Received: 19 September 2023 | Revised: 29 January 2024 | Accepted: 3 March 2024

DOI: 10.1002/mp.17026

MEDICAL PHYSICS DATASET ARTICLE

MEDICAL PHYSICS

A comprehensive lung CT landmark pair dataset for evaluating deformable image registration algorithms

Edward R. Criscuolo[1] | **Yabo Fu**[2] | **Yao Hao**[3] | **Zhendong Zhang**[1] | **Deshan Yang**[1]

[1]Department of Radiation Oncology, Duke University, Durham, North Carolina, USA

[2]Memorial Sloan Kettering Cancer Center, New York, New York, USA

[3]Washington University School of Medicine, St. Louis, Missouri, USA

Correspondence
Deshan Yang, Department of Radiation Oncology, School of Medicine, Duke University, 40 Duke Medicine Circle, Room 04212, 3640 DUMC, Durham, NC 27710, USA.
Email: deshan.yang@duke.edu

Funding information
National Institute of Biomedical Imaging and Bioengineering (NIBIB), Grant/Award Number: R01-EB029431

Abstract

Purpose: Deformable image registration (DIR) is a key enabling technology in many diagnostic and therapeutic tasks, but often does not meet the required robustness and accuracy for supporting clinical tasks. This is in large part due to a lack of high-quality benchmark datasets by which new DIR algorithms can be evaluated. Our team was supported by the National Institute of Biomedical Imaging and Bioengineering to develop DIR benchmark dataset libraries for multiple anatomical sites, comprising of large numbers of highly accurate landmark pairs on matching blood vessel bifurcations. Here we introduce our lung CT DIR benchmark dataset library, which was developed to improve upon the number and distribution of landmark pairs in current public lung CT benchmark datasets.

Acquisition and Validation Methods: Thirty CT image pairs were acquired from several publicly available repositories as well as authors' institution with IRB approval. The data processing workflow included multiple steps: (1) The images were denoised. (2) Lungs, airways, and blood vessels were automatically segmented. (3) Bifurcations were directly detected on the skeleton of the segmented vessel tree. (4) Falsely identified bifurcations were filtered out using manually defined rules. (5) A DIR was used to project landmarks detected on the first image onto the second image of the image pair to form landmark pairs. (6) Landmark pairs were manually verified. This workflow resulted in an average of 1262 landmark pairs per image pair. Estimates of the landmark pair target registration error (TRE) using digital phantoms were 0.4 mm $\pm$ 0.3 mm.

Data Format and Usage Notes: The data is published in Zenodo at https://doi.org/10.5281/zenodo.8200423. Instructions for use can be found at https://github.com/deshanyang/Lung-DIR-QA.

Potential Applications: The dataset library generated in this work is the largest of its kind to date and will provide researchers with a new and improved set of ground truth benchmarks for quantitatively validating DIR algorithms within the lung.

KEYWORDS
deformable image registration, ground truth dataset, lung motion

1 | INTRODUCTION

Deformable image registration (DIR) is an image processing task that involves finding the voxel-wise correspondence between two images.[1] The output of DIR is a DVF, which describes the spatial transformation between the two images. DIR is a key enabling technology for many important diagnostic and therapeutic tasks, for example, tumor definition,[2] image segmentation,[3,4] motion estimation,[5] evaluation of tumor response or

treatment,[6] and image-guided surgery.[7] DIR is applicable to most cancer patients, whose multiple image scans need to be evaluated and analyzed jointly to derive diagnosis or to support treatment decisions. While DIR can be used to register images from different patients or modalities, this study focuses on DIRs between image pairs of the same modality and patient.

Despite its importance, DIR errors frequently exceed clinical tolerances, particularly in cases with poor image quality or significant tissue deformation.[8,9] The verification of DIR accuracy before implementation in the clinic is crucial but poses several challenges. Common methods to evaluate DIR accuracy include; using manually defined landmark pairs and structure contours to evaluate target registration error (TRE), using similarity coefficients between deformed moving and target images, and the use of digital phantoms.[1] Landmark pairs are a simple and accurate method, however, defining these landmarks often requires time-consuming effort from expert observers and is subject to inter-observer variations. In addition, a large number of uniformly distributed landmarks is required to produce a valuable quantitative measure. Image similarity metrics are convenient, but not directly related to TRE,[10] therefore are not directly relevant to treatment target and organ motion. Phantoms can be useful in ideal conditions, however, patient-specific anatomy, image quality, artifacts, and DVF complexity from clinical cases often cannot be captured.[11]

Lung registration is one of the most tractable registration scenarios, with mean TREs of 0.7 to 2 mm.[12,13] This is in part due to the benchmark landmark pair datasets that currently exist within the lung that have supported algorithm development in the past decade. Of note are the POPI, the DIRLAB COPD gene, and the DIRLAB 4DCT datasets. The POPI-model dataset contains 10 4DCT phases of a single breathing cycle, with 40 labeled anatomical landmarks in each frame.[14,15] Effectively, this results in 9 CT image pairs with 40 landmarks per pair. In addition, 6 separate 4D CT images with 100 landmarks identified in the full inhale and exhale frame are available.[16] The DIRLAB 4DCT and COPD datasets from Castillo et al.[17,18] have been widely used to verify image registration algorithms in recent years. The 4DCT dataset consists of 300 manually identified landmark pairs in 10 end-exhalation and end-inhalation 4DCT image pairs, as well as 75 landmarks for the intermediate phase images. The COPD gene dataset similarly contains 300 manually identified landmark pairs for 10 inspiratory-expiratory CT image pairs from patients with chronic obstructive pulmonary disease (COPD).

Although these and similar datasets have been effective in spurring algorithm development in recent years, they are not sufficient for robust evaluation of current high-performance DIR methods. The error in the semi-manual landmark identification in the POPI validation dataset was estimated to be 0.5 ± 0.9 mm,[16] while the average inter-observer uncertainty of the DIRLAB landmarks was reported as ~0.88 ± 1.31 mm.[19] TREs of the most accurate DIR algorithms among the 26 assessed using the DIRLAB datasets were 0.91 ± 1.07 mm, as published on www.dir-lab.com. This indicates that the variability in the manual delineation of landmarks, both in the DIRLAB and POPI datasets, limits the evaluation of high-precision registration algorithms, as the inter-observer uncertainty is comparable to the sub-millimeter precision of the algorithms. In addition, the 300 public landmark pairs per 4DCT image pair in the DIRLAB dataset, as well as the 40 to 100 landmark pairs per image pair in the POPI dataset, are too scarce to accurately sample the entire lung volume. Consequentially, few landmarks in these datasets are available in the lower lung, where tissue motion is most significant. As such, the landmarks may not fully reflect DIR performance.[20] This presents a hurdle for researchers to robustly evaluate their own DIR algorithms. Despite this, the DIRLAB and POPI datasets have proved crucial tools in algorithm development and highlight the need for more comprehensive, public benchmark datasets.

To reduce the difficulty in manually identifying landmark pairs, a number of efforts have been made to automate this process. Yang et al. developed a method to automatically detect large numbers of landmark pairs using scale-invariant feature transforms (SIFT) and Harris-Laplacian corner detection algorithms.[21] However, this method is limited by the landmark matching done by the proposed Multiple-Resolution Inverse-Consistency Guided Matching (MRICGM) procedure, where the best available landmark in the second image may not truly correspond with the first. Werner et al. and Polzin et al.[22] used a Foerstner3D operator to detect landmarks on one image and then transferred the landmarks to the other image using a cross correlation-based block matching strategy. However, not only was the cross-correlation metric an inaccurate measure for images with significant deformation or distortion, only a small fraction of the detected landmarks could be reliably transferred, and the block-matching strategy accuracy was not verified in the study. The results of Fu et al.[19] offered an important step forward by detecting landmarks using the Harris-Stephens corner detection algorithm on a vascular probability tree, projecting these landmarks using an approximate registration, and refining the spatial positions using a deep learning network. This automatic process resulted in an average of 1886 landmarks across 10CT image pairs from the DIRLAB 4DCT dataset, with a mean TRE of 0.47 mm ± 0.45 mm, with 98% of landmark pairs having a TRE smaller than 2 mm for 10 digital phantom cases. However, these landmarks were not manually verified. In addition, although a vascular probability map was used, the landmarks were detected using the Harris Stephens corner detection algorithm, and therefore did not necessarily occur on any vessel bifurcations. As such, it is difficult to confidently

publish these results as a ground truth dataset. A similar method was recently described by Cazoulat et al.,[23] in which vessel bifurcations in DIRLAB images were identified on the skeleton of vessel trees segmented by vesselness thresholding. Landmark pair correspondence was then established via registration of the vessel trees.

Small vessels and vessel bifurcations are well suited for CT image DIR evaluation because they are stable through time, display high contrast, and can represent complex deformations of nearby tissues and organs. In addition, they can be used to verify any type of intra-patient DIR algorithm, including novel and state-of-the-art deep-learning-based approaches. The current need, therefore, is a large dataset that is easily accessible and publicly available, covers a variety of scanner types and clinical cases, and contains a high number of uniformly distributed landmark pairs on visible vessel bifurcations, with adequate landmark pairs in the lower and outer regions. Therefore, we developed a semi-automatic workflow based on the process developed by Fu et al.[19] to efficiently identify a large number of vessel-bifurcation landmark pairs in CT images. We then applied this workflow to 30 CT image pairs to develop a benchmark dataset library.

2 | ACQUISITION AND VALIDATION METHODS

2.1 | Data sources

Forty-four pairs of CT scans were selected from several publicly available image repositories as well as clinical data from Barnes Jewish Hospital (BJH). An image pair in this dataset references two CT images acquired on the same patient at different times, meaning eighty-eight total CT scans were acquired from 44 different patients, forming 44 pairs. The protected health information (PHI) of the patients in these pairs were removed from the headers of the DICOM files and identifiers were encoded as case numbers. The process for obtaining and encoding the data underwent IRB approval. The scan acquisition parameters or characteristics of the patients between image pairs may differ. For a full description of each scan, please refer to the Supplemental information or the original data repository.

Among the 44 cases, 21 cases were selected from the EMPIRE10 grand challenge.[24] Data from the EMPIRE10 challenge consists of pairs of chest CT scans. Each pair was obtained from a unique patient, sourced from a variety of institutes. Scan pairs were originally obtained on many different scanners in different voxel sizes and image quality. Scans may be taken at various phases in the breathing cycle (full inspiration, full expiration, phase from 4D breathing data), from healthy subjects or subjects with lung diseases.

Six CT scan pairs were sourced from The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) dataset from the NIH.[25,26] The TCGA-LUAD data collection contained clinical scans as well as genetic and pathological data for patients with lung adenocarcinoma, either from routine clinical care or as a part of research trials. Four CT scan pairs were sourced from The Cancer Genome Atlas Lung Squamous Cell Carcinoma Collection (TCGA-LUSC).[25,27] The TCGA-LUAD imaging data was obtained from The Cancer Imaging Archive (TCIA),[25] and was originally sourced from various institutions and scanners, and therefore also displayed varying voxel resolutions or image quality.

Thirteen CT scan pairs were obtained from clinical scans at BJH with IRB approval. Scans were from different patients, either as a part of routine care or from lung cancer screening and treatments.

After the application of the landmark detection workflow to the image pairs described above, four cases from EMPIRE 10, two cases from TCGA-LUAD, and eight cases from BJH were thrown out due to poor image quality and a lack of landmark pairs. The final result was a high number of landmark pairs identified for 30 CT scan pair cases; 17 from EMPIRE10, 4 from TCGA-LUAD, 4 from TCGA-LUSC, and 5 from BJH.

2.2 | Landmark pair identification

Due to the laborious task of manual landmark pair delineation, a semi-automatic landmark identification pipeline was developed. This process was based on the pathway developed by Fu et al.[19] The goal of this pipeline is to accurately and efficiently identify a high number of small vessel bifurcation landmarks pairs between two CT images.

First, image denoising and automatic segmentation of the lungs and vessels were performed for both images of a pair. Bifurcations were then detected directly on the skeleton of the segmented vascular tree from the higher-quality image in a pair. These landmarks were then projected onto the second image using the DVF acquired from a registration between the two images. These landmark pairs were then manually verified for accuracy and discarded or adjusted if necessary. The result was a high number of accurate landmark pairs uniformly distributed throughout the lung. The process is summarized below in Figure 1.

2.2.1 | Denoising and lung auto segmentation

The CT images were first passed to FFDNet, a flexible deep learning based denoiser.[28] This denoising helped to ensure less artifacts and noise were included in the subsequent vessel segmentation step. The strength of

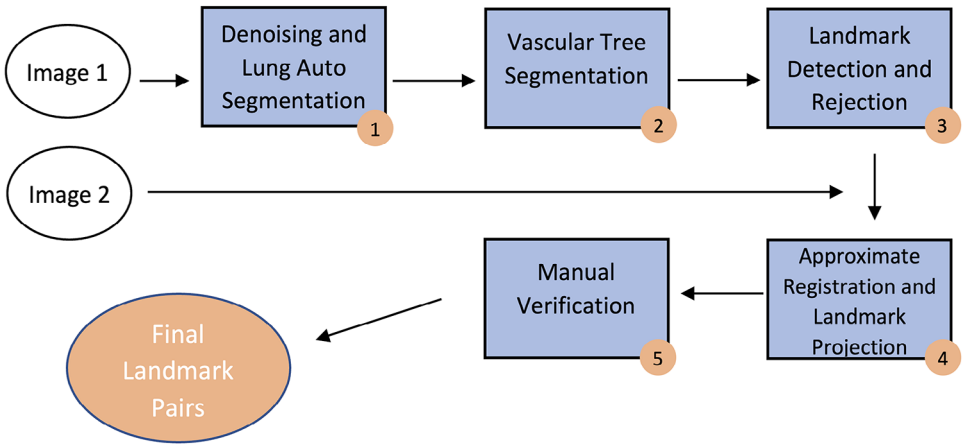

FIGURE 1 The landmark pair identification pathway. The depicted pathway was applied to each image pair in the dataset. Details of each step in the process can be found in the subsequent sections, numbered according to Figure 1.

denoising in FFDNet was controlled by a tunable noise level map input, which ranged from 0 (no noise) to 75 (highest noise). The denoising strength values were chosen empirically per case by visually checking if blood vessels began to be obscured or lost in the denoising images. To decide this, an image was denoised and the absolute difference between the denoised image and the original image was taken and displayed. High levels of denoising remove small vessel anatomy from the image, which is undesirable given our pathway. Therefore, if blood vessels were present on the noisy-denoised difference image, then there are vessels that appear in the original image that were removed during denoising. In this case, the denoising strength should be reduced. If no vessel anatomy was present in the difference image, then there is room to increase the denoising strength. This denoising strength selection process was repeated until the ideal strength value was found. This is visualized in Figure 2.

For each image pair, the lungs, airways, and fissures were all automatically segmented on both images after denoising using the pulmonary toolkit (PTK) from Doel et al.[29] The airways and fissure masks were expanded by 3 voxels and excluded from the subsequent vessel segmentation step to minimize the number of false landmarks, as the vessel segmentation often included artifacts near these regions. In addition, the vessel segmentation is not robust at the periphery of the lung due to the vast number of minor peripheral vessels, so the lung mask was also eroded by three voxels.

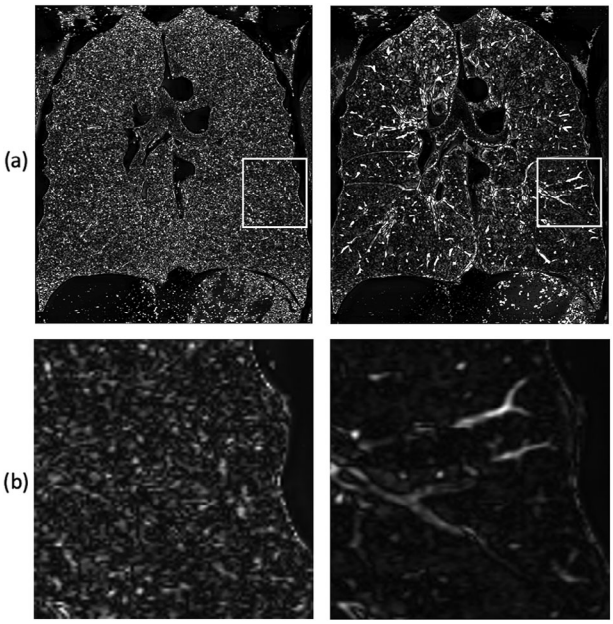

FIGURE 2 Choosing of denoising kernel for a single scan. The absolute difference between the noisy and denoised images for two different FFDNet noise level maps is shown above (left and right). The final noise map size should exist somewhere in between these two. (a) Difference image showing vessel anatomy not removed (left) by denoising and when it is (right). (Image from case 1 CT2, denoising kernels 10 and 60, respectively). (b) Zoomed in portion of difference images from (a). Left: The difference is mostly random noise. Right: Denoising kernel is too large, some of the vessel anatomy is removed and shows up in the difference image.

2.2.2 | Lung and vascular tree segmentation

To segment the vessel tree in each image, a vesselness map of all points in the image was calculated with the Frangi vesselness filter.[30] This vesselness is then used to segment the vessel tree either by an in-house code utilizing hysteresis thresholding (including contributions from the vesselness and image intensity), or by thresholding methods offered by PTK. The vessel tree segmented by these methods includes both arteries and veins, and the final bifurcation landmarks can occur on either. This should have little to no effect on DIR

FIGURE 3 Example of segmented lung, airways (blue), and vessels (red) from case 2 CT2. The vessel tree in this case was segmented using an in-house hysteresis thresholding method.

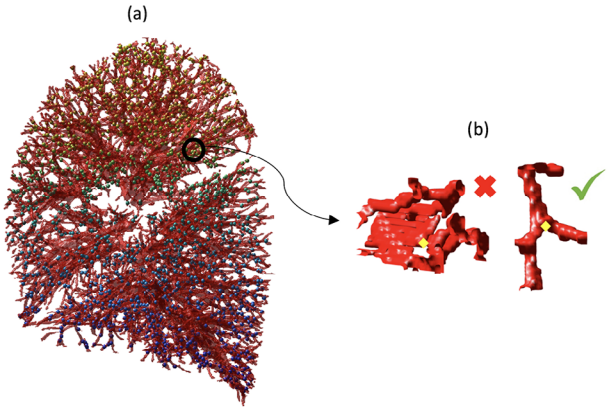

FIGURE 4 Segmented vessel tree with bifurcations followed by individual rejected landmarks. Segments are from case 21 CT1, and the vessel tree was segmented using in-house hysteresis thresholding. Bifurcations, identified on the skeleton of the tree displayed, are shown as colored nodes. (a) Example of a vessel map with bifurcations. (b) Two example bifurcation landmarks are labeled in yellow. (Left) A landmark was rejected due to density, likely the remnant of the lung fissure. (Right) A good bifurcation was included as a landmark.

algorithm validation, provided landmark pairs are confirmed to occur on the same anatomy, as both displayed with similar contrast in the lung.

To choose the thresholding method, internal or PTK, both were applied to each image case. Bifurcations are then detected directly on the skeletons of both segmented vessel trees using vessel thinning and branching point detection methods in MATLAB. A rule-based rejection (step 3) was then applied to eliminate bifurcations that are likely false in both trees. All of these steps are automated and can be performed rapidly. The segmented tree with more bifurcations after this step was considered superior and was used in all subsequent steps, and the tree generated using the other method was discarded. The superior thresholding method was not consistent between cases, and therefore this step was applied in each image pair case to ensure the maximum number of landmarks were generated. If the two trees had a very similar number of landmarks, the superior tree was chosen based on its visual quality as determined by artifacts and the vessel density in the outer lung. An example output of a segmented vessel tree is shown in Figure 3.

2.2.3 | Landmark detection and rejection

The vessel tree segmentation step was efficient but not robust, so the vessel tree and its skeleton often included artifact bifurcations that were not on true vessel anatomy. For example, a vessel tree ideally should develop outward, with progressively smaller vessels until they can no longer be identified given the image resolution. If the vessel tree forms a localized loop, that portion of the tree is likely segmented incorrectly. Secondarily, regions of the skeleton tree that are particularly dense (see Figure 4b), indicate that bifurcations from this region are likely artifacts of image noise or non-vessel anatomy.

In addition, short branches of the vessel tree, although potentially existing as true vessel anatomy, are often the result of segmentation noise or error. As such, it is most efficient to exclude these from the dataset as well. Finally, large vessels have bifurcations with poorly defined spatial positions, so automated landmark placement at these points may have unpredictable variations. They also most commonly occur in regions of little to no anatomical motion, making them less useful in DIR algorithm validation. Therefore, we applied a rule-based approach to eliminate bifurcations of these types. This step was also used in choosing the superior vessel tree segmentation method, as previously described in Step 2. This approach automatically discarded bifurcations on large vessels, that occur as part of localized loops or in high-density regions, or that contain short branches. The limit for the diameter of the large vessels was chosen to be 4 voxels, as this only eliminated an average of 5% of initially detected bifurcation landmarks and limited the centerline to a region of about a voxel. Short branches were empirically chosen to be those less than 5 voxels long to limit the number of false bifurcations to verify while maintaining a high number of total landmarks.

Loops were identified by creating a MATLAB graph from the vessel skeleton and flagging branches that

connect back to the same bifurcation in less than 3 graph nodes (bifurcations). All bifurcations in these loops were considered artifacts and were excluded. To identify small branches, the length of each branch connected to a bifurcation was calculated by iterating outward voxel by voxel from the bifurcation point in the vessel skeleton. If the number of iterations upon the branch terminating or connecting with a second bifurcation was less than 5, then the branch was considered too short and was excluded. This was followed by a density reduction step, where landmark spatial density was made uniform within local regions. Uniformity was set to include only the highest quality bifurcations, those with the highest local vesselness value, spaced by 5 mm.

The above steps, 1–3, were automated and applied to both images in an image pair. However, only the landmarks from the higher-quality of the two images were used in subsequent steps. Quality was determined visually and by the total number of bifurcations identified in each image.

2.2.4 | Approximate registration and landmark projection

The two images in a pair were then registered using pTVreg, a parametric image registration algorithm that performed highly on previous benchmarks.[15] The higher quality of the two images, in which the landmarks were detected, was used as the fixed image. Both images were denoised before registration, according to the steps outlined in Figure 2. Due to the limitations on system memory, the left and right lungs were registered separately, and then recombined after. The parameters of the registration are as follows: Image dissimilarity metric: local correlation. Metric param = [2.1, 2.1, 2.1]. Border Mask = 5 voxels. The number of pyramid levels was determined automatically based on image size. All other parameters were set as default according to pTVreg documentation.

Using the displacement information from the DVF from this registration, Landmarks detected on the fixed image were then projected to the second image, as shown in Figure 5. These detected landmarks and their projections form the landmark pairs of the dataset. Non-integer voxel locations of the second landmark are determined using linear interpolation.

2.2.5 | Manual verification

The landmark pairs were manually checked for accuracy using both CT image information around the bifurcation points and vessel segmentation from both images. Some landmark pairs, particularly in areas of significant image noise or distortion, were not properly projected to the second image. This was caused by image registration errors. These false pairs were manually identified and excluded from the final list of landmark pairs. Landmarks that were clearly not on vessel bifurcations despite the rule-based rejection were also removed. This process is demonstrated in Figure 6.

Finally, if a projected landmark was close to the proper bifurcation position but required minor adjustment, the position was corrected manually. This correction was only done for a small subset of landmarks (less than 1%), so contributions of intra-observer variability to the final landmark TRE should be minimal.

2.3 | Accuracy assessment

2.3.1 | TRE estimation

The manual verification to reject poor landmarks was the first step of quality control in this dataset. It ensured that any inaccurate landmark pairs resulting from errors in the approximate registration or by the bifurcation detection step were not included in the final dataset.

Despite the semi-automatic manner in which the landmark pairs were generated, there were still positional errors associated with the final projected landmark pairs. This error arises due to uncertainties in the vessel segmentation and landmark projection steps in our pathway. To assess the error magnitude, a series of digital phantom image pairs with known DVFs were created. DVFs were created by registering 5 randomly selected image pair cases in the dataset using the demon's DIR algorithm.[31] The demon's algorithm was used in generating the DVF to avoid bias towards the pTVreg algorithm which was used to project landmarks in the pathway. The generated DVFs were used to deform images from different patients to form new CT image pairs. The landmark identification pathway in Figure 1 was then performed on the initial and artificially deformed images, and ground truth knowledge from the demon's DVFs was used to estimate the error in the final landmark positions. This process results in 3562 phantom landmark pairs with ground truth information we can use to estimate the TRE. 85 outliers (2.4%) with errors greater than 3 mm that would be identified by the manual correction step were excluded from the final TRE calculation. The mean registration error and standard deviation after outlier exclusion was TRE = 0.4 mm $\pm$ 0.3 mm. Over 77% of the phantom landmark pairs had a TRE lower than 0.5 mm. This process is visualized below in Figure 7.

2.3.2 | Validation using landmarks labeled by experts

To verify our pathway with landmark pairs from expert observers, we applied our pathway to the first 5 cases

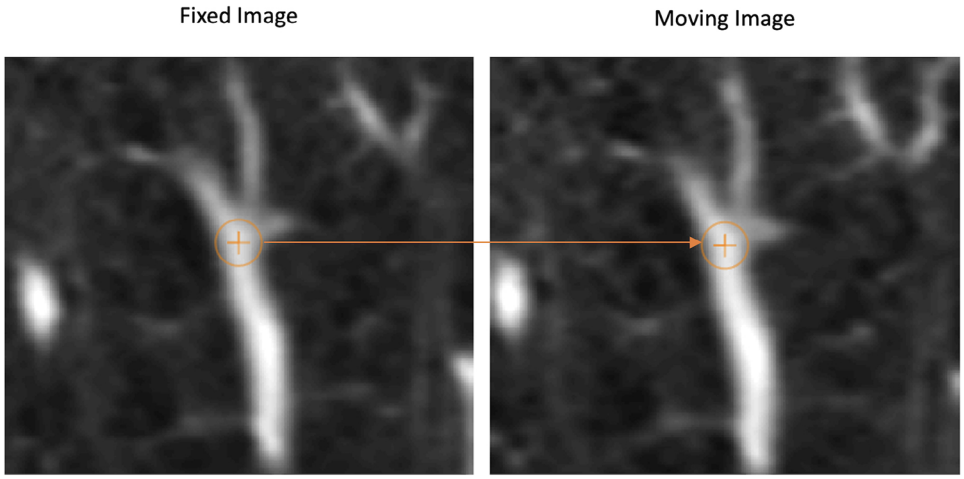

FIGURE 5 Example of Landmark projection. Landmarks (bifurcations) are detected on the left image and are projected onto the image on the right using the DVF from registration with pTVreg.

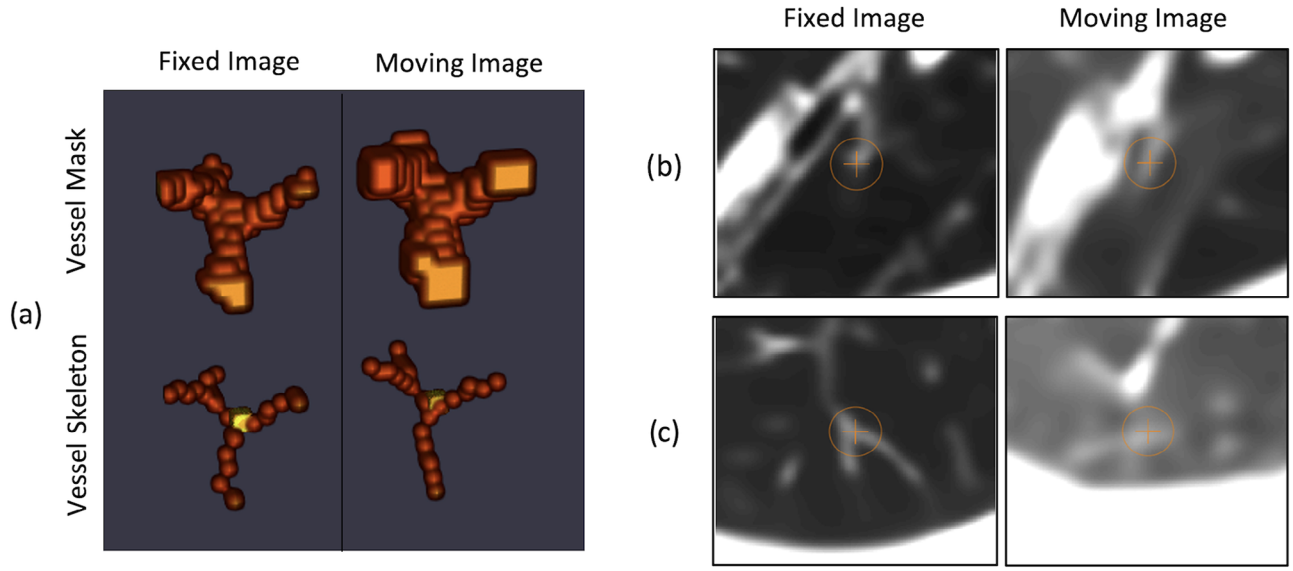

FIGURE 6 The two views were used to manually verify the landmark accuracy. In (a, b, c), landmarks detected on the fixed image (left) were projected onto the moving image (right). (a) Local vessel and skeletal segmentation of landmark pairs were used to provide 3D information. This landmark pair was verified as a good match. Bifurcation points on the skeleton are shown as yellow voxels. (b) A landmark pair was removed due to a projection error. The bifurcation point is shown by crosshairs. Visually, it was determined that the projected location of the landmark (right image) does not correspond with the same anatomy as where it was detected in the first image (left), and therefore the pair was removed. (c) Landmark pair removed for not being on true bifurcation. The bifurcation point is shown by crosshairs. Visually, this landmark point is determined to likely occur on non-vessel anatomy with a poorly defined bifurcation. This artifact of vessel segmentation was not filtered out by the rule-based approach.

from DIRLAB 4DCT dataset. Each case had 300 vessel bifurcation landmark pairs identified by experts in thoracic imaging.[17] On the first image in each pair, vessels were segmented and landmarks were identified as described in steps 1 and 2 of Figure 1. To maximize the number of landmarks available for verification, the automatic landmark rejection step in our pathway (step 3 of Figure 1) was omitted. The detected landmarks within 2 voxels of a DIRLAB expert-defined landmark were taken to be landmarks of the same bifurcation. This 2-voxel distance was to account for slight placement variations on the bifurcation points between our automated pathway and DIRLAB experts. Once the matching landmarks were found, DIRLAB landmarks were projected onto the second image using pTVreg in the same manner as steps 4 and 5 in Figure 1.

The distance between the projected DIRLAB landmarks and the expert landmark placements on the second image was calculated. This measured the registration accuracy of the subset of DIRLAB

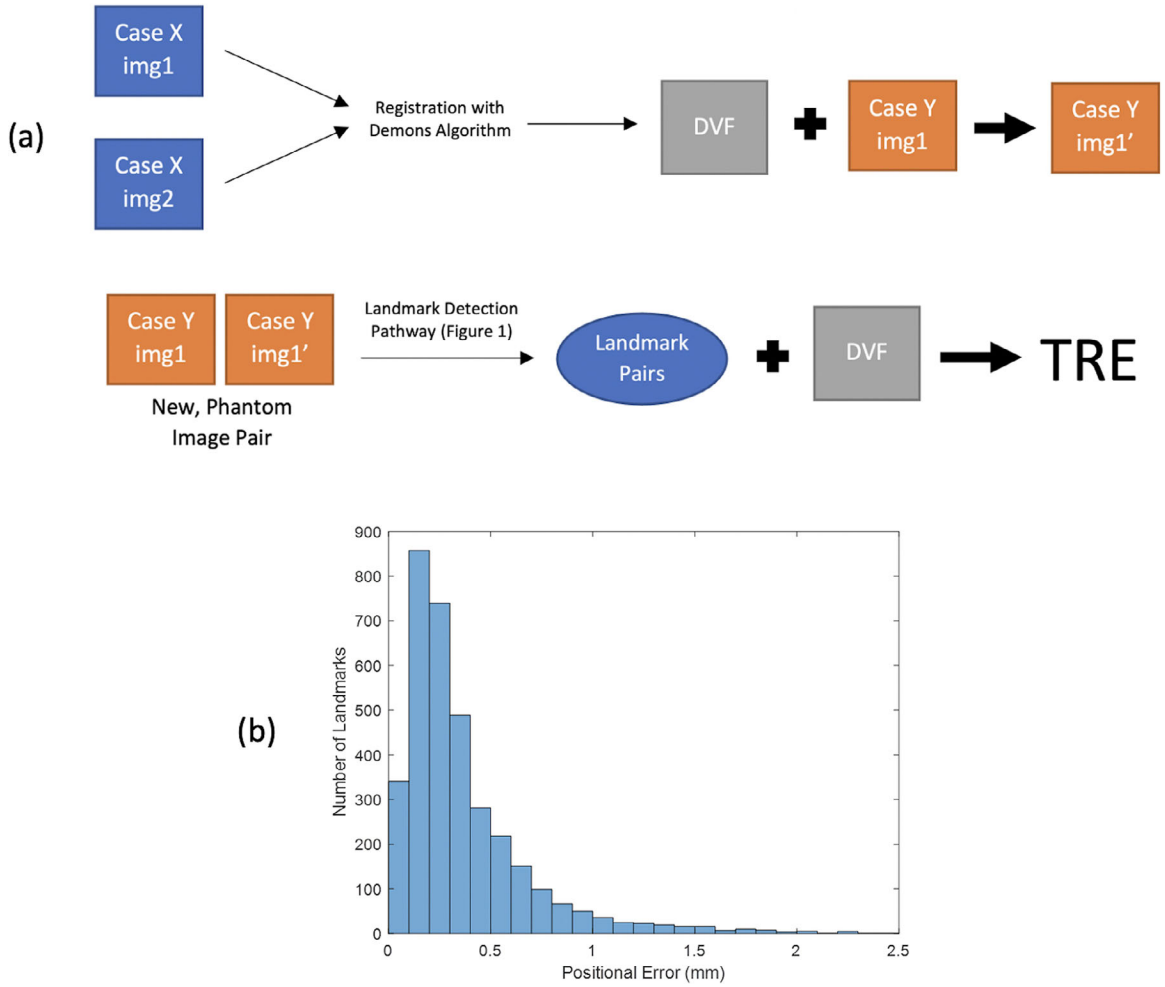

FIGURE 7 The TRE estimation pathway and the results of the digital phantom studies. (a) The process used to estimate TREs of landmark pairs. The ground truth deformation between the case Y image 1 and image 1′ was used to calculate the accuracy of the landmark pairs created using the pathway in Figure 1. Case X and Y were randomly selected from images in the dataset. (b) The histogram plot showing the estimated TRE errors in landmarks across all 5 phantom cases. Note that the 86 outliers greater than 3 mm were excluded by the manual outlier detection and rejection step.

landmarks that were also identified by our pathway. The average distance between 510 projected and expert-placed landmarks was 1.45 mm $\pm$ 1.43 mm. In voxels, this can be expressed as an average difference of $[x, y, z] = [0.60 \pm 0.21,\ 0.58 \pm 0.27,\ 0.36 \pm 0.13]$. Given the average isotropic voxel size of the images in our dataset, ~0.7 mm, this would suggest TRE errors in our landmarks on the order of 0.72 mm $\pm$ 0.77 mm. Additionally, the image quality was much better for our dataset than the DIRLAB dataset, which this could further reduce the TRE in our landmarks.

2.3.3 | Vessel order analysis

It is valuable to quantify the smallest vessels in our data on which bifurcation landmarks were identified, as this can inform future research segmenting even smaller vessels. In addition, the size of these vessels influenced the precision of the bifurcation placement. The smallest diameter of the segmented vessels was 2 voxels, or ~1.4 mm given the 0.7 mm isotropic resolution of the images.

The branching of pulmonary vessel trees can be quantified in generations by counting bifurcations from the lung hilus, or in orders by counting bifurcations from the lung periphery.[32] To classify the smallest vessels, we utilized a diameter-modified Strahler ordering system as described by Kassab et al.[33] and applied to human pulmonary vasculature by Huang et al.[34] Orders were chosen instead of generations because vessels of a given diameter can be distributed across a number of generations, making them hard to define.[32] We manually measured the sizes of our smallest segmented vessels. The bifurcations on these smallest vessels were within the 10th or 11th order arteries, or 9th or 10th order veins. Vessels of 4 voxels in diameter, given our largest voxel size, corresponded to 13th order arteries or 12th order veins. Since we excluded vessels greater than 4 voxels in diameter, our landmarks all fell on

TABLE 1 Processing parameters and final landmark numbers.

Patient	Number of landmark pairs	Denoising kernel size		Vessel segmentation method	Image landmarks detected on
		CT1	CT2		
1	891	10	10	Hysteresis	CT2
2	884	15	15	Hysteresis	CT2
3	1284	10	10	Hysteresis	CT2
4	1319	10	10	Hysteresis	CT2
5	1337	15	15	Hysteresis	CT2
6	1546	15	15	Hysteresis	CT2
7	1595	20	15	PTK	CT2
8	896	10	15	Hysteresis	CT2
9	1098	20	12	Hysteresis	CT2
10	1219	15	15	Hysteresis	CT2
11	1500	20	10	Hysteresis	CT2
12	907	20	10	Hysteresis	CT2
13	1210	15	15	Hysteresis	CT2
14	2135	10	10	Hysteresis	CT2
15	2090	10	10	Hysteresis	CT2
16	1292	20	10	Hysteresis	CT2
17	1658	20	10	Hysteresis	CT2
18	1204	5	5	Hysteresis	CT2
19	1244	10	15	Hysteresis	CT1
20	673	10	10	PTK	CT1
21	1125	10	10	Hysteresis	CT1
22	1381	10	10	PTK	CT1
23	878	10	10	PTK	CT1
24	1419	10	10	PTK	CT1
25	593	10	10	PTK	CT1
26	1197	15	15	PTK	CT2
27	1097	10	10	PTK	CT1
28	915	15	15	PTK	CT2
29	1889	10	10	PTK	CT1
30	1397	10	10	PTK	CT1

bifurcations of 10th–13th order arteries or 9th–12th order veins.

3 | DATA FORMAT AND USAGE NOTES

The final landmark data after the application of the pathway to the 30 selected CT image pairs is shown below in Table 1.

3.1 | Dataset overview

The data is stored in the Zenodo online data repository at https://doi.org/10.5281/zenodo.8200423. Instructions for getting started with the dataset can be found on our GitHub at https://github.com/deshanyang/Lung-DIR-QA. The data is saved in two forms. First, the image data was saved as NIfTI files, while the landmarks were saved as text files. The orientation and order of the image slices in the NIfTI file are in NIfTI format (sagittal, coronal, transverse), ordered from left to right, posterior to anterior, and inferior to superior, respectively. The voxel size of the images is stored in the header of the NIfTI file. Intensity values were stored as Hounsfield units shifted to start at 0, in little-endian byte order with 32 bits per voxel.

The image data, voxel sizes, and landmarks were also saved together in a second folder as .mat files for easy input to Matlab. Image data in the .mat files was organized in the order (coronal, sagittal, transverse), with the order of the slices in DICOM format: anterior to posterior, right to left, and inferior to superior, respectively. The landmark positions for the NIfTI and .mat files corresponded with the organization of the image slices in their respective format.

Voxel sizes and intensities of the images are saved in the format of the original scans, that is, no resampling or intensity thresholding was applied. This gives researchers flexibility in how they load and process the data. For information on scan parameters of images in the dataset, including in-plane resolution, slice thickness, kV, exposure, and more, please reference the Supplemental information. If the information available in this is not sufficient, please reference the original repository or contact us.

3.2 | Use

Code and usage instructions can be found on our GitHub, at https://github.com/deshanyang/Lung-DIR-QA. For loading the files into Matlab, we recommend utilizing the .mat files provided in the dataset. The NIfTI files coupled with the .txt landmarks should be generalizable to most other software or applications. To visualize the landmark pairs in the dataset, we recommend the use of MatchGui, a MATLAB-Based tool developed in-house. The program allows researchers to visualize the landmark pairs individually, as well as manipulate and flag them as they see fit. Instructions for installing and using the program can be found on our GitHub. MatchGui is not meant to be used for any advanced image analysis or interpretation. It is included so researchers can easily examine the data to determine if it is right for their research.

The images included in the dataset come from a variety of scanners with a range of image quality and voxel sizes, but the pathway in general was only effective in producing large numbers of landmarks in the higher-quality cases. Researchers may find the number of landmarks in some of the lower-quality images to be too limited for robust algorithm validation. To

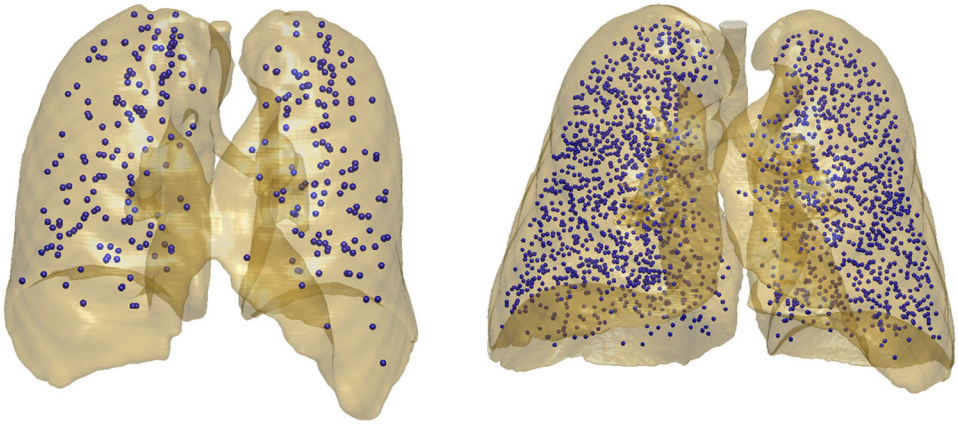

FIGURE 8 Current results improved spatial density of prior datasets. (Left) 300 DIRLAB landmarks. (Right) 2135 landmarks detected using the proposed method. Note: Landmarks detected on different images.

simulate more clinically challenging settings, we recommend implementing image augmentation and noise map addition to the higher quality cases with large numbers of landmark pairs. In this way, challenging DIR situations can be robustly tested.

4 | DISCUSSION

In this work, we have successfully developed a landmark pair dataset for real patient lung CT images. We have identified over 37 000 landmarks across 30 image pairs, with a high level of spatial accuracy and uniform density distribution, as shown in Figure 8. The average number of landmark pairs per case was 1260, a significant increase over other publicly available lung CT landmark pair dataset libraries. In addition, due to the automated process of landmark detection, the landmark precision is not subject to inter-observer variability. The estimates of TRE = 0.4 mm ± 0.3 mm, measured using digital phantoms, comfortably exceeded the current commonly accepted TRE goals for lung CT registrations. The distribution of landmarks is relatively uniform, which permits a high level of registration accuracy throughout the lungs.

We applied our pathway to 5 DIRLAB 4DCT dataset cases to compare our landmarks to those of expert observers. We found that landmarks identified both by DIRLAB expert observer and our pathway were projected with an accuracy of 1.45 mm ± 1.43 mm. However, for our dataset, given the typical voxel size of 0.7 mm, this distance would likely be more on the order of 0.72 mm ± 0.77 mm, which would be below the accuracy limits of current lung DIR algorithms. This value could be expected to be even lower, as the quality of our images is far greater than those in the DIRLAB, supporting more accurate registrations.

We also were able to estimate the order of our vessel bifurcation generations to be within 10th–13th order arteries or 9th–12th order veins. Future work involving CT scans with smaller voxel sizes could allow bifurcation identification on even smaller vessels, which would greatly increase the total number of available landmarks.

The pathway developed in this study can be generalizable to other anatomical sites where vasculature appears with reasonable contrast. Work is underway to generate a similar dataset within the vessels of the liver, abdomen, and head-neck. Implementation of a deep learning classifier may be useful in fully automating the pathway and removing the manual confirmation step, the most laborious step in the process, and lead to even more expansive datasets.

We hope this dataset provides a useful tool for DIR algorithm development and validation. The results of this work can be used to quantitatively verify any type of DIR algorithm, including novel methods that are not yet widely implemented. In doing so, we aim to spur development that will translate to clinical use. In particular, we see this dataset as being useful in evaluating highly precise registration algorithms, due to the low TRE of our pathway and the density of landmarks in the lower region of the lung, where distortion is the greatest.

This study will be part of a larger effort to develop landmark pairs dataset libraries across the body, which also will be made publicly available. In the future, we also hope to streamline the automated steps of the pathway developed in this study for application to patient-specific DIR-QA.

4.1 | Limitations

A major limitation of this pathway is that the number and accuracy of the landmark pairs are limited by pTVreg. This is in part corrected by the manual verification and adjustment. However, there are still landmark pairs in the more difficult clinical cases that had to be

excluded because the DVF from pTVreg did not properly project them onto the second image. In addition, the vessel segmentation step is only robust in high-quality image cases. Some of the lowest-quality images had to be thrown away because the vessel segmentation step failed to identify a sufficient number of landmarks. Verifying performance in these challenging cases, however, is important in supporting clinical care. Therefore, the inclusion of these types of scans in future versions of the dataset is important for robust algorithm validation. Alternatively, low-quality images can be synthesized by down-grading and down-sampling the high-quality and high-resolution images used in this study. Finally, the use of pTVreg in generating the landmark pairs may mean the dataset is biased to it or similar algorithms. Future work incorporating different or multiple algorithms into the workflow could help minimize this effect.

The manual verification is a key step in purporting the data as ground truth. Especially in cases of significant image noise or distortion, often in the periphery of the lung, the researcher had to use their best judgment in determining if a landmark pair is valid or not. As such, a small number of landmark pairs may not truly correspond with vessel anatomy. The number and the impact of these cases should be minimal, however, due to the several layers of processing applied to the initial vessel segmentation. In addition, the large number of landmark pairs in each case should minimize the effect of any outliers.

As previously noted, the pathway in Figure 1 does not separate arteries and veins during segmentation or subsequent steps. Implementing a method to do this in the future could intelligently inform landmark pair matchings while simultaneously increasing the overall spatial information and decreasing the number of false landmark pairs. It also could allow us to better define the order number of the vessels segmented, which could help quantitatively classify the landmarks.

Finally, the process used to estimate the TRE of the pathway operates on the assumption that the landmarks detected on the first image are the ground truth, which may not be completely accurate. Due to the high levels of precision of the landmark pairs, it is hard to say whether the bifurcation point of the skeleton is the ideal place to put the bifurcation of the vessel. There is likely some error associated with this step, in addition to any rounding that was applied to the landmark voxel positions during processing. This should not affect the final measurements of algorithm TRE, since the projected landmark positions were not rounded, but is worth noting for future applications of the pathway.

5 | CONCLUSIONS

In this work, we have developed a comprehensive landmark pair dataset in the lungs of a variety of clinical scans. This dataset is the most extensive of its kind to date and can be used for DIR algorithm validation and development. The number of, distribution, and accuracy of the landmark pairs will help support higher levels of precision than allowed by previous datasets. The dataset is publicly available and easy to access, which will enable researchers to verify their algorithms in real-time. In this, we hope it will support future implementation of novel DIR algorithms in the clinic.

ACKNOWLEDGMENTS

This research was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant R01-EB029431. The results here are in whole or part based upon data generated by the TCGA Research Network.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Jingu K, Kadoya N. Use of Deformable Image Registration for Radiotherapy Applications. *J Radiol Radiat Ther.* 2014;2(2):1042. doi:10.47739/2333-7095/1042
2. Zhang T, Chi Y, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. *Int J Radiat Oncol Biol Phys*. 2007;68(2):522-530. doi:10.1016/j.ijrobp.2007.01.038
3. Chao M, Xie Y, Xing L. Auto-propagation of contours for adaptive prostate radiation therapy. *Phys Med Biol*. 2008;53(17):4533-4542. doi:10.1088/0031-9155/53/17/005
4. Hof H, Rhein B, Haering P, Kopp-Schneider A, Debus J, Herfarth K. 4D-CT-based target volume definition in stereotactic radiotherapy of lung tumours: comparison with a conventional technique using individual margins. *Radiother Oncol*. 2009;93(3):419-423. doi:10.1016/j.radonc.2009.08.040
5. Boldea V, Sharp GC, Jiang SB, Sarrut D. 4D-CT lung motion estimation with deformable registration: quantification of motion nonlinearity and hysteresis. *Med Phys*. 2008;35(3):1008-1018. doi:10.1118/1.2839103
6. Lee C, Langen KM, Lu W, et al. Assessment of parotid gland dose changes during head and neck cancer radiotherapy using daily megavoltage computed tomography and deformable image registration. *Int J Radiat Oncol Biol Phys*. 2008;71(5):1563-1571. doi:10.1016/j.ijrobp.2008.04.013
7. Cleary K, Peters TM. Image-guided interventions: technology review and clinical applications. *Annu Rev Biomed Eng*. 2010;12:119-142. doi:10.1146/annurev-bioeng-070909-105249
8. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys*. 2017;44(7):e43-e76. doi:10.1002/mp.12256
9. Brock KK. Results of a multi-institution deformable registration accuracy study (MIDRAS). *Int J Radiat Oncol Biol Phys*. 2010;76(2):583-596. doi:10.1016/j.ijrobp.2009.06.031
10. Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans Med Imaging*. 2012;31(2):153-163. doi:10.1109/tmi.2011.2163944
11. Li S, Glide-Hurst C, Lu M, et al. Voxel-based statistical analysis of uncertainties associated with deformable image registration. *Phys Med Biol*. 2013;58(18):6481-6494. doi:10.1088/0031-9155/58/18/6481

12. Latifi K, Zhang G, Stawicki M, van Elmpt W, Dekker A, Forster K. Validation of three deformable image registration algorithms for the thorax. *J Appl Clin Med Phys*. 2013;14(1):3834. doi:10.1120/jacmp.v14i1.3834
13. Kadoya N, Fujita Y, Katsuta Y, et al. Evaluation of various deformable image registration algorithms for thoracic images. *J Radiat Res (Tokyo)*. 2013;55(1):175-182. doi:10.1093/jrr/rrt093
14. Vandemeulebroucke J, Sarrut D, Clarysse P. The POPI-Model, a point validated pixel-based breathing thorax model. *Proceeding of the XVth ICCR Conference, Toronto, Canada*. 2007.
15. Vishnevskiy V, Gass T, Szekely G, Tanner C, Goksel O. Isotropic total variation regularization of displacements in parametric image registration. *IEEE Trans Med Imaging*. 2017;36(2):385-395. doi:10.1109/TMI.2016.2610583
16. Vandemeulebroucke J, Rit S, Kybic J, Clarysse P, Sarrut D. Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs. *Med Phys*. 2011;38(1):166-178. doi:10.1118/1.3523619
17. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Phys Med Biol*. 2009;54(7):1849. doi:10.1088/0031-9155/54/7/001
18. Castillo R, Castillo E, Fuentes D, et al. A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive. *Phys Med Biol*. 2013;58(9):2861-2877. doi:10.1088/0031-9155/58/9/2861
19. Fu Y, Wu X, Thomas AM, Li HH, Yang D. Automatic large quantity landmark pairs detection in 4DCT lung images. *Med Phys*. 2019;46(10):4490-4501. doi:10.1002/mp.13726
20. Varadhan R, Magome T, Hui S. Characterization of deformation and physical force in uniform low contrast anatomy and its impact on accuracy of deformable image registration. *Med Phys*. 2016;43(1):52. doi:10.1118/1.4937935
21. Yang D, Zhang M, Chang X, et al. A method to detect landmark pairs accurately between intra-patient volumetric medical images. *Med Phys*. 2017;44(11):5859-5872. doi:10.1002/mp.12526
22. Werner R, Duscha C, Schmidt-Richberg A, Ehrhardt J, Handels H. Assessing Accuracy of Non-linear Registration in 4D Image Data using Automatically Detected Landmark Correspondences. *SPIE Proceedings*. 2013;8669:264-272. doi:10.1117/12.2002454
23. Cazoulat G, Anderson BM, McCulloch MM, Rigaud B, Koay EJ, Brock KK. Detection of vessel bifurcations in CT scans for automatic objective assessment of deformable image registration accuracy. *Med Phys*. 2021;48(10):5935-5946. doi:10.1002/mp.15163
24. Murphy K, van Ginneken B, Reinhardt JM, et al. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. *IEEE Trans Med Imaging*. 2011;30(11):1901-1920. doi:10.1109/tmi.2011.2158349
25. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
26. Albertina B, Watson M, Holback C, et al. Data from: the Cancer Genome Atlas Lung Adenocarcinoma Collection (TCGA-LUAD). *The Cancer Imaging Archive*. 2016;4. doi:10.7937/K9/TCIA.2016.JGNIHEP5
27. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525. doi:10.1038/nature11404
28. Zhang K, Zuo W, Zhang L. FFDNet: toward a fast and flexible solution for CNN-based image denoising. *IEEE Trans Image Process*. 2018;27:4608-4622. doi:10.1109/tip.2018.2839891 2018
29. Doel T, Matin TN, Gleeson FV, Gavaghan DJ, Grau V. Pulmonary lobe segmentation from CT images using fissureness, airways, vessels and multilevel B-splines. *9th IEEE International Symposium on Biomedical Imaging (ISBI) ,Barcelona, Spain*. 2012:1491-1494. doi:https://doi.org/10.1109/ISBI.2012.6235854
30. Frangi AF, Niessen WJ, Vincken KL, Viergever MA. *Multiscale Vessel Enhancement Filtering*. Springer; 1998:130-137.
31. Vercauteren T, Pennec X, Perchant A, Ayache N. Diffeomorphic demons: efficient non-parametric image registration. *Neuroimage*. 2009;45(Suppl 1):S61-S72. doi:10.1016/j.neuroimage.2008.10.040
32. Townsley MI. Structure and composition of pulmonary arteries, capillaries, and veins. *Compr Physiol*. 2012;2(1):675-709. doi:10.1002/cphy.c100081
33. Kassab GS, Lin DH, Fung Y. Morphometry of pig coronary venous system. *Am J Physiol Heart Circ Physiol*. 1994;267(6):H2100-H2113.
34. Huang W, Yen RT, McLaurine M, Bledsoe G. Morphometry of the human pulmonary vasculature. *J Appl Physiol (1985)*. 1996;81(5):2123-2133. doi:10.1152/jappl.1996.81.5.2123

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Criscuolo ER, Fu Y, Hao Y, Zhang Z, Yang D. A comprehensive lung CT landmark pair dataset for evaluating deformable image registration algorithms. *Med. Phys.*. 2024;1-12.
https://doi.org/10.1002/mp.17026