An approach to generate synthetic 4DCT datasets to benchmark
Mid-Position implementations

Firass Ghareeb a,1, Djamal Boukerroui b,1, Joep Stroom a,*, Edward Jackson c, Mariana Pereira a,
Mark Gooding b, Carlo Greco a

a Champalimaud Foundation, Department of Radiation Oncology, Lisbon, Portugal
b Mirada Medical Ltd., Oxford, United Kingdom
c Netherlands Cancer Institute, Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Mid-Position
Evaluation
Synthetic 4DCT
Deformable image registration

ABSTRACT

Purpose: The Mid-Position image is constructed from 4DCT data using Deformable Image Registration and can be used as planning CT with reduced PTV volumes. 4DCT datasets currently-available for testing do not provide the corresponding Mid-P images of the datasets. This work describes an approach to generate human-like synthetic 4DCT datasets with the associated Mid-P images that can be used as reference in the validation of Mid-P implementations.

Methods: Twenty synthetic 4DCT datasets with the associated reference Mid-P images were generated from twenty clinical 4DCT datasets. Per clinical dataset, an anchor phase was registered to the remaining nine phases to obtain nine Deformable Vector Fields (DVFs). These DVFs were used to warp the anchor phase in order to generate the synthetic 4DCT dataset and the corresponding reference Mid-P image. Similarly, a reference 4D tumor mask dataset and its corresponding Mid-P tumor mask were generated. The generated synthetic datasets and masks were used to compare and benchmark the outcomes of three independent Mid-P implementations using a set of experiments.

Results: The Mid-P images constructed by the three implementations showed high similarity scores when compared to the reference Mid-P images except for one noisy dataset. The biggest difference in the estimated motion amplitudes (-2.6 mm) was noticed in the Superior-Inferior direction. The statistical analysis showed no significant differences among the three implementations for all experiments.

Conclusion: The described approach and the proposed experiments provide an independent method that can be used in the validation of any Mid-P implementation being developed.

1. Introduction

Radiotherapy of lung cancer patients is challenging due to the presence of respiratory motion which induces geometrical uncertainties of both the target and normal organs in the thoracoabdominal region [1]. Recent advances in treatment delivery techniques such as stereotactic body radiotherapy (SBRT) have made it possible to deliver a more precise dose to the tumor while sparing the surrounding healthy tissue [2]. The state-of-the-art SBRT technique includes procedures to determine organ motion, such as the acquisition of four-dimensional computed tomography (4DCT) scans.

Two approaches are currently in use to account for 4DCT motion offline. The first is the internal target volume (ITV) approach based on defining the ITV that encompasses the clinical target volume (CTV) position in all phases of 4DCT, then adding extra margins to account for non-respiratory geometric uncertainties when expanding from ITV to planning target volume (PTV) [3]. The ITV approach (incorrectly) treats respiratory motion as systematic errors [3]. Although this approach has a high tumor control [4], it leads to too large margins to cover for the geometric uncertainties (especially for patients with considerable respiratory motion), causing more healthy tissue to be unnecessarily exposed to high radiation doses [5].

The other approach is to use either the Mid-ventilation (Mid-V) or Mid-position (Mid-P) strategies. Both are based on defining the gross
tumor volume (GTV), or CTV, on a single 3DCT image. In the case of Mid-V, the 4D phase closest to the time-weighted mean position of anatomy during the respiratory cycle [6] is chosen for treatment planning. Mid-P (which is a refinement of Mid-V with reduced noise and image artifacts) is constructed using deformable image registration (DIR) on 4DCT data to estimate a time-weighted mean position of the anatomy during the breathing cycle. In addition, the deformation vector fields (DVFs) resulting from the registration can be used to estimate the motion amplitudes for relevant regions of interest (ROIs) [7].

In both Mid-P and Mid-V, respiratory motion is taken into consideration as a random error which is added quadratically to other random errors when calculating the PTV margins. This means that the Mid-P and Mid-V approach yields smaller PTV volumes than the ITV approach, leading to more sparing of healthy tissue [8,9]. For instance, the Mid-P or Mid-V approach reduced the PTV volume (compared with the ITV approach) by an average of 23 %, 34 %, and 13.9 % respectively for the lung SBRT [10], liver SBRT [11], and pancreatic cancer RT [12]. Some recent work has been done to exploit the reduced PTV advantage of the Mid-P approach for MR-Liner planning [13,14].

The Mid-V approach has demonstrated an excellent two-year local tumor control of 98 % [5], but the results of a survey of practice conducted at the British thoracic oncology conference in 2016 amongst fifty radiotherapy centers showed that only 6 % of centers used the Mid-V or Mid-P approach [15]. One reason is that software implementations to create the Mid-P/Mid-V images are limited to a small number of centers, since (to the best of available knowledge) they are not yet commercially available. Given the advantages Mid-P (and Mid-V) provides without any extra complication of the workload, an increased number of institutions are interested in developing their own in-house Mid-P implementations.

The translation of any new image processing research software into a clinical application is usually limited by a lack of established validation methodology, and since DIR algorithms are based on complex mathematical models, there is no guarantee that anatomical changes between phases represented by the DVFs reflect the actual ones accurately [16]. For these reasons, a thorough testing process is required to assess the performance of any Mid-P implementation, and special attention should also be paid to the choice of the reference data to be used.

Several 4DCT publicly or commercially available datasets to test and benchmark DIR algorithms, such as the point-validated pixel-based breathing thorax model (POPI model) [17] and the DIR-Lab Thoracic 4DCT model [18] that can be used to validate the accuracy of the constructed Mid-P CT. Although the accuracy of the DVFs between phases for these datasets was verified by experts using anatomical landmarks, these data provide a limited number of cases, they are subjected to human anatomy [19,21] and thus lack of sufficient realism to reproduce some of the features seen in human anatomy [19,21]. Virtual phantom datasets can be created starting from real datasets and deformed using known simulated DVF to serve as “ground-truth” [16]. Such an approach has been used to benchmark commercially available DIR software on head-and-neck cancer patients [22,23].

However, none of the available testing datasets provides the associated Mid-P CT images. Indeed, the assessment of a Mid-P implementation should not be restricted to an accuracy assessment of the used DIR algorithm, but should consider the whole process of the Mid-P image construction and its associated motion amplitudes information. For example, the accuracy assessment of the DIR algorithm focuses on the motion estimation and does not include image resampling when applying the registration [22,23]. Thus, it is difficult to assess the correctness of a Mid-P implementation end-to-end because the true output is unknown. To overcome these issues, this work proposes an approach to create a synthetic 4DCT dataset (similar to clinical data, also referred to by virtual) with the associated Mid-P images that can be used as reference, using pre-defined DVFs derived from real patients’ data. In addition, a set of experiments to be used for assessment and benchmarking Mid-P implementations are suggested.

2. Materials and methods

2.1. Patient data

Twenty 4DCT scans for lung tumor patients (10 from Champalimaud Clinical Center (CCC) and 10 from the Netherlands Cancer Institute (NKI)) were acquired with patients’ consent and ethics committee approval. Left and right lung cases with various GTV volumes (0.2 to 283 cc) and motions (0 to 10 mm) were selected. The scans were performed during free breathing in helical mode, with the patient in the head-first supine position. The pitch was set according to the patient breathing pattern. For each patient, 10 respiratory phases were retrospectively reconstructed using phase sorting. Each phase represents 1/10th of the breathing period.

CCC scans were acquired using abdominal compression to reduce respiratory-induced motion. The CCC 4DCT datasets had the size of 512×512 pixels with a pixel spacing ranging from 0.84 × 0.84 to 1.14 × 1.14 mm and slice spacing of 2 mm for all 10 scans, while NKI 4DCT datasets had the size of 512×512 pixels with a pixel spacing of 0.976 × 0.976 mm and slice spacing of 3 mm for all 10 scans.

2.2. Mid-P implementations

Three independent implementations were used to construct the Mid-P CT images and estimate the mean peak-to-peak motion amplitudes within ROIs. The first (Implementation I) is an automated research prototype (developed by Mirada Medical Ltd, Oxford, UK) using Mirada’s CT Deformable Image Registration (DIR) algorithm which is a highly optimized derivative of Lucas-Kanade Optic Flow [24]. The second (Implementation II) is clinically used, called Wimp, and developed by NKI; it uses a Phase based DIR algorithm [25]. The last implementation (Implementation III) is an under-development research tool from CCC using home-written Python modules, 3D Slicer [26] (V.4.13.0, https://www.slicer.org/) to provide a suitable graphical user interface, and the Elastix registration toolbox [27,28] integrated into 3D Slicer software to perform the DIRs. Trilinear interpolation was used for image resampling by the three implementations in order to avoid any biases. The Mid-P method also requires a numerical inversion of the DVF. Implementation I used a solver based on a fixed-point approach [29,30]; implementation II applies an in-house method using a Nelder-Mead optimization [31]; implementation III used a method that minimizes a cost function to derive a parametric inverse as B-spline transformation [32].

2.3. Mid-P CT construction workflow

All of the three Mid-P implementations are using the same workflow described by Wolthus et al. [7] to construct the Mid-P CT image and to estimate the motion amplitudes for relevant ROIs. The workflow consists of two main steps.

- The first step involves deformably registering all the 4DCT phases to a reference (anchor) CT phase (usually maximum exhale), then computing the mean (DVF$_{mean}$) of the resulting (DVFs), and thus defining the mid-position location within the patient breathing cycle. All the CT phases are then deformed and resampled at the mid-Position location (using the inverse of DVF$_{mean}$ DVF$_{mean}^{-1}$) for the reference phase and the composition of the phase-to-phase DVFs
with the DVF\textsubscript{mean} for the remaining phases). Finally, the Mid-P CT image is constructed by calculating the voxel-wise median intensity from all the deformed CT phases.

- The second step involves contouring ROIs (usually GTV or CTV) on the constructed Mid-P CT image, then using the resulting DVFs from the first step to calculate the mean peak-to-peak motion amplitudes within the ROIs in the Left-Right (L-R), Anterior-Posterior (A-P), and Superior-Inferior (S-I) directions.

  The peak-to-peak motion amplitude at the mid-P location is defined as the range of tissue motion throughout the respiration cycle. Specifically, it is defined at every image voxel location of the reconstructed Mid-P image and is given by:

\[
MA(v, k) = \max_{i=0\ldots9} D_i(v, k) - \min_{i=0\ldots9} D_i(v, k)
\]


where \(D_i\) represents the DVF used to resample the \(i\)th CT phase to the mid-position location, \(v\) is a voxel coordinate and \(k = 1, 2, 3\) is an index representing the 3 spatial directions of the displacement vectors, L-R, A-P, and S-I, respectively.

2.4. Dataset generation

Starting from the acquired 4DCT scans, twenty synthetic 4DCT datasets with the associated reference Mid-P images and reference tumor locations (both to be used as gold standard for comparison), were generated. The underlying principle of the proposed generation method is to use real patient data to estimate all the needed spatial mappings for the synthetic data generation starting from one single original CT phase. Since the synthetic 4DCT data mimics the original data, the tissue motion in the synthetic data should be statistically comparable to true motion observed in patient population in the clinic.

2.4.1. Synthetic CT data

A schematic representation of the procedure to generate the synthetic data is illustrated in Fig. 1.

For each patient 4DCT scan, first an anchor phase was chosen as the maximum exhale phase because it is expected to have fewer imaging artifacts than other phases due to the minimal motion amplitudes and the higher reproducibility of the tumor position compared to the other phases [7].

Next, the anchor phase (moving image) was registered to the remaining nine phases (fixed image). This results in 9 DVFs (\(DVF_1, DVF_2, \ldots, DVF_9\)) where \(DVF_v\) represents the deformation vector fields needed to warp the anchor phase onto the \(i\)th phase (see Fig. 1a).

Then all the original phases (except the anchor) were deleted (Fig. 1b) and replaced by new synthetic CTS, generated by warping the anchor phase image with the 9 DVFs (Fig. 1c).

The associated Mid-P CT image of the synthetic 4DCT dataset can be constructed from the anchor phase CT by applying the appropriate averaged DVF to the anchor phase. However, the computed DVF cannot be averaged directly since these are obtained using different fixed images and thus the DVFs are tracking different image features (illustrated by different starting points of the arrows DVF representation in Fig. 1c). Thus, first the inverse of each DVF is computed (Fig. 1d), then they are averaged. Formally this can be expressed as:

\[
DVF^{-1}_{\text{mean}} = \frac{1}{10} \sum_{i=1}^{9} DVF^{-1}_{ri}
\]

Note the sum is normalized by the number of CT phases to account for the DVF representing an identity transformation of the registration of the anchor phase to itself. On Fig. 1e, the \(DVF^{-1}_{\text{mean}}\) is represented as an arrow pointing from the anchor phase to the Mid-Position location. This is the inverse of the required mapping, therefore the \(DVF^{-1}_{\text{mean}}\) is inverted back to compute the DVF\textsubscript{MidP}.

Finally, the associated Mid-P CT image of the synthetic 4DCT data can be constructed by warping the anchor phase image (Fig. 1f). This can be formally expressed as:

\[
I_{\text{MidP}}(v) = I_r(v + \text{DVF}_{\text{MidP}}(v))
\]

where \(I_{\text{MidP}}\) represents the constructed reference Mid-P CT image, \(I_r\) represents the anchor phase CT, and \(v\) is a voxel location (Fig. 1f).

2.4.2. Tumor reference mask

Similar to the synthetic 4DCT data and reference Mid-P images, reference tumor locations on all the generated CT data can be obtained by first segmenting the tumor on the anchor phase and then propagating the tumor ROI to the remaining phases and the reference Mid-P image, using the same mappings that were used to generate the synthetic CT data.

Specifically, GTV contours were first manually delineated on the anchor phase of the synthetic 4DCT datasets using Mirada RTx 1.8 (Mirada Medical Ltd, Oxford, UK); GTV volumes ranged from 0.23 up to 283.24 cm\textsuperscript{3} with a median volume of 4.09 cm\textsuperscript{3}. The GTV contours are then converted to binary image masks and then propagated as detailed above to create reference tumor masks for the reference 4DCT data and the reference Mid-P images. Trilinear interpolation was used for image resampling.

2.4.3. Image registration

All registrations for constructing the synthetic images were performed using MATLAB® (R2020b) Demons DIR algorithm [33,34] on a GeForce RTX 2080 SUPER GPU with 8 GB of dedicated random-access memory (RAM) and with 300 iterations for each registration. Trilinear interpolation was used for all image resampling. Numerical inversion of the DVF was also performed in MATLAB using a fast technique using a gaussian kernel splatting [35]. To avoid any biases, both the DIR and the DVF numerical inversion algorithms used to generate the test data were chosen to be independent of the methods used in the Mid-P implementations to be tested.

2.5. Assessment measures for benchmarking

2.5.1. CT image quality

For each synthetic 4DCT dataset, the Mid-P CT image was constructed using the three implementations (I, II, III) as described in section 2.3, using the same anchor phase. Then, the three constructed Mid-P CT images for each synthetic dataset were individually compared to the reference Mid-P CT image slice by slice. Two quantitative measures to assess synthetic CT image quality were used [36]. The Root Mean Square Error (RMSE) is more intuitive to interpret however it does not take into account the spread in the difference because it is a voxel-wise measure. The Structural Similarity Index (SSIM) index [37] is a perceptual structural measure that quantifies image quality degradation and carries information about inter-dependencies between pixels. The SSIM is a normalized measure upper bounded by the value 1 representing perfect equality. Note that the RMSE is a metric but the SSIM is not [38]. Image slices were cropped to the body contour to exclude any quantification from the background image pixel.

2.5.2. Tumor appearance

Mid-P tumor masks were generated using the mappings used to
construct the Mid-P CT images by the three implementations but applied
to the reference 4D tumor masks. Thus, the mask from each 4D phase
was warped too and reconstructed at the Mid-P location. These were
then fused into a single mask using a voxel-wise median to be compared
to the corresponding reference mask generated in section 2.4.2. Propa-
gated masks can be binarized or converted to contours by thresholding
halfway the initial foreground and background mask intensity values.

\( \text{GTV}_I, \text{GTV}_II \) and \( \text{GTV}_III \), refer to the constructed tumor contours by the
three implementations, and \( \text{GTV}_{\text{ref}} \) refers to the reference Mid-P contour
to be compared.

The spatial accuracy of the tumor location on the constructed Mid-P
images was assessed using the 3D Dice Similarity Coefficient (DSC). The
DSC scores and volumes were measured at voxel precision i.e. the full
voxel is considered in or out depending on the center voxel location. In
addition, the mean intensity values within \( \text{GTV}_I, \text{GTV}_II \) and \( \text{GTV}_III \) were
calculated and individually compared to the mean intensity value within
\( \text{GTV}_{\text{ref}} \) to validate the intensity consistency within the GTVs of the Mid-P
images.

2.5.3. Tumor motion amplitude

Since all the synthetic data were generated by a given set of DVFs, the
reference peak-to-peak motion amplitude can be obtained by replacing
\( D_i \) in Eq. (1) with the composition of \( \text{DVF}_i^1 \) and \( \text{DVF}_{\text{Mid-P}} \), for \( i = 1 \) to 9
and \( D_0 \) being \( \text{DVF}_{\text{Mid-P}} \) (illustrated in Fig. 1d and Fig. 1f).

The estimated mean peak-to-peak motion amplitudes by the three
implementations within the corresponding ROIs (\( \text{GTV}_I, \text{GTV}_II \) and
\( \text{GTV}_III \)) were individually compared to the mean reference peak-to-peak
motion amplitudes within the reference \( \text{GTV}_{\text{ref}} \) contour in the L-R, A-P,
and S-I directions.

3. Results

Fig. 2 shows an example of an original and the synthetic 4DCTs for
the same dataset at different phases and views. The generated datasets
were visually inspected and compared to the original 4DCT to ensure
that the anatomical features and motion characteristics of the original
datasets are accurately represented in the synthetic datasets.

Fig. 3 shows the reference mean peak-to-peak motion amplitudes
histograms for the 20 GTVs in the L-R, A-P, and S-I directions. The
motion amplitudes were up to 3.3, 5.3 and 9.9 mm in the L-R, A-P, and S-
I directions, respectively.

3.1. CT image quality

An illustrative example of image quality assessment on one synthetic
4DCT is shown in Fig. 4. The SSIM index and the RMSE for all the slices
for the three constructed Mid-P CT images when compared to the
reference Mid-P CT image are shown in Fig. 4. The three implementa-
tions showed similar behavior with higher SSIM (lower RMSE) values at
the first and last slices which are located at the upper lung and lower
abdominal regions, respectively. Image slices located in the diaphragm
region showed somewhat lower SSIM (higher RMSE) scores, which is
expected due to the larger breathing motion amplitudes. The same behavior (lower image quality in the lower lung and diaphragm region) was noticed in all cases studied. Interestingly a high correlation is observed between the SSIM and the RMSE. Fig. 5 shows images of the difference in HU between the reference Mid-P CT and the Mid-P CT created by implementations I, II, and III at two locations; slice number 20 located at the upper lung where high (low) values for the SSIM (RMSE) were reported, and slice number 91, located near the diaphragm where low (high) values for the SSIM (RMSE) were reported (indicated by the dashed lines on Fig. 4).

Boxplots of the mean image quality scores for both measures over all image slices are shown on Fig. 6, for the three implementations. The figure shows that overall implementations I and III have slightly better results than Implementation II for both measures. The outermost outlier for the three implementations corresponds to the same synthetic dataset. It was found that this dataset has significantly higher image intensity.
noise in comparison to the rest of datasets. For this particular dataset, both measures suggest that the constructed Mid-P CT image by Implementation I is more similar to the reference Mid-P CT image while Implementation III showed the least similar results.

Two-sided Kruskal-Wallis test was conducted to examine whether image quality scores are statistically different. No significant differences were found (p = 0.4467 for the SSIM and p = 0.4264 for the RMSE).

3.2. Tumor motion

Boxplots of the difference in motion amplitude for the 20 GTVs between the estimated values by the three implementations and the reference motion amplitude are shown in Fig. 7.

Fig. 7 shows that the difference in the estimated motion amplitudes by Implementation I was less than 1 mm for all cases and in all directions. For Implementation II, 95 %, 95 %, and 90 % of the estimated motion amplitudes difference were less than 1 mm in the L-R, A-P, and S-I directions, respectively. For Implementation III, 100 %, 95 %, and 80 % of the estimated motion amplitudes difference were less than 1 mm in the L-R, A-P, and S-I directions, respectively.

The overall mean difference in the motion amplitudes (± one standard deviation) by Implementation I, Implementation II, and Implementation III were 0.15 ± 0.24, 0.05 ± 0.36, 0.19 ± 0.28 mm in the L-R direction, 0.08 ± 0.27, −0.02 ± 0.46, 0.00 ± 0.56 mm in the A-P direction, and −0.09 ± 0.29, −0.40 ± 0.69, −0.32 ± 0.77 mm in the S-I direction, respectively.

The maximum difference in the motion amplitude was noticed in the S-I direction where larger motion amplitudes and larger voxel spacing

---

3. The minus sign indicates that the implementation underestimated the reference value.
(slice thickness) exist, in this direction the maximum differences detected by Implementation I, Implementation II, and Implementation III were 0.84, 2.64, and 2.56 mm, respectively.

A two-sided Kruskal-Wallis test indicated that no statistically significant differences were present, with p-value of 0.069, 0.757 and 0.254 in the L-R, A-P, and SI directions respectively.

3.3. Intensity consistency within tumor

The reference intensity values distribution within the 20 GTV$_{ref}$ in addition to the difference in the mean intensity values between the GTV$_{ref}$ and estimated ones by the three implementations within the corresponding GTVs (GTV$_I$, GTV$_II$, and GTV$_III$) for the entire 20 datasets are shown in Fig. 8.

The absolute difference in the intensity values within the GTVs was

---

**Fig. 7.** Boxplot representing the differences in the mean peak-peak motion amplitudes (within the GTV for all 20 GTVs) between the reference values and the estimated by the three implementations in the Left-Right (L-R), Anterior-Posterior (A-P), and Superior-Inferior (S-I) directions.

**Fig. 8.** The reference intensity distribution within the 20 GTV$_{ref}$ (a), the differences in the mean intensity value values (within the GTV for all 20 GTVs) between the reference values and the estimated ones by Implementation I (b), Implementation II (c), and Implementation III (d).
less than 20 HU for 95% of the dataset processed with Implementation I and less than 20 HU for 90% of the datasets evaluated by both Implementation II and Implementation III. The three implementations showed more than 50 HU of difference (overestimating) in the intensity values for the same dataset which has a small GTV volume (0.2 cm$^3$) and relatively large motion amplitude (4 mm in the S-I direction).

The mean intensity difference values within the 20 GTVs ($\pm$ one standard deviation) for the datasets processed with Implementation I, Implementation II, and Implementation III were 3.4 $\pm$ 15.1, 2.4 $\pm$ 17.8, and 9.6 $\pm$ 14.9 HU, respectively.

3.4. Tumor spatial accuracy

Fig. 9 shows an example of the difference between the GTV$_{ref}$ contour for one dataset and the propagated ones by the three implementations (GTV$_I$, GTV$_{II}$, and GTV$_{III}$) in the axial, coronal, and sagittal views. The DSCs for the implementations I, II, and III contours were 0.962, 0.958, and 0.941, respectively.

The results show that the GTVs propagated to the Mid-P CT by Implementation I match on average better (higher DSC values) with the reference GTVs than the other two implementations. A two-sided Kruskal-Wallis test was conducted to examine whether DSC scores are statistically different according to the used Mid-P implementation. No significant differences ($p = 0.54$) were found among the three implementations.

The tumor volume, as measured by the GTVs on the Mid-P images constructed by the three implementations, were compared to the reference GTV volumes using the Bland Altman plots [39] shown in Fig. 11. A two-sided Kruskal-Wallis test showed that there are no significant differences in tumor volume as measured by the three Mid-P implementations ($p = 0.98$). However, all three methods underestimated slightly the tumor volume (p-values of Wilcoxon signed-rank tests are less than 1.3e-5).

4. Discussion

Benchmarking is a process to evaluate the performance and the outputs of a given application using standardized tests. In this work, the outputs of three implementations for generating Mid-P CT images from the 4DCT data and estimating the motion amplitudes in regions of interest were evaluated. Since currently available 4DCT testing datasets do not provide the Mid-P CT images of the datasets, the primary objective was less about the comparison of different implementations used to generate the Mid-P CT images but more about the establishment of a clinically relevant procedure to generate reference data. In addition, we aimed to utilize a set of standard tests that can be used to assess the performance of any implementation being developed to calculate the Mid-P CT images from 4DCT data. Assessment data produced in such benchmarking process are valuable for comparative purposes [22].
For this purpose, a workflow for generating life-like but synthetic 4DCT datasets and their associated Mid-P CT images from a single original CT phase using predefined estimated DVFs per dataset was first provided. The same methodology was applied to also generate corresponding reference GTV contours starting from the contouring of the GTV on the one CT phase used for 4DCT data generation.

While Implementation I showed slightly better agreement with the reference data than the other two implementations, the statistical analysis showed that the observed differences are not statistically significant. This suggests that the mid-position method is robust to the choice of deformable image registration algorithm. However, all three implementations underestimated slightly the tumor volume. This is at least in part a consequence of the use of binary masks to represent the GTV segmentations. This representation is not robust to image interpolation, and the GTVs constructed by image warping resulted in eroded masks. Alternative implicit contour representation using signed distance transform that offers sub-pixel accuracy can be used. The image-based representation of the GTV contour has the advantage of using the same methods used to construct the Mid-P image to generate the Mid-P tumor contour without any manual delineations. The alternative is to use 3D points set representation of the contours to generate GTV reference dataset. Mapping this representation will, however, need a different software component to produce the tumor Mid-P contour. While a mask/image representation suffers from the image interpolation, it is helpful to confirm that implementation of image warping/resampling used by the Mid-P implementation is correct. A contour representation, as a set of polygons, will have greater volumetric accuracy in generating a “ground truth” segmentation but would not be useful to test the implementation of the Mid-P image creation software.

One limitation of the proposed approach is that deformable image registration is used to estimate the tissue motion as a means to generate synthetic 4DCT images with known and realistic motion. In doing so, there are a number of risks; that the motion is not plausible or that it introduces bias in any assessment. The generated motion field may misrepresent the correct motion where the deformable image registration is not able to fully capture the motion in the original 4DCT images. For instance, many DIR algorithms do not allow for sliding motion, therefore the motion field at the edge of the lung is likely to be incorrect in some way (not enough lung motion or “motion” of the ribs). This relates to the risk of bias; where the motion field used to create the synthetic 4DCT is one that a DIR algorithm is able to recover, it will appear to be more correct. Therefore, because a DIR algorithm has been used to generate the motion in the synthetic images, a DIR algorithm with a similar motion model is more likely to be able to recover that motion [40]. An attempt was made to mitigate some of these risks by using an independent DIR algorithm to generate the test dataset. Also, three implementations with different DIR algorithms were evaluated with similar results. However, the risk of some bias remains.

A challenge for using an existing patient image to define the motion field is that not all patients breath in the same way. Therefore, it is sensible to include patients from multiple institutions, as in this study, to represent a population and identify any weaknesses in implementation that may be patient specific. Furthermore, the acquisition of 4DCT varies between institutions regarding the used protocols and respiratory reduction techniques. Thus, the acquired 4DCT datasets from different institutions may have different image quality and different motion amplitude ranges. The limited size of the available datasets used in this study prevented a robust statistical analysis of any differences between the data from each institution being performed.

The proposed assessment procedure allows clinical physics teams to use their own 4DCTs to create custom reference data, to evaluate the accuracy of the constructed Mid-P CT images, to choose the DIR algorithm that matches better with the constructed reference data, and to optimize the registration parameters to achieve the best performance. The selected DIR algorithm should be independent from the one utilized by the evaluated implementation to eliminate any possible bias. And more importantly, it should be able to accurately capture the motion characteristics in the patients’ population, as represented by the original datasets.

Although two measures were used for an overall assessment of image quality, only the accuracy of the tumor anatomy was considered in this study, since the main focus of the mid-position approach is accurately defining the GTV and suitable margins for treatment. However, this method could be extended to other parts of the patient’s anatomy by performing an assessment of other structures, in addition to the GTV, required for treatment. The reported absolute difference in the intensity values within the GTVs (Fig. 8), along with the overall image difference within the body contour as assessed by the RMSE are good indicators on dosimetry impact arising from the observed differences [41]. According to [42,43], variations of ± 50, ±30, and ±100 HU in the CT number are required to produce 2 % dose errors using 6 MV photons for lung, soft tissue, and bone, respectively.

Although the generated synthetic 4DCT were visually inspected to ensure that the features and motion characteristics of the original data are correctly represented in the generated data, clinical Quality Assurance (QA) of any Mid-P implementation should be performed using original patients’ data. However, the proposed workflow provides a valuable methodology for commissioning any newly developed Mid-P implementation and perform software QA at regular intervals.

The proposed evaluation approach and set of experiments were able to detect implementation errors in the work-in-progress development of Implementation III when resampling the anchor phase to the Mid-P location and also in the DVFs composition when estimating the tumor peak-to-peak motion amplitude values. It is important to highlight that such errors would have not been detected with a DIR validation tool. Indeed, DIR validation tools focus on one registration between two images and often only on the motion estimation part and not on its application to generate the registered image (DVF interpolation + image
resampling). Furthermore, such tools do not include two important DIR related components in the Mid-P workflow, namely the computation of the inverse of a DIR and the composition of two DIRs. The detected errors were corrected before benchmarking the implementation, thus were not reported in the results. Neither of the other implementations were tested using this method prior to the evaluation performed.

The described work may serve as the first step in the validation process or commissioning of a Mid-P implementation for clinical use. The proposed set of assessment measures should be regarded as a bare minimum rather than a comprehensive set. Measures to include dosimetry impact of the computed images are of high importance. Additionally, efficiency measures or facility of organs at risk (OARs) contouring are also valuable differentiators to help identify the strengths and limitations of different solutions, allowing users to make informed choices based on their specific requirements.

5. Conclusion

In this work, an approach to generate synthetic 4D CT reference data that can be used in the assessment of Mid-P implementations was described. Three independent Mid-P implementations have been benchmarked using these reference data. The presented results and assessment procedure can be part of the commissioning of a Mid-P implementation for clinical use.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Joao Marques for coordinating the patient selection and for asking for their consent to participate in this study at Champalimaud Clinical Centre. Djamal Boukerroui thanks Professor Marcel Van Herk for valuable discussions on the Mid-Position method. This work may serve as the first step in the validation process or commissioning of a Mid-P implementation for clinical use.

References


