A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods

2016 Phys. Med. Biol. 61 2780

(http://iopscience.iop.org/0031-9155/61/7/2780)

IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology

Phys. Med. Biol. **61** (2016) 2780–2801 doi:10.1088/0031-9155/61/7/2780

A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods

Abhinav K Jha[1], Brian Caffo[2] and Eric C Frey[1]

[1] Division of Medical Imaging Physics, Department of Radiology and Radiological Sciences, Johns Hopkins University, Baltimore, MD 21218, USA
[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

E-mail: ajha4@jhmi.edu

Received 15 July 2015, revised 19 December 2015
Accepted for publication 14 January 2016
Published 15 March 2016

Abstract

The objective optimization and evaluation of nuclear-medicine quantitative imaging methods using patient data is highly desirable but often hindered by the lack of a gold standard. Previously, a regression-without-truth (RWT) approach has been proposed for evaluating quantitative imaging methods in the absence of a gold standard, but this approach implicitly assumes that bounds on the distribution of true values are known. Several quantitative imaging methods in nuclear-medicine imaging measure parameters where these bounds are not known, such as the activity concentration in an organ or the volume of a tumor. We extended upon the RWT approach to develop a no-gold-standard (NGS) technique for objectively evaluating such quantitative nuclear-medicine imaging methods with patient data in the absence of any ground truth. Using the parameters estimated with the NGS technique, a figure of merit, the noise-to-slope ratio (NSR), can be computed, which can rank the methods on the basis of precision. An issue with NGS evaluation techniques is the requirement of a large number of patient studies. To reduce this requirement, the proposed method explored the use of multiple quantitative measurements from the same patient, such as the activity concentration values from different organs in the same patient. The proposed technique was evaluated using rigorous numerical experiments and using data from realistic simulation studies. The numerical experiments demonstrated that the NSR was estimated accurately using the proposed NGS technique when the bounds on the distribution of true values were not precisely known, thus serving as a very reliable metric for ranking the methods on the basis of precision. In the realistic simulation study, the NGS technique was used to rank reconstruction

methods for quantitative single-photon emission computed tomography (SPECT) based on their performance on the task of estimating the mean activity concentration within a known volume of interest. Results showed that the proposed technique provided accurate ranking of the reconstruction methods for 97.5% of the 50 noise realizations. Further, the technique was robust to the choice of evaluated reconstruction methods. The simulation study pointed to possible violations of the assumptions made in the NGS technique under clinical scenarios. However, numerical experiments indicated that the NGS technique was robust in ranking methods even when there was some degree of such violation.

Keywords: no-gold-standard evaluation, quantitative nuclear imaging, evaluate reconstruction methods

(Some figures may appear in colour only in the online journal)

1. Introduction

In nuclear-medicine imaging, quantitative measurements obtained from patient images are often used to facilitate clinical decision making in both diagnostic and therapeutic procedures. For example, radiotracer uptake in organ and tumor measured from quantitative single-photon emission computed tomography (SPECT) images used to estimate absorbed doses for treatment planning for targeted radionuclide therapy (TRT) (Ljungberg *et al* 2002, Flux *et al* 2006, Dewaraja *et al* 2012, 2013, Bailey and Willowson 2013a, 2013b), biomarkers such as standardized uptake value (SUV), metabolic tumor volume and total lesion glycolysis obtained from quantitative positron emission tomography (PET) images to predict cancer therapy response (Naqa 2014), and myocardial blood flow measured using quantitative SPECT or PET images to diagnose cardiac diseases (Rahmim *et al* 2014, Petretta *et al* 2015). A variety of imaging systems and methods are being designed for these quantitative imaging applications (He *et al* 2005, Zaidi and Erwin 2007, Vanzi *et al* 2009, Song *et al* 2011). There is an important need for objective techniques to optimize and evaluate these imaging systems and methods on the basis of their performance on the clinically relevant task of measuring the true quantitative value. These objective evaluation techniques evaluate the reliability of the measurements obtained using the various methods in comparison to the true value, typically in terms of bias, standard deviation, and mean square error between the true and the measured values. Thus the evaluation is facilitated by the knowledge of the ground truth quantitative value.

Performing such objective evaluation with patient data is highly desirable, since these systems and methods are eventually meant to be used with patients. However, *in vivo* determination of the true quantitative value is often impossible, complicating the evaluation process. In some cases, a measure of the quantitative value from a gold-standard procedure is used as a surrogate for the true quantitative value to compare the different imaging methods (Sridhar *et al* 2014). However, such gold standards are often unavailable. Consequently, simulation (He *et al* 2008), physical-phantom (He *et al* 2005, Du and Frey 2009), and animal (Turco *et al* 2014) studies, where either the ground truth or some reliable measure of the ground truth is available, are often used as surrogates for the optimization and evaluation process. However, animal studies are not definitive since the organ sizes and geometries in animals are different from humans. Similarly, phantom studies often do not model anatomy, physiology, or patient variability well enough, and simulations may not model some aspects of the

biology or instrumentation and the results are often not accepted by clinical practitioners. Consequently, optimization and evaluation using these surrogates is often not recognized as definitive.

One approach to objectively evaluate quantitative imaging systems and methods with patient data in the absence of a gold standard is using statistical techniques that can evaluate methods without knowledge of the ground truth. In this context, a regression-without-truth (RWT) technique for objective evaluation of quantitative imaging modalities in the absence of ground truth was proposed by Kupinski *et al* (2002) and Hoppin *et al* (2002). The approach was used to compare segmentation methods (Lebenberg *et al* 2012) and software packages (Kupinski *et al* 2006) to estimate the cardiac ejection fraction (EF) in the absence of the true EF value. This method assumed that the true and measured quantitative values were related linearly, and that the true values were sampled from a unimodal probability distribution function (PDF) with known bounds. More specifically, the true quantitative values were assumed to be sampled from a beta or a truncated normal distribution, both of which were lower and upper bounded by 0 and 1, respectively. However, in several instances, such bounds are not known. For example, the bounds of the true PDF of activity concentration within a certain volume of interest (VOI) in a SPECT image or the the metabolic tumor volume in a PET image are often unavailable *a priori*. To address the case where *a priori* bounds on the distribution of true values are not available, in this paper, we develop a no-gold-standard (NGS) evaluation procedure. The procedure extends the RWT technique, and has similarities with another technique proposed for evaluating lesion-segmentation methods for diffusion-weighted magnetic resonance imaging (DWMRI) (Jha *et al* 2010, 2012). However, in that work, the objective was not to develop a technique for NGS evaluation of imaging methods for applications where the bounds on the PDF of true values are not known.

An issue with the basic RWT approach is that it requires images from many patients. For example, the evaluation of three cardiac-ejection-fraction estimation algorithms and three segmentation algorithms for DWMRI with the corresponding NGS approaches used data from 85 patients (Kupinski *et al* 2006) and 70 different simulated patient images (Jha *et al* 2012), respectively. It is often difficult to obtain such a large number of patient studies. However, for several quantitative tasks in nuclear-medicine imaging, we can exploit the fact that, for a given patient, more than one quantitative value can be obtained from each patient image. For example, for the task of organ activity estimation in nuclear-medicine imaging, from one patient image, we can get activity estimates from multiple VOIs that correspond to different organs or different tumors in the same patient. This could help reduce the required number of patient studies. The proposed NGS technique is designed to exploit this premise.

Another important feature of this paper is the validation of the proposed NGS technique with realistically simulated imaging data. The data were generated by simulating imaging of realistic anthropomorphic phantoms where the organ uptakes were based on actual patient data. Thus, the true quantitative values used in the validation study were obtained from patient data, and not from a synthetic parametric distribution as in previous validation studies (Hoppin *et al* 2002, 2003, Kupinski *et al* 2002, Jha *et al* 2012). Similarly, the measured quantitative values were not generated synthetically, but an outcome of realistic and rigorous simulations. Thus, the efficacy of the NGS approach was investigated for a very clinically realistic scenario. The study revealed practical issues that could affect the performance of the NGS technique. We conducted several numerical experiments, that provided insights into the effect of these issues on the performance of the NGS technique.

2. Methods

2.1. Theory

Consider a nuclear-medicine imaging system that images multiple patients. The patient images are obtained using K different imaging methods. In this context, an imaging method means any combination of image acquisition protocol, system or parameters and image reconstruction, processing, or analysis method used to obtain some quantitative value. From the images, a set of quantitative measurements are obtained. These measurements could, for example, be the mean activity concentration in different volumes of interest, or the volumes of different tumors in a patient. Let P denote the total number of quantitative measurements obtained from all the patients. Our objective is to use the measured values to rank the imaging methods based on the task of precisely measuring the true quantitative values, but without any knowledge of the true quantitative values. As mentioned above, the procedure we developed for this purpose builds upon the RWT approach. Our description of the developed approach will specifically be in the context of imaging applications where bounds on the true values are not known and where multiple measurements from the same patient image are used as inputs to the technique.

We begin with the assumption that there is a linear stochastic relationship between the true and measured quantitative values. The stochastic component of the relationship is assumed to be a zero-mean normally distributed noise term. Denote the true quantitative value for the pth case by a_p, and the corresponding measured value for this case using the kth imaging method by $\hat{a}_{p,k}$. Denote the slope, intercept and standard deviation of the noise term for the kth imaging method by u_k, v_k, and σ_k, respectively, where each of these parameters are assumed to be independent of a_p. Using the linearity assumption, we write the relationship between the true and measured quantitative values for the kth method as

$$\hat{a}_{p,k} = u_k a_p + v_k + \epsilon_{p,k}, \tag{1}$$

where $\epsilon_{p,k}$ denotes a random variable sampled from a zero-mean normal distribution with standard deviation σ_k. Denote the linear model parameters for all the K imaging methods, i.e. $\{u_k, v_k, \sigma_k, k = 1, 2 \ldots, K\}$, by the vector $\mathbf{\Theta}$. Also denote the entire set of measured quantitative values, i.e. $\{\hat{a}_{p,k}, p = 1, 2, \ldots P, k = 1, 2 \ldots K\}$, by the vector $\hat{\mathcal{A}}$.

We next assume that the true values a_p have been sampled from some unknown parametric distribution. This unknown distribution must be able to model a wide variety of shapes of the true distribution. Typically, values obtained in quantitative imaging applications are positive, so a distribution that incorporates this constraint is preferable. Further, since the bounds on the distribution of true values are not known, the assumed distribution should allow incorporating different values of these bounds. A distribution that satisfies several of these properties is the generalized beta distribution (GBD). This distribution can model several types of unimodal distributions, including symmetric, non-symmetric, negatively-skewed, strictly increasing, strictly decreasing, concave, convex and uniform distributions. Further, in addition to the two shape parameters (α and β), this distribution also has two other parameters, l and g, that specify the lower and upper bounds of the distribution, respectively. We assume that the true values a_p have been sampled from this distribution, so that the probability distribution function (PDF) for a_p, denoted by $\text{pr}(a_p|\alpha, \beta, g, l)$, is given by

$$\text{pr}(a_p|\alpha, \beta, g, l) = \frac{(a_p - l)^{(\alpha-1)}(g - a_p)^{(\beta-1)}}{B(\alpha, \beta)(g - l)^{(\alpha+\beta-1)}}, \tag{2}$$

where $B(\alpha, \beta)$ denotes the beta function. We denote the vector of parameters that characterize the GBD by the vector $\mathbf{\Omega}$.

The developed NGS approach estimates the values of the unknown parameters $\{\mathbf{\Theta}, \mathbf{\Omega}\}$ that maximize the probability of the measured values using all the methods, and is thus a maximum-likelihood (ML) technique. Denote the estimated quantitative values for the pth true value using the K imaging methods, i.e. $\{\hat{a}_p^k, k = 1, 2, \ldots K\}$ by $\hat{\mathcal{A}}_p$. Assuming that the normally distributed noise terms $\epsilon_{p,k}$ are independent for a given value of a_p, using equation (1) we can write

$$\text{pr}(\hat{\mathcal{A}}_p|\mathbf{\Theta}, a_p) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi{\sigma_k}^2}} \exp\left[\frac{-(\hat{a}_{p,k} - u_k a_p - v_k)^2}{2{\sigma_k}^2}\right]. \tag{3}$$

Define

$$S = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi{\sigma_k}^2}}. \tag{4}$$

Multiplying both sides of equation (3) by $\text{pr}(a_p|\mathbf{\Omega})$ and using the rules of conditional probability yields

$$\text{pr}(\hat{\mathcal{A}}_p, a_p|\mathbf{\Theta}, \mathbf{\Omega}) = S\text{pr}(a_p|\mathbf{\Omega}) \exp\left[-\sum_{k=1}^{K} \frac{(\hat{a}_{p,k} - u_k a_p - v_k)^2}{2{\sigma_k}^2}\right]. \tag{5}$$

Marginalizing equation (5) over the true value a_p yields

$$\text{pr}(\hat{\mathcal{A}}_p|\mathbf{\Theta}, \Omega) = S \int \text{d}a_p \text{pr}(a_p|\Omega) \exp\left[-\sum_{k=1}^{K} \frac{(\hat{a}_{p,k} - u_k a_p - v_k)^2}{2{\sigma_k}^2}\right]. \tag{6}$$

Note that after this marginalization step, the expression for the PDF of $\hat{\mathcal{A}}_p$ requires only a knowledge of the distribution of a_p, as defined in this case by equation (2). No knowledge of the actual values of a_p is required.

We next assume that the P true quantitative values are independent. This is a reasonable assumption when the values are from different patients. However, even when the true values are obtained from the same patient, there are often scenarios where these values could still be reasonably independent. For example, consider the case where multiple true mean activity concentrations are obtained from VOIs in the same patient, where these VOIs are in different organs. In this case, since the different organs have different physiologies, it is reasonable to assume that the true activity concentrations in the VOIs are independent. There are of course constraints such as the total activity in all organs cannot be more than the total activity administered, but these constraints are relatively weak, especially if the set of VOIs does not fill the entire region where there is non-zero activity. Further, as we will show in the Discussions section, some violation of this assumption does not affect the performance of the NGS technique. The assumption allows us to write the expression for the likelihood of the set of all measured quantitative values corresponding to all P true values, as denoted by $\mathcal{L}(\mathbf{\Theta}, \mathbf{\Omega}|\hat{\mathcal{A}})$, as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\Omega}|\hat{\mathcal{A}}) &= \mathrm{pr}(\hat{\mathcal{A}}|\boldsymbol{\Theta},\boldsymbol{\Omega}) \\ &= \prod_{p=1}^{P} \mathrm{pr}(\hat{\mathcal{A}}_p|\boldsymbol{\Theta},\boldsymbol{\Omega}) \\ &= \prod_{p=1}^{P} S \int \mathrm{d}a_p \mathrm{pr}(a_p|\boldsymbol{\Omega}) \exp\left[-\sum_{k=1}^{K} \frac{(\hat{a}_{p,k} - u_k a_p - v_k)^2}{2{\sigma_k}^2}\right]. \end{aligned} \tag{7}$$

where, in the last step, we have used the expression from equation (6).

The NGS method estimates values of $\{\boldsymbol{\Theta},\boldsymbol{\Omega}\}$ that maximize the likelihood function, or alternatively, the logarithm of the likelihood function. Denoting the ML estimates of $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\Omega}}$, using equation (7), we obtain

$$\begin{aligned}\{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\Omega}}\} &= \mathrm{argmax}_{\boldsymbol{\Theta},\boldsymbol{\Omega}} \log \mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\Omega}|\hat{\mathcal{A}}) \\ &= \mathrm{argmax}_{\boldsymbol{\Theta},\boldsymbol{\Omega}} P \log S + \sum_{p=1}^{P} \log \int \mathrm{d}a_p \mathrm{pr}(a_p|\boldsymbol{\Omega}) \exp\left[-\sum_{k=1}^{K} \frac{(\hat{a}_{p,k} - u_k a_p - v_k)^2}{2{\sigma_k}^2}\right]. \end{aligned} \tag{8}$$

Note in the above equation that the likelihood function does not depend on the values of a_p, and can thus be maximized without any knowledge of the actual values of a_p. Thus, the above equation presents a formalism to estimate the parameters that characterize the relationship between the true and the measured values without using any knowledge of the true values.

2.2. Implementation of the NGS technique

The developed NGS technique was implemented using Matlab (Mathworks, Natick, Mass). To determine the ML estimates $\{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\Omega}}\}$ (equation (8)), we used a constrained optimization technique based on the interior-point algorithm (Byrd *et al* 1999) that was designed to search between reasonable values of the parameters. The search space for the parameters should allow for modeling a large range of reasonable relationships between the true and measured quantitative values. However, it should also not be too large, to avoid the possibility of the optimization routine being trapped in local minima. The search spaces for the α and β parameters of the GBD were both [1, 20] to enable modeling a wide variety of shapes of distributions of the true value, as shown by the examples in figure 1. The search spaces for the upper and lower limits of the GBD, and parameters of the linear relationship depend on the range of the quantitative values and the imaging methods. An example of defining these search spaces for a quantitative SPECT application will be discussed later.

The optimization routine was initialized with different values to minimize the possibility of being trapped in a local minimum. The process was repeated until the estimated ML values and the likelihood function were stable, thus providing confidence that the global minimum had been reached. A flowchart describing the procedure to compute the ML estimates is shown in figure 2.

The estimated NGS parameters, in particular the estimated noise standard deviation and slope terms, denoted by $\hat{\sigma}_k$ and $\hat{u}_k$, respectively, were used to compute the noise-to-slope ratio (NSR) $\hat{\sigma}_k/\hat{u}_k$ for each method. This figure of merit was, as in previous literature, used to rank the methods on the basis of precision (Hoppin *et al* 2002, Kupinski *et al* 2002, 2006, Jha *et al* 2012). The basic idea is that once the slope is estimated, it could be used to calibrate the different methods. However, the use of this multiplicative calibration leads to amplification of the random component of the difference between the true and measured values, i.e., the standard deviation of the kth method σ_k, by a factor $1/\hat{u}_k$, as is evident from equation (1). Thus, the NSR can rank the calibrated outputs from each of the methods on the basis of precision.

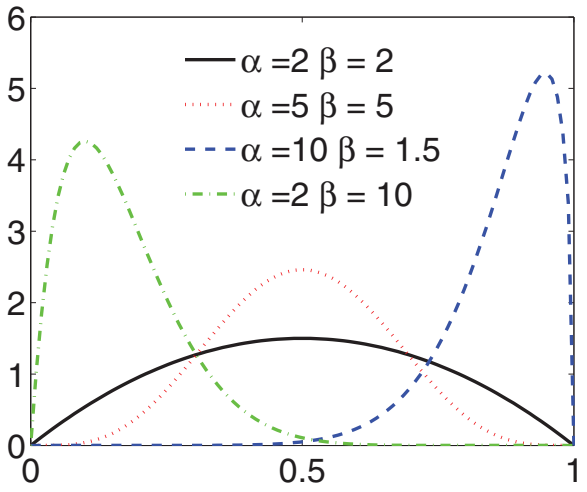

Figure 1. The shapes of the GBD with different values of α and β. The upper and lower values of the GBD were fixed to 1 and 0, respectively, in this illustration.

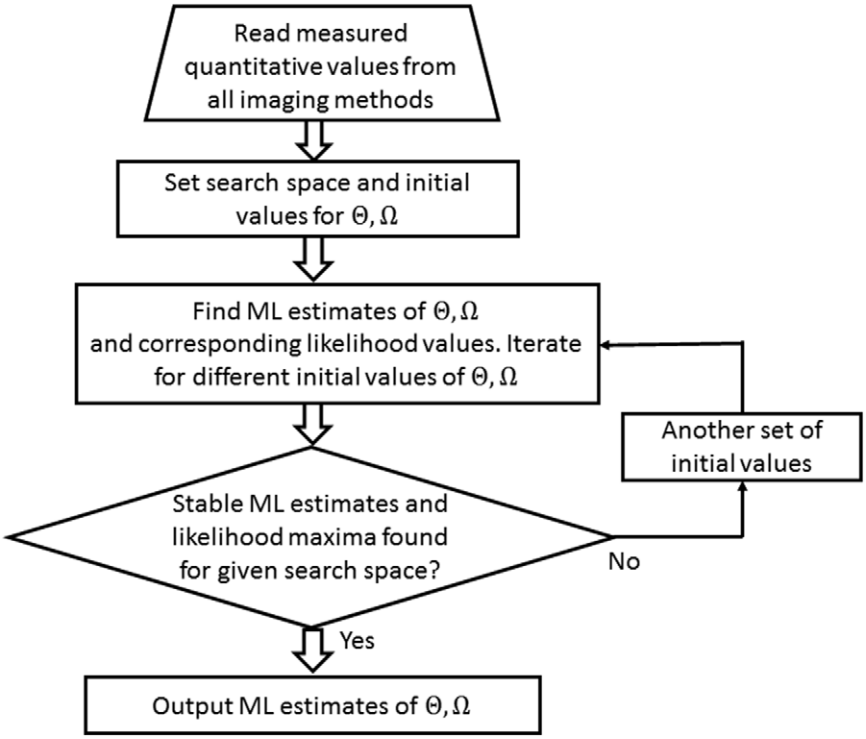

Figure 2. A flowchart describing the procedure to compute the ML estimates of the linear relationship and the true distribution parameters

2.3. Evaluation of the NGS technique

2.3.1. Numerical experiments. The performance of the developed NGS technique was evaluated numerically for cases where the upper and lower limits of the true distribution were not

known. We sampled 200 true values from a GBD for various values of the upper and lower limits. The values of the upper limit, g, and the lower limit, l, were varied between 4.5 and 5, and 0 and 0.5, respectively, over an evenly spaced 10×10 grid. Using the sampled true values, noisy synthetic data were generated for three hypothetical methods, each yielding outputs that were related to the true value by a slope, bias, and noise term. The NGS technique was applied to this noisy synthetic data. In the NGS technique, the search spaces for the slope, bias, and noise standard deviation parameters for all the imaging methods were [0.75, 1.25], [−0.4, 0.4], and [0.01, 0.5], respectively. Also, as mentioned above, the search spaces for the α and β parameters were both [1, 20]. Finally, the search spaces for the upper and lower limits of the GBD were set to [4.5, 5] and [0, 0.5], respectively. The NSR was computed from the parameters estimated with the NGS technique. The experiment was repeated for 50 different noise realizations for each set of true values and linear relationship parameters. From the results, the reliability of the parameters estimated using the NGS technique was evaluated.

The experiment was repeated for true values sampled from GBDs with three different values of the $\{\alpha, \beta\}$ parameters, namely, $\{1.5, 5\}$, $\{3, 3\}$ and $\{5, 1.5\}$. Note that for each choice of $\{\alpha, \beta\}$, 100 combinations of g and l values were considered. Further, for each of the 300 combinations of $\{\alpha, \beta, g, l\}$, 50 different noisy realizations of the measured data were considered. Thus, the NGS technique was used to estimate the rankings of the considered set of methods for a total of 15 000 trials where in each trial, either the parameters that characterized the distribution of true values were different, or we had a different noise realization of the synthetic measured data for a given set of true values.

For each set of sampled true values, the experiment was performed with two sets of methods. In the first set of three methods, we set the values of slope to $\{1.1, 0.9, 1.05\}$ and $\{0.1, 0.2, 0.3\}$ and noise standard deviations to $\{0.03, 0.05, 0.08\}$. These methods had smaller values of the noise standard deviation terms in comparison to the range of the true quantitative value. To evaluate the performance of the method when the noise was higher, another set of three methods was considered that had higher values for the noise standard deviations. The values of slope, bias and noise standard deviations for these methods were $\{1.1, 0.1, 0.2\}$, $\{0.9, 0.2, 0.3\}$ and $\{1.05, 0.3, 0.45\}$, respectively.

2.3.2. Realistic simulation studies. The proposed NGS approach was next evaluated using projection data generated from the realistic simulation of a SPECT system. The projection data were reconstructed using four different methods. From the reconstructed images, the mean activity concentrations in the different organs were measured. The NGS approach was used to rank the four reconstruction methods based on how precisely they estimated the true activity concentration in the absence of the ground truth activity concentration values. These rankings were compared to the rankings computed with the knowledge of the ground truth, providing insight into the efficacy of the NGS method to rank these methods in terms of quantitative performance. The procedure to perform the realistic simulation study is described in the following.

A SPECT imaging system was simulated that imaged the uptake of an I-131 labeled anti-CD20 antibody, I-131 tositumomab, used for radionuclide therapy of non-Hodgkin's lymphoma. The details of the simulation procedure are described in Song *et al* (2011), Song (2001), but we will summarize the important features here. Patient anatomy was modeled using the 3D digital Nurbs-based Cardiac-Torso (NCAT) phantom (Segars *et al* 2001). The activities in the organs in the phantom were based on those from quantitative SPECT/CT reconstructions of a 5 mCi injection of In-111 ibritumomab tiuxetan 24 hours post-injection. The biodistribution of In-111 ibritumomab tiuxetan is a reasonable surrogate for that of I-131 tositumomab since both radiopharmaceuticals target the same CD-20 membrane protein of the

B cells. Similarly, the sizes of the organs in the phantom were adjusted to match the anatomies from these five patients. Combining each of the five sets of activity uptakes with each of the five anatomic sets yielded a phantom population of 25 patients, each with a different uptake/anatomy combination.

A Philips Precedence SPECT system with a 9.525 mm thick NaI crystal and a high-energy general-purpose (HEGP) collimator was simulated. Low-noise projections of eight regions, i.e. heart, lung, liver, kidney, spleen, pelvic marrow, blood vessels and whole-body remainder, were generated from photons with emission energies of 364, 637, and 722 keV and the appropriate abundances. The Simulation System for Emission Tomography (SimSET) software, in conjunction with angular response functions (ARF) tables, which accurately modeled the collimator and detector effects (Song *et al* 2011), was used to simulate the projections. The low-noise projections were scaled and summed according to the emission abundance and activity distributions in the 25-patient phantom population, yielding 25 low-noise projection datasets. Subsequently, 50 independent noisy projection datasets were generated from each low-noise dataset using a Poisson pseudo-random-number generator. This dataset allowed us to test the performance of the NGS procedure 50 times, once for each noise realization. This provided 50 sets of ranking values with which to test the ability of the NGS technique to consistently and accurately rank the reconstruction methods.

The simulated projection data were each reconstructed using four different reconstruction methods, each of which provided different combinations of compensation for the image-degrading processes in SPECT, and were implemented using the ordered subsets expectation maximization (OS-EM) algorithm. The four methods provided compensation for attenuation and scatter (AS), AS and the geometric collimator response (AGS), AS and the collimator-detector response (ADS), and ADS plus an explicit compensation for down-scatter from high-energy photons (ADS.DWN) (Song *et al* 2011). Scatter compensation used the effective scatter source estimation (ESSE) scatter model (Frey and Tsui 1996) and full collimator-detector response compensation included the geometric, septal-penetration, septal-scatter, detector intrinsic, and detector-scatter components. The activities in eight different VOIs, corresponding to the eight different regions mentioned earlier, were measured assuming knowledge of the true organ VOIs.

The activity values obtained using the four reconstruction methods were input to the developed NGS technique. From our simulation results, we had observed that the measured values were not substantially different from the true values. Further, the measured values were in the range [2, 57] kBq cc^{-1}. Considering these two observations, the search spaces for the slope, bias, and noise standard deviation parameters for all the imaging methods were set to [0.75, 1.25], [−4, 4] kBq cc^{-1}, and [0.5, 5] kBq cc^{-1}, respectively. Based on the measured values of activity concentrations using the different methods, the search space for the upper and lower limits of the GBD were set to [55, 60] and [1, 5] kBq cc^{-1}, respectively. Also, as mentioned above, the search spaces for the α and β parameters were both [1, 20]. The parameter values determined using the NGS technique were used to compute the NSR values for the four reconstruction methods, which were then used to rank the four methods. The experiment was also repeated for different combinations of three methods.

The next task was to rank the methods when the ground truth values, i.e. the true mean activity concentration values, were known. Note that this data was generated using an imaging simulation study and not using mathematical models. Thus, to compute an estimate of the slope, bias, and noise standard deviation parameters that characterized the linear relationship between the true and measured activity concentration values for each method, we used a linear-regression technique. We first assumed that the estimated and true mean activity concentration values were linearly related, as defined by equation (1). The linear-model parameters

were computed using a least-squares (LS) approach. These parameters were used to compute the NSR values, which were then used to rank the reconstruction methods. We refer to the rankings obtained with the knowledge of the ground truth as the true rankings, although it must be noted that these rankings were obtained under assumptions of linearity between the true and estimated mean activity concentration values.

3. Results

3.1. Numerical experiments

We first present the results for the set of methods that had noise standard deviation values of $\{0.03, 0.05, 0.08\}$. The mean and standard deviation values of the slope, bias, noise standard deviation terms and the NSR, averaged over all 50 noise realizations and the different GBDs, for this set of methods are summarized in table 1. The slope and the standard deviation terms, and consequently the NSR values, were estimated relatively accurately. This is illustrated in more detail in figure 3, which shows a plot of the mean values of the estimated NSR averaged over the 50 noise realizations as a function of the upper and lower limits of the GBD from which the true values were sampled. These plots show that the mean value of the estimated NSR was different for each combination of α, β, g, and l value.

More importantly the NSR values estimated using the NGS approach yielded the correct rankings in 14 988 of the 15 000 total applications of the NGS technique. More specifically, the true rankings using the NSR metric were that method 1 was the most precise, and method 3 was the least precise, and the NGS technique yielded NSR values that predicted the same rankings in more than 99.9% of the 15 000 trials. However, the bias values were not predicted as accurately as the NSR.

The means and standard deviations of the slope, bias, noise standard deviation and NSR values, averaged over all noise realizations and the different GBDs, for the set of methods that had noise standard deviation values of $\{0.25, 0.4, 0.45\}$ are summarized in table 2. It was again observed that the slope and standard deviation terms, and hence the NSR, were estimated relatively accurately. Thus, the NSR estimate was relatively accurate, as shown in more detail in figure 4. Consequently, the NSR values estimated using the NGS approach yielded the same rankings as the true rankings in 14 818 of the 15 000 total trials. More specifically, the true rankings based on the NSR metric were that method 1 was the most precise and method 3 was the least precise, and the NGS technique predicted the same rankings for close to 99.9% of the 15 000 trials. However, again, the bias values were not estimated as accurately as the NSR.

3.2. Realistic simulation studies

The performance of the developed NGS technique in ranking the four reconstruction methods was evaluated for all 50 noise realizations. In figure 5, the NSR values obtained using the NGS technique and that obtained when the ground truth values were known are plotted. The ranking of the methods determined using the NSR metric when the ground truth was known was the same as the rankings obtained using the NGS technique for 49 out of 50 noise realizations. More specifically, when the ground truth was known, the AS method was ranked as the least precise (highest NSR value) and the ADS.DWN method was ranked the most precise (lowest

Table 1. The mean and standard deviation of the NSR values estimated using the NGS technique.

Method index	True slope	Est. slope	True bias	Est. bias	True std. dev.	Est. std. dev.	True NSR	Est. NSR
1	1.10	1.10 ± 0.01	0.10	0.04 ± 0.03	0.03	0.03 ± 0.01	0.03	0.03 ± 0.01
2	0.90	0.90 ± 0.01	0.20	0.15 ± 0.03	0.05	0.05 ± 0.01	0.06	0.06 ± 0.01
3	1.05	1.05 ± 0.01	0.30	0.23 ± 0.03	0.08	0.08 ± 0.01	0.08	0.08 ± 0.01

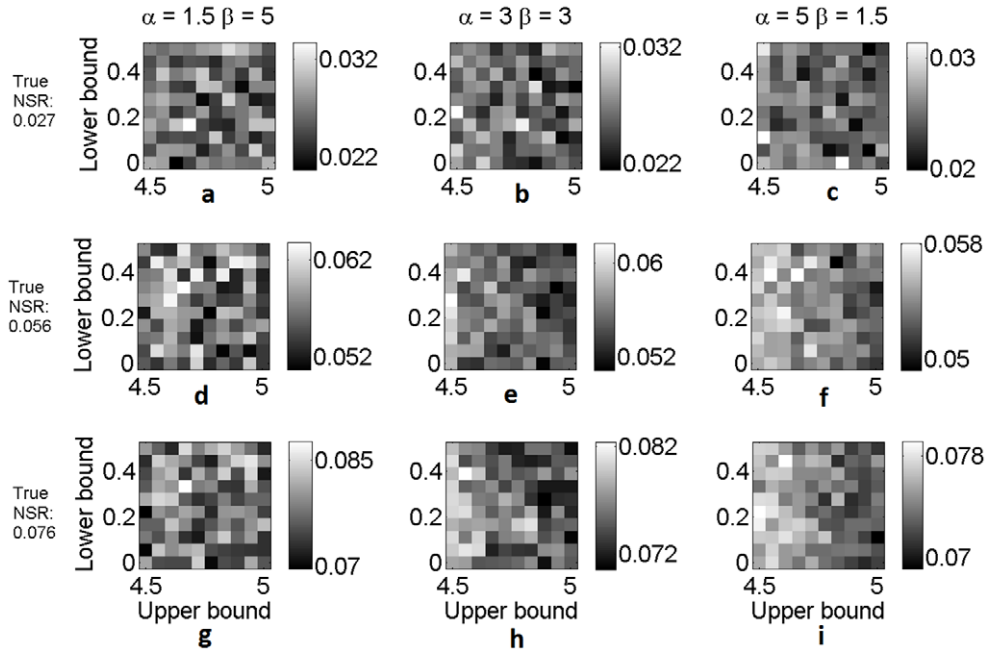

Figure 3. The value of estimated NSR averaged over the 50 noise realizations when the true values were sampled from GBDs with different upper and lower limits. The upper and lower limits were varied between $\{4.5, 5\}$ and $\{0, 0.5\}$, respectively. Panels (a)–(c), (d)–(f) and (g)–(i) correspond to the methods with slope, bias and standard deviation values of $\{1.1, 0.1, 0.03\}$, $\{0.9, 0.2, 0.05\}$ and $\{1.05, 0.3, 0.08\}$, respectively. The true NSR with each method is given at the start of the row. The different columns correspond to GBDs with indicated values of α and β.

NSR value). This was the same as the rankings obtained using the NGS technique for 49 out of the 50 noise realizations.

The experiment was repeated for different combinations of three reconstruction methods. In figures 5(c)–(d) the rankings obtained when evaluating AS, ADS, and ADS.DWN methods using the NGS technique are shown. The rankings estimated using the NGS technique were the same as the true rankings for all 50 noise realizations. Next the AGS, ADS, and ADS.DWN methods were compared using the NGS technique, as shown in figures 5(e)–(f). Again it was observed that, for all 50 noise realizations, the rankings obtained using the NGS technique were the same as the true rankings. Thus, in this study, the NGS technique was not very sensitive to the choice of the reconstruction methods being evaluated.

Table 2. The mean and standard deviation of the NSR values estimated using the NGS technique.

Method index	True slope	Est. slope	True bias	Est. bias	True std. dev.	Est. std. dev.	True NSR	Est. NSR
1	1.10	1.11 ± 0.04	0.10	0.01 ± 0.12	0.2	0.19 ± 0.04	0.18	0.18 ± 0.04
2	0.90	0.91 ± 0.04	0.20	0.12 ± 0.11	0.3	0.30 ± 0.02	0.33	0.33 ± 0.03
3	1.05	1.06 ± 0.05	0.30	0.22 ± 0.11	0.45	0.45 ± 0.03	0.43	0.43 ± 0.03

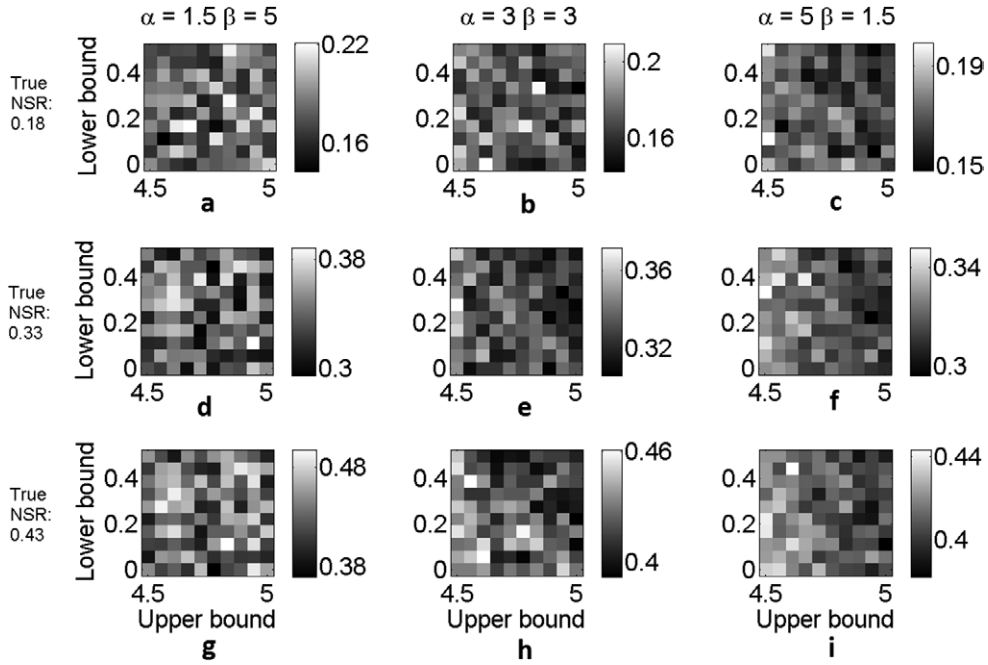

Figure 4. The value of estimated NSR averaged over the 50 noise realizations when the true values were sampled from GBDs with the upper and lower limits varying between $\{4.5, 5\}$ and $\{0, 0.5\}$, respectively. Panels (a)–(c), (d)–(f) and (g)–(i) correspond to the methods with slope, bias and standard deviation values $\{1.1, 0.1, 0.2\}$, $\{0.9, 0.2, 0.3\}$ and $\{1.05, 0.3, 0.45\}$, respectively. The true NSR with each method is mentioned at the start of the row. The different columns correspond to GBDs with indicated values of α and β.

4. Discussions

4.1. Evaluating effect of correlation between true values

A major issue with evaluation in the absence of ground truth is the requirement of a large number of patient studies (Kupinski *et al* 2006, Jha *et al* 2012). This requirement arises because of the relatively large number of parameters that must be estimated to evaluate imaging methods using this technique. For example, evaluating three methods requires estimating 13 parameters. The difficulty and cost of obtaining a sufficient number of patient studies may restrict the clinical applicability of NGS evaluation. In this manuscript, we explored the use of multiple quantitative values from the same patient to overcome this requirement. We found that,

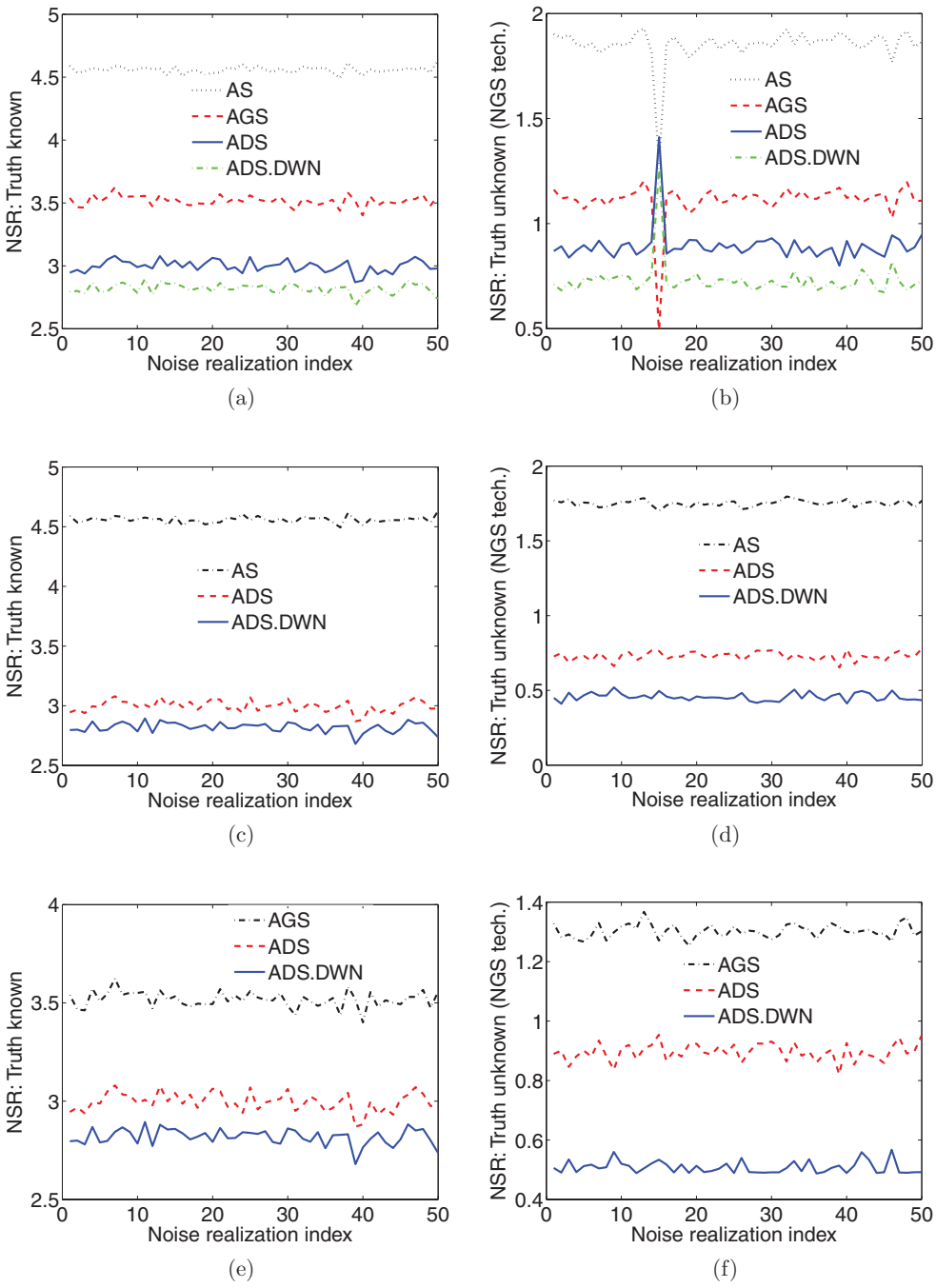

Figure 5. The NSR for each of the 50 noise realizations when ((a)–(b)) evaluating all four reconstruction methods, ((c)–(d)) evaluating AS, ADS and ADS.DWN methods and ((e)–(f)) evaluating AGS, ADS and ADS.DWN methods. The plots on the left side show rankings with known ground truth and on the right side are rankings using the NGS technique.

with this strategy, the proposed NGS technique predicted the rankings accurately with a limited patient dataset for the realistic quantitative SPECT study. However, an assumption made in the derivation of the NGS technique is that the true quantitative values are independent of each other. Our argument was that since the true activity concentration across different organs could be considered reasonably independent, given the different physiology of each organ, this assumption was reasonable. Despite the fact that the method accurately predicted the rankings, we investigated the sensitivity of the method to this assumption. For this purpose, we conducted numerical experiments where the true values of the activity concentrations were correlated with each other.

To generate these correlated variables, we first sampled a set of values independently from a GBD with parameters $\{\alpha, \beta, g, l\} = \{1, 5, 55, 2\}$. These parameters were chosen to replicate a distribution that was similar to the actual distribution of the true values in the realistic simulation study. To simulate correlations that might be present between the true activity values from different organs in a given patient, the sampled values were grouped into sets of eight, corresponding to the eight organs for each patient. Each set was considered as an instance of a random process. To introduce correlation between these eight sampled values, we filtered the random process using a Gaussian filter with various standard deviations. The filtering operation led to a correlated random process where the correlation was characterized by the standard deviation of the Gaussian filter (Barrett and Myers 2004).

Using the filtered dataset as the set of true values, synthetic estimated values were generated for four different methods, with the values of slope of $\{0.75, 0.95, 0.8, 0.8\}$, bias of $\{4.5, 3.0, 1.0, 0.5\}$ kBq cc^{-1} and noise standard deviation of $\{3.5, 3.3, 2.4, 2.0\}$ kBq cc^{-1} for the four methods. The values of the linear-relationship parameters for methods 1–4 were chosen similar to the linear-relationship parameters for the AS, AGS, ADS and ADS.DWN reconstruction methods, respectively. Using these estimated values as input to the proposed NGS technique, the values of $\{\hat{u}_k, \hat{v}_k, \hat{\sigma}_k\}$ were estimated. These values were then used to compute the NSR values for the different methods. The experiment was repeated with different values for the standard deviation of the Gaussian filter. For each value of the standard deviation, the experiment was repeated for 50 different noisy realizations of the synthetic data.

The mean and standard deviation of the NSR values were computed and plotted as a function of the standard deviation of the Gaussian filter, as shown in figure 6. In general, it was observed that while the estimated NSR value varied with an increase in the correlation between the true values, this variation did not have a detrimental effect on the accuracy of the rankings predicted using the NGS technique for this experiment, especially when the degree of correlation was small. This observation indicated that the independence of true values might not be critical for the proposed NGS method, and some degree of correlation between the true values can be tolerated. More studies will be required before the effect of correlation between the true values on the performance of the NGS approach can be firmly established. However, the observation encourages further studies where, for example, different regions within the same organ could be considered as different VOIs and yield separate inputs to the NGS technique. This could help further decrease the required number of patient studies for the NGS technique.

The use of different quantitative values from different VOIs of the same patient as inputs to the NGS technique relies on the assumption that the regions corresponding to the different VOIs have been reconstructed with the same method. This is often the case in SPECT and PET studies (He *et al* 2009, Dewaraja *et al* 2014, Palmedo *et al* 2014). However, in some applications, the optimal reconstruction parameters or method could be different for different organs (Cheng *et al* 2013, 2014), and advanced imaging procedures could incorporate this feature. In those cases, to evaluate a particular reconstruction procedure, only VOIs that were reconstructed using the considered procedure must be considered.

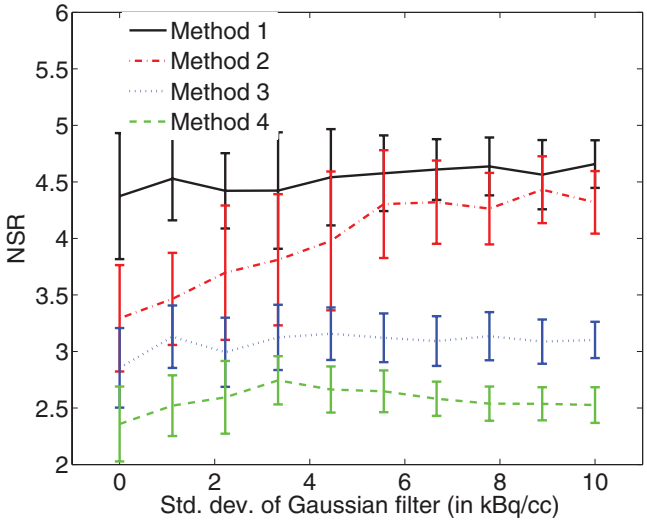

Figure 6. Plot of the NSR as a function of the standard deviation of the Gaussian filter, which parameterizes the degree of correlation between the true values sampled from a patient. The error bars denote the standard deviation of the NSR values.

4.2. Effects of model mismatch on the NGS evaluation technique

The results in figure 5 demonstrated that, with realistic simulated data, the NGS method accurately ranked different reconstruction methods. However, there was a difference between the value of NSR estimated using the NGS technique (ground truth unknown) and the LS technique (ground truth known), as observed in figure 5. This result was unlike the results from the numerical experiments where the NGS method was able to estimate the NSR fairly accurately. In deriving the NGS technique, certain assumptions were made about the models of the true and measured values. The error in estimating the NSR values in the SPECT simulation study could be because the simulated data, given its realistic nature, did not follow these assumptions, or, in other words, there was a model mismatch. We investigated some important sources of this model mismatch and their effect on the performance of the NGS technique in detail, as discussed below.

4.2.1. Assumption of linearity between true and measured quantitative values. In applying the NGS technique in the quantitative SPECT study, it was implicitly assumed that true and measured activity concentration values using the different reconstruction methods were linearly related. The same assumption was made in the validation study in section 2.3 to compute the values of the figures of merit when the ground truth was known. The validity of this linearity assumption could be argued based on the premises that the SPECT imaging system is described by a linear operator, reconstruction methods that sufficiently compensate for image-degrading processes can be considered approximately linear (as we have observed and describe in detail below), and, finally, computing the mean activity concentration from either the object or the reconstructed image is another linear operation. However, several widely used reconstruction methods are non linear. Further, in the simulation study, the mean activity concentrations were obtained from different organs that each had different sizes. The scatter and partial-volume effects thus differed for the different organs, and thus errors were not proportional solely to the activity. These effects could lead to non-linearity between the true and

Table 3. The Δ_i values for different polynomial orders with different reconstruction methods.

	Polynomial order		
Recon. method	1st order	2nd order	3rd order
AS	10.59 ± 0.49	4.36 ± 0.49	0.00 ± 0.00
AGS	0.88 ± 0.48	0.75 ± 0.51	0.03 ± 0.10
ADS	0.00 ± 0.00	1.68 ± 0.29	2.12 ± 0.72
ADS.DWN	0.33 ± 0.46	0.27 ± 0.33	1.42 ± 0.66

measured activity concentration values. To investigate the nature of the relationship between the true and measured mean activity concentration values using the different reconstruction methods, we performed a LS fit between these values for different polynomial orders.

To determine the best model for the relationship, we used the Akaike information criterion (AIC) (Burnham and Anderson 2004). The AIC measures the relative goodness of fit of a statistical model by quantifying the information loss that occurs when the model is used to describe the measured data. Since we had a finite amount of data, we used the AIC with the second order correction term, denoted by $\mathrm{AIC_c}$. Denote the number of coefficients to be estimated by Q and the residual sum of squares (RSS) in the LS fit by δ^2. The value of $\mathrm{AIC_c}$ for normally distributed noise is given by

$$\mathrm{AIC_c} = P\log(\delta^2) + 2Q + \frac{2Q(Q+1)}{P-Q-1}. \tag{9}$$

For each reconstruction method, the value of $\mathrm{AIC_c}$ was computed for different polynomial orders using the above equation. The minimum AIC value over all polynomial orders was then obtained. Finally, the computed AIC values were subtracted from this minimum AIC value, yielding a term Δ_i for each polynomial order, where the index i denotes the polynomial order. According to the AIC, the best model for the data will have $\Delta_i = 0$, and a model that has $\Delta_i < 2$ has substantial support in modeling the data (Burnham and Anderson 2004). The AICs were computed for the different reconstruction methods and the different polynomial-order models for all fifty noise realizations. The mean and standard deviation values of Δ_i for the different reconstruction methods and different models are shown in table 3. We observed that, while the linear relationship was indeed suitable for the AGS, ADS, and ADS.DWN method, for the AS method the linear relationship had a very high value of Δ_i, and thus could not be considered a good fit for the data. For this reconstruction method, a quadratic or third-order relationship could be considered a suitable fit using the AIC measure. This is plausible since the AS method did not include compensation for geometric or collimator-detector response. Thus, partial-volume effects were larger with this method, and consequently the errors between the true and measured activity concentration values depended more strongly on the organ size.

To study the effect of non-linearity on the performance of the NGS approach, we conducted a numerical study where the relation between the true and estimated quantitative values was quadratic for one of the methods. To replicate the true value distribution in the simulation study, we sampled 200 true values from a GBD. The GBD was characterized by parameters $\{\alpha, \beta, g, l\} = \{1, 5, 55, 2\}$. As mentioned previously, these parameters were chosen to yield a distribution of true values similar to the actual distribution of true values in the realistic simulation study. The synthetic data were generated from the true values for

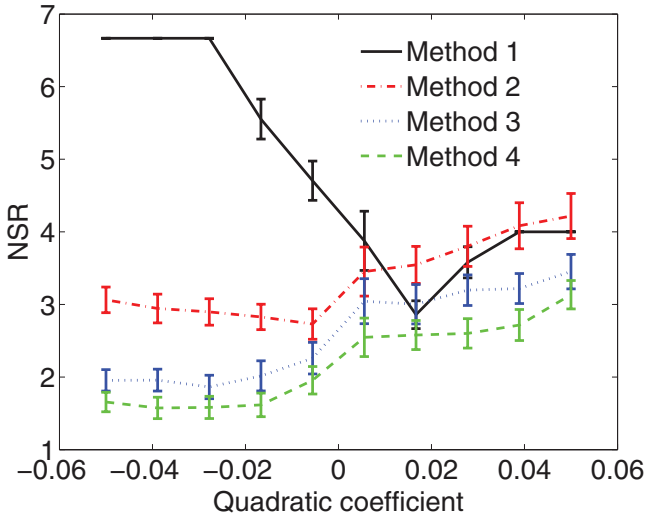

Figure 7. Plot of the estimated NSR as a function of the quadratic coefficient, which parameterizes the degree of deviation from the assumed linear relationship between true and measured values. The error bars denote the standard deviation of the NSR values.

three different methods. For methods 1–4, we set the values of slope to $\{0.75, 0.95, 0.8, 0.8\}$, bias to $\{4.5, 3.0, 1.0, 0.5\}$ and noise standard deviation to $\{3.5, 3.3, 2.4, 2.0\}$. Method 1, which corresponded to the AS method, was quadratic and the second-order coefficient of the quadratic relationship was varied from -0.05 to 0.05. This range was chosen since it was observed that the second-order coefficient for the AS method was about 0.01. The experiment was repeated for 50 different noise realizations for each set of true values. The variation in the mean NSR as a function of the value of the quadratic coefficient is shown in figure 7. We observed that the mean NSR was different from the true NSR as the quadratic component of the relationship increased, a trend that was very clear for method 1. Thus, this study showed that non-linearity in the estimated and true values affected the performance of the NGS technique. However, when the quadratic coefficient was 0.01 (the coefficient with the AS method), the rankings obtained with the NGS technique were the same as the true rankings, which was the same result as observed in our SPECT simulation study.

4.2.2. Assumption of sampling true values from unimodal distribution. In the quantitative SPECT simulation study, we assumed that the true mean activity concentration values were sampled from a generalized beta distribution. However, in reality, these true values were obtained from patient data and not sampled from any known distribution. Thus, modeling the distribution of true values using a uni-modal distribution could be simplistic and inaccurate. For example, in our simulation study, the bone, spleen, pelvis, and kidney had low values of activity (0–10 MBq), but the liver and the background generally had higher activities (10–50 MBq). This could cause multimodal distributions of true values, leading to a mismatch between the true and assumed distributions.

To investigate the effect of this mismatch on the performance of the NGS technique, we performed a numerical experiment where the distribution of true values was perturbed from the assumed distribution by varying amounts. For this experiment, the unperturbed or assumed

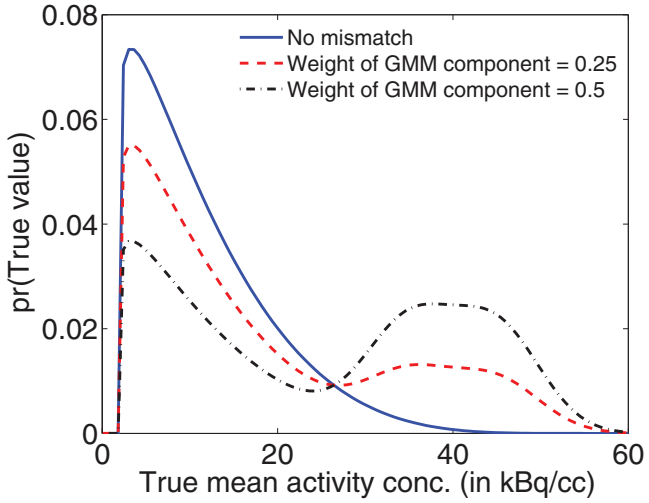

Figure 8. Representative plots of the distribution function of true values for different amounts of mismatch from the GBD.

distribution of true values was a GBD. The perturbed distribution was a mixture model consisting of the GBD and a Gaussian mixture model (GMM), i.e.

$$\mathrm{pr}(a_p|\mathbf{\Omega}, \mu_m, \sigma_m) \sim k_\beta \mathcal{B}(\mathbf{\Omega}) + \sum_{m=1}^{M} \pi_m \mathcal{N}(\mu_m, \sigma_m^2), \tag{10}$$

where $\mathcal{B}(\mathbf{\Omega})$ denotes a GBD characterized by parameter vector $\mathbf{\Omega}$, $\mathcal{N}(\mu_m, \sigma_m^2)$ denotes a normal distribution with mean μ_m and standard deviation σ_m, and k_β and π_m denote the relative weights of the GBD and the mth component of the GMM, respectively. We chose this form since it allowed quantifying the effect of the mismatch. Further, the distribution could model clustering of the mean activity concentrations of the different organs. The amount of mismatch between the true and assumed distributions was quantified by the relative weight w of the Gaussian components in the mixture model. The relative weight w is given by

$$w = \frac{\sum_{m=1}^{M} \pi_m}{k_\beta + \sum_{m=1}^{M} \pi_m}. \tag{11}$$

The performance of the proposed NGS method was studied as a function of this relative weight.

In this study, the GBD in the mixture model had parameters $\{\alpha, \beta, g, l\} = \{1, 5, 55, 2\}$. The mixture model consisted of two Gaussian components, with means equal to $\{35, 45\}$ and standard deviations equal to $\{5, 5\}$. A total of 200 true values was generated for each instance of the mismatch, i.e. for each value of w. An example of the mismatch between a pure GBD and the multimodal distribution arising with this configuration of the mixture model is shown in figure 8. Synthetic data were generated from the true values for four different methods. For the four methods, we set the values of slope to $\{0.75, 0.95, 0.8, 0.8\}$, bias to $\{4.5, 3.0, 1.0, 0.5\}$ and noise standard deviation to $\{3.5, 3.3, 2.4, 2.0\}$. Using the NGS technique, the values of $\{\hat{u}_k, \hat{v}_k, \hat{\sigma}_k\}$, were estimated, which were then used to compute the NSR. For a given value of mismatch, i.e. for each w, the experiment was repeated for 50 different noise realizations.

The mean and standard deviation of the NSR as a function of the relative weight of the mismatch are shown in figure 9. In general, it was observed that while the NSR value varied

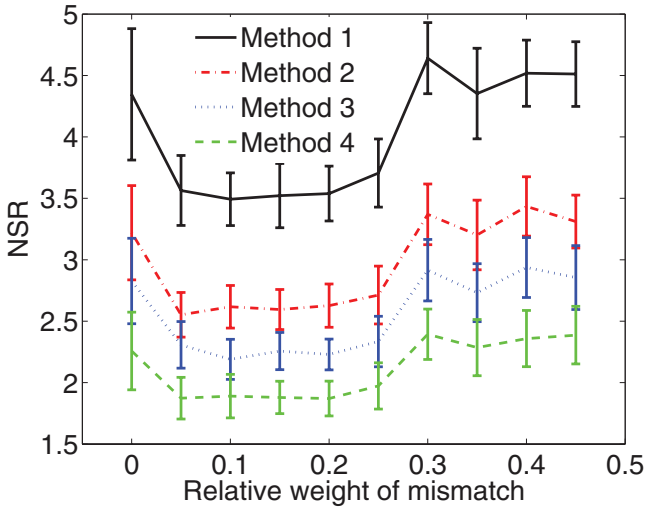

Figure 9. Plot of the estimated NSR as a function of the relative weight of the Gaussian components, which parameterizes the mismatch between the assumed and true distributions. The error bars denote the standard deviation of the NSR values.

with an increase in the mismatch between the true and assumed distributions, this variation did not have any systematic effect on performance of the NGS technique in ranking the different methods. Similar results were observed in the realistic simulation study (figure 5), where, while the estimated values of the NSR were inaccurate, they yielded the same rankings as the true NSR values.

The above studies suggest that practical issues such as model mismatch, which could arise with patient data, could bias the NSR values. Nevertheless, in several cases, the rankings obtained with the NGS technique under model mismatch were the same as the true rankings. To improve the estimates of NSR under model mismatch, enhancements to the proposed NGS technique would be helpful.

Another issue with the proposed NGS evaluation technique, as observed in the numerical experiments, was the inaccurate estimate of the bias term. This inaccurate estimate was not observed when the bounds on the distribution of true values were known *a priori* (Hoppin *et al* 2002), and arises since the upper and lower limits of the distribution of the true values are not known. Accurate estimates of the bias terms would allow accurate ranking of the methods on the basis of accuracy. Further, the estimated bias term could help in correcting for the bias in the different imaging methods. Thus, estimating the bias term accurately is highly desirable, but would require improvements to the NGS technique.

5. Conclusions

We have developed a NGS method for evaluating imaging methods for quantitative nuclear-medicine imaging in the absence of ground truth when the bounds on the distribution of true value are not known. The method assumes that the true and measured quantitative values are linearly related and that the true values have been sampled from a unimodal distribution. Using numerical experiments, we demonstrated that the proposed technique yielded a reliable value for the noise-to-slope ratio metric. Using this metric, an accurate ranking for the imaging methods on the basis of precision was obtained for more than 99 % of the numerical experiments. The approach was also evaluated using data from a realistic simulated limited-patient

dataset, where the objective was to rank four reconstruction methods for quantitative SPECT on the task of estimating the activity concentration within a VOI. In this evaluation, activity concentrations from eight VOIs corresponding to different organs from the same patient image were considered as inputs to the NGS technique. It was observed that, with data from just 25 patient images, the NGS approach was able to rank the four reconstruction methods accurately in all 50 trials. Further, similar results were obtained when different combinations of three of these four reconstruction methods were ranked using the NGS technique, thus demonstrating the robustness of the NGS technique to the choice of combination of the reconstruction methods. The realistic simulation study also pointed to some practical model-mismatch issues such as possible non-linearity between the true and measured quantitative values and multimodal distribution of the true quantitative value. We conducted numerical experiments to study the effects of these issues. The experiments indicated the robustness of the rankings predicted by the technique to some degree of model mismatch, especially the possible multimodal distribution of the true quantitative values. However, incorporating more accurate models of the relationship between true and measured values in the NGS technique would be definitely desirable.

The developed NGS technique is general and could be used to evaluate other imaging methods using patient data, such as segmentation (Zaidi and Erwin 2007) or parameter-estimation (Vanzi *et al* 2009, Jha and Frey 2015) methods in nuclear-medicine imaging, under the assumption of a linear relationship between the true quantitative value and the quantitative value measured using these techniques. Finally, while the investigations on the performance of the NGS technique in the Discussions section were performed in the context of quantitative nuclear-medicine imaging where the end task was estimating the activity concentration, these investigations are general and thus relevant to applications of the NGS technique in other quantitative imaging modalities.

Acknowledgment

This work was supported by National Institute of Health under grant numbers R01-EB016231, R01-CA109234 and U01-CA140204. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Drs Na Song, M Kupinski, E Clarkson, H Barrett, I Buvat, and J Brankov for helpful discussions.

Disclosure

A portion of the reconstruction code used in this work has been licensed to GE Healthcare for inclusion in a commercial product. Under separate licensing agreements between the General Electric Co. and the Johns Hopkins University and the University of North Carolina at Chapel Hill and GE Healthcare, Dr Frey is entitled to a share of royalty received by the universities on sales of products described in this article. The terms of this arrangement are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

References

Bailey D L and Willowson K P 2013a An evidence-based review of quantitative SPECT imaging and potential clinical applications *J. Nucl. Med.* **54** 83–9

Bailey D L and Willowson K P 2013b Quantitative SPECT/CT: SPECT joins PET as a quantitative imaging modality *Eur. J. Nucl. Med. Mol. Imaging* **41** 17–25
Barrett H H and Myers K J 2004 *Foundations of Image Science* 1st edn (New York: Wiley)
Burnham K P and Anderson D R 2004 Multimodel inference *Sociol. Methods Res.* **33** 261–304
Byrd R H, Hribar M E and Nocedal J 1999 An interior point algorithm for large-scale nonlinear programming *SIAM J. Optim.* **9** 877–900
Cheng L, Hobbs R F, Segars P W, Sgouros G and Frey E C 2013 Improved dose-volume histogram estimates for radiopharmaceutical therapy by optimizing quantitative SPECT reconstruction parameters *Phys. Med. Biol.* **58** 3631–47
Cheng L, Hobbs R F, Sgouros G and Frey E C 2014 Development and evaluation of convergent and accelerated penalized SPECT image reconstruction methods for improved dose-volume histogram estimation in radiopharmaceutical therapy *Med. Phys.* **41** 112507
Dewaraja Y K, Frey E C, Sgouros G, Brill A B, Roberson P, Zanzonico P B and Ljungberg M 2012 MIRD pamphlet no. 23: quantitative SPECT for patient-specific 3-dimensional dosimetry in internal radionuclide therapy *J. Nucl. Med.* **53** 1310–25
Dewaraja Y K *et al* 2013 MIRD pamphlet no. 24: Guidelines for quantitative 131I SPECT in dosimetry applications *J. Nucl. Med.* **54** 2182–8
Dewaraja Y K *et al* 2014 Tumor-absorbed dose predicts progression-free survival following (131) I-tositumomab radioimmunotherapy *J. Nucl. Med.* **55** 1047–53
Du Y and Frey E C 2009 Quantitative evaluation of simultaneous reconstruction with model-based crosstalk compensation for 99mTc/123I dual-isotope simultaneous acquisition brain SPECT *Med. Phys.* **36** 2021–33
Flux G, Bardies M, Monsieurs M, Savolainen S, Strands S E and Lassmann M 2006 The impact of PET and SPECT on dosimetry for targeted radionuclide therapy *J. Med. Phys.* **16** 47–59
Frey E and Tsui B 1996 A new method for modeling the spatially-variant, object-dependent scatter response function in SPECT *IEEE Nuclear Science Symp.* vol 2 pp 1082–6
He B, Du Y, Song X, Segars W P and Frey E C 2005 A Monte Carlo and physical phantom evaluation of quantitative In-111 SPECT *Phys. Med. Biol.* **50** 4169–85
He B, Wahl R L, Du Y, Sgouros G, Jacene H, Flinn I and Frey E C 2008 Comparison of residence time estimation methods for radioimmunotherapy dosimetry and treatment planning–Monte Carlo simulation studies *IEEE Trans. Med. Imaging* **27** 521–30
He B *et al* 2009 Comparison of organ residence time estimation methods for radioimmunotherapy dosimetry and treatment planning-patient studies *Med. Phys.* **36** 1595–601
Hoppin J W, Kupinski M A, Kastis G A, Clarkson E and Barrett H H 2002 Objective comparison of quantitative imaging modalities without the use of a gold standard *IEEE Trans. Med. Imaging* **21** 441–9
Hoppin J W, Kupinski M A, Wilson D W, Peterson T E, Gershman B, Kastis G, Clarkson E, Furenlid L and Barrett H H 2003 Evaluating estimation techniques in medical imaging without a gold standard: experimental validation *Proc. SPIE* **5034** 230–7
Jha A K and Frey E C 2015 Estimating ROI activity concentration with photon-processing and photon-counting SPECT systems *Proc. SPIE* **9412** 94120R
Jha A K, Kupinski M A, Rodriguez J J, Stephen R M and Stopeck A T 2010 Evaluating segmentation algorithms for diffusion-weighted MR images: a task-based approach *Proc. SPIE* **7627** 762701
Jha A K, Kupinski M A, Rodriguez J J, Stephen R M and Stopeck A T 2012 Task-based evaluation of segmentation algorithms for diffusion-weighted MRI without using a gold standard *Phys. Med. Biol.* **57** 4425–46
Kupinski M A, Hoppin J W, Clarkson E, Barrett H H and Kastis G A 2002 Estimation in medical imaging without a gold standard *Acad. Radiol.* **9** 290–7
Kupinski M A, Hoppin J W, Krasnow J, Dahlberg S, Leppo J A, King M A, Clarkson E and Barrett H H 2006 Comparing cardiac ejection fraction estimation algorithms without a gold standard *Acad. Radiol.* **13** 329–37
Lebenberg J *et al* 2012 Nonsupervised ranking of different segmentation approaches: application to the estimation of the left ventricular ejection fraction from cardiac cine MRI sequences *IEEE Trans. Med. Imaging* **31** 1651–60
Ljungberg M, Sjogreen K, Liu X, Frey E, Dewaraja Y and Strand S E 2002 A 3-dimensional absorbed dose calculation method based on quantitative SPECT for radionuclide therapy: evaluation for (131)I using Monte carlo simulation *J. Nucl. Med.* **43** 1101–9

Naqa I 2014 The role of quantitative PET in predicting cancer treatment outcomes *Clin. Transl. Imaging* **2** 305–20

Palmedo H *et al* 2014 Whole-body spect/ct for bone scintigraphy: diagnostic value and effect on patient management in oncological patients *Eur. J. Nucl. Med. Mol. Imaging* **41** 59–67

Petretta M *et al* 2015 Quantitative assessment of myocardial blood flow with SPECT *Prog. Cardiovasc. Dis.* **57** 607–14

Rahmim A, Tahari A K and Schindler T H 2014 Towards quantitative myocardial perfusion PET in the clinic *J. Am. Coll. Radiol.* **11** 429–32

Segars W P, Tsui B M, Lalush D S, Frey E C, King M A and Manocha D 2001 Development and application of the new dynamic Nurbs-based cardiac-torso (NCAT) phantom *J. Nucl. Med.* **42** 7P

Song N, Du Y, He B and Frey E C 2011 Development and evaluation of a model-based downscatter compensation method for quantitative I-131 SPECT *Med. Phys.* **38** 3193–204

Song N 2001 Development and validation of quantitative imaging methods for patient-specific targeted radionuclide therapy dosimetry *PhD Thesis* Johns Hopkins University

Sridhar P, Mercier G, Tan J, Truong M T, Daly B and Subramaniam R M 2014 PDG PET metabolic tumor volume segmentation and pathologic volume of primary human solid tumors *Am. J. Roentgenol.* **202** 1114–9

Turco A, Duchenne J, Nuyts J, Gheysens O, Voigt J, Claus P and Vunckx K 2014 Validation of anatomy-enhanced cardiac FDG-PET imaging: an ex vivo sheep study *IEEE Medical Imaging Conf.*

Vanzi E, Genovesi D and Di Martino F 2009 Evaluation of a method for activity estimation in Sm-153 EDTMP imaging *Med. Phys.* **36** 1219–29

Zaidi H and Erwin W D 2007 Quantitative analysis in nuclear medicine imaging *J. Nucl. Med.* **48** 1401