IOP Publishing Phys. Med. Biol. 65 (2020) 035011 (11pp) https://doi.org/10.1088/1361-6560/ab652a

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

# Localised delineation uncertainty for iterative atlas selection in automatic cardiac segmentation

Robert Finnegan[1,2,9], Ebbe Lorenzen[3,4], Jason Dowling[5,6], Lois Holloway[1,2,6,7,8], David Thwaites[1] and Carsten Brink[3,4]

[1] Institute of Medical Physics, School of Physics, University of Sydney, Sydney, Australia
[2] Ingham Institute for Applied Medical Research, Liverpool, Australia
[3] Institute of Clinical Research, University of Southern Denmark, Odense, Denmark
[4] Laboratory of Radiation Physics, Odense University Hospital, Odense, Denmark
[5] The Australian e-Health and Research Centre, CSIRO Health and Biosecurity, Herston, Australia
[6] School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, Australia
[7] South Western Sydney Clinical School, University of New South Wales, Sydney, Australia
[8] Centre for Medical Radiation Physics, University of Wollongong, Wollongong, Australia
[9] Author to whom any correspondence should be addressed.

**E-mail:** robert.finnegan@sydney.edu.au

**RECEIVED**
26 August 2019

**REVISED**
25 November 2019

**ACCEPTED FOR PUBLICATION**
23 December 2019

**PUBLISHED**
4 February 2020

**Keywords:** atlas-based segmentation, heart contouring, whole heart segmentation, medical image processing

## Abstract

The heart is an important organ at risk during thoracic radiotherapy. Many studies have demonstrated a correlation between the mean heart dose and an increase in cardiovascular disease. Different treatments result in significant dose variation within the heart and individualised dose estimation increasingly requires more attention to delineation of various cardiac structures. Automatic segmentation tools are critical for consistent and accurate delineation of organs at risk in large, retrospective studies, however the challenge of ensuring a robust method must be addressed.

In a multi-atlas based segmentation framework the uncertainty in delineation can be modelled over the surface of the heart. We extend this concept with an iterative atlas selection procedure designed to remove inconsistent atlas contours, in turn improving the reliability of the segmentation.

Two independent datasets comprising 15 and 20 planning computed tomography (CT) images of Danish and Australian breast cancer patients, respectively, had the whole heart and left anterior descending coronary artery (LADCA) delineated. Using a cross-validation strategy, where each dataset is used as an atlas set to segment each image in the other, we assess segmentation performance qualitatively and quantitatively, using the dice similarity coefficient (DSC), mean surface-to-surface distance (MASD) and Hausdorff distance (HD).

After using the iterative atlas selection procedure, every segmentation error was removed. For the whole heart, the resulting segmentation achieved a DSC, MASD and HD of $0.937 \pm 0.009$, $1.66 \pm 0.336$ mm, and $13.4 \pm 4.54$ mm.

## 1. Introduction

Breast cancer patients treated with radiotherapy benefit from reduced rates of local recurrence and overall mortality (Darby *et al* et al 2011, Cutuli *et al* 2014, EBCTCG (Early Breast Cancer Trialists' Collaborative Group) 2014). However, this treatment also engenders inevitable radiation dose to nearby organs at risk. As a result, radiotherapy of the breast is associated with increased mortality from cardiovascular disease decades after the treatment (Hooning *et al* 2007, Darby *et al* 2005). Retrospective studies of dose to the heart have primarily focused on the mean heart dose (Darby *et al* 2010); time-consuming contouring of cardiac substructures precludes obtaining doses to individual parts of the heart on a large scale. However, with variations in the dose distribution within the heart as a result of different treatments, there is a gap in studies evaluating the dose to cardiac substructures (Gagliardi *et al* 2010). The left anterior descending coronary artery (LADCA) also represents a significant organ at risk, particularly during left sided breast cancer radiotherapy (Correa *et al* 2007).

Dose inhomogeneity within the heart is associated with increased risk of late cardiac effects for mediastinal radiotherapy (Hahn *et al* 2017), so the development of accurate and consistent measurements of the dose to cardiac substructures is pertinent.

In recent years there have been a number of atlas-based approaches to perform automatic cardiac segmentation in clinical radiotherapy. The utility of these tools stems from the accurate delineation, time-saving, and reduced contouring variability (Eldesoky *et al* 2016, Ciardo *et al* 2017). For accurate dose estimation, critical for large-scale data analysis, automatic cardiac segmentation has been a topic of interest for several years (Lorenzen and Brink 2012). Recently, Kaderka *et al* (2018) have tested the geometric and dosimetric accuracy of automatic cardiac segmentation using a commercial system, finding good agreement between manual and automatic delineations, and high correlations in dose estimates using these volumes. In moving towards a segmentation approach that is able to be utilised consistently across clinics an open-source software solution is required. Zhou *et al* (2017) developed and validated an atlas for automatic segmentation of cardiac substructures, with promising results demonstrating clinical feasibility. The recent work by Morris *et al* (2019) uses a combined magnetic resonance imaging (MRI) and x-ray computed tomography (CT) cardiac atlas to generate reliable cardiac segmentation. Despite this recent work, a challenge remains: how can segmentation errors be detected, and hopefully, removed?

In developing an open-source multi-atlas based automatic segmentation (MABAS) package that can be distributed between clinics, we aim for a robust framework that can be applied with no need for local optimisation, with the goal of enabling error-free segmentation. The primary source of errors in atlas-based segmentation is mis-registration of atlas images to the target image. A potential solution to improve segmentation performance is to select only the most suitable atlases, or to remove inaccurately registered atlases before combining the atlas labels to generate the final segmentation. There has been substantial effort in such atlas selection techniques, with many methods using image similarity between atlas and target images to define the optimal atlas set (Klein *et al* 2008, Aljabar *et al* 2009, Ou and Doshi 2012, Doshi *et al* 2016). Studies into atlas selection prior to the time-consuming deformable image registration step have the potential to improve segmentation performance and efficiency (van Rikxoort *et al* 2010, Langerak *et al* 2013). Investigation into iterative methods for selecting the best atlases for segmentation generation have been explored in the past (Langerak *et al* 2010, Antonelli *et al* 2019), however these algorithms rely on image similarity to quantify registration accuracy. Since registration algorithms use image similarity as a metric to compute deformation fields between atlas and target images, it follows that reliance on this similarity to quantify performance is problematic.

The aim of this work is to develop a procedure to measure the quality of individual atlases. Our goal is to provide a cross platform tool that is usable in all clinics, and is able to robustly collect information from large, retrospective datasets. We evaluate the effect of using this tool, in an iterative atlas selection procedure, on segmentation errors and overall accuracy for the heart and LADCA.

## 2. Materials and methods

### 2.1. Imaging data

Two independent datasets of breast cancer patients treated in Odense, Denmark and Liverpool, Australia were retrospectively obtained. In a cross-validation strategy each dataset was segmented using the other as an atlas set.

The Odense dataset consists of 15 patients, with a single observer providing manual delineations of the whole heart and LADCA contoured following local guidelines based on recommendations by Feng *et al* (2011). Imaging was acquired in axial slices, with in-plane resolution of 0.97 mm × 0.97 mm and slice thickness of 3 mm. Patients were imaged in the treatment position, laying supine on an inclined breast board and with their arms raised. The CT scans were acquired on a Philips Brilliance Big Bore scanner, without the use of contrast enhancement. These data were previously used as input to a commercial system (ABAS, Elekta AB) for segmentation of the heart in radiotherapy (Lorenzen *et al* 2014).

The Liverpool dataset consists of 20 patients, manually contoured by three independent observers, also following the guidelines based on the Feng *et al*. In addition to the whole heart and LADCA, contouring of the cardiac chambers, great vessels, cardiac valves and coronary arteries was performed. Manual contours are combined into a probabilistic segmentation, as described in Finnegan *et al* (2019), to enable the propagation of contour variability throughout the segmentation pipeline. In order to enable comparisons between manual and automatic delineations, we define the ground truth whole heart volume using a majority vote of the three observers' contours. For the LADCA, where the manual contours may have little to no overlap, we use a simple splining procedure to define the location of the LADCa on each axial slice as the area-weighted average position of the three contours, and generate a spline between these points that is expanded to a tube of diameter 4 mm, following published data on coronary artery diameters (Dodge *et al* 1992). This tube is then sampled into image space and defines the ground truth LADCA delineation. Imaging was acquired in 2 mm axial slices with in-plane resolution of 0.97 mm × 0.97 mm. Patients were imaged in the supine position on a flat table with arms raised above the

head to simulate treatment, and the CT scans were acquired on a Philips Brilliance Big Bore scanner, without the use of contrast enhancement.

For both datasets imaging was acquired without breath hold nor cardiac gating techniques, and imaging quality is of the clinical standard in radiotherapy.

### 2.2. Image registration

Automatic segmentation of the whole heart in thoracic CT scans was performed using an in-house multi-atlas framework (Finnegan *et al* 2019). This software is open-source and available to download[10] The input to this system is an atlas set consisting of images from a number of patients contoured by one or more observers, which are independently registered to the target image using rigid alignment with subsequent demons-based deformable image registration. The resulting deformation fields are used to deform the atlas labels and propagate them onto the target image. Rigid image registration was performed using multi-resolution strategy, with three stages consisting of down-sampling the target and atlas images by factors of 8, 2, and 1, and smoothing with a Gaussian filter with a scale of 4 mm, 2 mm, and 0 mm (no smoothing). The sum of square differences was used as a registration metric, with a fixed sampling rate of 0.1, optimised using a limited memory Broyden–Fletcher–Goldfarb–Shannon algorithm (Fletcher 1987). Deformable registration is performed with a multi-resolution symmetric diffeomorphic demons algorithm (Vercauteren *et al* 2009), with 4 stages consisting of down-sampling the target and atlas images by factors of 8, 4, 2, and 1, with smoothing with a Gaussian filter with a scale of 8 mm, 4 mm, 2 mm and 1 mm. The displacement and update fields are both smoothed using a Guassian filter with a scale of 1.5 voxels in each stage.

Throughout this work, we follow standard conventions (Langerak *et al* 2013) to define an atlas $A = (I, L)$ as containing a 3D image, $I$, and a contour (or label) $L$ for each structure that has been delineated[11]. An atlas set, then, is simply a collection of atlases.

We use a cross-validation strategy in this work, where the Liverpool data is used as an atlas set to delineate the Odense data, and vice versa. This means each patient image in the Odense cohort will have 20 atlases, while each patient image in the Liverpool cohort will have 15. Previous studies have shown that optimal segmentation accuracy is reached when using more than approximately 10 atlases (Aljabar *et al* 2009, van Rikxoort *et al* 2010), and so we do not expect the difference in the number of atlases used to produce variation in the segmentation accuracy between datasets.

### 2.3. Iterative atlas selection

The accuracy of atlas-based segmentation depends on the precise registration of each atlas image to the target image. For some atlases this registration fails, for example due to large anatomic variation or the failure of the underlying optimisation algorithm. As a result, using mis-registered atlases when generating the final segmentation can at best introduce noise, and at worst create gross errors in volume definition.

When testing the automatic segmentation framework before the introduction of the iterative atlas selection method, as described in this paper, we observed segmentation errors related to mis-registration of the atlas image with the target image. Detection of these mis-registrations is difficult based on an image similarity measure as the registration algorithm is designed to maximise image similarity, which is the cause of these errors. The aim was thus to create an algorithm, independent of the image similarity measures, which could detect incorrect registrations using the information from the registration of the entire atlas set and not just the individual atlas.

Considering variability in the boundary surface of the individual atlas labels used to generate a segmentation, we build on the work by Lorenzen *et al* (2014) by mapping surface-to-surface distances over the heart surface to quantify segmentation variability.

The projected difference between two volumes is defined as the distance between the boundary surfaces along a ray traced from the centre of mass of the reference volume. This centre of mass is calculated from the consensus segmentation, defined by the region where the propagated contours from all atlases overlap. By tracing a ray outwards to each point on the surface, the projected difference can be mapped onto a sphere. Throughout this work we use an equal-area Eckert IV projection. The process of generating a map of projected surface difference is illustrated in figure 1. The advantage of this measure of uncertainty is that it is possible to make comparisons between different patients or different atlas sets, while preserving spatial information.

This mapping of projected distance is used to identify outliers in the registered set of atlases. The common grid enables spatially localised modelling of segmentation uncertainty. For a given target image we assess the quality of each atlas in the following way:

[10] Software is available at https://github.com/rnfinnegan/simpleseg

[11] In this context labels can be either binary, where voxels have value 1 if they are within the contour and 0 if not, or probabilistic, where a value between 0 and 1 indicates the relative likelihood of each voxel belonging to a structure.

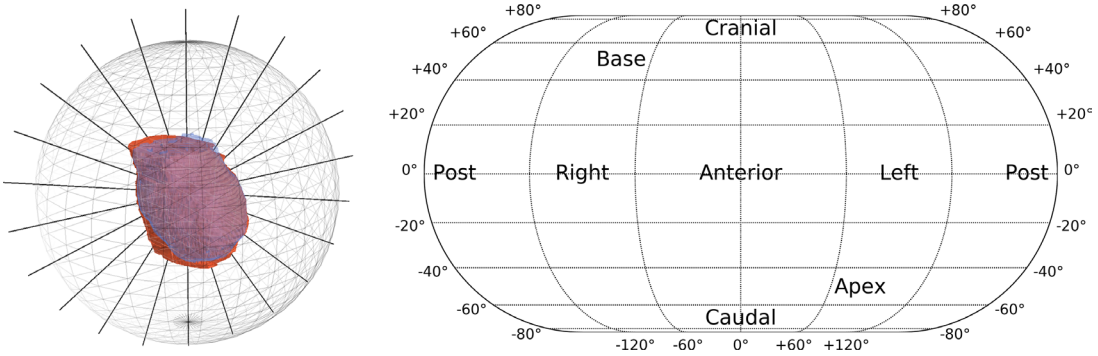


**Figure 1.** Left: surface differences between two delineations of the heart volume are projected onto a sphere using radial rays from the centre of the volume. Right: the Eckert IV projection, used to map values surface differences measured on the surface of the heart to a sphere.

1. A consensus segmentation is generated, which is defined by the region where the propagated contours from all atlases overlap.

   In order to form a reference volume to which each atlas contour can be compared, a consensus segmentation is generated. This is defined by the region where the propagated contours from all atlases overlap. This consensus segmentation is a minimal volume, and is not an appropriate delineation of the heart itself.
2. The projected difference from the consensus segmentation to each atlas contour is computed, and the values are mapped onto a common spherical grid using bilinear interpolation. This projection is designed for convex surfaces, where each ray from the centre of mass intersects the boundary surface of a volume once. In situations where an atlas has been mis-registered, it is possible that this condition is not met, and in this situation the interpolation procedure retains the most distant point of intersection to ensure errors are detected.
   At each point on the spherical grid, $\boldsymbol{r} = (\theta, \phi)$ (where $\theta$ and $\phi$ are the polar and azimuthal angles, respectively), we define the projected difference from each atlas contour (denoted by subscript $i$) to the reference surface as $\boldsymbol{x}_i(\boldsymbol{r})$.
3. By quantifying the relative variation of each atlas, outliers can be identified and removed. This procedure, illustrated in figure 2, proceeds as follows:

(a) In order to measure the spatially localised uncertainty the median absolute deviation was calculated. For each surface point on the sphere shown in figure 1, we calculate the median consensus-to-label distance, $\widetilde{x}(\boldsymbol{r})$, and the median absolute deviation in this distance, $MAD(\boldsymbol{r}) = median\,(|x_i(\boldsymbol{r}) - \widetilde{x}(\boldsymbol{r})|)$. This is illustrated in figure 3.

(b) To compare the variation of each atlas contour, we calculate a scaled distance over the surface which accounts for the spatially localised uncertainty:

$$\hat{x}(\boldsymbol{r}) = \frac{x(\boldsymbol{r}) - \widetilde{x}(\boldsymbol{r})}{MAD(\boldsymbol{r})}. \tag{1}$$

(c) The distribution of scaled distances, $f(\hat{x})$, is used to compute a single figure of merit indicative of the variation of the atlas contour over the entire surface.

The expected distribution of scaled distance was determined using an independent dataset (Lorenzen *et al* 2013), where 9 observers delineated the whole heart in 15 CT images. The averaged distribution of the scaled distances for manual contours was found to be approximately normal.

To quantify the deviation from the expected delineation variability, the measured distribution is compared to the non-linear least squares Gaussian curve fit to the distribution, $f_{\text{fit}}(\hat{x})$. We define the figure of merit, $Q$, as:

$$Q = \int_{-\infty}^{\infty} |f(\hat{x}) - f_{\text{fit}}(\hat{x})| \cdot (\hat{x})^2 \, d\hat{x} \tag{2}$$

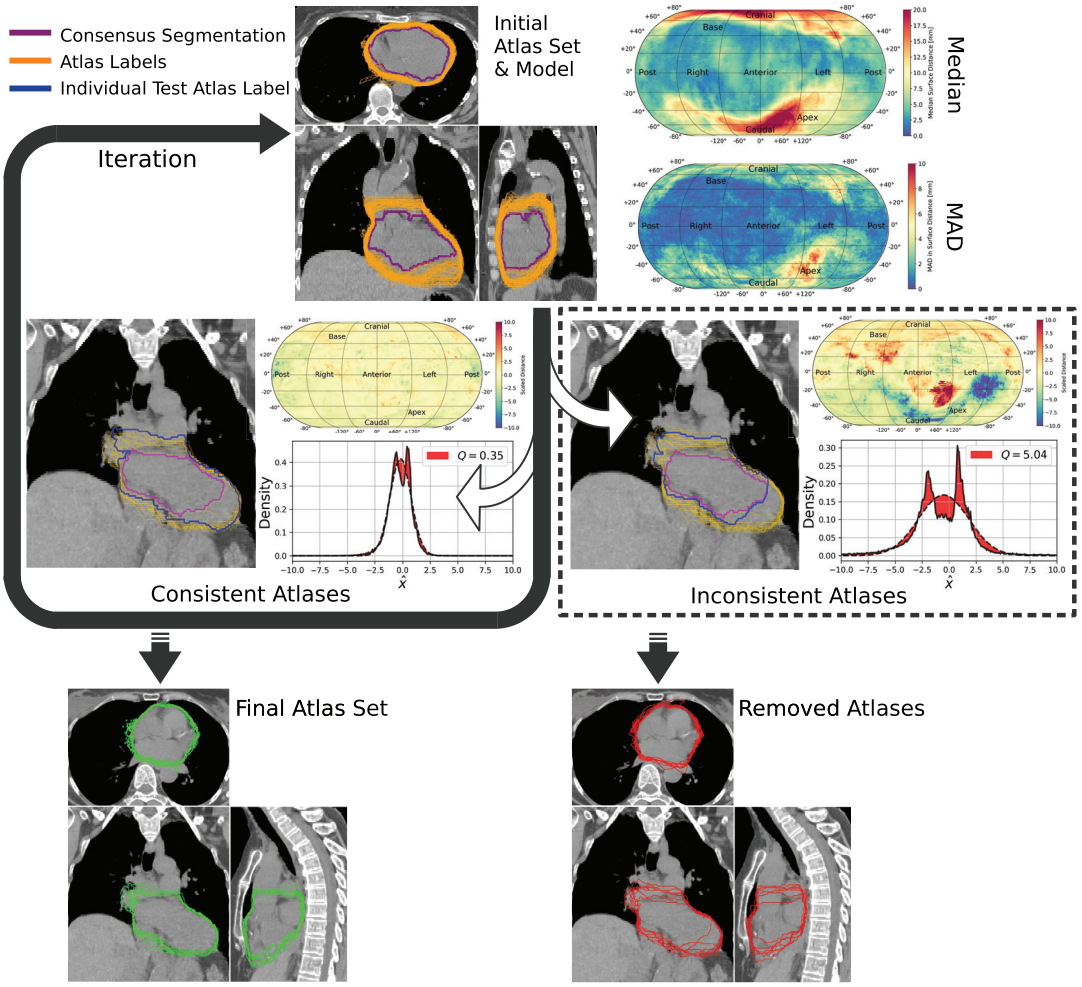


**Figure 2.** The iterative atlas selection process begins with the initial set of propagated atlas labels, which are used to generate a model of the scaled distance from the consensus segmentation to each label (top row). Next, for each atlas the overall variation is quantified, using the proposed figure of merit, $Q$. A histogram of these $Q$-values is shown for a consistent atlas (middle, left) and inconsistent atlas (middle, right), where red regions correspond to regions of larger deviation. These $Q$-values are aggregated over the remaining atlas set, outliers are identified and removed (middle row). This entire procedure is then repeated until there are no outliers detected, resulting in a subset of atlases that are kept for generating the final segmentation, as well as a subset of atlases that are removed (bottom row).

where the weighting factor of $(\hat{x})^2$ penalises variations from the idealised normal distribution that occur at larger absolute values of $\hat{x}$, which correlate with more extreme deviations from typical surface distances.

4. For each atlas contour, $Q$ increases with larger deviations from typical surface-to-surface distances. We hypothesise that this value can be used as a indication of the performance of the individual atlas labels to segment the image, and consequently remove any atlases for which the corresponding value of $Q$ is an outlier, defined as greater than 1.5 interquartile range above the 75th percentile (thus outside the whiskers in a standard boxplot).

**End** Iterate steps 1–5 until there are no outliers identified, and return the remaining atlases for subsequent processing.

This iterative atlas selection allows for the detection of atlas contours which are subject to registration error, an issue we identified as substantially contributing to errors in segmentation (Finnegan *et al* 2019). Importantly, this procedure has no dependence on any manual delineations, and takes into account spatially-dependent segmentation uncertainty as measured independently for each patient.

### 2.4. Segmentation generation

Prior to iterative atlas selection, there is a set of contours from each of the registered atlases. After this process we have a reduced set of contours, where outlier atlases have been removed. These remaining atlases are combined

using a locally weighted voting regime based on the inverse squared difference in image intensity (Išgum *et al* 2009). This method generates a probabilistic label map, where each voxel is assigned a value from 0 to 1 indicating the relative likelihood it belongs to the heart. To generate a binary segmentation from the resulting probabilistic label maps we use a single, fixed threshold for each atlas set. This threshold was calculated using a published optimisation procedure (Finnegan *et al* 2019) designed to minimise differences between manual and automatic volumes without *a priori* delineations. For the Odense atlas set and Liverpool atlas set the optimal threshold was determined to be 0.32 and 0.45, respectively.

In general, atlas based methods are not suitable for generating delineations of small structures such as the LADCA due to a lack of overlap in propagated atlas delineations. To overcome this shortcoming we have developed a simple vessel splining regime. On each axial slice of the target image we compute the centres of mass of the atlas LADCA labels; these points are used to define a connected spline, which is expanded by a predefined radius to generate a tube. This tube is voxelised into image space to produce the final LADCA segmentation. For this work we used a fixed radius of 4 mm, following published data on coronary artery diameters (Dodge *et al* 1992) and prior cardiac contouring guidelines for radiotherapy (Duane *et al* 2017).

The result is an automatic segmentation of the heart and LADCA on each target image.

### 2.5. Evaluation of segmentation performance

Comparison of manual and automatic delineations was performed using the dice similarity coefficient (DSC), the mean absolute surface-to-surface distance (MASD) and the Hausdorff distance (HD, the largest surface-to-surface distance). We compare the manual delineation to the automatic segmentations generated both before and after applying the iterative atlas selection procedure, to assess the impact on segmentation accuracy. To assess the potential statistically significant improvement in these metrics the Wilcoxon signed-rank test is used. A qualitative visual comparison between the boundary surface of the automatic segmentation and manual delineation is used to assess the overall quality and check for segmentation errors.

For the Liverpool dataset we use a majority voting regime to generate a gold-standard manual contour from the three observers in order to make comparisons to the automatic segmentations. Only the whole heart and LADCA are evaluated, as they were delineated in both datasets.

## 3. Results

Geometric measures of the segmentation performance are illustrated in figure 4. These quantitative metrics are averaged over all 35 patients and presented in table 1. Considering inter-observer contouring variation, the automatic segmentation of the whole heart and LADCA was performed accurately. For the whole heart the iterative atlas selection algorithm improved this accuracy, however segmentation accuracy was not improved for the LADCA.

There were several patients for which segmentation errors were observed, most of which were observed in the Liverpool data (with the Odense data serving as an atlas). These segmentation errors are attributable to mis-registration of the heart volume into the tissues at the inferior border of the heart, however after applying the iterative atlas removal procedure all of these observed errors were removed (figure 5).

The number of atlases remaining after applying iterative removal varies significantly between patients; the fraction of atlases remaining to the total number in the atlas set is $0.42 \pm 0.18$ (mean $\pm$ standard deviation). This suggests that, on average, approximately half of the initial atlases are removed using this procedure.

The entire segmentation framework is computationally efficient: pre-processing, registration, iterative atlas selection, label fusion and binary segmentation processing takes around 6 minutes per patient running on a 10-node cluster, with the deformable image registration step constituting the majority of this. Since the MABAS process requires registration of each atlas image to the patient image this results in 20 and 15 independent deformable registration tasks for the Liverpool and Odense atlas sets, respectively.

## 4. Discussion

This work represents a beneficial addition to atlas-based segmentation techniques. The proposed iterative atlas selection procedure is able to remove segmentation errors and produce accurate and consistent segmentations of the whole heart. This procedure is efficient, robust and simple, and can be easily incorporated into existing atlas-based segmentation frameworks.

The results of this study have several clinically relevant implications. Firstly, the ability to detect and correct segmentation errors is crucial in the analysis of large, retrospective datasets, where not only is it unfeasible for manual review and editing of each segmentation, but this may introduce additional observer bias. The independence of our iterative atlas selection procedure on manual contours is beneficial in situations where there are no existing contours, but also when contouring follows different guidelines or local protocols. It is known that inter-

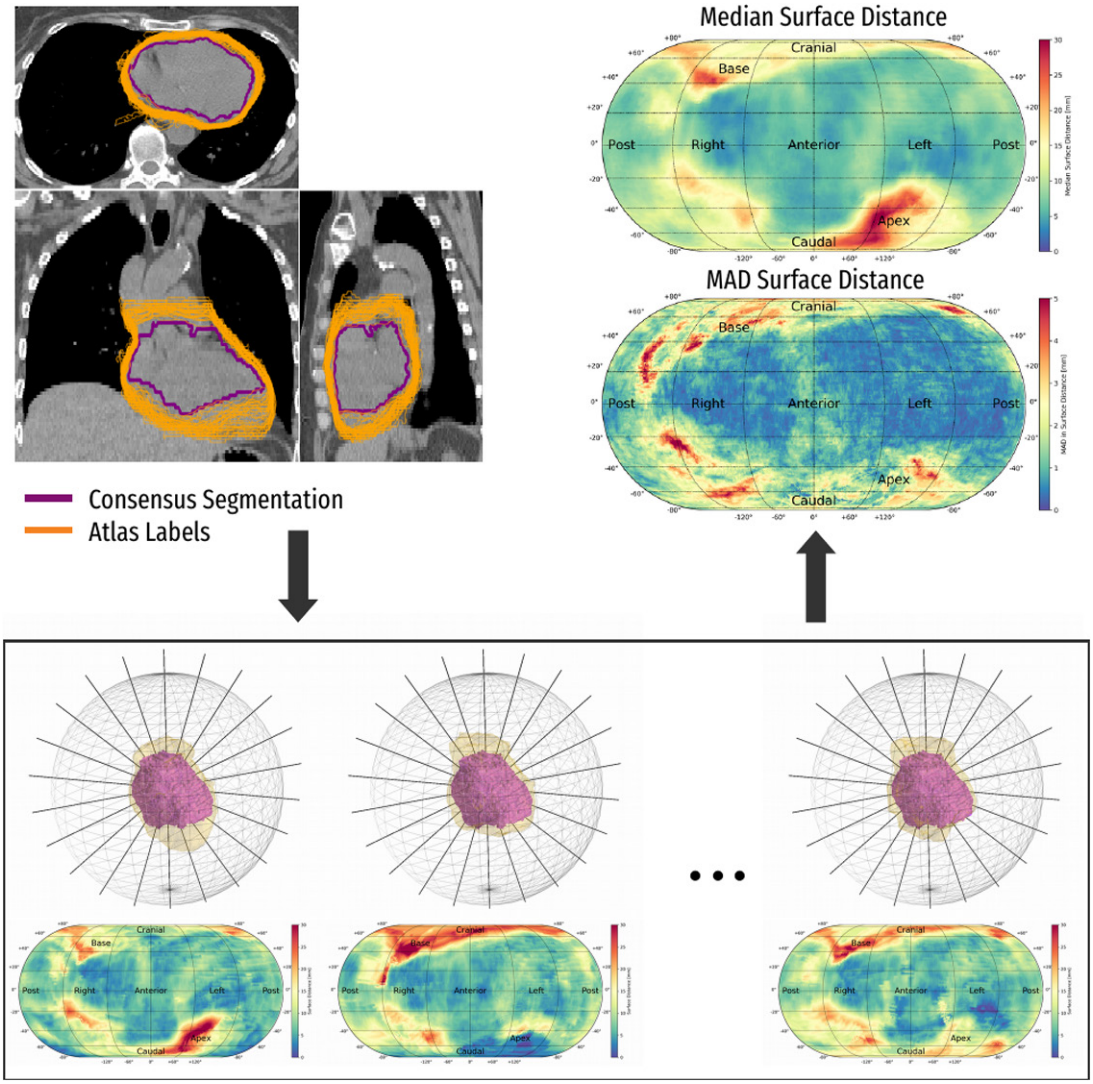


**Figure 3.** Generating the patient-specific model of segmentation variability: the surface maps of the surface-to-surface distance between each propagated atlas label and the consensus segmentation are combined to generated a map of the median and median absolute deviation (MAD) in surface distance.

observer contouring variability contributes to differences in dose estimates of the heart (Li *et al* 2009, Feng *et al* 2011, Cui *et al* 2015), and LADCA (Lorenzen *et al* 2013). While this variability can be reduced for prospective patients, for example by following contouring guidelines (Wennstig *et al* 2017), using existing manual delineations in retrospective data may increase the uncertainty of dose estimates.

Significant inter-patient variation in the heart dose, even for similar treatments, makes dose prediction difficult (Taylor *et al* 2015). Small differences in delineations of the heart and larger substructures have minimal impact on the dosimetric parameters, supporting the case for utilising automatic segmentation techniques in radiotherapy (Luo *et al* 2018), and highlighting the importance of detecting large segmentation errors in a minority of patients. In turn this accurate dose prediction enables analysis of the risk of radiation-induced cardiotoxicity for thoracic radiotherapy patients, which up until now has been limited at least in part due to the difficulties in providing accurate and consistent delineation of the heart and in particular cardiac substructures. In large, retrospective studies and data mining projects involving automated segmentation it is necessary to have consistent and error-free delineations, a challenge we aim to address with this method.

Furthermore, while it is clear that dose to the heart is associated with increased mortality from cardiovascular disease 10–20 years after treatment (Darby *et al* 2005, Hooning *et al* 2007), studies are needed to evaluate the dose to particular cardiac substructures to better understand the mechanisms behind this effect (Stam *et al* 2017). Ensuring the atlases used to generate segmentations are of as high quality as possible will enable precise substructure dosimetry.

The accuracy of the automatic segmentation for the whole heart was excellent, particularly considering the expected inter-observer contouring variability. We found that the segmentation of the Liverpool dataset images was in general slightly less precise and more prone to errors than that of the Odense dataset. This could be as a

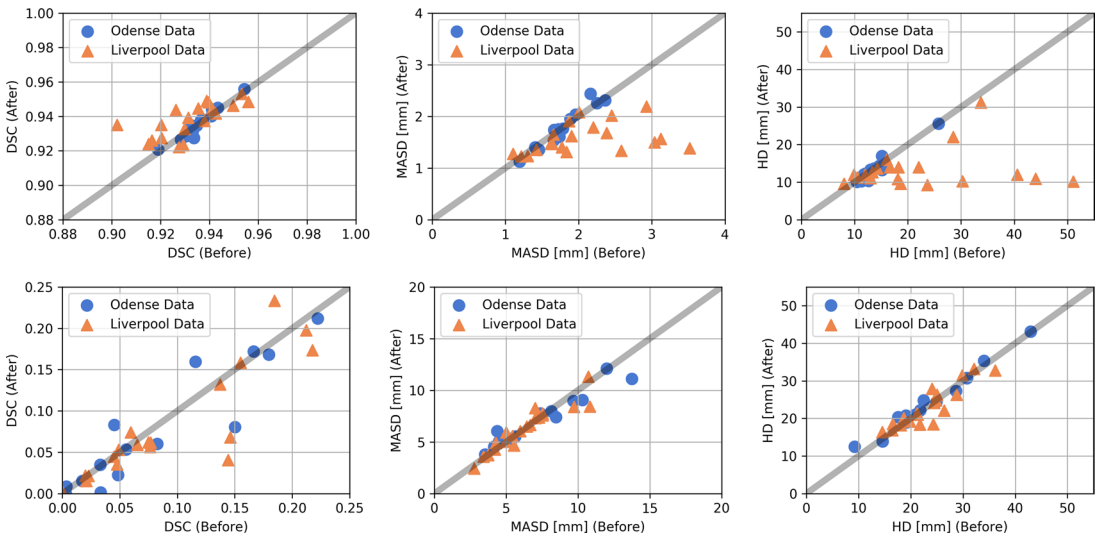


**Figure 4.** The effect of the iterative atlas selection procedure is illustrated by comparing the manual delineations to the automatic segmentations generated both before and after the procedure is applied. The dice similarity coefficient (DSC), mean absolute surface-to-surface distance (MASD) and Hausdorff distance (HD) are used to assess changes in segmentation performance. The results for the Odense data (using the Liverpool data as an atlas set) and Liverpool data (using the Odense data as an atlas set) are plotted separately. Top row: whole heart segmentation, bottom row: left anterior descending coronary artery segmentation.

**Table 1.** Evaluation of automatic segmentation performance for the whole heart and left anterior descending coronary artery (LADCA). Values presented are mean ± standard deviation over all 35 patients in both the Liverpool and Odense datasets.

| | Whole heart | | | LADCA | | |
|---|---|---|---|---|---|---|
| | Before | After | Inter-obs.[a] | Before | After | Inter-obs.[a] |
| DSC | 0.934 ± 0.011 | 0.937 ± 0.009[b] | 0.939 ± 0.011 | 0.086 ± 0.069 | 0.076 ± 0.068 | 0.172 ± 0.086 |
| MASD (mm) | 1.96 ± 0.567 | 1.66 ± 0.336[b] | 1.59 ± 0.356 | 6.71 ± 2.69 | 6.63 ± 2.28 | 4.34 ± 1.30 |
| HD (mm) | 18.7 ± 10.3 | 13.4 ± 4.54[b] | 14.9 ± 3.21 | 23.6 ± 6.91 | 23.6 ± 6.61 | 24.3 ± 5.49 |

[a] The inter-observer variation is calculated using pairwise comparisons of the three sets of contours on images in the Liverpool dataset.
[b] Statistical significant improvement in the metric, given as $p < 0.05$ using a one-sided Wilcoxon signed-rank test.

result of using a smaller atlas set, since only 15 atlases were available when segmenting the 20 Liverpool atlases. Additionally, the Odense imaging had lower image resolution, which may reduce the accuracy of the deformable image registration and hence decrease segmentation accuracy. In using the automatic segmentation framework we intentionally performed no optimisation of any processes in order to better simulate a situation where there are no manual contours available for comparison. Analysis of the geometric accuracy of the whole heart segmentation highlights the importance of using multiple metrics, as measures based only on volumetric overlap (such as the DSC) are insensitive to segmentation errors that do not significantly change the overlapping volume of delineations, such as that shown in figure 5. For this reason, the improvement in segmentation accuracy is not clear when only considering the DSC, however surface-based metrics, such as the MASD and particularly the HD, are more sensitive to such errors, and indicate a marked improvement in segmentation accuracy.

For the LADCA, the use of the iterative atlas selection tool did not significantly change the accuracy of the segmentations. Automatic delineations of the LADCA were generated with a vessel splining procedure, which measures the position of the centre of mass of the LADCA label from each atlas on each slice and connects them with a tube of fixed radius. We expect reasonable registration accuracy along most of the left anterior interventricular sulcus due the sharp boundary between the heart and lung tissue, and thus do not necessarily expect the removal of atlases to change the LADCA definition. Moreover, as the accuracy of our automatic method is limited to the inter-observer contouring variation of the LADCA, we can consider these delineations to be approaching the limit of what is achievable in clinical, non-contrast CT imaging.

The novel measure of the consistency of an individual atlas, the $Q$ value, is an attempt to encapsulate the variation in the boundary surface of an atlas label relative to the remaining atlas set. In generating the model of surface variability we chose to use the median and MAD as measures of typical distance and variation, respectively, as these are generally more robust and resilient to outliers than the mean and standard deviation. When mapping surface distances onto a common grid we chose to use a spherical map projection. In addition to convenient visualisation, this also makes it possible to compare between patients, for example to build a population average of surface segmentation variability. This could serve as a useful tool for evaluating the consistency of contouring,

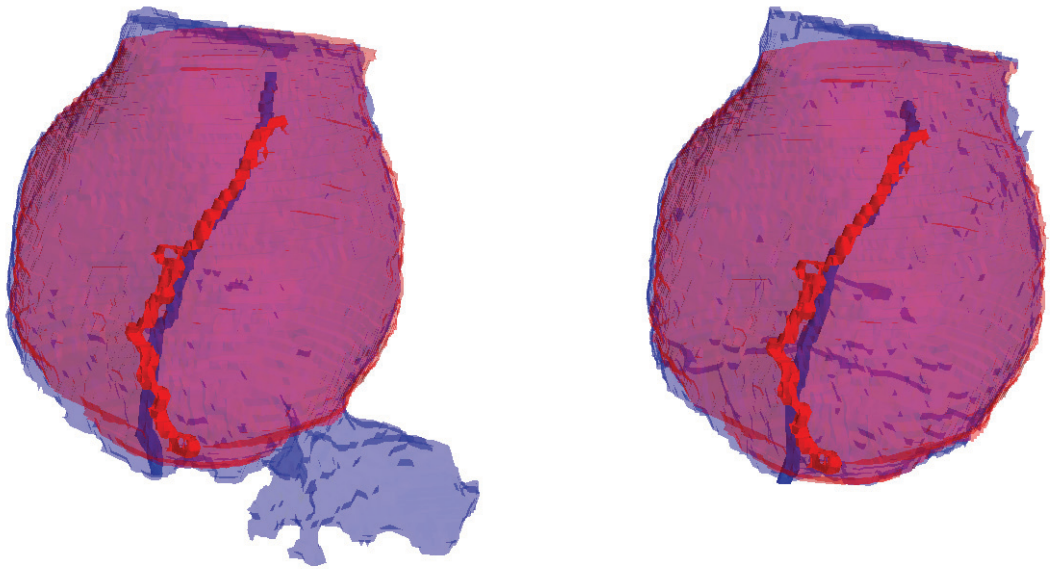

**Figure 5.** The effect of the iterative atlas selection is illustrated using an example from the Liverpool dataset. Left: the automatic segmentation is subject to an error introduced by mis-registered atlases. Right: after iterative atlas selection is applied these mis-registered atlases are removed, resulting in a significant improvement in segmentation performance. Surface renderings of the automatic segmentations of the whole heart and LADCA are shown in blue, with manual delineations shown in red.

for example in clinical trial quality assurance where analysis of contouring differences is difficult (Jameson *et al* 2010, Vinod *et al* 2016).

In the iterative atlas selection procedure, the calculation of the scaled distance effectively removes the effect of the consensus volume as a reference. Although the reference volume is thus not required for this calculation, it is useful for several reasons. Firstly, computing the surface variation relative to a consistent surface provides a measure of the physical distance between contours, and so the surface variation between different patients can be directly compared. Secondly, this provides additional flexibility in adapting this framework, for example by scaling the surface variation, where the physical distance to a reference surface is more independent of the structure size and thus more appropriate to use. Lastly, and most importantly, we are currently in the process of extending this procedure so that it can be used on structures of any shape. For non-convex structures, the projection onto a common grid is replaced with measuring the distance between each atlas contour and a reference surface directly, and following a similar iterative selection process from there. This differs from the current procedure in that distance evaluation would occur on the reference surface, rather than the atlas surface, and because of this the surface variation is not evaluated at every point on each atlas surface, meaning it is possible that errors in atlas contours may be missed. For this reason we have used the more robust method as presented, which requires interpolation onto a common grid.

In evaluating whether an atlas should be removed we introduce the only free parameter in the procedure: the threshold for atlas removal. If this threshold is too low the process could remove atlases of reasonable quality. The use of larger atlas sets has been shown to improve segmentation accuracy generally (Aljabar *et al* 2009), as well as specifically for whole heart segmentation in CT imaging (Zhuang *et al* 2015), with measures of segmentation accuracy reaching a maximum when approximately 10 atlases are used and a plateau in quality with the use of additional atlases. The removal of too many atlases may reduce segmentation accuracy. Conversely, a threshold set too high could potentially introduce atlases subject to mis-registration with the target image and produce errors in the final segmentation. The optimal threshold will also vary with the desired application: in the context of unsupervised segmentation of large, variable datasets the removal of all errors at the cost of a potential reduction in overall accuracy is preferred. The given threshold, derived from the standard definition of an outlier (1.5 inter-quartile ranges above the 75th percentile), was able to correct segmentation errors in the data used in this study with no overall reduction in measures of segmentation accuracy; future work could address the effect of different thresholds.

This study used cross-validation with independent datasets to examine the effect of differences in patient set-up, image acquisition parameters and on the accuracy of automatic segmentation and iterative atlas selection. A limitation of our presented iterative atlas selection procedure is that the unique mapping of points of the heart surface relies on a convex shape. For more complicated structures this condition will not hold, however, a simple solution would be to calculate the scaled distance on the surface of the consensus segmentation, rather than projecting this distance onto a grid. Imaging data used in this study was collected without gating techniques, which

leads to blurring of the images and hence the delineations can be thought of as a time-averaged volume. The dose that would be calculated for these volumes would thus also be a time-averaged estimation, without taking into account anatomical motion. While we expect this effect to be small it may affect the accuracy of dose measurements using both automatic segmentations and manual contours.

Atlas-based cardiac segmentation has been used in the context of cardiac imaging (Išgum *et al* 2009, Zhuang *et al* 2010, Bai *et al* 2014), where the motivation for accurate segmentation stems from characterising cardiac function or disease. The presented iterative atlas selection procedure is independent of the imaging modality, and is therefore suitable to be included in existing atlas-based segmentation frameworks designed for cardiac imaging.

## 5. Conclusion

This work serves as a promising step for the accurate and precise cardiac segmentation in retrospective datasets. Future studies to ensure the dosimetric consistency of automatic segmentations are necessary, and using a multi-observer dataset (Lorenzen *et al* 2014) we plan to validate our framework considering the variability in dose as a result of contouring differences. Further investigation to assess the impact of the iterative atlas selection procedure on the segmentation accuracy of cardiac substructures is desirable for situations where accurate substructure delineation is critical.

The iterative atlas selection algorithm leverages the *a priori* power of atlases, and with optimised segmentation thresholds, underpins accurate, consistent and robust automatic segmentation of the heart. The independence of imaging modality makes this approach useful in radiotherapy, cardiac imaging, and potentially for segmentation of many other anatomical structures. We intend to apply our framework in the analysis of large, retrospective breast cancer studies, where precise dose estimation is necessary to understand the effect of radiation on the heart at a more detailed level.

## Acknowledgments

CB acknowledges support from AgeCare (Academy of Geriatric Cancer Research), an international research collaboration based at Odense University Hospital, Denmark.

We would like to thank Eng-Siew Koh, Simon Tang, Geoff Delaney, Vikneswary Batumalai, Carol Luo, Pramukh Atluri, and Athiththa Satchithanandha at Liverpool and Maccarthur Cancer Therapy Centres, for providing delineations of the Liverpool CT data and helpful discussions.

We are grateful for the help of Marianne Ewertz, Carolyn W Taylor, Maja Maraldo, Mette H Nielsen, and Birgitte V O'ersen, Maria R Andersen, Dean O'Dwyer Lone Larsen, Sharon Duxbury, and Baljit Jhitta in providing contouring of the Odense CT data.

We would also like to thank James Otton for providing cardiology expertise.

## ORCID iDs

Robert Finnegan https://orcid.org/0000-0003-4728-8462

## References

Aljabar P, Heckemann R A, Hammers A, Hajnal J V and Rueckert D 2009 Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy *NeuroImage* **46** 726–38

Antonelli M, Cardoso M J, Johnston E W, Appayya M B, Presles B, Modat M, Punwani S and Ourselin S 2019 GAS: a genetic atlas selection strategy in multi-atlas segmentation framework *Med. Image Anal.* **52** 97–108

Bai W, Shi W, Ledig C and Rueckert D 2014 Multi-atlas segmentation with augmented features for cardiac MR images *Med. Image Anal.* **19** 98–109

Ciardo D *et al* 2017 Atlas-based segmentation in breast cancer radiotherapy: evaluation of specific and generic-purpose atlases *Breast* **32** 44–52

Correa C R, Litt H I, Hwang W T, Ferrari V A, Solin L J and Harris E E 2007 Coronary artery findings after left-sided compared with right-sided radiation treatment for early-stage breast cancer *J. Clin. Oncol.* **25** 3031–7

Cui Y *et al* 2015 Contouring variations and the role of atlas in non-small cell lung cancer radiation therapy: analysis of a multi-institutional preclinical trial planning study *Pract. Radiat. Oncol.* **5** 67–75

Cutuli B, Bernier J and Poortmans P 2014 Radiotherapy in DCIS, an underestimated benefit? *Radiother. Oncol.* **112** 1–8

Darby S C *et al* 2010 Radiation-related heart disease: current knowledge and future prospects *Int. J. Radiat. Oncol. Biol. Phys.* **76** 656–65

Darby S C, McGale P, Taylor C W and Peto R 2005 Long-term mortality from heart disease and lung cancer after radiotherapy for early breast cancer: prospective cohort study of about 300 000 women in US SEER cancer registries *Lancet Oncol.* **6** 557–65

Darby S *et al* (Early Breast Cancer Trialists' Collaborative Group (EBCTCG)) 2011 Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials *Lancet* **378** 1707–16

Dodge J T, Brown B G, Bolson E L and Dodge H T 1992 Lumen diameter of normal human coronary arteries. Influence of age, sex, anatomic variation, and left ventricular hypertrophy or dilation *Circulation* **86** 232–46

Doshi J, Erus G, Ou Y, Resnick S M, Gur R C, Gur R E, Satterthwaite T D, Furth S and Davatzikos C 2016 MUSE: MUlti-atlas region segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection *NeuroImage* **127** 186–95

Duane F *et al* 2017 A cardiac contouring atlas for radiotherapy *Radiother. Oncol.* **122** 416–22

EBCTCG (Early Breast Cancer Trialists' Collaborative Group) 2014 Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: meta-analysis of individual patient data for 8135 women in 22 randomised trials *Lancet* **383** 2127–35

Eldesoky A R *et al* 2016 Internal and external validation of an ESTRO delineation guideline dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer *Radiother. Oncol.* **121** 424–30

Feng M *et al* 2011 Development and validation of a heart atlas to study cardiac exposure to radiation following treatment for breast cancer *Int. J. Radiat. Oncol. Biol. Phys.* **79** 10–8

Finnegan R *et al* 2019 Feasibility of multi-atlas cardiac segmentation from thoracic planning CT in a probabilistic framework *Phys. Med. Biol.* **64** 085006

Fletcher R 1987 *Practical Methods of Optimization* 2nd edn (New York: Wiley-Interscience)

Gagliardi G, Constine L S, Moiseenko V, Correa C, Pierce L J, Allen A M and Marks L B 2010 Radiation dose volume effects in the heart *Int. J. Radiat. Oncol. Biol. Phys.* **76** S77–85

Hahn E, Jiang H, Ng A, Bashir S, Ahmed S, Tsang R, Sun A, Gospodarowicz M and Hodgson D 2017 Late cardiac toxicity after mediastinal radiation therapy for hodgkin lymphoma: contributions of coronary artery and whole heart dose-volume variables to risk prediction *Int. J. Radiat. Oncol. Biol. Phys.* **98** 1116–23

Hooning M J, Botma A, Aleman B M P, Baaijens M H A, Bartelink H, Klijn J G M, Taylor C W and van Leeuwen F E 2007 Long-term risk of cardiovascular disease in 10-year survivors of breast cancer *JNCI J. Natl Cancer Inst.* **99** 365–75

Išgum I, Staring M, Rutten A, Prokop M, Viergever M A and Van Ginneken B 2009 Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans *IEEE Trans. Med. Imaging* **28** 1000–10

Jameson M G, Holloway L C, Vial P J, Vinod S K and Metcalfe P E 2010 A review of methods of analysis in contouring studies for radiation oncology *J. Med. Imaging Radiat. Oncol.* **54** 401–10

Kaderka R *et al* 2019 Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients *Radiother. Oncol.* **131** 215–20

Klein S, van der Heide U A, Lips I M, van Vulpen M, Staring M and Pluim J P W 2008 Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information *Med. Phys.* **35** 1407–17

Langerak T R, Berendsen F F, Van Der Heide U A, Kotte A N and Pluim J P 2013 Multiatlas-based segmentation with preregistration atlas selection *Med. Phys.* **40**

Langerak T R, Van Der Heide U A, Kotte A N T J, Viergever M A, Van Vulpen M and Pluim J P W 2010 Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE) *IEEE Trans. Med. Imaging* **29** 2000–8

Li X A *et al* 2009 Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG multi-institutional and multiobserver study *Int. J. Radiat. Oncol. Biol. Phys.* **73** 944–51

Lorenzen E and Brink C 2012 Automatic segmentation of heart evaluated with multi-institution interobserver variation *Int. J. Radiat. Oncol. Biol. Phys.* **84** S800

Lorenzen E L, Ewertz M and Brink C 2014 Automatic segmentation of the heart in radiotherapy for breast cancer *Acta Oncol.* **53** 1366–72

Lorenzen E L *et al* 2013 Inter-observer variation in delineation of the heart and left anterior descending coronary artery in radiotherapy for breast cancer: a multi-centre study from Denmark and the UK *Radiother. Oncol.* **108** 254–8

Luo Y, Xu Y, Liao Z, Gomez D, Wang J, Jiang W, Zhou R, Williamson R, Court L E and Yang J 2018 Automatic segmentation of cardiac substructures from noncontrast CT images: accurate enough for dosimetric analysis? *Acta Oncol.* **58** 1–7

Morris E D, Ghanem A I, Pantelic M V, Walker E M, Han X and Glide-Hurst C K 2019 Cardiac substructure segmentation and dosimetry using a novel hybrid magnetic resonance and computed tomography cardiac atlas *Int. J. Radiat. Oncol. Biol. Phys.* **103** 985–93

Ou Y and Doshi J 2012 Multi-atlas segmentation of the prostate: a zooming process with robust registration and atlas selection *MICCAI Grand Challenge: Prostate MR Image Segmentation* pp 1–7

Stam B, Peulen H, Guckenberger M, Mantel F, Hope A, Werner-Wasik M, Belderbos J, Grills I, O'Connell N and Sonke J J 2017 Dose to heart substructures is associated with non-cancer death after SBRT in stage III NSCLC patients *Radiother. Oncol.* **123** 370–5

Taylor C W, Wang Z, Macaulay E, Jagsi R, Duane F and Darby S C 2015 Exposure of the heart in breast cancer radiation therapy: a systematic review of heart doses published during 2003 to 2013 *Int. J. Radiat. Oncol. Biol. Phys.* **93** 845–53

van Rikxoort E M, Isgum I, Arzhaeva Y, Staring M, Klein S, Viergever M A, Pluim J P W and van Ginneken B 2010 Adaptive local multi-atlas segmentation: application to the heart and the caudate nucleus *Med. Image Anal.* **14** 39–49

Vercauteren T, Pennec X, Perchant A and Ayache N 2009 Diffeomorphic demons: efficient non-parametric image registration *NeuroImage* **45** S61–72

Vinod S K, Jameson M G, Min M and Holloway L C 2016 Systematic review Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies *Radiother. Oncol.* **121** 169–79

Wennstig A K, Garmo H, Hållström P, Nyström P W, Edlund P, Blomqvist C, Sund M and Nilsson G 2017 Inter-observer variation in delineating the coronary arteries as organs at risk *Radiother. Oncol.* **122** 72–8

Zhou R *et al* 2017 Cardiac atlas development and validation for automatic segmentation of cardiac substructures *Radiother. Oncol.* **122** 66–71

Zhuang X, Bai W, Song J, Zhan S, Qian X, Shi W, Lian Y and Rueckert D 2015 Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection *Med. Phys.* **42** 3822–33

Zhuang X, Rhode K S, Razavi R S, Hawkes D J and Ourselin S 2010 A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI *IEEE Trans. Med. Imaging* **29** 1612–25