Assessment of bootstrap resampling performance for PET data

P J Markiewicz1,2, A J Reader3,4 and J C Matthews2

1 Translational Imaging Group, CMIC, University College London, 3rd Floor, Wolfson House, 4 Stephenson Way, London NW12HE, UK
2 Imaging Sciences, Wolfson Molecular Imaging Centre, MAHSC, University of Manchester, Oxford Rd, Manchester M13 9PL, UK
3 Division of Imaging Sciences and Biomedical Engineering, Department of Biomedical Engineering, King’s College London, St. Thomas’ Hospital, London WC2R 2LS, UK
4 The Brain Imaging Centre, Montreal Neurological Institute, McGill University, 845 Rue Sherbrooke Ouest, Montreal, QC, Canada

E-mail: p.markiewicz@ucl.ac.uk

Received 30 May 2014, revised 28 October 2014
Accepted for publication 5 November 2014
Published 9 December 2014

Abstract
Bootstrap resampling has been successfully used for estimation of statistical uncertainty of parameters such as tissue metabolism, blood flow or displacement fields for image registration. The performance of bootstrap resampling as applied to PET list-mode data of the human brain and dedicated phantoms is assessed in a novel and systematic way such that: (1) the assessment is carried out in two resampling stages: the ‘real world’ stage where multiple reference datasets of varying statistical level are generated and the ‘bootstrap world’ stage where corresponding bootstrap replicates are generated from the reference datasets. (2) All resampled datasets were reconstructed yielding images from which multiple voxel and regions of interest (ROI) values were extracted to form corresponding distributions between the two stages. (3) The difference between the distributions from both stages was quantified using the Jensen–Shannon divergence and the first four moments.

It was found that the bootstrap distributions are consistently different to the real world distributions across the statistical levels. The difference was explained by a shift in the mean (up to 33% for voxels and 14% for ROIs) being proportional to the inverse square root of the statistical level (number of counts). Other moments were well replicated by the bootstrap although for very low statistical levels the
estimation of the variance was poor. Therefore, the bootstrap method should be used with care when estimating systematic errors (bias) and variance when very low statistical levels are present such as in early time frames of dynamic acquisitions, when the underlying population may not be sufficiently represented.

Keywords: resampling, bootstrap, positron emission tomography, image analysis, statistical uncertainty

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

1. Introduction

Medical imaging such as positron emission tomography (PET), while being used for estimation of some biological parameters of interest (e.g. tissue metabolism for normal or diseased physiology), is always subject to statistical errors at more than one level. For example, at one level, when a set of such parameters are sought to examine or diagnose pathology in a given diseased population, such diagnosis or examination will be subject to sampling errors due to the usually limited size subsets (samples) of the population. At another level, sampling errors are present in the form of measurement errors due to random errors (e.g. variance due to the limited number of annihilation photon pairs recorded by each pair of detectors in PET) and/or systematic errors (bias) in the image acquisition and reconstruction. Although, sampling errors cannot generally be determined, they can be estimated for example using the bootstrap method provided that any given population sample (or the number of events recorded during the PET data acquisition) is large enough and representative of the studied population of detection events resulting from the underlying radiotracer distribution.

Such estimation of sampling errors is important for identifying optimal methods for extracting a meaningful signal (information) in the presence of noise caused by random sampling among other limiting effects. The bootstrap method (Efron and Tibshirani 1994), provides a means of estimating the accuracy (bias) and precision (standard error) of any statistic (an estimate of a population parameter) through resampling with replacement from the observed dataset. For example, the bootstrap has been applied to a pool of recruited participants in order to assess robustness of covariance patterns extracted from a dataset of a cross-sectional study on diseased and healthy participants as well as estimating the prediction accuracy of the model used in such analysis (Habeck et al 2008, Markiewicz et al 2009, Markiewicz et al 2011) (note that in these cases the considered stochastic process is not the measurement but the sampled participants). Further, the bootstrap has been used in the kinetic modelling applied to the dynamic PET datasets in order to estimate the errors of the estimated kinetic parameters while the resampling was performed in the pool of residuals estimated by the prediction of the kinetic model (Turkheimer et al 1998, Kukreja and Gunn 2004) with the assumption that the residuals are independent, identically distributed and have zero mean.

Bootstrap resampling has also been successfully applied to single measurements from nuclear medicine imaging techniques such as PET or SPECT (single-photon emission computed tomography). In such cases the resampling is performed either on the histogrammed PET data formed into sinograms or on the raw and uncorrected PET list-mode data in order to assess the statistical properties of PET images. An example of the former is Buvat (2002) where a set of empirical sub-sinograms is generated in a gated fashion from the original data in order to estimate the unknown distribution of the sinograms (note that the sinogram bins might not be Poisson distributed due to preprocessing steps). Any bootstrap sinogram is then
generated by resampling with replacement every sinogram row from the set of corresponding empirical sub-sinogram rows. Examples of the latter resampling of raw and unhistogrammed PET data have also been reported (Haynor and Woods 1989, Dahlbom 2002, Groiselle and Glick 2005, Verhaeghe et al 2010). Such bootstrap resampling was evaluated when applied to estimating PET image noise yielding similar performance to that of independent realisations of phantom scans (Dahlbom 2002) or Monte Carlo simulations (Groiselle and Glick 2005). Whereas Verhaeghe et al (2010) used the bootstrap resampling of the list mode data in the procedure developed for task-based selection of image reconstruction methodology where the mean square error estimated with the bootstrap was minimised.

In the resampling of PET data two methods of the bootstrap are used: (i) the non-parametric bootstrap which does not require any knowledge of the underlying distribution from which the PET data are randomly drawn (e.g. Buvat (2002)) and (ii) a parametric bootstrap resampling which assumes that the observed data is taken from a defined distribution (e.g. the Poisson distribution for the raw measured PET data (Haynor and Woods 1989) or the multivariate Gaussian distribution in the image space of reconstructed PET scans (Maitra 1998)). The parameters for the distributions used in the resampling are obtained from the original dataset.

Different bootstrap resampling approaches (non-parametric and parametric bootstrap methods applied to list-mode and sinogram data) were compared by Lartizien et al (2010) suggesting that non-parametric and parametric methods produce similar results whether the resampling is done in list-mode or sinogram data. However, it has been shown that resampling based on larger samples (more than one list-mode and sinogram dataset available for resampling) produces more accurate estimates particularly of the mean but also of the variance.

Since the computational and storage costs have significantly lowered today, the method of bootstrap resampling is increasingly gaining a momentum providing greater insights into the statistical analysis of medical images. Therefore, in this work the performance of the more popular non-parametric bootstrap as applied to the raw and uncorrected PET list-mode data is assessed in a novel way using systematic and rigorous comparison of distributions generated through resampling at different statistical levels. In particular, the Jensen–Shannon divergence between two corresponding distributions is used for the assessment of how the bootstrapped distributions diverge from reference distributions here generated and treated as a gold standard. The reference distributions came from long PET acquisitions of one slice of the novel phantom (Markiewicz et al 2011) ensuring very high statistics for the slice voxels whilst accurately simulating the spatial distribution of either the FDG and raclopride in the brain. A real raclopride PET dataset was also used to confirm the findings from the assessment of the bootstrap resampling using the phantom data. The non-parametric bootstrap was chosen as it does not require any knowledge or assumption about the distribution which gave the observed data and therefore the non-parametric bootstrap is less restricted for different applications (Buvat 2002).

The proposed assessment of the bootstrap method can be carried out in more than one way and on many levels but this work focuses on the bootstrap resampling as applied to the data from which medical images or parameters are formed. It is possible though to apply the results of this study to most bootstrap resampling schemes beyond medical imaging.

2. Theory and methods

In order to assess the bootstrap resampling a set of reference realisations (treated as the gold standard) was generated to form sampling distributions with which corresponding bootstrap distributions can be compared. The reference distributions of this study have to be such as to
allow a wide range of statistical levels to be investigated. This potentially could be achieved in three ways: (i) through multiple scan realisations of the same object with realistic distributions of the same radioactivity fixed at exactly the same position for all scans; (ii) accurate Monte Carlo simulations generating such independent realisations; or (iii) using one very long list-mode acquisition of a realistic activity distribution with high radioactivity concentration to ensure a very high statistical level and good representation of the underlying distribution. Based on this list-mode acquisition, which takes the role of the underlying distribution, multiple reference realisations (corresponding to real world multiple measurements) can be generated by sampling with replacements and with the statistical level varied by changing the size of the sampled realisations. This is similar to bootstrap resampling with the difference that the size of newly generated samples changes according to the desired statistical level as indicated in Haynor and Woods (1989) and Groiselle and Glick (2005).

Points (i) and (ii) for the assessment of the bootstrap in this work are impractical as generating such data through physical scans is close to impossible whereas using Monte-Carlo simulations would take an even longer time than the already long acquisition times required for multiple realisations (Buvat 2002). Therefore in this work strategy (iii) has been adopted.

2.1. Long acquisition scans

As a basis for the generation of the reference distributions two long PET scans of a simulated brain slice were acquired on the high-resolution research tomograph (HRRT, Siemens, (de Jong et al 2007)) using a specially designed phantom (Markiewicz et al 2011). This phantom enables printed planar sources of any radioactivity distribution to be imaged in transaxial or axial planes. Two digital phantoms based on the realistic digital MRI brain phantom (Aubert-Broche et al 2006) publicly available at www.bic.mni.mcgill.ca/brainweb/ are here considered, that is, two discrete models are used based on subjects #18 and #20 from the public database. The structural discrete models of grey matter (GM), white matter (WM), the meninges and striatal regions (putamen, caudate) are used to define two digital PET phantoms with values corresponding to two distinct spatial distributions of two common radiotracers: \( ^{18}\text{F}\)FDG (subject #18 used) and \( ^{11}\text{C}\)raclopride (subject #20). The two planar distributions (see figure 1) after a long acquisition of 6 h (with the total radioactivity of 21.6 MBq in the plane at the beginning of the scan) and 5 h (with the total radioactivity of 7.9 MBq), respectively, were reconstructed with the ordinary Poisson ordered subsets expectation maximisation algorithm (OP-OSEM, (Comtat et al 2004)) with inclusion of a model of the point spread function (PSF) of the scanner (Reader et al 2003).

One of the reasons why such planar phantoms are used is a greater radioactivity concentration in one image plane (and thus good statistical representation within that plane) as opposed to volumetric phantoms for which such high radioactivity concentration in each plane could cause a reduction in count-rate performance of the detectors or even damage them. Further, with the digital planar phantom it is possible to investigate the impact that different radioactivity distributions may have on the accuracy of bootstrap resampling.

The findings based on the two planar phantoms are then compared with a real human raclopride scan (figure 1(c)). It is worth noting the supreme statistical quality of the reconstructed images from the two phantoms compared to the real human scan shown in the figure.

2.2. The bootstrap

The bootstrap is a tool used in statistical inference to ascertain the precision and/or accuracy of a statistic of interest \( \theta \) by generating a set of \( K \) statistically equivalent datasets based on one
or more original datasets. \( \theta \) can be a voxel value, the mean voxel value of any region of interest (ROI) or any statistic based on the reconstructed image of the measured PET data \( x \), i.e.

\[
\theta = s(x),
\]

where \( x = \{x_1, x_2, ..., x_n\} \) and \( n \) is the data size (number of recorded events in the list-mode data). One bootstrap sample \( x' \) of size \( n \) is generated by resampling with replacement from the original data \( x \), regarded as representing the underlying distribution. It is then possible to estimate some parameters of the sampling distribution such as the standard error or bias of \( \theta \) from \( K \) bootstrap replicates \( \theta' = s(x') \).

2.3. Two stage resampling

One of the key aspects of this work is the two stage non-parametric resampling in order to assess the accuracy of the bootstrap resampling. The first stage resampling (with replacement) plays the role of the real world where multiple realisation scans of the same object would be performed. The second stage resampling is the bootstrap which takes one of the realisations from the first stage to generate an equivalent bootstrap distribution to that of the first stage.
2.3.1. Stage I. World of multiple realisations. The resampling in this stage is very similar to the bootstrap (see section 2.3.2) with the only difference that the number of events in the resampled dataset is lower than in the original long scan (apart from one case used for comparison where the resampled datasets in this stage are of the same size as the original dataset). The scheme of the resampling is presented in figure 2 in which the probability model $P$ representing the population with its underlying probability distribution of the infinite number of events required to describe the spatial distribution of the radiotracer gives the observed PET data of the long acquisitions. A crucial step in both resampling stages is the estimation of $P$ using $\hat{P}$, the empirical distribution function constructed using the measured list-mode data and putting probability $1/n$ on each recorded event (true, scatter, random and delayed coincidences recorded in the list-mode data), where $n$ is the number of all recorded events in the list-mode data. Hence, a new sample $(x^*)$ can be generated from the estimated probability model $\hat{P}$ with replacement exactly the same way as the original data was generated from $P$. In this way a number of new list-mode datasets can be generated with different total numbers of events $m$ (see figure 2) corresponding to six statistical levels $S$ (or sizes of generated datasets) as the percentage of the original long list-mode dataset used in the assessment ($S = \{0.01\%, 0.1\%, 1\%, 10\%, 20\%, 100\% \}$). For the statistical level of 100% the total number of detected true coincidence events was $8.8 \times 10^9$ for the FDG phantom and $3.1 \times 10^9$ for the raclopride phantom. For the chosen image slice of the measured clinical raclopride scan the estimated number of true coincidence events was $28 \times 10^6$.

The resampling of the events as recorded in the observed list-mode data is done in a non-parametric way in short time-periods (of approximately 1 s) of the list-mode data in order to maintain the dynamic changes of the underlying distribution of the radiotracer. Such resampling of the statistically very rich list mode dataset can be viewed as repeated scans of the same object and here it is repeated $2^K = 2 \times 100$ times (the reason for the 2 is that the so formed distributions based on 100 samples are then compared to each other in order to see any differences for the reference distributions). Also it is worth noting that any division of the long acquisition data into smaller independent datasets would be suboptimal for this assessment as such breaking up of the data would destroy the original probability model $P$ in a systematic way.
2.3.2. Stage II. World of bootstrap. For each of the six statistical levels, a single realisation from stage I is chosen for further use in the second stage (II) of resampling to generate $K$ bootstrap datasets which are statistically equivalent to the one chosen from stage I. Importantly, in this stage (II) of bootstrap resampling, the probability model $P$ (figure 2) is estimated based on the data of the chosen realisation from the first resampling stage (I), and, crucially, not the empirical distribution which gave the realisations of stage I ($\tilde{P}$) which is known based on the rich measured dataset. Therefore, in the bootstrap resampling (stage II) the underlying distribution is estimated by $\tilde{P}$ through assigning a probability of $1/n$ to each event of chosen list-mode data $k$ of stage I ($\tilde{x}^k$). A number $K$ of bootstrap replicates is then generated through random sampling with replacement from $\tilde{P}$ of exactly the same number of events $n$ as in $\tilde{x}^k$ ($m = n$ in figure 2). The bootstrap resampling is carried out for all six statistical levels $S = \{0.01\%, 0.1\%, 1\%, 10\%, 20\%, 100\%\}$ of stage I. This stage (II) of bootstrap resampling is repeated five times by choosing five independent realisation of stage I in order to facilitate more robust comparison of the distributions of the two stages.

Thus, two statistically equivalent sets of list-mode data are generated in the two stages differing in that the reference datasets (stage I) are resampled from a high statistics dataset whereas the bootstrap datasets (stage II) are resampled from the reference datasets of lower statistics.

2.4. Processing and analysis of the generated datasets

The pathway of data analysis for the two resampling stages is presented graphically in figure 3. Each dataset of stage I and II are histogrammed into sinograms and corrected the same way...
as in Markiewicz et al (2011). Since the phantom generates negligible scatter (below 2% of all events) the scatter correction is not applied apart from the dataset of the human brain scan with raclopride. The random coincidence events are estimated using measured delayed coincidences (which for the bootstrap replications are also resampled). The list-mode datasets (recordings of individual PET events) are histogrammed and then reconstructed with 10 iterations and 16 subsets of OP-OSEM with the modelled PSF of the HRRT scanner (Hong et al 2007, Comtat et al 2008). In this assessment single voxels and circular ROIs at 9 different locations were considered (the locations are different for the FDG and raclopride scans) as shown in figure 1. For the raclopride image (phantom and human datasets) three ROIs are considered in the striatal regions, three in the grey matter (GM) and three in the white matter (WM). For the FDG image four GM ROIs, four WM ROIs, and one ROI in the meninges (close to the skull) are considered. Nine voxels are considered for each image at the centre of each of the nine ROI. The three steps of histogramming, reconstruction with corrections, and analysing the ROIs and voxels are included in the operator $s$ in equation (1).

2.4.1 Jensen–Shannon divergence.

The metric of Jensen–Shannon (J–S) divergence $D_{JS}$ was used in order to quantify how different or ‘distant’ from each other are the equivalent sampling distributions from stage I ($R$, based on the full dataset) and stage II ($B$, based on bootstrap resampling). The J–S divergence ($D_{JS}$) is derived from the Kullback–Leibler (K–L) divergence ($D_{KL}$) and is defined by

$$D_{JS}(R \parallel B) = \frac{1}{2}(D_{KL}(R \parallel M) + D_{KL}(B \parallel M)),$$

where

$$D_{KL}(R \parallel M) = \sum_i R_i \log \frac{R_i}{M},$$

and $M = \frac{1}{2}(R + B)$. The reason behind the J–S divergence is that it is symmetric as opposed to K–L divergence, i.e. $D_{JS}(R\parallel B) = D_{JS}(B\parallel R)$ whereas $D_{KL}(R\parallel B) \neq D_{KL}(B\parallel R)$. The proprieties of the JS divergence are such that when two distributions $R$ and $B$ are totally different (e.g. there is no overlap) $D_{JS}$ is 1 (100%) and when the distributions are exactly the same $D_{JS}$ is 0 (0%).

2.4.2 Two methods of histogramming.

In order to apply the J–S divergence to the sampling distributions $R$ and $B$ of the two stages, the distributions of the ROI and voxel values have to be histogrammed (note that it is a different histogramming procedure to that of histogramming PET coincidence events into sinograms). There are two possible ways of achieving this: (i) with constant binning where the bin size for all statistical levels (data sizes) are kept constant and (ii) adaptive binning where the bin size is varying according to the size of the datasets at each statistical level and therefore the effects of the data size are normalised out in the distribution comparison. In particular, the bin size $b_{size}$ in the adaptive histogramming is

$$b_{size} = r_{max} \sqrt{S} / b_n,$$

where $r_{max}$ is the maximum range of $\theta$-values in the sampling distributions for different statistical levels expressed in percentages of the original data size, i.e. $S = \{0.01\%, 0.1\%, 1\%, 10\%, 20\%, 100\% \}$ and $b_n$ is the number of bins. This way the product $b_{size}\sqrt{S}$ is constant for all statistical levels. The difference between the two histogramming methods is depicted in figure 4.

2.4.3 Analysis of moments.

For greater insight into the potential differences between the distributions of stages I and II, further measures of the distributions are used. These are the mean
(the first raw moment), coefficient of variation (CoV, related to variance—the second central moment), skewness (the third standardised moment) and kurtosis (the fourth standardised moment measuring the peakedness of the distributions). The moments were investigated in two ways, i.e. with and without the normalisation by the statistical level in the considered datasets. The two-sample t-test was applied to the estimated moments to test if the bootstrap-reference and reference-reference absolute differences of the moments are significantly different. If the differences were statistically insignificant it would indicate that the bootstrap well replicates the parameters matching the performance of the reference distributions.

2.4.4. Optimal number of bootstrap samples and bins. For optimal analysis using the J–S divergence, the impact of the number of bootstrap samples, number of bins used in the histogramming, the statistical level of the datasets and the choice of ROIs/voxels on the variability of the J–S divergence is assessed. First, a linear model was fitted to the responses of the J–S divergence explained by the above predictors: the number of bootstrap samples, histogram bins and statistical level of the datasets were treated as continuous variables whereas the ROIs/voxels were treated as a categorical variable. Analysis of variance based on the linear regression was performed to gain a rough estimate of the factor mostly influencing the J–S divergence. Next, the optimal number of bootstrap samples and bins is chosen based on investigation of the specific relationships between the J–S divergence and the number of bootstrap samples as well as histogram bins.

For the investigation of the impact of the number of bootstrap datasets (samples) $K$ on the J–S divergence, $K$ was varied linearly from 10 to 100 samples with steps of 10. Similarly, the number of histogram bins $b_n$ was varied from 2 to 513 in exponential steps ($b_n = 2^n + 1, n = 1\ldots 9$). The statistical level of the datasets was $S = \{0.01\%, 0.1\%, 1\%, 10\%, 20\%, 100\% \}$ and the choice of ROIs was treated as a categorical variable with 9 categories corresponding to the 9 ROI positions (figure 1).
3. Results and discussion

The resampling of the two stages (see section 2.3) generates two distributions for each chosen statistical level (data size) and each ROI and voxel (figure 1). The resampling in the first stage generates the reference distributions which here play the role of multiple realisations acquired by scanning the same object many times. Therefore, they are considered to be a very good approximation of the gold standard. The bootstrap resampling of the second stage takes one of the reference realisations and resamples it with replacement to generate statistically equivalent datasets whose distribution is expected to estimate the reference sampling distribution. Figure 5 shows the two equivalent distributions plotted next to each other (the reference distribution are represented by black boxplots whereas the bootstrap distributions are represented by grey boxplots) for the nine ROIs/voxels and two data sizes (0.01%, bottom and 1% top). The position of the ROIs and voxels are shown in figure 1(a).

The differences between the reference and bootstrap distributions as noticed in figures 5 and 6 can be measured using the J–S divergence by assigning a value to the ‘distance’ between...
any two equivalent distributions. Before such a measure can be made the sampling distributions of the statistic (voxel and mean ROI values) have to be histogrammed to a common set of bins. Two binning methods were proposed (see section 2.4.2): (i) constant for which the bin size remains constant for all statistical levels and (ii) adaptive whose bin size adapts to the statistical level of a given dataset. Figure 7 plots the J–S divergence for the distributions generated for the raclopride phantom and for two methods of binning the distributions. For each binning method (top and bottom graphs) there are four J–S divergence curves plotted against the six statistical levels (data sizes expressed in percentages of the original dataset). Two curves are plotted for the J–S divergence between two equivalent reference distributions of size $K = 100$ (therefore there were $2K$ reference realisations generated) for ROIs and voxels. This curves (denoted as ‘Reference’ in the figure) represent the reference point to which the other two curves (denoted as ‘Bootstrap’ in the figure) should come close. These two ‘bootstrap’ curves represent the divergence between the equivalent bootstrap and reference distributions. In order for the ‘Bootstrap’ curves to be smoother the bootstrap resampling in stage II was repeated five times and the J–S curves for each resampling procedure were found and then averaged over the five repetitions. Similarly, figures 8 and 9 present the same plots of J–S divergence but applied to the other two datasets: FDG planar phantom (figure 8) and raclopride human scan (figure 9).

It becomes clear that there is a systematic error present in the bootstrap resampling which appears constant when the effect of the statistical level in the datasets are taken out through the automatic normalisation present in the adaptive histogramming. Further, when constant binning is used, and thus the effects of statistical level are present in the analysis, the J–S
divergence, as expected, decreases and all the curves come closer to each other which is especially apparent for the raclopride phantom dataset (figure 7). Note that the pattern of difference between the bootstrap and reference curves is present for the three original datasets. The reason that the difference is still observable even for the statistical level of 100% is that the bootstrap realisations in stage II are based on one noisy realisation of stage I (it is true for all statistical levels), and hence, bootstrap realisations are more representative of that particular realisation of stage I than of the remaining realisations of stage I. Furthermore, since stage I realisations diverged from the original data, with the amount of divergence relative to the number of counts resampled, consequently the J–S divergence will remain unchanged if the bin size is changed according to the statistical level S.

For further assessment why such differences in the J–S divergence exist for the bootstrap distributions, the difference in moments (1st–4th) of the reference and their corresponding bootstrap distributions were analysed. It was found that in terms of the third and fourth moments—the skewness and kurtosis respectively—there was no statistically significant difference for the equivalent distributions of stages I and II and the chosen binning resolution.
used for generation of the distributions. However, when the first raw moment (the mean) was considered for the striatal regions of the brain, an interesting effect was found in the absolute difference of the mean being similar in shape to that found in the analysis of the J–S divergence. That is, the two curves of the difference of the mean (unnormalised to the statistical level) for the bootstrap-reference and reference-reference distributions have similar behaviour to the two curves of J–S divergence for bootstrap-reference and reference-reference distributions generated with constant binning. Figure 10 shows the difference between the means of the sampling distributions generated in the two resampling stages for ROIs (left figure) and voxels (right figure) for the raclopride phantom dataset, similarly to the presentation of the J–S divergence (e.g. figure 7).

The percentage change of the mean difference for the bootstrap and reference distributions for the striatal regions (an important measure in raclopride studies) was observed to be up to 14% for the ROIs and 33% for the voxels for the low statistical level (0.01%, which for the count level of the phantom measurement, it corresponds to the real count levels in some time frames of dynamic acquisitions).

For further analysis, the difference of the mean for the six statistical levels was normalised by the statistical level represented by the percentage of the original dataset (the top of
Although the richness of the dataset represented at the higher statistical levels makes the distributions very close to each other in absolute terms (bottom plots), in relative terms they are different which is confirmed by the t-test indicating that the curves are statistically different with $p \ll 0.01$. Therefore, it is more advantageous to look at the difference normalised by the statistical richness (top graphs of figure 10) for which the difference remains constant similarly to that of the J–S divergence with adaptive binning. The normalised mean difference drops slightly for the very small data sizes, possibly due to very poor statistical representation and the non-negativity constrain in the OSEM image reconstruction.

Since the bootstrap resampling is based on one limited dataset drawn from a given population (here one realisation from stage I) the bootstrap estimated mean converges to the value of that one realisation. Hence, the estimated mean is more representative of that realisation and may be significantly distant to the population mean from which the noisy realisation was taken depending on how statistically rich the realisation is.

Similar to the above analysis of the mean, the coefficient of variation (CoV) for the bootstrap and reference distributions is shown in figure 11. Both normalised and unnormalised absolute difference of CoV was presented for the chosen statistics of ROI and voxel values.
The two-sample *t*-test was applied to the six points of the two CoV curves to investigate whether the points representing (1) the absolute difference between bootstrap and reference distributions are significantly different from (2) the points representing the difference between two independent reference distributions for the six statistical levels. Therefore, it is possible to infer when the differences between (1) and (2) are not related to limited data. The *p*-values of the test are the same for the normalised and unnormalised data and are shown at the top of the plots for each of the six points. Small *p*-values (e.g. below 0.05) indicate that it is highly likely that the points are significantly different. Note that the raw and normalised CoV differences presented in figure 11 correspond to the J–S divergence for adaptive and constant binning respectively as shown in figures 8–9.

The presented test data shows that the absolute CoV differences of bootstrap-reference distributions can be considered significantly different from the reference-reference distributions in cases of statistically poor acquisitions (data size of 0.01%) for voxels and ROIs. Since the voxels are considerably smaller than ROIs it is more likely for the voxels to exhibit significant differences (see the right plots for the data size of 10%). The cases of significant differences show that the bootstrap estimation of CoV is likely to fail as the absolute difference between bootstrap and reference distributions cannot match the absolute CoV differences of the gold standard. This is especially evident for the data size of 0.01% which for this raclopride phantom measurement it is around $0.3 \times 10^6$ counts. Importantly, this is in the same order of magnitude as the number of true coincidence counts in early time frames emitted from the image slice of the human raclopride scan.

Figure 10. The absolute difference of the mean for the equivalent distributions from stages I and II of resampling (Bootstrap–Reference) and stage I as a reference point (Reference–Reference) plotted against the six statistical levels of the original dataset of the raclopride phantom. The top graphs represent the difference in the mean normalised by the statistical level of the dataset. Average ROI (left) and voxel (right) values were considered at the positions as indicated in figure 1(a).
From the foregoing analysis of the J–S divergence and the difference of the mean and coefficient of variation it is clear that the bootstrap cannot fully replicate the sampling distribution of a statistic of interest for the PET data across the different datasets. The main reason for that seems to be due to a systematic shift of the mean with the bootstrap distributions relative to the reference distributions. This effect has also been observed by other researchers (D’Asseler et al 2003, Groiselle et al 2003, Lartizien et al 2010).

The reason for this effect is due to the fact that only one noisy realisation (measurement or simulation) is used in bootstrap resampling where the estimated mean is representative of that single noisy realisation (Dahlbom 2002, D’Asseler et al 2003) and not of the underlying population. Furthermore, such single realisations are rarely representative of the population but are nevertheless playing the role of the underlying population distribution (because this is all that is available for analysis). Therefore, the estimated mean based on the reference distributions which are sampled from a much richer dataset will be different and more accurate.

Groiselle et al (2003) found that only for very low statistical levels of their dataset the means of the given statistic were significantly different to the means found using Monte Carlo simulations. This would agree with the results shown in figure 10 (bottom) for the unnormalised mean difference where for higher statistical levels the performance of the bootstrap is very close to the reference distributions in absolute terms. Similarly, Lartizien et al (2010)
reported that the mean estimated based on bootstrap resampling of low statistics dataset does not match the mean estimated based on multiple realisations of Monte Carlo simulation. However, the mean as estimated by the bootstrap matched closely the Monte Carlo estimated mean if the bootstrap resampling was based on more than one datasets thus increasing the statistical level of the resampled dataset. Nevertheless, as shown in this work, the difference in the mean can be observed in relative terms for all statistical levels (data size) when the statistical level is normalised out. Therefore, it is important to note that if in a given study higher statistics measurements are used to observe effects of a smaller size, the difference in the mean would start to become evident again as shown using the J–S divergence with adaptive binning and the normalised difference of the mean.

Comparatively, when estimating the variance the bootstrap performs much better achieving good accuracy which has also been reported when estimating noise properties of image reconstruction (Buvat 2002, Dahlbom 2002, D’Asseler et al 2003, Groiselle et al 2003, Groiselle and Glick 2005, Lartizien et al 2010).

It is interesting to note that the shift of the bootstrap distributions in terms of the mean is relative to the statistical richness of the dataset, that is, the distributions become narrower with higher statistical levels and so they become close to each other in absolute terms but in relative terms of the difference to the statistical level it is approximately a constant error. Also, it is worth noting that when performing resampling of the datasets with 100% statistical level the bootstrap still fails in estimating the mean relative to the statistical level (as presented in the J–S divergence with the adaptive binning and in the analysis of the mean and CoV with the normalisation to the data size) although in absolute terms the error decreases as seen with the J–S divergence and constant binning as well as in the analysis of the mean and CoV without such normalisation.

3.1 Optimisation of the assessment of the bootstrap

It was found through a simple linear regression and the analysis of variance (ANOVA, table 1) that for the adaptive binning the numbers of bins $b_n$ has a considerably greater impact on the J–S divergence than the number of bootstrap datasets $K$; whereas for constant binning the statistical level plays a greater role as expected. All results were highly significant ($p \ll 0.01$). The reason for the number of bins $b_n$ (and hence their width) having such a significant impact on the divergence lies in the fact that two different distributions can be made very similar by binning them with a small $b_n$ (wide bins) for which the J–S divergence of the two distributions will tend to be small. However, if $b_n$ is chosen large (narrow bins) the histogram representation of the two distributions will be dissimilar capturing more differences and hence the

### Table 1. Sum of squares attributable to the number of bins $b_n$, bootstrap replications $K$, statistical level $S$ and ROI position in the linear regression of the J–S divergence.

|                | Adaptive Binning | Constant Binning |
|----------------|------------------|------------------|
| SumSq          | DF               | SumSq            | DF               |
| $b_n$          | 1422.5           | 473.4            |
| $K$            | 185.73           | 65706            |
| $S$            | 0.59848          | 129.79           |
| ROI            | 4.0495           | 3.0242           |
| Error          | 956.14           | 845.65           |
|                | 26988            | 26988            |
J–S divergence will be comparatively larger. The impact of $b_n$ on the J–S divergence is even greater for adaptive binning as, in addition to the above effect, the bin size varies according to the statistical level. The statistical level $S$ has a little impact on the divergence for the adaptive binning due to the bin size being changed according to $S$ and hence normalising it out in the analysis of variance. On the other hand, $S$ has a considerable impact for the constant binning since there is no such normalisation and for low statistical levels the distributions tend to be more dispersed than for higher statistical levels (see figures 4 and 5). It is worth noting that although the above relationship may not be linear for the chosen predictors, the ANOVA for the linear regression provides a useful overview of the relationship.

The relationship between the J–S divergence and the number of bins used in the adaptive binning is shown in figure 12. The dispersion of $D_{JS}$ due to other factors such as the statistical level and ROI position is shown using boxplots. The relationship is presented for two cases: (i) the J–S divergence between two independent reference distributions generated in stage I (black boxplots) (this is why $2K$ random samples were generated in stage I); (ii) the J–S divergence between the reference (stage I) and bootstrap (stage II) distribution (grey boxplots). As expected, the figure confirms that the fewer the number of bins are used the more similar will be the distributions (low value for the J–S divergence). In order to better differentiate the reference distributions from the bootstrap distributions the number of bins would be best chosen between 10 and 30. In this work for adaptive binning the number of bins was chosen to be 20. Similarly for constant binning the number can be estimated but since the size of the bins is kept constant the number was chosen to be greater (50) in order to capture the variability across all statistical levels.

In order for a valid assessment of the bootstrap resampling to be conducted the number of bootstrap samples has to be large enough for accurate estimation of the properties of the sampling distribution. The impact of the number of bootstrap replicates (samples) on the J–S divergence is plotted in figure 13. As can be seen there is considerably less change in the J–S divergence when close to 100 bootstrap samples. It is possible that the divergence would drop even more but it would happen at a much larger computational costs. In this study the number of bootstrap replications was chosen $K = 100$ for the whole assessment.
4. Conclusion

The method of bootstrap resampling is gaining momentum due to the lower computational and storage costs today as well as its significant aid for medical image analysis (in research and in clinical practice) in estimating errors, robustness, noise properties, etc of reconstructed images and other estimates obtained from the observed data.

The key aspects of this work are the generation of the reference distributions through the resampling with replacement of the statistically very rich datasets obtained by scanning the specifically designed phantoms. The rich dataset then was resampled into a set of datasets of different statistical levels. It was appropriately assumed that the long acquisitions to be treated as the population, that is, the underlying true distribution from which all other samples can be taken.

If bootstrapping is to be used for characterisation of the distribution of a given statistic derived from PET data (equally it can be any other nuclear medicine imaging modality) there are errors that occur, primarily in terms of the mean but also in terms of the absolute distribution as measured by the J–S divergence. The extent to which the error of the mean will affect further statistical inference will depend on the effect size which is being investigated in a given PET or SPECT study—if the observed effect size is significantly larger than the error (which depends on the statistical richness of the dataset), then the bootstrap resampling will produce acceptably accurate estimates.

In this assessment reference distributions were used as a gold standard to which bootstrap distributions were compared. Based on the comparison, it was found that the bootstrap should be used with caution due to its limitations especially when estimating the first moment (mean). For higher order moments, e.g. the variance, the bootstrap method can be accurate if the datasets are reasonably statistically rich. Hence, in the early time frames of dynamic acquisitions, the bootstrap can fail to estimate the variance accurately due to the poor statistical level in such frames. Therefore, when estimating some aspects of sampling distributions care has to be taken in order to avoid wrong findings supported by the bootstrap used in an inappropriate way.
Acknowledgments

This work was supported by a grant from the UK Engineering and the Physical Sciences Research Council (EPSRC EP/F028695/1). Dr Reader acknowledges the support of the Canada Research Chairs programme.

References

Aubert-Broche B, Evans A C and Collins L 2006 A new improved version of the realistic digital brain phantom NeuroImage 32 138–45
Buvat I 2002 A non-parametric bootstrap approach for analysing the statistical properties of SPECT and PET images Phys. Med. Biol. 47 1761
Comtat C, Bataille F, Michel C, Jones J, Sibomana M, Janeiro L and Trebossen R 2004 OSEM-3D reconstruction strategies for the ECAT HRRT IEEE Nuclear Science Symp. Conf. Record (Rome, Italy, 16–22 October 2004) vol 6, pp 3492–6
Comtat C, Sureau F, Sibomana M, Hong I, Sjoholm N and Trebossen R 2008 Image based resolution modeling for the HRRT OSEM reconstructions software IEEE Nuclear Science Conf. Record (Dresden, Germany, 19–25 October 2008) pp 4120–3
Dahlbom M 2002 Estimation of image noise in PET using the bootstrap method IEEE Trans. Nucl. Sci. 49 2062–6
D’Asseler Y, Groiselle C, Gifford H, Vandenberghe S, Van de Walle R, Lemahieu I and Glick S 2003 Evaluating human observer performance for list mode pet using the bootstrap method IEEE Nuclear Science Symp. Conf. Record (Portland, Oregon, USA, 19–25 October 2003) vol 5, pp 3070–3
de Jong H W A M, van Velden F H P, Buijs F L, Boellaard R and Lammertsma A A 2007 Performance evaluation of the ECAT HRRT: an LSO-LYSO double layer high resolution, high sensitivity scanner Phys. Med. Biol. 52 1505
Efron B and Tibshirani R 1994 An Introduction to the Bootstrap (London: Chapman and Hall)
Groiselle C, D’Asseler Y, Gifford H and Glick S 2003 Performance evaluation of the channelized hotelling observer using bootstrap list-mode PET studies IEEE Nuclear Science Symp. Conf. Record (Portland, Oregon, USA, 19–25 October 2003) vol 4, pp 2511–5
Groiselle C and Glick S 2005 Using the bootstrap method to evaluate image noise for investigation of axial collimation in hybrid PET IEEE Trans. Nucl. Sci. 52 95–101
Habeck C, Foster N L, Perneckzy R, Kurz A, Alexopoulos P, Koepp R A, Drzezga A and Stern Y 2008 Multivariate and univariate neuroimaging biomarkers of alzheimer’s disease NeuroImage 40 1503–15
Haynor D and Woods S 1989 Resampling estimates of precision in emission tomography IEEE Trans. Med. Imaging 8 337–43
Hong I, Chung S T, Kim H K, Kim Y, Son Y D and Cho Z 2007 Ultra fast symmetry and SIMD-based projection-backprojection (SSP) algorithm for 3D PET image reconstruction IEEE Trans. Med. Imaging 26 789–803
Kukreja S L and Gunn R N 2004 Bootstrapped DEPICT for error estimation in PET functional imaging NeuroImage 21 1096–104
Lartizien C, Aubin J B and Buvat I 2010 Comparison of bootstrap resampling methods for 3D PET imaging IEEE Trans. Med. Imaging 29 1442–54
Maitra R 1998 An approximate bootstrap technique for variance estimation in parametric images Med. Image Anal. 2 379–93
Markiewicz P J, Angelis G I, Kotasidis F, Green M, Lionheart W R, Reader A J and Matthews J C 2011 A custom-built PET phantom design for quantitative imaging of printed distributions Phys. Med. Biol. 56 N247
Markiewicz P J, Matthews J C, Declerck J and Herholz K 2009 Robustness of multivariate image analysis assessed by resampling techniques and applied to FDG-PET scans of patients with Alzheimer’s disease NeuroImage 46 472–85
Markiewicz P J, Matthews J C, Declerck J and Herholz K 2011 Robustness of correlations between PCA of FDG-PET scans and biological variables in healthy and demented subjects NeuroImage 56 782–7
Reader A, Julyan P, Williams H, Hastings D and Zweit J 2003 algorithm system modeling by image-space techniques for PET reconstruction *IEEE Trans. Nucl. Sci.* **50** 1392–7

Turkheimer F, Sokoloff L, Bertoldo A, Lucignani G, Reivich M, Jaggi J L and Schmidt K 1998 Estimation of component and parameter distributions in spectral analysis *J. Cereb. Blood Flow Metab.* **18** 1211–22

Verhaeghe J, Gravel P and Reader A J 2010 Task-oriented quantitative image reconstruction in emission tomography for single- and multi-subject studies *Phys. Med. Biol.* **55** 7263