Published text:

The use of PET images for radiotherapy treatment planning: an error analysis using radiobiological endpoints

R. Perrin, P.M. Evans, S. Webb and M. Partridge
Joint Department of Physics, Institute of Cancer Research and Royal Marsden NHS Foundation Trust, Downs Road, Sutton, Surrey. SM2 5PT United Kingdom

ABSTRACT

Introduction
There is significant current interest in the use of biological-image guidance in radiotherapy planning. In lung-cancer treatment, tumor motion due to respiration is known to be a limitation. This is particularly true in PET, where image data are collected over a number of minutes. An in-house-developed 4D PET acquisition mode is described and an analysis is presented of the effects of acquisition parameters on reconstructed image quality. The potential impact of the resulting biological image quality on radiotherapy planning is then quantified in terms of tumor control probability.

Method
Data were acquired using a human torso phantom comprised of a hot $^{18}$F-filled spheroidal “tumor” (40 mm diameter) suspended in an air-filled “lung” cylinder and surrounded by a warm $^{18}$F-filled background. Two different sphere-to-background (S/B) ratios were used. The ‘tumor’ was connected to a 3-axis computer-controlled motion stage and could be moved during PET data acquisition. Images were acquired with a range of count statistics, motion-blurring and CT attenuation correction (CTAC) misalignment. The voxel value distributions in the images were assessed using histograms. Four simple models were proposed for the assignment of clonogenic cell density according to voxel value. The impact of image artifacts was then assessed by calculating the tumor control probability (TCP), which is the probability that no clonogenic tumor cell remains after a given dose of radiation. TCP was calculated for a uniform dose distribution in the tumor.
Results
Reduced count statistics and misaligned CTAC images had the most detrimental impact on image fidelity. It was found that in both cases the images became less intense, demonstrated by smaller number of voxels at the maximum values. The maximum TCP difference between images with the least and most noise was 3.4% (S/B = 3) and with weakest and strongest CT misalignment artifacts it was 3.2% (S/B = 10). Motion-blurring only contributed weakly to TCP imprecision at 1.7% (S/B = 10) between best- and worst-case images. However, the model-calculated TCP showed increasing differences from the ground truth as the complexity of model increased (maximum difference ~ 8% (Model 3)), which could be attributed to the partial volume effect.

Conclusions
Based on the results of this study, it is believed that simple techniques of biologically-guided radiotherapy planning for lung cancer should be feasible at intermediate contrast levels (tumor-to-background ~ 10) with the clinically achievable image quality.

Key words: tumor control probability, radiation therapy, treatment planning, PET/CT

I. INTRODUCTION

\(^{18}\)F-FDG-PET has had significant success in diagnosis and staging of lung cancer due to its high sensitivity and specificity\(^1\). It has more recently been indicated as a means to aid gross tumor volume (GTV) delineation. However, the exact correspondence of FDG-PET uptake to histology from lung tissue specimens is difficult to establish\(^2\), and thus the ability of FDG-PET images to image the full spread of malignant cells is still unclear. Nevertheless some groups have reported \(^3,4\) considerable benefit of PET imaging over CT alone for the precision of GTV delineation in interobserver investigations, largely due to non-inclusion of collapsed lung in the GTV.

Several groups have developed methods for GTV segmentation from PET images\(^5-11\). The thresholding techniques involve using a calibration curve of appropriate threshold for full volume-recovery as a function of source-to-background\(^6,9\), or contrast-to-background\(^11\), generated by a sphere phantom. Volume segmentation meets additional
problems in lung-cancer-PET images due to tumor motion during the long-data acquisition times.

Respiratory gating of PET acquisitions (4DPET) in lung cancer has been proposed as a method for reducing motion-blurring. 4DPET has received considerable attention in the literature with several groups documenting its implementation and potential limitations \(^{12-20}\). Limitations include the coregistration of a 4DPET frame to an appropriate CT attenuation-corrected image\(^{16,21,22}\), irregular breathing causing misregistration of data over extended acquisition times\(^{13,18}\), correlation of respiratory surrogate signals to internal motion\(^ {23}\) and decreased signal-to-noise levels in image frames containing a fraction of the image data\(^ {24}\). To improve the latter methods have been proposed to increase the signal-to-noise ratio by elastically transforming raw image data onto a reference frame using prior knowledge of the motion field\(^ {25}\). This seems a promising solution but this method has hitherto only been tested on idealized phantom data and is still in the research domain.

A different paradigm has been suggested for treatment planning based on the premise that tumors are not homogeneous in function. Treatment plans, rather than concentrating on a single volume irradiated uniformly, have an inhomogeneous dose distribution according to intra-tumoral differences in radiosensitivity and tumor burden with biological image guidance\(^ {26-28}\) (dose-painting). Strategies for optimization and evaluation of treatment plans with explicitly inhomogeneous dose distributions using radiobiological objectives have been given in various publications\(^ {28-35}\). Several groups have investigated the technical feasibility of this approach\(^ {36-40}\) in planning studies. However, dose painting with biological images requires accurate quantitation of function across the tumor. Since quantitation on the voxel level is known to be challenging in PET imaging, due to partial-volume effect, motion-blurring, incorrect CT attenuation correction and image noise, this approach must be further validated before application to patients. The relative impact of each source of PET image artifact on image-guided dose-painting has not been systematically performed.

The goal of this study is to assess the feasibility of using 4D PET images for dose painting in lung cancer. Sets of experiments are presented using a human torso phantom. The phantom contained clinically realistic \(^ {18}\)F activity concentrations in a moving lung ‘tumor’ insert and the background volume. 4D images were generated by phase-binning the respiratory-encoded list-mode data. The impact of artifacts on
dose painting was assessed for two different sphere-to-background ratios. This included the effects of residual motion within a phase bin, misregistration of CT data used for attenuation correction and poor count statistics. As a comparative measurement, tumor control probability (TCP), which is the probability that no clonogenic tumor cell remains after a given dose of radiation, was calculated for all artifactual or variable-noise cases. This was done by assigning a biological property, clonogenic cell density, to image voxels. While FDG-PET image values are not likely to directly correlate to clonogenic cell density alone, we used the assumption as a way of comparing the impact of different artifacts on the estimation of TCP. We shall show that the artifacts and noise commonly present in PET images do not alter the TCP significantly.

II. MATERIALS AND METHODS

II.A. Image Data Acquisition

All data presented were acquired using a torso-shaped phantom. It contained a tumor-shaped (spheroidal) bottle (diameter 40 mm, volume 30 cm³) and air cavities for lungs. The tumor-shaped bottle was deliberately designed to be aberrant from a regular sphere so that it was more representative of a lung tumor (see Figure 1b). The air cavities were contained within a sealed torso-shaped cylinder (see Figure 1a). Both the spheroidal tumor-shaped bottle and the torso-shaped cylinder were filled with solutions of ¹⁸FDG at different concentrations. Two sphere-to-cylinder (or sphere-to-background, S/B) ratios were investigated: 3 and 10. These values were chosen since a S/B = 3 represents the limit of contrast below which the tumor would be obscured by background noise, and S/B = 10 is a likely contrast surveyed from patient data. The sphere activity concentrations were 9 kBq cm⁻³ and 16 kBq cm⁻³ respectively. The second activity concentration was measured to give a S/B = 10 with the background having decayed in the phantom since the first phantom set-up with S/B = 3.

The scanner used was a Philips Gemini GS PET/CT system. This consists of a modified Allegro 3D PET camera in line with an Mx8000 EXP Dual-Slice CT Imaging System. The Allegro PET camera is a fully 3D system (no septa) with 576 mm and 180 mm transaxial and axial fields-of-view respectively. A 409-665 keV
energy window was used and the acquisition time was 15 min for a single bed position. The list-mode data were subsequently rebinned into sinograms that were corrected for randoms. They were corrected for scatter using the Philips Single Scatter Simulation algorithm\textsuperscript{42} and reconstructed onto a 144x144 image voxel grid with 4mm slices using a fully 3D iterative reconstruction algorithm (3D-RAMLA). The reconstruction protocol used for all image reconstructions had a ‘blob’ radius (Bessel-function shaped basis functions) of 2.5 pixels, relaxation parameter of 0.045 and one iteration. To investigate residual motion the tumor oscillated in a one-dimensional trajectory with a $\cos^4 \theta$ shape\textsuperscript{43}. Two different amplitudes of motion were tested at 20 and 25 mm. An acquisition with the sphere stationary was also performed as a control. The phantom signaled to the PET system via the cardiac gating interface at the start of each oscillation by generating a gating pulse (TTL). This simulated a gate signal being produced at the start of ‘inhale’ and thus Phase 1 corresponded to the ‘start inhale’ phase (see §II.B). The gating pulses detected by the system at the cardiac gate interface were written into the list-mode data during the PET emission acquisitions. After image data resorting, the rebinned sinograms were reconstructed and CTAC was performed using an attenuation correction image recorded for a stationary ‘tumor’ at centered at the mean position of the pertinent phase. For the CT attenuation correction artifact investigation, several low-dose CT scans (120 kVp, 50 mAs/slice) were acquired at the same table position while the sphere was shifted in its axial position between scans. The displacements ranged up to the maximum translation of 22 mm at 4 mm intervals. The sphere was stationary during the CT scan. An in-house listmode sorter was developed in IDL (ITT Visual Information Solutions). The sorter read in the original list-mode file and used the timing information of the coincident counts relative to the gating pulses to re-sort the counts contained in the list. This enabled

1) the number of events sorted from the original listmode file to be varied, to produce shorter list-mode files with fewer counts as in the count statistics investigation, and/or

2) the number of phase bins to be varied, so lists corresponding to different time lengths of the oscillation could be produced.
II.B. Image Data Post-processing

The data were reprocessed to perform investigations of different artifacts.

(a) Count statistics investigation

When list-mode data are phase-binned, the count statistics will necessarily be reduced in each image that represents a different phase of the respiratory cycle (or oscillation in the present study). To investigate this effect list-mode data were acquired from a stationary sphere in a 16-minute acquisition in which trigger pulses were generated at regular intervals. Using the in-house developed list-mode sorter (§II.A) the list-mode data were retrospectively re-sorted into smaller list-mode sections using the trigger pulses as references. This enabled the single 16 min listmode data (for each S/B) acquisition to serve as a ‘parent’ listmode file that could be used to generate a family of smaller listmode files with fewer count statistics. The smaller list-mode files that contained a range of count statistics between 2 - 25 Mcounts for S/B = 3, and 1 - 10 Mcounts for S/B = 10. Images were reconstructed using a CT image for attenuation correction in which the sphere was stationary in the same position as for the PET emission acquisition.

(b) Residual motion investigation

For a moving object, residual motion will inevitably be present in some phase bins, depending on the motion trajectory and the number of phase bins into which the respiratory cycle is split. To investigate this, list-mode data from a moving sphere (for which the acquisition was described previously) were sorted into a variable number of phase bins (2-5). This enabled images to be reconstructed with a range of known residual motion (13 mm, 14 mm, 15 mm, 18 mm, 20 mm and 25 mm). For residual motion comparison Phase bin 1 of a list-mode file sorted into 2 bins (phase 1 of 2 bins) was reconstructed for both the 20 and 25 mm peak-to-peak amplitude acquisitions. Other residual motion trajectories were also compared: phase 2 of 4 bins, phase 1 of 3 bins and phase 2 of 5 bins (see Figure 2 for details of each motion trajectory). The count statistics per phase bin were kept constant by varying the total number of cycles included in the reconstruction. To ensure that CTAC errors were not present in this set of experiments, an appropriate CT image was used where the position of the sphere matched that in the given PET phase bin.
(c) **CT attenuation correction artifact investigation**

Since the Gemini Dual GS PET/CT system does not have the capability to acquire 4DCT image data to match the gated PET acquisition, the effect of a possible misalignment in the two image data sets was measured experimentally. A series of CT scans of a sphere at different axial positions was used to correct PET emission sinograms acquired with the sphere stationary at a known position. The Philips Gemini GS PET/CT system’s own CT attenuation correction method was used. A range of misalignments \( d_{\text{mis}} \) up to 25 mm was investigated (8 mm, 12 mm, 18 mm, 22 mm, 25 mm), where the misalignment was defined as the distance between the centre of the sphere in a static PET emission acquisition and that in the CT scan. Since the same PET emission sinogram was used for all reconstructions in this image series the count statistics were maintained in all images in the series, and was the same as that used for the residual motion investigation (3 Mcounts per image for S/B = 10; 7 Mcounts per image for S/B = 3).

(d) **Replicate image generation for error analysis**

To estimate the uncertainties in each of the 3 investigations, repeat images were generated by splitting the original list-mode data into equal and consecutive time lengths generating 2-6 replicate list-mode data files per artifact level. For the count statistics investigation all replicates at all levels of count statistics were corrected with a CTAC image for which the sphere was stationary at the start position of the oscillation, matching the stationary position of the spheroid in the PET acquisition. For the residual motion series, each replicate list-mode file was re-sorted to perform phase binning, and then the selected phase (e.g. phase 1 of 3 bins) reconstructed into an image (with the appropriate CTAC image for the spheroid mean position in the phase of the oscillation, as described previously). Each replicate list-mode file after phase-binning for all residual motion levels contained the same level of count statistics. For the CTAC misalignment series 3 replicate listmode files were produced for each contrast acquisition (i.e. S/B = 3 and 10) for which the sphere was stationary. Each replicate list file had the same count statistics. Then each replicate was corrected with a CTAC image that had a given mismatch relative to the PET acquisition sphere.
position. This was repeated for all 5 CTAC images acquired with a range of mismatch.

II.C. Image analysis

To quantify the noise present in the series of sphere images that were reconstructed from sinograms with differing count statistics, a volume of interest (VOI)-based method was used. A spheroidal VOI (of volume = 10 cm$^3$ or around a third of the total spheroid volume) was manually delineated on the 4 central slices of the best-quality PET image of the count statistics investigation (i.e. the lowest level of artifact) using an image processing package. The VOI was used throughout to calculate the average and standard deviation of voxel intensities within the VOI for each of the replicate sphere images with different count statistics. The coefficient of variation (C.V) (standard deviation/mean) was then calculated and used to compare noise-to-signal levels between the images.

For the TCP analysis each image voxel was normalized to the standard deviation of background counts in the torso cavity that contained a warm background. This was necessary, especially in the count statistics investigation, since the inhouse-developed sorter included all of the time stamps from the parent file in the smaller daughter list mode files. As a consequence the scanner, which reports voxel intensity in terms of counts/sec, incorrectly calculated the absolute voxel intensities in the reconstruction step. To measure this quantity a large VOI was manually delineated in the torso (volume = 110 cm$^3$) on the slices containing the spheroid. The same background VOI was utilized for all images to determine the standard deviation. This normalization is implicit in ‘voxel intensity’ (I) used throughout to assign clonogenic cell density.

Since the maximum voxel value is heavily influenced by noise in the image, a ‘smoothed maximum’ image value was calculated by finding the voxels greater than or equal to 90% of the absolute maximum image voxel, and then calculating the mean of these voxel values. This value is subsequently referred to as the image maximum value, $I_{\text{max}}$. The important threshold levels for the subsequent TCP analysis were 70% and 30% of $I_{\text{max}}$. The 70% of the $I_{\text{max}}$ threshold level was chosen arbitrarily to illustrate the effects of uncertainty when defining a volume of increased uptake for dose-boosting. On inspection of the images it was found to cover roughly half of the
volume of the uniform-activity concentration sphere in the lowest artifact image. The 30% of $I_{\text{max}}$ threshold level was found for the S/B=10 image with 3 Mcounts (C.V~0.01) to yield the true volume of the sphere.

II.D. Radiobiological evaluation of segmented images

To evaluate the impact of image artifacts on biological treatment planning with PET images, the PET voxel values were used to infer a biological parameter. For the purposes of this study we use clonogenic cell density, since FDG uptake has been linked to the cell density in some studies\textsuperscript{44-46}. The ensemble of voxels making up the sphere image was designated the ‘tumor’. Four maps of clonogenic cell density were generated per image at the same resolution as the PET image using four sets of rules. These represented a range of planning strategies that may be attempted while making use of biological signal from PET images (see §II.E for further details).

The tumor control probability (TCP) for a uniform irradiation of the apparent ‘tumor’ was calculated for each of the error-prone images. This study therefore highlighted the errors in TCP that would be found simply due to artifactual variation in voxel values to which a biological significance is attached, for an ideal treatment plan. A uniform irradiation also ensured the integral dose stayed constant for all of the planning scenarios used in this investigation. The mechanistic Poisson-linear-quadratic radiobiological model with a Gaussian spread of radiosensitivity parameter $\alpha$, representing the patient-population average of radiosensitivity, was used to predict dose-response\textsuperscript{30}. The parameters used in the model were chosen for a typical lung tumor\textsuperscript{47} and are given in Table 1. Briefly, the method involved calculating TCP for each voxel for a given $\alpha$ and then finding the total TCP for the image under investigation (for a given artifact and a given S/B) by multiplication of the TCPs of all constituent voxels. This was then repeated for all $\alpha_i$ values within 2 standard deviations of the mean $\alpha$, and the total TCP derived from the image found by a weighted average of all of the total image TCPs for given $\alpha_i$, with weighting factors determined by the normal distribution. The TCPs were evaluated for a range of total doses (delivered in 2 Gy fractions) for each image artifact, S/B combination and for each given set of planning rules. Plots were generated of TCP against the level of artifact for each case and each set of planning-rules (see §II.E) for a dose that gave 50% tumor response.
II.E. Treatment planning rules/assumptions concerning clonogenic cell density

All modeling examples were carried out using images of the ‘tumor’. Baseline clonogenic cell density values were proposed based on estimated values from literature of \(10^7 \text{ cm}^{-3}\). Four image-driven planning models were considered for this analysis and are outlined below and in Figure 3. Clonogenic cell density values were assigned to ‘tumor’ background-normalized voxel intensity (see §II.C). These were compared to a ground truth case for which the known uniform activity volume was assigned a single clonogenic cell density. The ‘tumor’ volume was segmented at the 30\% \(I_{\text{max}}\), \(I_{30}\), in all 4 planning models. This threshold level had previously been found to render the correct ‘tumor’ volume for the high-quality image case for our imaging system and with this phantom set-up. The 70\% \(I_{\text{max}}\), \(I_{70}\), was chosen to represent the ‘average’ intensity across the ‘tumor’, and was assigned a clonogenic cell density according to Equation (1).

II.E.1. Ground truth case

The ‘tumor’ was filled with known uniform activity concentrations, and so the “true” clonogenic cell density was, by definition, uniform within the tumor and zero elsewhere. The ‘true’ volume of the tumor was also known from measurement. For the case of the ‘tumor’ with a S/B = 10, a maximum clonogen density of \(10^8 \text{ cm}^{-3}\) was used, based on the assumption that subvolumes of the tumor with a ten-fold increase in \(^{18}\text{FDG}\) uptake had a factor of ten more clonogens than the rest of the tumor. For the case of the tumor with a S/B = 3, a \(\rho_{\text{max}}\) of \(3 \times 10^7 \text{ cm}^{-3}\) was used. This ensures that \(\rho_{\text{max}}\) reflects the different absolute measured activity concentration in the tumor in the two different S/B cases.

II.E.2. Planning assumptions for image-driven cases

Model 1 (see Figure 3(1)): Uniform clonogenic cell density. Each voxel in the ‘tumor’ volume was assigned the same clonogenic cell density, \(\rho_{\text{repr},i}\), which was calculated using Equation (1) using the 0.7\(I_{\text{max},i}\) value for the given image case (\(I_{70,i}\)).

\[
\rho_{\text{repr},i} = \rho_0 + (\rho_{\text{max}} - \rho_0)/(I_{\text{max},S/B=10} - I_{30,S/B=3}) \times (I_{70,i} - I_{30,S/B=3}) - (1)
\]

where,

\(\rho_{\text{repr},i}\) = representative clonogenic cell density assigned in image case i
\[ \rho_{\text{max}} = 10^8 \text{ cm}^{-3} \]
\[ \rho_0 = 10^7 \text{ cm}^{-3} \]

\[ I_{30, \text{S/B}=3} = 30\% \text{ of the maximum I of highest count statistics image of the S/B = 3 case} \]
\[ I_{\text{max}, \text{S/B}=10} = \text{maximum I of the highest quality image of the S/B = 10 case} \]
\[ I_{70,i} = 70\% \text{ of the maximum I of image case } i. \]

**Model 2** (see Figure 3(2)): Two levels of clonogenic cell density were assigned to voxels within given intensity ranges, and thus assumed that there was a subvolume of the ‘tumor’ that had a higher cell density. Voxels with \( I > I_{70} \) were assigned a maximum cell density and voxels with \( I < I_{70} \) were assigned \( \rho_{\text{repr},i} \), as in Case 1 for S/B case \( i \).

In the S/B=10 case the higher-intensity voxels were assigned \( \rho_{\text{max}} = 10^8 \text{ cm}^{-3} \), and in the S/B=3 case in higher-intensity voxel value range were assigned a cell density scaled by the ratio of \( I_{\text{max}} \) for the highest quality images in the S/B=3 and S/B=10 cases.

**Model 3** ((see Figure 3(3)): Clonogenic cell density was assigned on a voxel-wise basis, and thus assumed that the voxel value was linearly proportional to the underlying cell density. The cell density per voxel was calculated according to Equation (2).

\[
\begin{align*}
\rho_{l,m,n,i} &= \rho_0 + (\rho_{\text{max}} - \rho_0) / (I_{\text{max}, \text{S/B}=10} - I_{30, \text{S/B}=3}) \times (I_{l,m,n,i} - I_{30, \text{S/B}=3}) - (2) \\
I_{l,m,n,i} &= \text{I value in matrix position } l,m,n \text{ of image S/B case } i \\
\rho_{l,m,n,i} &= \text{clonogenic cell density assigned to voxel in matrix position } l,m,n \text{ in image S/B case } i
\end{align*}
\]

**Model 4** (see Figure 3(4)): This case assumed that the voxels on the periphery of the ‘tumor’ were influenced by the partial volume effect, but that the remaining voxel intensities reflected real voxel-to-voxel differences in clonogenic cell density. The lower-valued voxels (\( I < I_{70} \)) of the ‘tumor’ were assigned \( \rho_{\text{repr},i} \) as defined in Equation (1). For the remaining voxels clonogenic cell density per voxel was calculated as a linear function of \( I \) according to Equation (2).
II.F. Error analysis

The error in boost volume segmented using thresholding at the 70% $I_{\text{max}} (I_{70})$ level in Models 2 and 4 was evaluated, since this had an impact on the number of voxels assigned to the highest clonogenic cell densities in Models 2, 3 and 4. The standard error (standard deviation / $\sqrt{\text{number of replicates}}$) in volume thresholded over all of the replicates for a given artifact level was used to estimate the error.

To test the difference in volume thresholded at the $I_{70}$ level between images that had a given level of artifact and minimal artifact, the mean and standard deviation of the volumes were used in the 1-sided t-test. The results were assigned statistical significance if the probability that sample means were the same was <0.05.

Secondly the error in TCP calculated for each artifact level was estimated by calculating the standard deviation in TCP over all replicates. Error bars on Figure 4, Figure 5 and Figure 6 represent the standard error of the TCP calculated over all replicates for a given artifact level, since not all data points had the same number of replicates (range 2 – 8).

III. RESULTS

III.A. Comparison of different sources of artifact

TCP differences were used to compare the impact of the artifact on the TCP precision. Since the differences were small the maximum differences found were reported (see Table 2). ‘best-to-worst case’ measurements were a comparison of TCP calculated for images in which the artifact present was minimal (best case) and maximal (worst case). ‘Ground-truth-to-best case’ measurements were a comparison of TCP calculated for the ‘ground truth’ and the images in which the artifact present was minimal. It thus gave a measure of how the image formation process under the best imaging conditions introduced inaccuracies in the image quantitation that impacted on the TCP when applying the different planning models.

(a) Count statistics

The lower count statistics images were less intense and lost edge definition compared to the higher statistics images (Figure 7(a), top row). It was found that the S/B = 3
contrast case required higher count statistics than the S/B =10 case to achieve a similar noise level. $I_{70}$ is particularly pertinent in the TCP analysis of Model 2 and 4, as it is used to delineate the boost volume. Smaller volumes were thresholded at the $I_{70}$ level as the count statistics decreased (Table 3). Furthermore, there was a steeper fall off in volume thresholded at $I_{70}$ with reducing count statistics or increasing coefficient of variation (C.V) at S/B =3 than at S/B = 10 (see Figure 8(a) ). TCP was found to increase slowly with increasing C.V. in Models 2, 3 and 4 for both contrast cases (see Figure 4). For Model 2 the ‘best-to-worst case’ TCP differences for S/B=3 and 10 were 0.011 (2.2%) and 0.008 (1.7%) respectively (see Table 2) demonstrating that the S/B = 3 images had a higher sensitivity to noise due to reduced count statistics. Figure 7(a) shows the line profiles through centre of ‘mass’ of PET image (side panel) and distribution of clonogenic cell density values (bottom row) assigned (under Model 4) to the maximum and minimum count statistics images from the count statistics investigation. They show that the clonogenic cell density values assigned were lower overall with lower count statistics due to the lower intensity of image voxels, which accounts for the increase in TCP (due to a lower total initial clonogen number).

(b) Residual motion

In Figure 7(b) the PET image (in top row) and line profile through centre of ‘mass’ of PET image (side panel) demonstrate the impact of the residual motion artifact (in terms of voxel intensity distribution and profile shape) between the ‘no motion’ image and the residual motion = 25 mm image at S/B = 10. The ‘tumor’ reduced in intensity and had a more peaked profile. This artifactual effect was weaker at S/B = 3. Consequently, larger differences were detected for thresholding at $I_{70}$ with increasing residual motion at S/B = 10 compared to S/B = 3 (Figure 8(b)). For residual motion displacements < 18mm at S/B = 3 the difference in volume thresholded at $I_{70}$ was not statistically significant (t-test p < 0.05), whereas for the S/B = 10 case significant differences in $I_{70}$ volume were found for displacements > 12 mm (see Table 4).

While differences in volume were detectable in the levels of artifact detailed above, TCP was only affected with increasing residual motion in Model 3 at S/B = 10 (see
Figure 5), where there was an increase in TCP between 0 mm and 25 mm of residual motion of 0.008 (1.7%).

(c) **CT attenuation correction error**

CTAC misalignment was shown to strongly influence the number of voxels thresholded at the $I_{70}$ level for misalignments of the CT image >12 mm (see Appendix for Table 5 containing results of t-test for statistical significance). Also the lower contrast case was more strongly affected by the CTAC error than the higher contrast case, with significant differences in $I_{70}$ found for $d_{mis} >8$mm. Figure 7(c) (top row and line profile in side panel) shows the image blurring introduced by a misaligned CTAC image for S/B = 3. An upward trend in TCP was found as $d_{mis}$ increased (see Figure 6). It had a stronger effect on the S/B = 3 images compared to S/B = 10 images as was seen in the $I_{70}$ results. For Model 3, the ‘best-to-worse’ TCP difference was 0.015 (2.7%) and 0.016 (3.2 %) respectively at S/B = 3 and 10 (Table 2) which was comparable to the TCP differences found for the count statistics investigation.

**III.B. Comparison of the ‘best-to-worst case’ image calculations and ‘ground truth’ TCP calculation**

Planning Model 3 assigns clonogenic cell density level on a voxel-wise basis, including voxels on the periphery of the object subject to the partial volume effect (PVE). Thus, making a comparison of the ‘ground truth-to-best case’ and ‘best-to-worst case’ TCP differences for each artifact compares the differences in the image precision due to the PVE with those due to the artifact being investigated (noise, CTAC, motion). The ‘ground truth-to-best case’ TCP differences for Model 3 were 5.9% and 7.9% for S/B = 3 and 10 respectively. For the count statistics series the ‘best-to-worst case’ TCP difference with Model 3 was 3.4% and 2.4% for S/B= 3 and 10 at a C.V. = 0.23 and 0.13 respectively. This shows that the effect of noise on TCP calculation was weaker than the PVE. For CTAC image misalignment the ‘best to worse case’ TCP difference for Model 3 was 2.7% and 3.2% for S/B = 3 and 10 respectively, which again shows that the maximum effect of CTAC misalignment on TCP outcome was weaker that the impact of the partial volume effect. Similarly, the
results also showed that the motion-blurring artifact was much less significant than the PVE.

IV. DISCUSSION

In the present study the assignment of voxel values to a radio-biological property was used to assess the robustness of PET imaging in the guidance of radiobiological treatment planning. To our knowledge, this has not been previously reported. Four types of work-up to a ‘dose-painting’ style approach were proposed where clonogenic cell density was assigned to image voxels depending on the voxel intensity in varying complexity. TCP calculations were performed on clonogenic cell distributions derived from images containing a range of artifacts. The relative errors in TCP derived from images with a range of artifact levels and sources were compared. The investigations reported here cover the range of artifacts likely to occur in patient images.

IV.A. Comparison of different sources of artifact

(a) Count statistics

By using a phantom emulating realistic lung-patient scatter conditions with clinically relevant $^{18}$F activity concentrations, noise propagation from tomographic projections to reconstructed images should also be clinically relevant. It was found that images became less intense and had less distinct edges as noise increased with fewer and fewer counts in the reconstruction. From inspecting Figure 4 it was found that 0.1 was roughly a threshold C.V. for which the TCP started to deviate from the ‘best-case’ image TCP level. From the most naïve model (Model 3) for TCP analysis, that assigned individual cell density values to each voxel, increasing noise from C.V. = 0.10 at 2.4 Mcounts to a C.V. = 0.13 at 1.2 Mcounts at a contrast of S/B = 10 worsened TCP precision (relative to the best case) by 2.2%. Similarly a reduction in C.V from 0.13 to 0.19 was found at S/B = 3 when the count statistics were reduced from 5 Mcounts to 3 Mcounts, yielding a worsening of TCP precision of 1%. If the
volume thresholded at $I_{70}$ for each image is examined (Figure 8), for the same afore-
mentioned reduction in count statistics, at $S/B = 10$ the volume reduced by ~30%, and
at $S/B = 3$ by ~50%. These large differences in volume-thresholded at $I_{70}$ only
translate to small differences in TCP. This weak trend in TCP on count statistics,
while the trend in volume thresholded at $I_{70}$ with count statistics was strong, can be
explained since the TCP is proportional to the negative exponential of the initial
number of clonogenic cells, which is a slowly decreasing function.

This investigation found that the lowest count statistics level that can be tolerated
without introducing detrimental quantities of noise was ~5.5 Mcounts and ~2.5
Mcounts at $S/B = 3$ and $S/B = 10$ respectively, based on the shoulder of the $I_{70}$ against
count statistics plot (Figure 8), where the volume thresholded at $I_{70}$ starts to reduce or
significantly from the baseline. The same cut-off in count statistics could be derived
from the TCP vs. C.V. plot (Figure 5). However the shoulder is less distinct in this
figure due to the weak dependence of TCP on initial clonogen number.

These numbers, while being obtained through a phantom, are representative of the
clinical context, since clinically observed activity concentrations were used in a
patient-sized phantom. However, counts from other parts of the patient’s body that
absorb the tracer, e.g. the myocardium and the liver, are likely to give higher numbers
of count statistics for the same tumor and background activity concentrations.

Acquisitions in the clinical setting where the tumor moves significantly may require
respiratory-correlated acquisition protocols where the list-mode data is sorted into
several phase bins. The acquisition time per gated frame required for good adequate
image quality at low contrast would be around 1.1 min, and this would have to be
multiplied by the number of gated frames required to sufficiently reduce the motion
artifact. Given the results from the residual motion investigation, which showed that
motion-blurring artifacts were not a dominant effect for medium-sized tumors (~30
cm$^3$), priority should be given to acquiring sufficient count statistics to achieve the
best image possible, as opposed to increasing the number of gated frames, especially
at low contrast.

(b) Residual motion

The image became elongated and less intense as the residual motion within a phase
bin increased. The distribution of voxel values shifted downwards in the higher
motion cases as well as the volume being smeared out in the direction of motion. This was shown in the statistical analysis of the volume thresholded at $I_{70}$ where, for the higher contrast case, the volume reduced significantly for residual motion $> 11$ mm ($p < 0.05$) compared to that with no artifact. For the lower contrast case, significant differences were found between images with no artifact and images with artifacts produced by residual motion $> 18$mm. This was not reflected by the TCP differences calculated in Models 1, 2 and 4. Significant differences in TCP (above error margins) were found with Model 3 at motion $> 18$mm for the higher contrast case only (maximum difference of 1.7%).

There could be several reasons for the lower-than-expected impact of motion on the images. Firstly, since a $\cos^4 \theta$ trajectory representative of a lung-cancer-patient breathing pattern (Lujan et al$^{43}$) was used, it contained a ‘rest’ period in which the spheroid did not move significantly, at the start and end of the oscillation. Consequently, for these phases (e.g. phase 1 of 2 bins) the counts forming the image have a significant proportion originating from time points when the spheroid was at rest. Translating this to the patient setting, if the end-expire or start-inspire phase is used for planning guidance, the residual motion, including the effect of ‘rest’ time, will only have a small impact on the voxel value distribution.

Secondly, the CT attenuation correction image was captured for a stationary sphere at the mean position of the oscillation due to CT scanner limitations. Thus voxels into which the sphere moved from its mean position along its trajectory would not have been corrected by matched attenuation correction information. The attenuation coefficient pertaining to these voxels would be for a lower density (air) and thus inadequate correction would have been applied, resulting in artificially lower voxel intensities. This meant that the result of thresholding at the 30% of image maximum was less likely to pick up the excursion of the object due to residual motion.

Thirdly, motion-blurring has also been found in the literature to be a more dominant effect for smaller spheres at a given amplitude$^{49}$ since it is the relative magnitude of the motion amplitude compared to the tumor diameter which governs the degree of blurring. Nevertheless, the size of spheroid used to simulate a tumor in the present study was 40 mm in maximum diameter and thus was of a comparable size to stage T3 (AJCC staging guidelines) lung tumors found in the clinic. Medium-to-large–sized tumors are more pertinent to this study since these are the ones which may contain inhomogeneous subvolumes benefiting from a dose-painting approach. The maximum
amplitude studied here is similar to the maximum amplitude of mobile tumors found in the clinic. Thus the magnitude of artifact found for this particular spheroid and its impact on the TCP end-point is highly relevant to that which may be encountered when performing IMRT planning with PET image guidance.

(c) **CT attenuation correction (CTAC) error**

Attenuation-correcting a PET image with a CT image for which the “tumor” was in a different position was found to have a significant effect on the volume thresholded at the I$_{70}$ level over the clinically feasible range of artifact level investigated. The t-tests for volume thresholded at the I$_{70}$ level showed statistically significant differences in volume between a perfectly-aligned CTAC image and a d$_{\text{mis}}$ > 8 mm and > 18 mm for S/B = 3 and 10 respectively. This was also reflected in the TCP calculated with increasingly misaligned CTAC images (Figure 6). While the TCP differences were small, there was a clear upward trend as d$_{\text{mis}}$ increased. The maximum artifact level at d$_{\text{mis}}$ = 25mm, was the maximum misalignment distance expected for a tumor that had a peak-to-peak amplitude of 25mm, as found in literature in lung tumors. Comparable TCP differences to the count statistics investigation were found between images with maximum and minimum artifact. However, unlike the motion investigation, the S/B = 3 contrast case TCPs increased at a higher rate with d$_{\text{mis}}$ than the S/B = 10 contrast case. This is plausible since the attenuation correction would have boosted voxels into which the “tumor” had moved in the CT acquisition and would not have adequately corrected the voxels belonging to the position of the tumor in the PET acquisition. Therefore the degree of image alteration with CTAC image misalignment depended both on the voxel values present in the background as well as the misalignment in volumes.

Boosting values containing noise decreased the signal-to-noise ratio and thus may cause the object edge to be difficult to pick out from the noise, especially in lower contrast images. In the present study, the air background did not contain activity, whereas in the patient case, lungs would be perfused with a low blood concentration of $^{18}$F-FDG. As a result background counts in the present study are likely to be lower than the patient case. However the air background has a lower density than lung tissue and thus the CTAC factor will be higher for the air density voxels compared to the
lungs-density voxels. This may have led to an exaggerated effect in the present investigation.

IV.B. Comparison of planning models

The differences in TCP precision between the planning models were subtle. Model 1, where all voxels with $I > I_{30}$ were given the same clonogenic cell density, was the least sensitive to noise and artifacts, as expected, since this case simply depends on the placement of the thresholded edge of the object. Model 1 was the safest approach since it implicitly assumes that the imaging system cannot supply sufficiently quantitative images to define accurate contours of increased uptake, and so all voxels are assigned an “average” clonogenic cell density. However the TCP overestimations (due to a smaller initial clonogen number) were 5.1 / 7.1% ($S/B = 3 / 10$) for the best image-driven case compared to the ground truth. Model 2 performed better in this comparison where differences of 3.4 / 5.4% ($S/B = 3 / 10$) were found. This was due to the voxel compartment given by $I > I_{70}$, which was approximately half of the voxels, being assigned the ground truth clonogenic cell density (i.e. the maximum clonogenic cell density = $10^8 \text{ cm}^{-3}$). Model 2 was potentially more sensitive to error, since the $I_{70}$ threshold level was used to delineate a boost volume at the maximum clonogenic cell density. TCP differences in the ‘best-to-worst case’ image comparison of the CTAC investigation were 1.8 – 2.8% for Model 2 compared to 1.2 -1.9% for Model 1 (both contrast cases), showing a difference in sensitivity. These differences in the ‘best-to-worst case’ image comparison are small but are likely to be significant since the overall sensitivity of TCP to artifact was found to be low.

The results from the calculations based on Models 3 and 4 demonstrate a significant limitation of PET-image-guided treatment planning, namely the partial-volume effect in PET images. Models 3 and 4 more closely emulated a scenario working up to the dose-painting by numbers approach (DPBN). Out of all of the planning models, Model 3 was predictably the most sensitive to image noise and artifacts (a maximum TCP difference of 3.4% for the ‘best-to-worse case’ comparison) and gave a 7.9% TCP overestimation in the ‘ground truth-to-best case’ difference ($S/B = 10$). This could be attributed to the effect of the partial-volume effect (PVE) on the voxel intensity. However, avoiding the partial-volumed voxels in Model 4, defined as those of intensity $< I_{70}$, with voxel-wise assignment of clonogenic cell density to the
remainder, interestingly reduced the ‘ground truth-to-best case’ TCP difference to ~6.1% (S/B = 10). The sensitivity to the other artifacts with Model 4 was comparable to Model 2. These results, and those from Models 1 and 2 ‘ground truth-to-best case’ differences, suggested that the PVE causes 4 - 8% error in absolute TCP, depending on the number of clonogenic cell density levels allowed in tumor subvolumes. Error margins should be considered if the absolute TCP is required.

TCP differences found in this modeling study were small compared TCP differences that may be found due to the expected range in the magnitude of clonogenic cell density and intrinsic radiosensitivity (represented by $\alpha$) between tumors. The magnitude of clonogenic cell density is not precisely known and estimates ranging from $10^6$ cm$^{-3}$ - $10^7$ cm$^{-3}$ have been used in literature$^{30,48}$. $\alpha$ has also been found to vary between lung cancer cell lines, and for non-small cell lung cancers the range was found to be 0.18 – 0.89 Gy$^{-1}$ in vitro$^{47}$. With dose of 44 Gy (as given in the present study) the TCP would range between 1.0 - 0.1 respectively for this range of $\alpha$. Compared to this range of TCP, due to possible range of $\alpha$, the TCP differences found in the present study (<0.08) is small.

The choice of threshold level to be $I_{70}$ for the boost volume in Models 2 and 4 was arbitrary in the design of the models discussed in the present study. This threshold level was used to identify a subvolume containing the highest signal voxels. Use of a different threshold level at > 70% $I_{\text{max}}$ may be expected to change the results found in the present study. This was investigated separately but is not reported here. The maximum threshold that could be sensibly used to threshold the high intensity voxels was 90% $I_{\text{max}}$ ($I_{90}$). While the difference in volume between thresholded at $I_{70}$ and $I_{90}$ was large (8-10 cm$^3$), the volume had a similar decreasing trend with artifact level. While the total TCP per image case was ~1% higher (due to a lower initial clonogen number), the ‘best-to-worst case’ differences in TCP did not change by more than 0.7% between TCP calculated in Model 2 utilizing the $I_{70}$ or $I_{90}$ threshold. This shows that the modeling methods used to evaluate artifact impact on TCP are not sensitive to choice of threshold level.

Out of the four the planning models used in the present study for assignment of clonogenic cell density, only methods similar to Model 2 have been planned and delivered in patient cases reported in the literature$^{51,52}$. This is probably due to the knowledge that quantitative PET has its limitations and it is only justifiable to use the
simplest image-guidance method conceivable. The voxel-wise assignment in Model 3 has to our knowledge only been performed in two treatment planning studies \(^{36,38}\) to assess the dosimetric feasibility of the technique. The impact of artifact was not considered in these studies, and the results of the present study should help to inform the best practice for use of PET images in IMRT treatment planning in clinical cases.

To summarize, the investigations reported here show that the range of artifacts likely to occur in patient images only produce small differences in TCP as a function of clonogenic cell density, and so the impact would be small when using TCP in a biological objective function used to optimize plans in inverse-planning techniques. However, all of the artifacts investigated had a strong influence on the size and contour of the high-intensity voxel volume for their clinically-plausible maximum artifact level. Clearly with incorrect delineation resulting in a geographic miss of clonogenic tumor cells, the TCP would be zero. While this was beyond the scope of this paper (i.e. the geographic location of the spheroid image), it should be the topic of further investigations.

### IV.C. Limitations

One possible limitation of this study was that a uniform-activity-concentration object was employed to generate biological images, whereas in the end-application of this technique tumors will have variable intra-tumoral radiosensitivity or tumor burden as evidenced by variable image voxel intensity. In a tumor, subvolumes with higher uptake that are sized below 3 times the image resolution (3 × ~7 mm) would be subject to the PVE, and this effect would render them indistinguishable from larger volumes of a relatively lower uptake not subject to the PVE. If the rules applied in model 4 were applied to this scenario, then the smaller, higher-uptake subvolumes, giving a lower image intensity (< 70% of the maximum image intensity), would be assigned an artificially low radiosensitivity and thus not dose-boosted appropriately. Nevertheless, the impact of artifactual variation in voxel intensity could be evaluated in the present study based on the simplest ground truth case. Thus, the findings here are relevant when considering how much a TCP calculation may be in error due to known PET imaging limitations, without the complex interaction of the sources of error with a variable activity concentration ground truth.
A further limitation is that differences in FDG uptake are likely to be caused by a complex combination of biological changes in the cell environment resulting in changes in the rate of metabolism of glucose\textsuperscript{53-56}. Other possible indications are hypoxia and increased proliferation, which will be the subject of future investigation. Hypoxia may result in larger differences in TCP between best-to-worst case images, since oxygenation modifies radiosensitivity by up to a factor of 3. Nevertheless, a linear relationship between image intensity and clonogenic cell density is a good first approximation to evaluate the imprecision of TCP calculations that may result from treatment planning with PET images.

Finally, it was assumed that the radiation field covered all of the thresholded volume at I\textsubscript{30}. Further work should include an evaluation of the penalty in TCP due to imprecision of the spatial position of the tumor relative to a prescribed treatment field.

V. CONCLUSION

This study considered artifacts in 4D respiratory-correlated PET images due to i) poor count statistics, ii) mis-registration of the attenuation correction CT image and iii) residual motion within a phase bin. The effects of these artifacts on radiotherapy planning with biological objectives were quantified by calculating TCP assuming a varying clonogenic cell density. While the relative effects of these artifacts were small, the partial volume effect produced large absolute differences in TCP from the ground truth calculation (up to 8%), which suggests that this approach should not be used without making steps to reduce this artifact.

Poor count statistics and CT attenuation correction image mismatch are likely to be the dominant sources TCP error, while residual motion at tumor sizes that may benefit from dose-painting led to negligible TCP differences. In terms of the impact of artifact on the volume of the high intensity compartment, amplitudes greater than a third of the tumor diameter were shown to reduce this volume. CT attenuation correction artifacts also led to significant differences in high-voxel-value compartment volume for misalignment between PET and CT greater than a third of the tumor diameter. Assigning clonogenic cell density on a voxel-by-voxel basis (Model 3) gave the largest TCP difference (out of all of the models), and this was 3.2\% between images with severe CTAC artifact and minimal artifact and 3.2\%
between images with low and high noise (3.2%). This difference is small compared to TCP differences that may be found due to the expected range in the magnitude of clonogenic cell density and intrinsic radiosensitivity between tumors. Overall, the impact of image artifacts on TCP (with variable clonogenic cell density) was small. However, techniques involving a voxel-by-voxel parameter assignment give large errors in TCP due to the partial volume effect. Based on these results, we believe that simple techniques of biologically-guided radiotherapy planning with PET image guidance for lung cancer should be feasible with the current clinically achievable image quality.

Acknowledgements We acknowledge NHS funding to the NIHR Biomedical Research Centre and the EPSRC for studentship funding.

GLOSSARY

3D-RAMLA – 3D Row Action Maximum Likelihood Algorithm

TTL – Transistor-transistor Logic. A TTL input signal is defined as "low" when between 0V and 0.8V with respect to the ground terminal, and "high" when between 2.2V and 5 V.

IDL – Interactive Data Language™, ITT Visual Information Solutions

CTAC – CT attenuation correction

d_{mis} - the distance or misalignment between the centre of the sphere in a static PET emission acquisition and that in the CT scan
**Residual motion** - the distance swept out by the tumour within the time of the phase bin

**phase X of Y bins** - phase bin number X of a list-mode file sorted into Y bins

**C.V** - coefficient of variation = standard deviation/mean)

**S/B** - sphere-to-background ratios

**I_{max}** – mean of voxels with values > 90% of the maximum voxel value in the ‘tumor’ image

**I_{70}** – Voxel value at 70% of **I_{max}**

**I_{30}** - Voxel value at 30% of **I_{max}**

**Planning model** – set of rules or assumptions concerning assignment of clonogenic cell density to image voxels.

**Contrast case** – image for which the S/B was either 3 or 10

‘**Best-to-worst case**’ – Comparison of TCP calculated for images in which the artifact present was minimal (best case) and maximal (worst case).
‘Ground truth case’ – Calculation of TCP for a uniform clonogenic cell density ‘tumor’ of volume (number of voxels) equivalent to that of the spheroid used to simulate the ‘tumor’

‘Ground truth-to-best case’ - Comparison of TCP calculated for the ‘ground truth’ and the images in which the artifact present was minimal.
FIGURES AND TABLES

Figure 1: a) Features of the torso-shaped phantom used in the experiments described b) Tumor-shaped bottle used in the investigations within the torso-shaped phantom.

Motion of ‘tumor’ - $(\cos \theta)^d$ along axis; time period = 5.3 s – up to 2.5 cm amplitude

‘tumor’ = spheroid bottle, FDG-filled; 30 cm$^3$, 4 cm diameter

Low activity concentration in shell
Figure 2: The motion trajectories pertaining to each of the residual motion magnitudes investigated. Variable numbers of phase bins, with the list-mode acquisitions of 20 mm and 25 mm peak-to-peak amplitude oscillations, were used to produce the residual motion. 1) Motion = 12 mm from Phase 1 of 3 bins (20 mm amplitude), 2) Motion = 13 mm from Phase 2 of 5 bins (20 mm amplitude), 3) Motion = 15 mm from Phase 2 of 4 bins (20 mm amplitude), 4) Motion = 18 mm from Phase 2 of 4 bins (25 mm amplitude), 5) Motion = 20 mm from Phase 1 of 2 bins (20 mm amplitude), 6) Motion = 25 mm from Phase 1 of 2 bins (25 mm amplitude)
Figure 3: Assignment of clonogenic cell density as a function of voxel intensity: Ground truth, clonogenic cell density versus ‘tumor’ volume (V) (top-center), Model 1 (top-left), Model 2 (top-right), Model 3 (bottom-left) and Model 4 (bottom-right). $I_{\text{max}} = \text{smoothed maximum (mean of voxels with } I > 0.90 \times \text{maximum voxel intensity)}$, $I_{30} = 30\% I_{\text{max}}$, $I_{70} = 70\% I_{\text{max}}$. $\rho_{\text{repr}}$ was the clonogenic cell density calculated using the $I_{70}$ value of the image. $\rho_{\text{max}} = 6.4 \times 10^5$ per PET voxel ($1 \times 10^8 \text{ cm}^3$), $\rho_0 = 6.4 \times 10^4$ per PET voxel ($1 \times 10^7 \text{ cm}^3$), and for the ground truth case, $\rho_{\text{max}, S/B = 10} = \rho_{\text{max}}$ and $\rho_{\text{max}, S/B = 3} = \rho_{\text{max}, S/B = 10} / 3.8$, since this was the exact relative activity concentration measured from assay.

**Ground truth**

![Graph showing clonogenic cell density versus tumor volume for ground truth, Model 1, Model 2, Model 3, and Model 4.](image)

**Model 1**

![Graph showing cell density for Model 1.](image)

**Model 2**

![Graph showing cell density for Model 2.](image)

**Model 3**

![Graph showing cell density for Model 3.](image)

**Model 4**

![Graph showing cell density for Model 4.](image)
Table 1: Radiobiological parameters using in the modeling study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Dose</td>
<td>44 / 46 Gy for S/B=3 / S/B=10</td>
</tr>
<tr>
<td>Fraction size</td>
<td>2 Gy</td>
</tr>
<tr>
<td>Maximum clonogenic cell density</td>
<td>$10^8$ cm$^{-3}$</td>
</tr>
<tr>
<td>Minimum clonogenic cell density</td>
<td>$10^7$ cm$^{-3}$</td>
</tr>
<tr>
<td>Radiosensitivity parameter, $\alpha$</td>
<td>0.36 Gy$^{-1}$</td>
</tr>
<tr>
<td>Standard deviation of $\alpha$, $\sigma_\alpha$</td>
<td>0.11 Gy$^{-1}$</td>
</tr>
<tr>
<td>Radiosensitivity parameter, $\beta$</td>
<td>0.036 Gy$^{-2}$</td>
</tr>
</tbody>
</table>

Figure 4: TCP calculated for image-guided plans with a uniform dose of 44 Gy using variable count statistics images. ‘Error bars = SEM’ indicates that Standard Error of the Mean is the error bar length.
Figure 5: TCP calculated for image-guided plans with a uniform dose of 44 Gy using images containing motion-blurring artifact. Images were reconstructed from data collected at different phases 20 mm and 25 mm peak-to-peak amplitude oscillations. Variable bin numbers (2-5) were used to vary the residual motion per bin. ‘Error bars = SEM’ indicates that Standard Error of the Mean is the error bar length.
Figure 6: TCP calculated for image-guided plans with a uniform dose of 44 Gy using increasingly misaligned CT attenuation correction images. ‘Error bars = SEM’ indicates that Standard Error of the Mean is the error bar length.
Figure 7: PET image (grayscale - top), clonogenic cell density maps after applying Model 4 (bottom) and line profiles through centre of ‘mass’ of PET image (side panel) at the minimum and maximum parameter investigated for: a) the count statistics experiment with S/B = 10 case, b) the residual motion experiment for the S/B = 10 case and c) the CTAC error series for S/B = 3 case. \( N_{cs} \) = total number of counts in the image.
**Figure 8:** Volume thresholded at 70% maximum voxel intensity ($I_{70}$) as a function of artifact level as follows: a) $I_{70}$ with increasing noise by reducing the count statistics; b) $I_{70}$ with increasing residual motion in the image; c) $I_{70}$ with increasing mismatch between spheroid position in the PET image and that in the CT attenuation correction image (measured in terms of $d_{mis}$, which is the distance between the centre of the real object in the PET and CT acquisitions).
Table 2: Absolute and percentage differences in TCP between a) Ground truth and the best quality image of the count statistics error series (reconstructed counts = 25 and 10 Mcounts, S/B=3 and 10 respectively, from the stationary acquisition), and b) the best and worst quality images in the given artifact investigation, for all four models, for each artifact and S/B contrast case (% difference in brackets)

### a) Ground Truth to ‘Best Case’ image TCP differences

<table>
<thead>
<tr>
<th>S/B</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0256±0.0006</td>
<td>0.0168±0.0004</td>
<td>0.0298±0.0003</td>
<td>0.0206±0.0003</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(5.1±0.10)</td>
<td>(3.4±0.08)</td>
<td>(5.9±0.06)</td>
<td>(4.1±0.06)</td>
</tr>
<tr>
<td>10</td>
<td>0.0374±0.0004</td>
<td>0.0251±0.0003</td>
<td>0.0370±0.0000</td>
<td>0.0288±0.0001</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(7.0±0.07)</td>
<td>(5.4±0.06)</td>
<td>(7.9±0.00)</td>
<td>(6.1±0.02)</td>
</tr>
</tbody>
</table>

### b) ‘Best Case’ image to ‘Worst Case’ image TCP differences

#### i) Count Statistics image series

<table>
<thead>
<tr>
<th>S/B</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.005±0.001</td>
<td>0.011±0.001</td>
<td>0.018±0.002</td>
<td>0.009±0.001</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(1.0±0.2)</td>
<td>(2.2±0.1)</td>
<td>(3.4±0.4)</td>
<td>(1.8±0.2)</td>
</tr>
<tr>
<td>10</td>
<td>0.005±0.001</td>
<td>0.008±0.000</td>
<td>0.013±0.000</td>
<td>0.010±0.001</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(1.0±0.2)</td>
<td>(1.7±0.0)</td>
<td>(2.4±0.0)</td>
<td>(2.0±0.2)</td>
</tr>
</tbody>
</table>

#### ii) Motion image series

<table>
<thead>
<tr>
<th>S/B</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.002±0.002</td>
<td>0.001±0.003</td>
<td>0.003±0.003</td>
<td>0.004±0.002</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(0.3±0.3)</td>
<td>(0.1±0.3)</td>
<td>(0.6±0.6)</td>
<td>(0.8±0.4)</td>
</tr>
<tr>
<td>10</td>
<td>0.001±0.001</td>
<td>0.004±0.002</td>
<td>0.008±0.003</td>
<td>0.002±0.002</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(0.2±0.2)</td>
<td>(0.8±0.4)</td>
<td>(1.7±0.2)</td>
<td>(0.5±0.5)</td>
</tr>
</tbody>
</table>

#### iii) CTAC misalignment image series

<table>
<thead>
<tr>
<th>S/B</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.006±0.004</td>
<td>0.009±0.004</td>
<td>0.015±0.005</td>
<td>0.007±0.004</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(1.2±0.5)</td>
<td>(1.8±0.8)</td>
<td>(2.7±0.9)</td>
<td>(1.4±0.8)</td>
</tr>
<tr>
<td>10</td>
<td>0.010±0.002</td>
<td>0.013±0.002</td>
<td>0.016±0.001</td>
<td>0.012±0.002</td>
</tr>
<tr>
<td>(%diff)</td>
<td>(1.9±0.4)</td>
<td>(2.8±0.4)</td>
<td>(3.2±0.2)</td>
<td>(2.5±0.4)</td>
</tr>
</tbody>
</table>


Appendix 1: Volumetric data and t-test results for each artifact thresholded at $I_{70}$

Table 3: Volumes thresholded at $I_{70}$ (±standard error) for S/B = 3 and 10 for different count statistics images with the t-test results for significant differences between the maximal count statistics image and lower count statistics images. The coefficient of variation (C.V.) was also quoted for each image in the count statistics investigation. Some points had no replicates over which to calculate a standard deviation and thus the error is quoted ‘n.k’ i.e. not known.

<table>
<thead>
<tr>
<th>Count statistics per image (Mcounts)</th>
<th>25.8</th>
<th>12.23</th>
<th>6.23</th>
<th>3.14</th>
<th>2.06</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. V.</td>
<td>0.08</td>
<td>0.01</td>
<td>0.12</td>
<td>0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>$I_{70}$ volume (cm$^3$)</td>
<td>13.35±n.k</td>
<td>10.89±0.00</td>
<td>10.41±0.65</td>
<td>4.63±0.71</td>
<td>2.08±0.31</td>
</tr>
<tr>
<td>($p = 1x10^{-6}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count statistics per image (Mcounts)</th>
<th>9.63</th>
<th>4.37</th>
<th>3.05</th>
<th>2.18</th>
<th>1.06</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. V.</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>$I_{70}$ volume (cm$^3$)</td>
<td>17.04±n.k</td>
<td>15.60±0.22</td>
<td>15.02±0.58</td>
<td>13.39±1.55</td>
<td>10.54±0.48</td>
</tr>
<tr>
<td>($p = 1x10^{-3}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Volumes thresholded at I$_{70}$ (±standard deviation, St.D) for S/B = 3 and 10 for different levels of residual motion with the t-test results for significant differences between no artifact and each artifact level.

<table>
<thead>
<tr>
<th>Motion (mm)</th>
<th>0</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>18</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/B = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I$_{70}$ volume ±St.D (cm$^3$)</td>
<td>10.10±1.17</td>
<td>9.16±1.07</td>
<td>10.19±1.20</td>
<td>9.53±2.36</td>
<td>10.25±2.28</td>
<td>5.90±2.90</td>
<td>6.75±3.19</td>
</tr>
<tr>
<td>p-value</td>
<td>0.09</td>
<td>0.45</td>
<td>0.30</td>
<td>0.45</td>
<td>0.004</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Motion (mm)</th>
<th>0</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>18</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/B = 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I$_{70}$ volume ±St.D (cm$^3$)</td>
<td>14.61±1.09</td>
<td>13.66±1.79</td>
<td>11.84±1.44</td>
<td>11.80±1.78</td>
<td>11.21±2.11</td>
<td>8.86±2.06</td>
<td>7.99±0.43</td>
</tr>
<tr>
<td>p-value</td>
<td>0.16</td>
<td>0.007</td>
<td>0.01</td>
<td>0.01</td>
<td>1x10$^{-4}$</td>
<td>4x10$^{-5}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Volumes thresholded at I$_{70}$ (± a standard deviation) for S/B = 3 and 10 for different levels of CTAC misalignment with the t-test results for significant differences between no artifact and each artifact level.

<table>
<thead>
<tr>
<th>CTAC misalignment (mm)</th>
<th>0</th>
<th>8</th>
<th>12</th>
<th>18</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/B = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I$_{70}$ volume ±St.D (cm$^3$)</td>
<td>9.72±0.79</td>
<td>9.93±1.30</td>
<td>9.29±2.04</td>
<td>7.39±1.77</td>
<td>6.34±1.06</td>
<td>5.27±0.44</td>
</tr>
<tr>
<td>p-value</td>
<td>0.22</td>
<td>0.02</td>
<td>2x10$^{-4}$</td>
<td>2x10$^{-4}$</td>
<td>1x10$^{-5}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CTAC misalignment (mm)</th>
<th>0</th>
<th>8</th>
<th>12</th>
<th>18</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/B = 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I$_{70}$ volume ±St.D (cm$^3$)</td>
<td>14.61±1.09</td>
<td>13.92±1.67</td>
<td>12.92±1.17</td>
<td>10.55±0.58</td>
<td>9.12±0.26</td>
<td>7.50±0.28</td>
</tr>
<tr>
<td>p-value</td>
<td>0.37</td>
<td>0.32</td>
<td>0.007</td>
<td>5x10$^{-5}$</td>
<td>1x10$^{-7}$</td>
<td></td>
</tr>
</tbody>
</table>