

Comparison of 3D PET data bootstrap resampling methods for numerical observer studies

C. Lartizien, I. Buvat

Abstract—Bootstrap methods have been proposed to estimate the statistical properties of PET and SPECT images by generating multiple statistically equivalent data sets from few data samples. Bootstrap methods might be very helpful for detection performance studies, whose aim is to evaluate detectability based on large series of statistically equivalent images. However, previous reports regarding bootstrap approaches suggest different results. The goal of this work was to compare the accuracy of three bootstrap methods, namely the list-mode based method used by Dahlbom and the parametric and non-parametric sinogram-based methods proposed by Haynor and by Buvat respectively, for predicting the moments of order 1 and 2 (mean and variance) of reconstructed images in 3D PET. We used simulated data generated with the GATE simulation tool and compared the mean and variance images estimated from 100 repeated scans and from series of 100 bootstrap resampled data generated with the three bootstrap methods. Results indicate that the non-parametric bootstrap method by Buvat based on a small number of statistically equivalent data samples seems to correctly estimate the mean of reconstructed images unlike the other two methods based on one original scan only. The comparison of variance images also indicates significant discrepancies between the method by Buvat and the other two bootstrap methods. However, the variance image from the 100 repeated scans, which served as a gold standard, was too noisy to allow us to determine which bootstrap method was the most accurate.

I. INTRODUCTION

Bootstrap methods [1] have been proposed to estimate the statistical properties of PET and SPECT images [2-6]. The bootstrap approach consists in generating multiple statistically equivalent data sets from few samples of data. A first group of bootstrap approaches in PET and SPECT [2, 4] is based on a unique list-mode file or a unique sinogram. The methods based on the list-mode file consist either in randomly choosing events from this file with replacement to produce a number of new list-mode files of the same size as the original file [4], or using a parametric approach which assumes that data from which resampling is performed follows a Poisson distribution [2]. A second group of method uses a set of statistically equivalent sinograms that can be obtained either by simulation or by gated acquisitions [3, 5]. Bootstrap sinograms are then produced either by randomly choosing sinogram bins from the set of original sinograms [3] or by assuming that data follows a Poisson distribution

the parameter of which is estimated from the series of statistically equivalent sinograms.

Haynor et al [2] produced resampled list-mode and sinogram PET data sets from one original file using the parametric bootstrap approach assuming a Poisson distribution. They showed that this method allowed accurate estimation of the variance in the final reconstructed image. More recently, Dahlbom [4] generated list-mode data files from one original 2D list-mode ECAT HR+ PET data set by choosing events at random and with replacement. Considering different reconstruction algorithms, he showed that the standard deviation images derived from the bootstrap list-mode files closely agreed with the standard deviation images derived from repeated scans. D'Asseler et al [6] used the list-mode bootstrap method for 2D simulated PET data and found that the mean background of the bootstrap realizations and the mean background of the noisy realizations were different.

Kim et al [5] used a sinogram-based bootstrap technique for pre-corrected 2D GE Advance PET data in which each distance-angle bin of the resampled sinogram was uniformly drawn from subsets of experimental and statistically equivalent sinograms. They found equivalent mean values between images reconstructed from the resampled data and from repeated acquisitions, but some discrepancies in variance values. Finally, Buvat [3] generated resampled data by randomly choosing one row (instead of one bin) from a set of statistically equivalent original sinograms, in order to potentially account for noise correlation within a row. This method was shown to produce accurate estimation of moments of order 1 to 3 and of the one-dimensional local covariance on simulated SPECT and real 2D PET data.

Bootstrap methods might be very helpful for observer detection performance studies [7], where the aim is to evaluate detectability based on large series of images with and without a signal. Observer studies indeed require a large number of statistically equivalent samples, which is hard to achieve both with experimental data and with simulated data because of the long duration of realistic Monte-Carlo simulations. Some of the previous reports regarding bootstrap approaches applied to PET and SPECT images suggest contradictory results. For instance, D'Asseler et al [6] showed that bootstrapped images could be used to evaluate human observer performance provided that the image background was smoothed, which is not consistent with the fact that the mean values in the reconstructed images are preserved by the bootstrap approach [3-5].

C. L. is with the CREATIS UMR5515 and INSERM U630, Lyon, France (e-mail: carole.lartizien@creatis.insa-lyon.fr)

I.B. is with the INSERM UMR 678 – UPMC lab, Paris, France. (e-mail: buvat@imed.jussieu.fr)

The goal of this work was thus to compare three bootstrap methods for 3D PET data, namely the list-mode based method used by Dahlbom [4], the sinogram-based parametric method proposed by Haynor [2] and the sinogram-based non-parametric method by Buvat [3]. One question we address is whether we can generate accurate resampled PET data series, each of a fixed number N of events, from a unique list-mode file or sinogram of the same number N of events. A positive answer to this question would validate the use of the first group of bootstrap methods described above based on a unique original PET data file. A related question is whether methods from the second group, based on a sub-series of original PET data, produce accurate resampled data, and how many statistically equivalent data sets of N events each are required to accurately estimate 1st and 2nd order moments.

This study uses simulated data obtained with the GATE Monte Carlo simulation tool [8] for a scanner geometry equivalent to the Concorde MicroPET R4 [9].

II. MATERIALS AND METHODS

A. 3D PET Data Monte Carlo Simulations

The GATE Monte Carlo simulation tool used in this study allows for modeling most of the phenomena encountered in PET acquisitions including scattered and random components of the PET signal, dead-time effects and contamination from activity outside the field-of-view [8]. This tool has already been validated for different PET and SPECT scanner geometries.

3D PET data were simulated for a scanner geometry equivalent to that of the small animal Concorde MicroPET R4 scanner [9]. The axial and transverse fields-of-views of this scanner are 78mm and 91mm respectively. For this study, we did not attempt to accurately reproduce the electronic processing responsible for dead-time effects for instance. The phantom geometry consisted of three 2cm-diameter water cylinders of uniform activities located in an 8cm-diameter water cylinder. The ratios between the small cylinder activity and the background activity were 10:1, 15:1 and 20:1 respectively. All cylinders were 2cm-long and the activity in the 8cm-diameter cylinder was 4MBq. The acquisition time was set to 5sec which corresponded to 1.88 million detected coincidence events.

B. Bootstrap Resampling

One hundred and fifty-two statistically equivalent list-mode files (each containing about 1.88 million detected coincidences) were generated with GATE using the LMF list-mode format proposed by the Crystal Clear Collaboration [10].

One of the list-mode files was used to derive 100 resampled list-mode files based on the method proposed by Haynor [2] and used by Dahlbom [4]. In this technique, events are chosen at random and with replacement among the original list-mode events. One event from the original file may thus be selected more than once in a bootstrap data set. This method is referred to as 'Boot_LMF' in the following.

Another list-mode file was first rearranged into 3D sinograms using a program developed by Morel et al within the Crystal Clear Collaboration [10] that will be soon incorporated in the STIR (Software for Tomographic Image Reconstruction) library [11]. This sinogram was then used to derive 100 resampled 3D PET sinogram based on the parametric bootstrap approach proposed by Haynor [1] that consists in drawing each bin of the resampled sinogram from a Poisson distribution with parameter equal to the corresponding bin value in the original sinogram. This method is referred to as 'Boot_Sino_Poisson' in the following.

Fifty of the original list-mode files were rearranged into 3D sinograms. Three series of 100 resampled sinograms each were then sampled from three original sets of 10, 30 and 50 of these sinograms respectively using the method proposed by Buvat [3]. This method consists in randomly choosing each row of a bootstrap sinogram among the $k=\{10,30,50\}$ rows corresponding to the same projection angle in the k original sinograms. This method is referred to as 'Boot_Sino_krep' in the following, where $k=\{10,30,50\}$ is the number of original repeated scans.

The remaining 100 repeated scans were used as a noisy gold standard and are referred to as 'GS' in the following.

C. Data Reconstruction

List mode data were rearranged into 3D sinograms with a span of 3 and a maximum ring difference of 31. The number of coincidence events after rebinning was about 8.2×10^5 . Raw data, i.e., without any correction for attenuation, scatter, random or geometrical effects were reconstructed using the FBP3DRP algorithm provided by the STIR library [11]. FBP was applied with a Ramp filter (cut-off frequency: 0.5 voxel^{-1}).

D. Statistical analysis

Mean and variance of reconstructed images were computed from the five series of 100 bootstrap images (Boot_LMF, Boot_Sino_Poisson, Boot_Sino_10rep, Boot_Sino_30rep, Boot_Sino_50rep) and from the series of 100 repeated scans that can be considered as a noisy gold standard. Profiles through these images were plotted to compare the accuracy of the three bootstrap methods.

A non-parametric Friedman analysis of variance [12] was performed on the mean images estimated from the reference data and from the five bootstrap data series. A similar analysis was conducted on the variance images. The Friedman analysis of variance determines whether the sums of the ranks of each tested method are similar. When the Friedman analysis demonstrated a global significant difference among the different methods, a post-hoc test for multiple comparisons was performed using the method of the smallest significant difference (SDD) [12].

A non-parametric global test for variance homogeneity among the six methods was performed on the mean and variance image using the Box's test [12].

All statistical tests were applied on all non-zero pixels of the central reconstructed planes of the 3D PET image volumes. Two-sided tests were used and a p-value < 0.05 was considered statistically significant.

III. RESULTS

A. Statistical properties of the resampled data

Figure 1 shows the mean images from the 100 repeated scans (Gold Standard) and from the series of 100 resampled data using the two bootstrap methods based on one original file (Boot_LMF and Boot_Sino_Poisson) and the non-parametric sinogram-based approach based on 10 (Boot_Sino_10rep), 30 (Boot_Sino_30rep) and 50 (Boot_Sino_50rep) repeated scans. Figure 2 shows horizontal profiles through these images corresponding to the line indicated in Figure 1. These profiles compare the GS mean image to each of the evaluated bootstrap methods (data for Boot_Sino_30rep are not shown). Visual analysis of Figure 1 indicates that the noise level in the mean images corresponding to Boot_LMF and Boot_Sino_Poisson is higher than the noise level of the GS image. The three images corresponding to non-parametric sinogram-based approach based on 10, 30 and 50 repeated scans better reproduce the noise texture of the gold standard. This qualitative analysis is confirmed by the comparisons of the horizontal profiles in Figure 2 indicating that higher discrepancies are observed with Boot_LMF and Boot_Sino_Poisson than with Boot_Sino_krep.

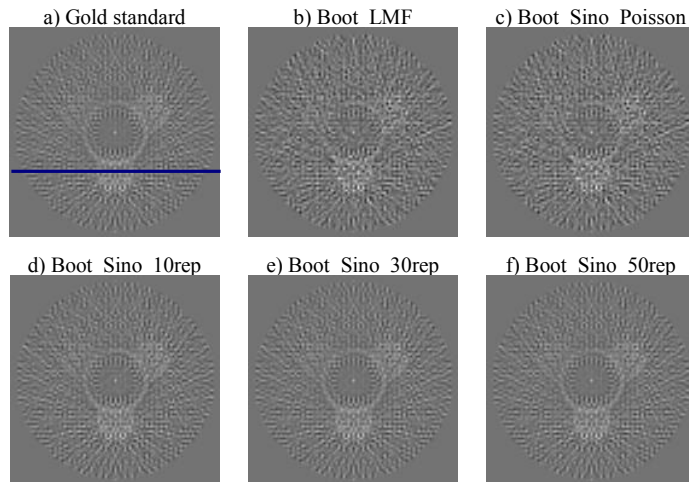


Fig. 1 Mean images computed from 100 images using (a) the repeated scans (b,c) the two bootstrap methods based on one original file and (d-f) the non-parametric sinogram-based approach using 10 (d), 30 (e) and 50 (f) repeated scans.

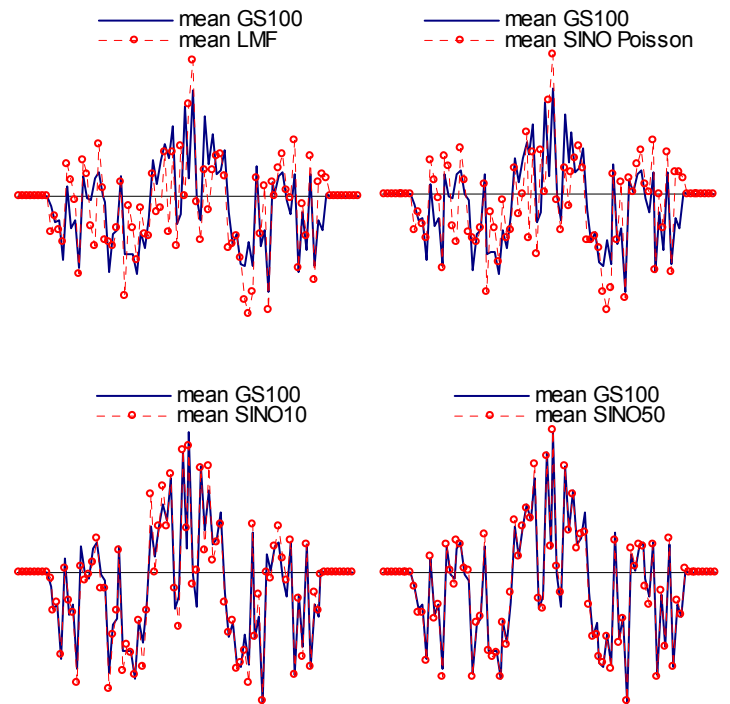


Fig. 2 Horizontal profiles through the mean images at the level of the line reported in Figure 1.

The overall Friedman non-parametric analysis of variance indicated that the mean rank sums of the six tested distributions were not significantly different. The Box's test evaluating the overall variance homogeneity however indicated that the variances of the six mean images were not similar. Figures 1 and 2 suggest that the variance of the Boot_LMF et Boot_Sino_Poisson reconstructed images were higher than that of the GS and Boot_Sino_krep images.

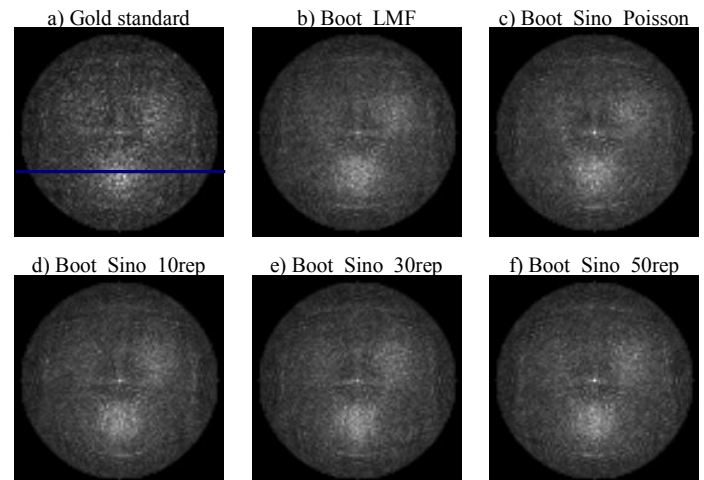


Fig. 3 Variance images computed from 100 images using (a) the repeated scans (b,c) the two bootstrap methods based on one original file and (d-f) the non-parametric sinogram-based approach using 10 (d), 30 (e) and 50 (f) repeated scans.

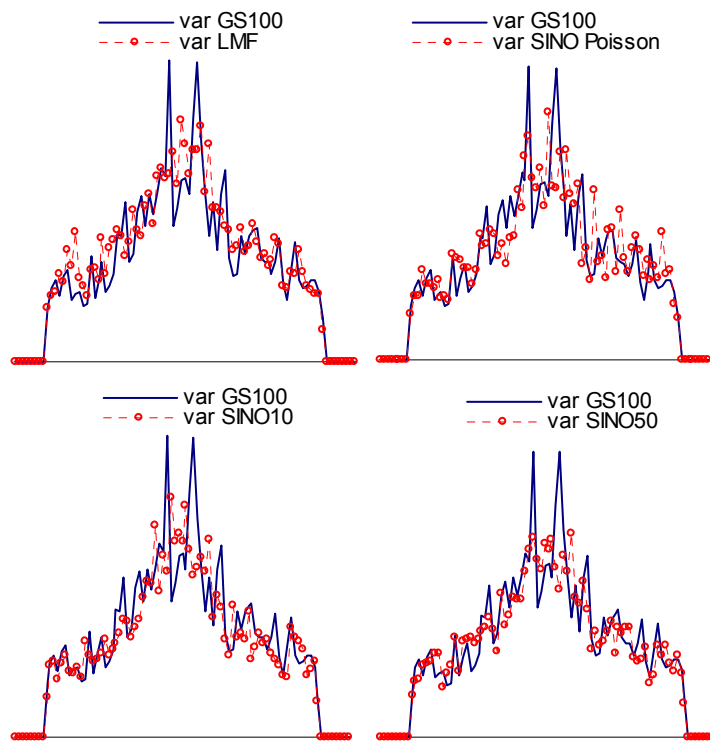


Fig. 3 Horizontal profiles through the variance images at the level of the line reported in Figure 3.

Figure 3 and Figure 4 show the same as Figure 1 and Figure 2 respectively but for variance images. Visual analysis does not indicate significant differences among the bootstrap methods, but they all seem less noisy than the GS variance image.

The overall Friedman non-parametric analysis of variance indicated that the mean rank sums of the six variance image distributions were significantly different. Results from the multiple comparisons based on the smallest significant difference (SSD) are presented in Table 1. The smallest significant difference for the comparison of the 6 images corresponding to a 0.05 confidence level is 369, thus indicating that two images are significantly different if the difference of their rank sums is higher than this SSD. Table 1 shows that all bootstrap methods produced significantly different variance images from the gold standard based on the sum of the ranks of their distribution. Comparison of Boot_LMF and Boot_Sino_Poisson indicated no significant difference. Similarly, comparison between Boot_Sino_30rep and Boot_Sino_50rep indicated no significant difference. The non-parametric sinogram-based approach bootstrap was found to be significantly different from the two bootstrap methods based on one original file (Boot_LMF and Boot_Sino_Poisson).

The Box's test evaluating the overall variance homogeneity yielded a significant conclusion, thus indicating that the variances of the sixth variance images were not similar.

TABLE 1
MULTIPLE COMPARISONS OF THE VARIANCE IMAGES BASED ON THE
SMALLEST SIGNIFICANT DIFFERENCE (SSD)

	GS	LMF	Sino_Poi	Sino_10	Sino_30
LMF	1078.5*				
Sino_Poi	1227*	148.5			
Sino_10	4135*	5214*	5362.5*		
Sino_30	584*	1662.5*	1811*	3551.5*	
Sino_50	615.5*	1842.5*	1842.5*	3520*	31.5

Reported values correspond to the difference of the sums of the ranks of the two images.

* Statistical test is significant when the difference of the sums of the ranks of two conditions is greater than the SSD of 369.

IV. DISCUSSION AND CONCLUSION

This study compares the accuracy of three bootstrap methods for 3D PET data, namely the list-mode based method used by Dahlbom [4] and the sinogram-based methods proposed by Haynor [2] and by Buvat [3], for predicting the variance and mean of reconstructed images in 3D PET.

Visual comparison of the mean images in Figure 1 and plots in Figure 2 indicate that the methods based on one original sample (Boot_LMF and Boot_Sino_Poisson) did not yield accurate estimates of the mean images unlike the method based on a small series of statistically equivalent samples. This result is predictable since these images from Boot_LMF and Boot_Sino_Poisson represent the mean of one noisy realization of the activity distribution, as underlined by D'Asseler et al [6], whereas the mean image estimated from the non-parametric sinogram-based approach better estimates the true activity distribution. The statistical analysis confirmed that the overall distributions of the mean images were significantly different based on the variance homogeneity test.

Visual analysis of the variance images in Figure 3 indicates no significant differences among the three bootstrap methods, whereas the gold standard variance image contains a higher noise level. This result suggests that data series based on 100 repeated scans was not sufficient to accurately estimate the GS variance image. The overall Friedman variance analysis and the variance homogeneity test both indicated a significant difference among the variance images. Multiple comparisons did not allow determining if one of the three bootstrap methods allowed a good prediction of the GS variance, possibly due to the high noise level in this reference image. However, multiple comparisons of the bootstrap methods underlined that the non-parametric sinogram-based method (Boot_Sino for 10, 30 and 50 sinograms) was significantly different from the methods based on one original sample (Boot_LMF and Boot_Sino_Poisson), whereas these two latter methods were not statistically different. Finally, Boot_Sino_10rep was statistically different from Boot_Sino_30rep and Boot_Sino_50rep whereas Boot_Sino_30rep and Boot_Sino_50rep were not significantly different. This suggests that a minimum number of repeated scans (higher than 10 for this specific activity

distribution and scan) is required to accurately estimate the variance.

As a conclusion, the non-parametric bootstrap method by Buvat [3] based on a small number of statistically equivalent data samples seems to correctly estimate the first order moment of reconstructed images unlike the two methods based on one original scan [2,4]. This result has not been confirmed yet by the statistical analysis. The comparison of variance images also indicates significant discrepancies between the method by Buvat and the two methods based on one original sample. The GS variance image from the 100 repeated scans was too noisy to determine which bootstrap method was the most accurate. More scans will be simulated to get a more appropriate gold standard and conclude at the accuracy of the three bootstrap methods for predicting the variance and mean of reconstructed images in 3D PET. Further investigation is also required to improve the statistical analysis.

V. REFERENCES

- [1] B. Efron and R. J. Tibshirani, *An introduction to the Bootstrap*. New York, 1993.
- [2] D. R. Haynor and S. D. Woods, "Resampling estimates of precision in emission tomography," *IEEE Transactions on Medical Imaging*, vol. 8, pp. 337-343, 1990.
- [3] I. Buvat, "A non-parametric bootstrap approach for analysing the statistical properties of SPECT and PET images," *Physics in Medicine and Biology*, vol. 47, pp. 1761-1775, 2002.
- [4] M. Dahlbom, "Estimation of image noise in PET using the Bootstrap method," *IEEE Transactions on Nuclear Science*, vol. 49, pp. 2062-2066, 2002.
- [5] J.-S. Kim, R. S. Miyaoka, R. L. Harrison, P. E. Kinahan, and T. K. Lewellen, "Detectability comparisons of image reconstruction algorithms using the channelized Hotelling observer with bootstrap resampled data," presented at IEEE Nuclear Science Symposium and Medical Imaging Conference, Norfolk, USA, 2002.
- [6] Y. D'Asseler, C. J. Groiselle, H. C. Gifford, S. Vandenberghe, R. Van de Walle, I. L. Lemahieu, and S. J. Glick, "Evaluating numerical observer performance for list mode PET using the bootstrap method," presented at IEEE Nuclear Science Symposium and Medical Imaging Conference, Portland, USA, 2003.
- [7] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," *Proc. Natl. Acad. Sci. USA*, vol. 90, pp. 9758-9765, 1993.
- [8] S. Jan, G. Santin, D. Strul S. Staelens et al, "GATE: a simulation toolkit for PET and SPECT," *Physics in Medicine and Biology*, vol. 49, pp. 4543-4561, 2004.
- [9] C. Knoess, S. Siegel, A. Smith, D. Newport, N. Richerzhagen, A. Winkeler, A. Jacobs, R. N. Goble, R. Graf, K. Wienhard, and W.-D. Heiss, "Performance evaluation of the microPET R4 scanner for rodents," *European Journal of Nuclear Medicine*, vol. 30, pp. 737-747, 2003.
- [10] Crystal Clear Collaboration, (<http://crystalclear.web.cern.ch/crystalclear/>).
- [11] STIR, "<http://stir.hammersmithmanet.com/>."
- [12] S. A. Glantz. "Primer of Biostatistics", Fourth Edition, Mc Graw-Hill, 1997.