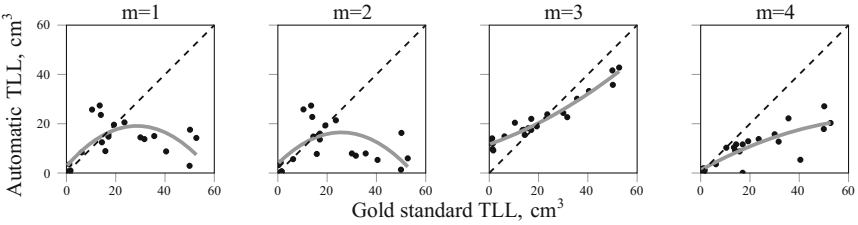

Fig. 1. Least-squares regression of quadratic polynomial to TLL values extracted from MRI by four automatic methods, whereas consensus TLL represent the gold standard.

Experiments. Two sets of experiments were performed: first involved the proposed modeling of systematic and random errors, while the second involved an assumption of independence of random errors, equivalent to constraining R in (7) to the identity matrix. The second experiment represents a model previously used for reference-free regression [8], thus it will be used as a baseline for comparison. In the following we will refer to the respective experiments as "proposed" and "control". Both experiments used the following priors: $b_{0m} \sim \mathcal{N}(0, 55/3)$ [cm^3], $b_{1m} \sim \mathcal{N}(1, 0.5)$, $b_{2m} \sim \mathcal{N}(0, 1/55)$ [cm^{-3}], $\sigma_m \sim 1/\sigma_m$ [cm^3], R_{ij}[1] $\sim \mathcal{U}(-1, 1)$, $x_{pt} \sim \mathcal{U}(0, 55)$ [cm^3]. Note that 55 corresponds to maximum TLL value in the gold standard rounded up to the nearest five. The estimated error model parameters $b_{km} \in B$, $\sigma_m \in S$ and true values x_{pt} were compared to the corresponding LS estimates obtained with respect to the gold standard TLL values.

For MCMC we were using an ensemble affine-invariant sampler with parallel tempering provided in Python package `emcee` [3]. Parallel tempering was setup with a ladder of 20 temperatures so as to provide a 25% replica exchange acceptance rate for Gaussian proposal distributions. For each temperature, ensemble sampler with sample size four times the number of parameters (44 and 38

[1] Not applicable to the "control" experiment.

for the "proposed"and "control"experiments, respectively) was initialized with a draw from a uniform distribution defined as follows: $b_{0m} \sim \mathcal{U}(0, 55)$ [cm^3], $b_{1m} \sim \mathcal{U}(1/3, 3)$, $b_{2m} \sim \mathcal{U}(-50, 50)$ [cm^{-3}], $\sigma_m \sim \mathcal{U}(0, 55)$ [cm^3], $R_{ij} \sim \mathcal{U}(-1, 1)$, $x_{pt} \sim \mathcal{U}(0, 55)$ [cm^3]. Sampling was allowed to run for at least 700000 iterations. The sampler positions from the last 100 iterations were pooled and analyzed.

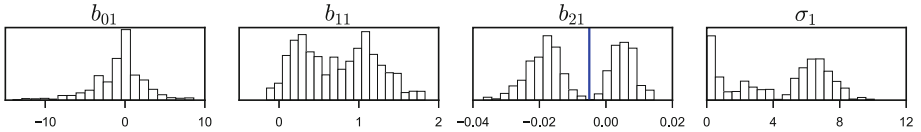

Fig. 2. Histograms of (marginal) posterior distribution of error model parameters of the first method ($m = 1$). *Blue line* indicates the mode split.

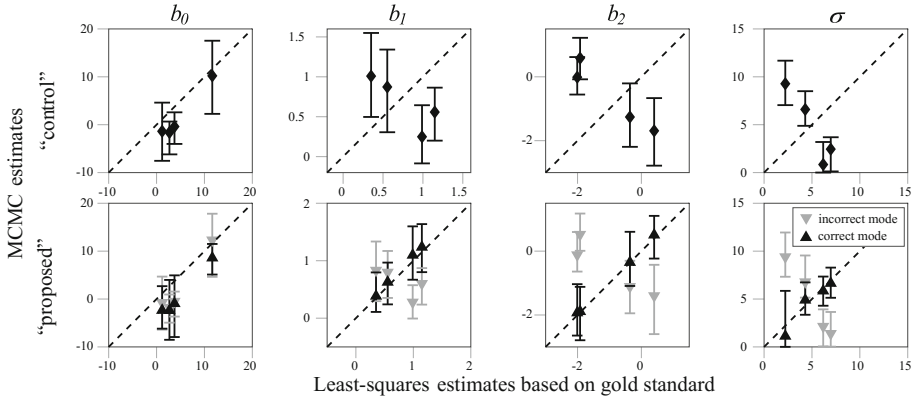

Fig. 3. Error model estimates obtained in "control"and "proposed" experiments (*top* and *bottom*, respectively) versus least-squares estimates based on gold standard.

Results. The posterior distribution obtained in the "proposed" experiment contained two well separated modes (cf. b_{21} histogram in Fig. 2), indicating two possible solutions. Based on visual assessment of b_{21} histogram in Fig. 2, 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. 53% versus 47% of the sample), the solutions corresponding to the left and right modes were designated as *correct* and *incorrect*. In the "control" experiment, the posterior contained only one mode.

Figure 3 and Table 1 show the reference LS based error model estimates and those obtained by the proposed framework. Parameters b_{1m}, b_{2m} and σ_m are in good agreement, while b_{0m} are slightly offset. This is expected as a small error in b_{1m} and b_{2m} estimates has a large impact on b_{0m}, but the overall fit is still comparable according to the similarities of σ_m. As mentioned earlier the sample can be used to estimate the true values of the TLL, as shown in Fig. 4.

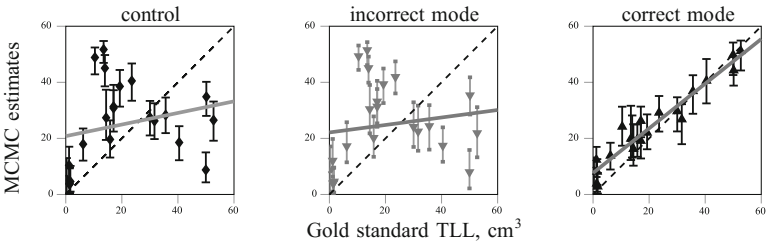

Fig. 4. Estimated true values x_{pt} of the measurand, where the *grey line* indicates corresponding linear trendline and *dashed line* the gold standard TLL.

Methods may be benchmarked and ranked according to the estimated precision σ_m or according to estimated RMSE accuracy shown in Table 1. According to the MCMC estimates the best precision and accuracy were achieved by method $m = 3$, which was the supervised method based on random forest classification [7]. This result is in agreement with the LS estimates (Table 1). High precision and accuracy of method $m = 3$ is also apparent to Fig. 1. Note that both the LS estimates and Fig. 1 require the GS, whereas the MCMC estimates were obtained without the GS.

Table 1. Error model estimates and root mean square error (RMSE) of the estimated true TLL values obtained with the proposed MCMC based method. The estimates obtained with LS regression to the gold standard are shown for comparison.

m	LS estimates					MCMC estimates				
	b_{0m} cm^3	b_{1m} 1	b_{2m} $10^{-3}\mathrm{cm}^{-3}$	σ_m cm^3	RMSE cm^3	b_{0m} cm^3	b_{1m} 1	b_{2m} $10^{-3}\mathrm{cm}^{-3}$	σ_m cm^3	RMSE cm^3
1	3.7	0.99	−19.2	7.0	17.6	−0.9	1.10	−18.9	6.6	19.6
2	2.7	1.15	−20.1	6.2	23.0	−2.4	1.24	−19.3	5.8	20.1
3	11.6	0.35	4.0	2.2	8.4	8.6	0.38	5.1	1.1	6.5
4	1.2	0.55	−3.6	4.3	15.0	−2.3	0.63	−3.5	4.9	16.0

In both Figs. 3 and 4 it is evident that in the "proposed" experiment one of the modes corresponds to the correct estimates, while the other corresponds to independent random errors (cf. "control" experiment). The correct mode had a higher maximal value of the posterior probability and higher mass hinting that it represents the mechanism underlying the data.

4 Discussion

A reference-free framework for benchmarking a group of measurement methods was presented. Benchmarking is provided though the estimation of systematic

and random error model parameters for each method, then the methods can be ranked according to precision based on random error dispersion estimate (σ_m) or RMSE accuracy derived from the complete error model.

The framework was validated against a gold standard in the context of QIB (brain lesion volume) measurement from MRI dataset of real patients. Such a validation is among first in the literature, to the best of our knowledge. Although inspired by RWT [8], important novel methodology was introduced in this work, such as the improved error model, a joint posterior probability criterion and the use of MCMC to find the estimates. The most significant contribution is the modeling of statistical dependence of the random error between different methods (2). The lack of such modeling is likely to lead to results not consistent with the estimates based on GS. This might be a possible reason that RWT was not yet validated on real datasets.

Analysis of the joint posterior provided by the MCMC allows to estimate the unknown true values (Fig. 4). 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 are possibly more meaningful than any of the individual measurements.

Acknowledgments. This work supported by Slovenian Research Agency under grants J2-5473 and P2-0232.

References

1. Grand Challenges in Biomedical Image Analysis (2017). https://grand-challenge.org/All_Challenges/. 24 Feb 2017
2. Barnard, J., McCulloch, R., Meng, X.L.: Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica **10**, 1281–1311 (2000). http://www.jstor.org/stable/24306780?seq=1#page_scan_tab_contents
3. Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J.: emcee: the MCMC hammer. Publ. Astron. Soc. Pac. **125**(925), 306 (2013)
4. Galimzianova, A., Lesjak, Z., Likar, B., Pernus, F., Spiclin, Z.: Locally adaptive MR intensity models and MRF-based segmentation of multiple sclerosis lesions. In: Proceedings of SPIE International Society Optics Engineering, vol. 9413, p. 94133G, 20 March 2015
5. Galimzianova, A., Pernus, F., Likar, B., Spiclin, Z.: Stratified mixture modeling for segmentation of white-matter lesions in brain MR images. NeuroImage **124**(Pt A), 1031–1043 (2016)
6. Jain, S., Sima, D.M., Ribbens, A., et al.: Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. NeuroImage: Clin. **8**, 367–375 (2015)
7. Jerman, T., Galimzianova, A., Pernuš, F., Likar, B., Špiclin, Ž.: Combining unsupervised and supervised methods for lesion segmentation. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 45–56. Springer, Cham (2016). doi:10.1007/978-3-319-30858-6_5

8. Kupinski, M.A., Hoppin, J.W., Clarkson, E., Barrett, H.H., Kastis, G.A.: Estimation in medical imaging without a gold standard. Acad. Radiol. **9**(3), 290–297 (2002)
9. Obuchowski, N.A., Reeves, A.P., Huang, E.A.: Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. Stat. Methods Med. Res. **24**(1), 68–106 (2015)