

Inter-observer Variability Analysis of Automatic Lung Delineation in Normal and Disease Patients

Luca Saba¹ · Joel C. M. Than² · Norliza M. Noor³ · Omar M. Rijal⁴ · Rosminah M. Kassim⁵ · Ashari Yunus⁶ · Chue R. Ng² · Jasjit S. Suri^{7,8,9}

Received: 3 February 2016 / Accepted: 18 April 2016 / Published online: 25 April 2016
© Springer Science+Business Media New York 2016

Abstract Human interaction has become almost mandatory for an automated medical system wishing to be accepted by clinical regulatory agencies such as Food and Drug Administration. Since this interaction causes variability in the gathered data, the inter-observer and intra-observer variability must be analyzed in order to validate the accuracy of the system. This study focuses on the variability from different observers that interact with an automated lung delineation system that relies on human interaction in the form of delineation of the lung borders. The database consists of High Resolution Computed Tomography (HRCT): 15 normal and

81 diseased patients' images taken retrospectively at five levels per patient. Three observers manually delineated the lungs borders independently and using software called ImgTracer™ (AtheroPoint™, Roseville, CA, USA) to delineate the lung boundaries in all five levels of 3-D lung volume. The three observers consisted of Observer-1: lesser experienced novice tracer who is a resident in radiology under the guidance of radiologist, whereas Observer-2 and Observer-3 are lung image scientists trained by lung radiologist and biomedical imaging scientist and experts. The inter-observer variability can be shown by comparing each observer's tracings to the automated delineation and also by comparing each manual tracing of the observers with one another. The normality of the tracings was tested using D'Agostino-Pearson test and all observers tracings showed a normal *P*-value higher than 0.05. The analysis of variance (ANOVA) test between three observers and automated showed a *P*-value higher than 0.89 and 0.81 for the right lung (RL) and left lung (LL), respectively. The performance of the automated system was evaluated using Dice Similarity Coefficient (DSC), Jaccard Index (JI) and Hausdorff (HD) Distance measures. Although, Observer-1 has lesser experience compared to Observer-2 and Observer-3, the Observer Deterioration Factor (ODF) shows that Observer-1 has less than 10 % difference compared to the other two, which is under acceptable range as per our analysis. To compare between observers, this study used regression plots, Bland-Altman plots, two tailed T-test, Mann-Whitney, Chi-Squared tests which showed the following *P*-values for RL and LL: (i) Observer-1 and Observer-3 were: 0.55, 0.48, 0.29 for RL and 0.55, 0.59, 0.29 for LL; (ii) Observer-1 and Observer-2 were: 0.57, 0.50, 0.29 for RL and 0.54, 0.59, 0.29 for LL; (iii) Observer-2 and Observer-3 were: 0.98, 0.99, 0.29 for RL and 0.99, 0.99, 0.29 for LL. Further, CC and R-squared coefficients were computed between observers which came out to be 0.9 for RL and LL. All three observers however

This article is part of the Topical Collection on *Education & Training*

✉ Jasjit S. Suri
jasjit.suri@atheropoint.com; jsuri@comcast.net

- ¹ Azienda Ospedaliero Universitaria (A.O.U.) di Cagliari – Polo di Monserrato, Università di Cagliari, s.s. 554 Monserrato, Cagliari 09045, Italy
- ² UTM Razak School of Engineering and Advanced Technology, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
- ³ Department of Engineering, UTM Razak School of Engineering and Advanced Technology, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
- ⁴ Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
- ⁵ Department of Diagnostic Imaging, Kuala Lumpur Hospital, Kuala Lumpur, Malaysia
- ⁶ Institute of Respiratory Medicine, Kuala Lumpur, Malaysia
- ⁷ Global Biomedical Technologies, Inc., Roseville, CA, USA
- ⁸ AtheroPoint™ LLC, Roseville, CA, USA
- ⁹ Department of Electrical Engineering (Affl.), Idaho State University, Pocatello, ID, USA

manage to show the feature that diseased lungs are smaller than normal lungs in terms of area.

Keywords Lung, CT, Automated delineation · Manual delineation · Inter-observer · Variability · Stability · Reliability · Statistical tests

Introduction

Lung diseases such as tuberculosis, lower respiratory infections and chronic obstructive pulmonary disease, and lung cancer are among the top ten factors of death worldwide [50]. There is an estimate of 9.5 million deaths related to lung disease which is one sixth of total deaths in the world annually [50]. In the United States, lung disease related deaths are on the rise, ranking third and continuing to increase [5]. An estimate of 400,000 Americans die from lung disease related cases annually [6].

Interstitial Lung Disease (ILD) is a broad category of diseases that share common physiologic and radiologic properties, and these could be caused by exposure to hazardous materials [31, 40, 42]. These disorders mostly have a common trait of progressive scarring or fibrosis of the lung tissue which decrease oxygen gas exchange at the lungs. In ILD, lung volumes can be reduced, decreasing the lungs diffusing capacity [34]. ILD is diagnosed based on the interpretation of the High Resolution Computed Tomography (HRCT) from CT images.

Computer Aided Diagnosis (CADx) systems serve as a complementary role to radiologists. CADx by providing a “second-opinion” to a radiologist [10]. Manual evaluations of large database are laborious, subjective to error and variability in observations, so, radiologists are therefore interested in CADx-based systems that can assist and provide as a second-opinion on the manual decision making process. CADx systems can help in classification of diseased areas and quantification of their masses [21, 46]. Radiologists serve to validate and test the system. Current medical regulatory requirements mandate human interaction to be present in a medical system for it to be clinically implemented [27].

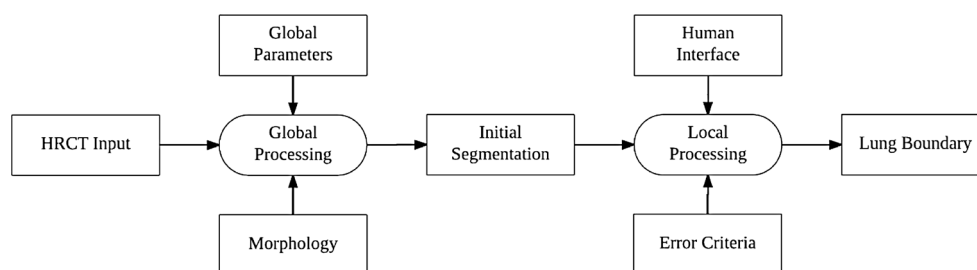
Automated delineation or segmentation is one of the preliminary and crucial steps in the development of a CADx

system to help radiologists [26]. Most automated lung delineation systems are evaluated through a comparison of their segmented regions with lung delineations or manual tracings done by a lung expert. In a previous study, an automated lung delineation system paired with a human interaction system is proposed as shown in Fig. 1 [30]. This system utilizes an initial automated lung delineation using threshold-based strategy combined with morphology and is termed as a global processing. A human interaction is introduced in the form of the manual tracing done by a trained individual termed as an observer. The human interaction is then compared with the initial segmentation output. This helps to catch large deviations and correct them by using a secondary automated lung delineation based on texture paradigm termed as local processing. The texture paradigm mentioned here involves the usage of the texture property to study the entropic behavior of the lung tissues. An entropy-based method is therefore adapted in this segmentation paradigm. Other segmentation techniques used widely include but are not limited to thresholding [17], active contours or snakes [32, 48], region growing [8] and texture based methods [9].

Since the human interaction mentioned thus far is in itself a source of error, there could exist two sources of variability [36]. The first variability is the intra-observer variability where the same observer may be inconsistent in the measurement of the lung [35]. The second variability is the inter-observer difference that the system shows when more than one observer is introduced. The observers’ manual tracing is considered the gold standard. As the system is not definitive, it is especially important to study the variability from both sources. In addition the process of tracing can be tedious and time consuming, which leaves room for variability from one observer to another [4]. For this study, there are three observers enlisted to perform manual tracings. Odd number of observers (such as three in this case) was used because they can easily reach a consensus decision while evaluating the CADx system. Ideally, the more observers available would be better, however for this study there are three observers available to do manual tracings.

The focus of this study is inter-observer variability. This is to explore the effects of different observers in the ALDS segmentation system on the segmentation performance yielded. Inter-observer variability is a growing factor to be studied

Fig. 1 Flow diagram of automated lung delineation system (ALDS)



signified by the numerous studies in different fields where the manual tracing relies on an input by an expert. Examples of such studies include studies on ventricular wall motion [43], carotid intima media thickness [38], pulmonary nodules [20, 51] lesions [1, 46], lung “honeycombing” [49] and tumor [12, 16, 37]. However there are also various other studies that showed inter-observer variability on lung borders produced from automated lung delineation [22, 28, 29, 39].

Thus, the goal of this study is to investigate the inter-observer variability analysis of an automated lung delineation system. The first observer (Obs-1) is a novice radiologist pursuing his medical residency with lesser experience, whereas the second (Obs-2) and third (Obs-3) are both trained biomedical image scientists and have more experienced. The inter-observer variability is studied using various methods to see both visually and numerically the difference of having multiple observers. The general statistics of the observers’ tracings and the analysis of variance (ANOVA) is demonstrated in this study. Next, the performance of the automatic segmentation system compared to the three observers was demonstrated using the following performance measures: Dice Similarity Coefficient (DSC), Jaccard Index and Hausdorff Distance. To determine the acceptability of an observer that has lesser experience, the Observer Deterioration Factor (ODF) was computed which was shown to be less than 10 %, which was an acceptable criteria for our analysis. This means that the observer being investigated is 90 % in agreement when compared against an experienced tracer. We show the derivation and rationale for these acceptance criteria in the performance evaluation section under the subsection observer deterioration factor (ODF) and interpretation of error for evaluation of the observers.

One approach to observe the difference between observers figuratively is by using the Bland-Altman (BA) Plot. BA plot used the Bland-Altman method which demonstrated the level of agreement between two methods measuring the same variable [23]. Statistically the difference between observers was shown using T-test, Mann–Whitney and Chi-Squared test [15, 19]. The correlation and regression test were also performed to show the agreement between observers. In Table 1, the abbreviations used in this study are displayed for the ease of understanding. These abbreviations listed here are the most commonly used throughout the text.

The limitation of this paper is that it only evaluates the inter-observer variability and not the intra-observer variability. This is due to limited resources and time in this study. However when more resources are readily available, the authors would like to investigate the intra observer variability as well. For future works, the authors would like to evaluate the observer variability in a three dimensional (3D) perspective since this study was limited to 2D analysis.

Table 1 Abbreviations used in this study and their definitions

Abbreviations	Definition
Obs	Observer
Obs-1	Observer 1
Obs-2	Observer 2
Obs-3	Observer 3
DSC	Dice Similarity Coefficient
JI	Jaccard Index
ANOVA	Analysis of Variance
HD	Hausdorff Distance
ALDS	Automatic Lung Delineation System
BA	Bland-Altman
ILD	Interstitial Lung Disease

Materials and method

Data acquisition and patient demographics

HRCT Thorax images were obtained from the Department of Diagnostic imaging of Kuala Lumpur Hospital with ethical clearance granted for 96 patients. Images were recorded using Siemens Somatom Plus4 CT scanner from May 2011 to June 2012. Each slice was attained at 10–30 mm intervals of patients in supine position during full suspended inspiration. All images were processed to be in the size of 512×512 pixels. For each patient a senior radiologist determined the five slices to represent the disease based on anatomical landmarks. The anatomical landmarks for the five levels were level 1: aortic arch, level 2: trachea carina, level 3: pulmonary hilar, level 4: pulmonary venous confluence and level 5: 1 to 2 cm above the dome of right hemi-diaphragm. The 96 patients consist of 15 healthy (normal cases) and 81 diseased cases. The diseased cases consisted of Interstitial Lung Disease patients (ILD cases) and non-ILD lung related diseases (non-ILD cases). There were 48 male patients and 48 female patients aging from 18 to 90 years old.

Automatic lung delineation system

The automatic lung delineation system (ALDS) was developed by members of Advanced Diagnostics And Progressive Human Care (Diagnostics) Research Group in UTM Razak School in collaboration with Global Biomedical Technologies, Inc., Roseville, CA, USA as shown in the flow-chart shown in Fig. 1 [30]. The system involves a global processing and a local processing. The global processing consists of two types of threshold based and morphology operations that include dilation and erosion. The first type of threshold method used is Otsu’s threshold method that finds the optimum threshold to separate the image into two classes for the variance between the classes to be minimum [33]. This

threshold is applied to remove the body pixels from the surrounding. Secondly, an empirical threshold is used to separate the lung pixels from the body pixels. The global parameters include the threshold value of -324 HU, the structure element used for morphology operation which is a “square” that is 3×3 pixels in size. This global processing yields the initial segmentation.

When the error criterion is exceeded, local processing is done on the initial segmentation. The error criterion is the error difference between the initial segmentation and the lung delineation done by the lung image experts. Large errors would indicate poor segmentation which shows that the lung analyzed is diseased. This adds the ability to detect possible erroneous segmentations and easily correct them. The local processing is a refined segmentation, based on the texture property, entropy of the image, and morphology operations of dilation and erosion. This refined segmentation will deal with the segmentation that was not able to be done through the global processing. Local parameters here are the structure element used, which is a 3×3 pixel ‘square’. Thus the system is robust, as it has not only the ability to spot large errors and correct them based on the local processing.

Manual delineation

In this study, manual tracings by three observers were used to study the effect of having different tracers. The first observer is a novice trainer who is a resident in radiology under the guidance of radiologist (L.S) with less experience. The second and third observers (C. R. N and J.C.M.T) are lung image scientists trained by lung radiologist (A.Y) and biomedical imaging scientist and experts (J.S.S and N.M.N). All tracings were done independently using the same software called

ImgTracer™, from AtheroPoint™, Roseville, CA, USA/ Global Biomedical Technologies, Inc., Roseville, CA, USA as shown in Fig. 2.

Inter-observer analysis

Statistical analysis was done to compare the lung area from the tracings of observers and the lung area segmented using the ALDS. The lung area is counted for each lung using segmentation borders and ground truth borders from the observers. The length of each pixel is obtained from each DICOM header. The area is counted as below (Eq. 1);

$$A = h \times l \quad (1)$$

Where h = height of the pixel (mm) and l = length in (mm). Area (A) is in mm^2 . The system used for the variability analysis in this research is shown in Fig. 3. The inputs of this system proposed are the output from the ALDS system mentioned before and the three manual delineations of the lungs from the three observers. These four inputs are analyzed with summary statistics and the performances of the segmentation based on all three observers are compared. The summary statistics are calculated for both right lung and left lung area. This includes the mean, variance, standard deviation (SD), relative standard deviation (RSD), standard error of the mean (SEM), minimum, maximum and median. After this is completed, the error plots were drawn. The error plots included Bland-Altman plots, Regression plots and overlays of the segmentation by the ALDS system together with manual delineations by the three observers. The next step of the system was to evaluate the inter-observer variability with T-test and Mann-Whitney test. These two tests were used to show the

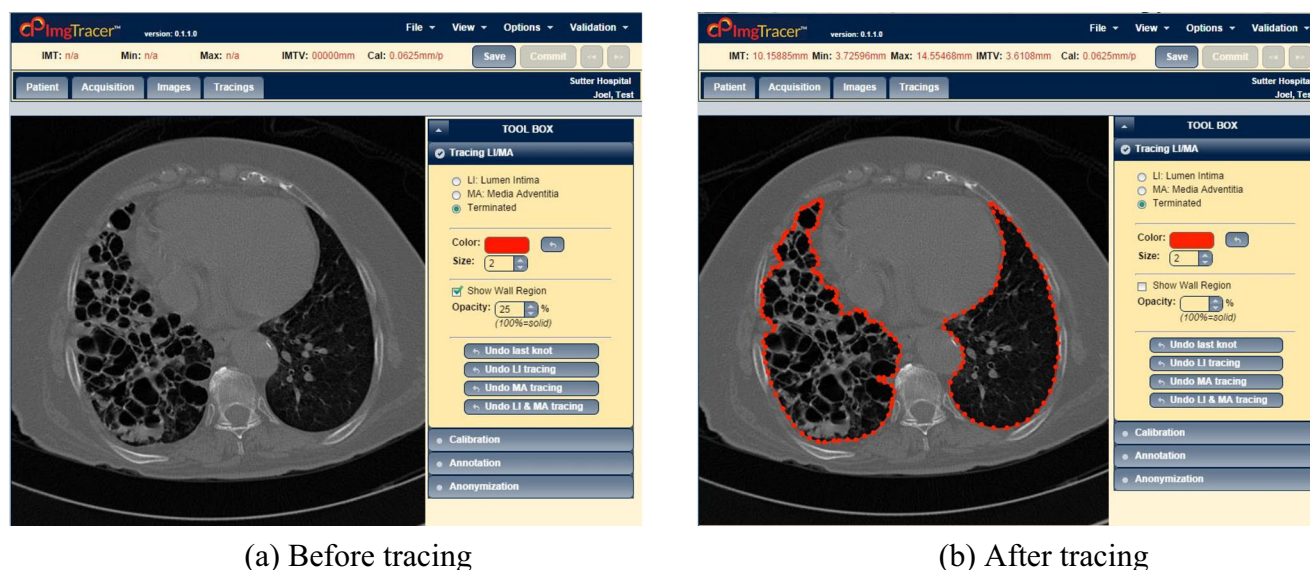
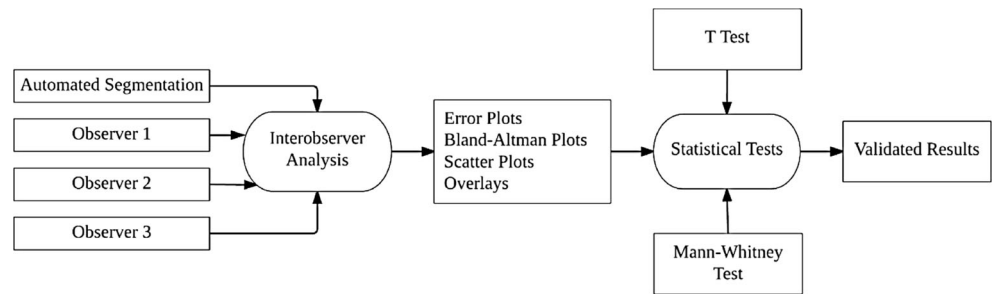


Fig. 2 ImgTracer™ system used for tracing the manual lung borders. (a) web-based ImgTracer™ has loaded the CT lung image, (b) boundary traced for lungs (Courtesy of AtheroPoint™, Roseville, CA, USA)

Fig. 3 Inter-Observer variability system overview



variability of the observers when compared to one another for both right and left lungs. Eventually, this leads to the validation of the tests on the variability of the observers.

Results

The results of the inter-observer analysis are shown in four subsections. The general analysis subsection shows summary statistics of the left lung and right lung; the ALDS segmentation performance compared with all three observers; comparison of the three ground truths where inter-variability is shown in the form of Bland-Altman plots where one ground truth represents one axis; and the classification capability of the ALDS segmentation system to identify normal and diseased lungs based on the delineated ground truth lung area by all observers.

General analysis

Area of L/R lung traced by observers

Tables 2 and 3 show the variability between all three observers and the summary statistics of the right lung and left lung for all three observers (Obs). The mean for Obs-2 and Obs-3 were very similar at 10427 and 10421 mm² for right lung and 7954.35 mm² compared to 7955.56 mm² for left lung respectively. Obs-1 had a higher mean than Obs-2 and Obs-3 with 10542 mm² as a mean area for the right lung and 8049.17 mm² for the left lung. The other parameters listed also differed in a similar manner.

Table 2 Summary statistics of the manual tracing done by three observers for the right lung

Obs	Area of right lung (1000 mm ²)								
	Mean	95 % CI	Var	SD	RSD	SEM	Med	Min	Max
Obs-1	10.542	10.261–10.823	9461.206	3.076	0.029	0.143	10.459	1.123	19.491
Obs-2	10.427	10.151–10.702	9113.126	3.019	0.029	0.140	10.334	1.117	19.773
Obs-3	10.421	10.147–10.696	9041.789	3.007	0.029	0.140	10.295	1.116	19.727

Obs indicates observer, *CI* confidence interval, *Var* variance, *SD* standard deviation, *RSD* relative standard deviation, *SEM* standard error of the mean, *Med* Median

Data normality test for three observer tracings

The normality of the tracings of the three observers was also tested using the D’Agostino-Pearson test. The test computes a *P*-value based on the coefficients of skewness and kurtosis. Obs-1 yielded a *P*-value of 0.08, Obs-2 yielded a *P*-value of 0.10, whereas Obs-3 yielded a *P*-value of 0.08 for both right and left lungs. The *P*-values of all three observers were above 0.05 that suggested normality for all three observers.

ANOVA test between observers and automated system (ALDS)

Further, Analysis of Variance (ANOVA) test was also performed on the segmented lung area corresponding to the three observers and the automated system is shown in the Tables 4 and 5, respectively. The ANOVA test results show the *P*-value (0.893) for the right lung (Table 4) and the *P*-value for the left lung 0.811 (Table 5) is above 0.05, respectively.

Performance evaluation of ALDS against three observers

Performance of ADLS using Jaccard index and dice similarity against three observers

Next the performance of the ALDS segmentation is shown when compared with three different observers. This is done by comparing the output of the ALDS segmentation with three different observer’s manual tracings. The similarity of the ALDS segmentation compared to the ground truth from the observers is measured to evaluate the ALDS system performance. To do this, the study used dice similarity coefficient (DSC) and Jaccard’s Index for both left lung and right lung.

Table 3 Summary statistics of the manual tracing done by three observers for the left lung

Obs	Area of left lung (1000 mm ²)								
	Mean	95 % CI	Var	SD	RSD	SEM	Med	Min	Max
Obs-1	8.04917	7.831–8.268	5704.412	2.388	0.0297	0.111	7.939	1.231	14.008
Obs-2	7.95435	7.740–8.169	5515.769	2.349	0.0295	0.109	7.867	1.350	14.020
Obs-3	7.95556	7.742–8.169	5454.474	2.335	0.0294	0.108	7.869	1.327	13.991

Obs indicates observer, *CI* confidence interval, *Var* variance, *SD* standard deviation, *RSD* relative standard deviation, *SEM* standard error of the mean, *Med* Median

The DSC and Jaccard’s Index were calculated and tabulated in Tables 6 and 7.

For the right lung, the DSC mean values were 97.25, 98.58 and 98.53 %, and the Jaccard’s Index mean values were 94.69, 97.24, 97.15 % for Obs-1, Obs-2 and Obs-3 respectively. For the left lung, the DSC mean values were 96.70, 98.21, 98.26 % and the Jaccard’s mean values Index were 92.75, 96.52 and 96.62 % for Obs-1, Obs-2, and Obs-3 respectively. The results show that for both the left lung and right lung, Obs-2 and Obs-3 gave a higher similarity. This suggests that there is Obs-2 and Obs-3 managed to give relatively comparable performance of segmentation due to the reason that they are more similar and have less variability between them compared to Obs-1.

Performance of ALDS using Hausdorff distance against three observers

We adapted Hausdorff Distances (HD) as a metric for computing the performance of ALDS. Here, we compute the HD between the automated lung borders and the manually traced lung borders using three observers. Since HD in principal extracts the maximum distance from one point of the automated lung border to another point of the manual traced lung border, any difference between these two borders are amplified. When comparing the two lung borders say *A* and *B*, HD is mathematically expressed as $H(A,B)$ and computed mathematically according to Eq. 2 [14, 18, 25]:

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{2}$$

Table 4 ANOVA calculation for the right lung area traced using three observers and the automated lung area

Source	DF	Adj SS	Adj MS	F-value	P-value
Obs	3	5.65E+06	1.88E+06	0.204	0.893
Error	1844	1.70E+10	9.22E+06		
Total	1847	1.70E+10			

Obs indicates observers, *DF* degrees of freedom, *Adj SS* adjusted sum of squares, *Adj MS* adjusted mean square

and $h(A,B)$ is expressed as (Eq. 3):

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{3}$$

where, $\|a - b\|$ is the underlying Euclidean distance between point *a* and point *b*. Point *a* is any point along the border *A*, and point *b* is any point along the *B* border. $h(A,B)$ in essence ranks each point of border *A* based on its nearest point on the border *B*, and uses the largest distance or highest rank as the distance. Note that $h(B,A)$ is computed the same way as $h(A, B)$. Using the above formulation, HD was computed for the left and right lungs, shown in Table 8.

It can be seen that Level 1 showed the lowest HD for all three observers for both right and the left lungs. This suggests that Level 1 has the highest segmentation quality. This concurs with the high DSC values represented in Level 1 (Tables 6 and 7). This was just opposite in Level 4, where HD was largest suggesting highest difference between the three observers and automated ALDS system. This again concurs with previous DSC values presented in Tables 6 and 7. Obs-2 and Obs-3 have least amount of difference in the HD values and are relatively consistent for all levels (L1 to L5) compared to Obs-1. Obs-1 showed the largest difference between Obs-1 and ALDS for all 5 levels for both left and right lungs. All this is very consistent with our assumptions because Obs-1 is least trained and is a medical resident compared to the experienced observers 2 and 3.

Observer deterioration factor (ODF) and interpretation of HD error for evaluation of the observers

On further inspection of the observers’ HD values, 90 % of tracings from Obs-2 and Obs-3 showed HD less than 20 mm,

Table 5 ANOVA calculation for left lung area traced using three observers and the automated lung area

Source	DF	Adj SS	Adj MS	F-value	P-value
Obs	3	5.40E+06	1.80E+06	0.319	0.811
Error	1844	1.04E+10	5.64E+06		
Total	1847	1.04E+10			

Obs indicates observers, *DF* degrees of freedom, *Adj SS* adjusted sum of squares, *Adj MS* adjusted mean square

Table 6 Performance evaluation of ALDS using right lung

Level	Right lung					
	Dice similarity coefficient			Jaccard's index		
	Obs-1 - ALDS	Obs-2 - ALDS	Obs-3 - ALDS	Obs-1 - ALDS	Obs-2 - ALDS	Obs-3 - ALDS
L1	97.26	98.55	98.43	94.79	97.20	97.02
L2	97.35	98.66	98.58	94.87	97.37	97.21
L3	96.95	98.43	98.39	94.11	97.02	96.85
L4	96.75	98.39	98.44	93.74	96.86	96.95
L5	97.94	98.87	98.83	95.98	97.78	97.70
Mean	97.25	98.58	98.53	94.69	97.24	97.15

whereas Obs-1 showed HD less than 30 mm. The deviation of 10 mm corresponded to the lack of experience of Obs-1 compared to Obs-2 and Obs-3. To understand and quantify the tracing performance of the Obs-1, we compute the degradation factor of Obs-1 with respect to the other two observers: Obs-2 and Obs-3. The ODF is mathematically expressed as the variability between the observer's HD errors (HDE) per unit lung maximum length. The maximum lung length (L_{max}) was computed as the maximum distance between any two pair of points along the lung boundary. The variability between the observers HDE is the difference of HDE between of observer which is being evaluated (say A) against reference observer (say B). Mathematically, it is expressed as (Eq. 4):

$$ODF(A, B) = \frac{|HDE(A) - HDE(B)|}{L_{max}} \times 100\% \quad (4)$$

where, HDE(A) is the HDE between ALDS borders and borders taken from Observer-A, while HDE(B) is correspondingly the HDE between ALDS borders and the borders from Observer-B. L_{max} is the mean of the maximum span of the lung space over all the images in the database. Using this concept, we can express ODF

between observer-1 and observer-2, ODF(1,2) and between observer-1 and observer-3, ODF(1,3) as follows (Eqs. 5 and 6):

$$ODF(1, 2) = \frac{|HDE(Obs-1) - HDE(Obs-2)|}{L_{max}} \times 100\% \quad (5)$$

$$ODF(1, 3) = \frac{|HDE(Obs-1) - HDE(Obs-3)|}{L_{max}} \times 100\% \quad (6)$$

where, L_{max} is the maximum span of the lung space over the entire database, ODF (1,2) represents the ODF of Obs-1 against Obs-2 and ODF(1,3) represents the ODF of Obs-1 against Obs-3. One can now compute the stability of the observer's ability by giving the threshold value on ODF as per the assumption adapted in medical industry for performance evaluation. Here, the stability of the system can be defined in regulatory spirit, where the medical device can be considered stable under average conditions. Such an assumption leads us to assume that a typical degradation of the performance should be less than 10 %, which implies an accuracy of 90 %. This means the ODF should be less than 10 % for a system to be accepted as stable, however has a potential for improvement under further training. Since ODF (1,2) and ODF (1,3) are both less than threshold deterioration of 10 %, this shows that Obs-1 is acceptable.

Table 7 Performance evaluation of ALDS using left lung

Level	Left lung					
	Dice similarity coefficient			Jaccard's index		
	Obs-1 - ALDS	Obs-2 - ALDS	Obs-3 - ALDS	Obs-1 - ALDS	Obs-2 - ALDS	Obs-3 - ALDS
L1	97.54	98.71	98.70	95.23	97.47	97.46
L2	97.02	98.48	98.49	94.25	97.01	97.04
L3	96.30	98.30	98.37	92.91	96.67	96.80
L4	96.10	97.64	97.79	92.53	95.44	95.73
L5	96.59	97.96	97.99	93.48	96.06	96.09
Mean	96.70	98.21	98.26	93.66	96.52	96.62

Table 8 Hausdorff distance performance of ALDS on the right and left lung

Level	Hausdorff distance (mm)					
	Right lung			Left lung		
	Obs-1 - ALDS	Obs-2 - ALDS	Obs-3 - ALDS	Obs-1 - ALDS	Obs-2 - ALDS	Obs-3 - ALDS
L1	7.33	3.73	4.26	8.62	4.80	3.59
L2	18.31	11.43	12.71	15.54	8.92	9.55
L3	19.70	10.42	11.48	19.45	9.65	9.12
L4	26.62	11.87	12.02	19.83	11.23	11.06
L5	15.79	9.13	10.00	14.13	8.98	9.79
Mean	17.55	9.32	10.09	15.51	8.72	8.62

Inter-observer variability of ground truth lung area

Coefficient of correlation and regression test between observers

Next, to demonstrate the similarity between the observers, we implemented the correlation and regression tests. This was to understand the relationships between observers: Obs-1 and Obs-2, Obs-1 and Obs-3, and Obs-2 and Obs-3. The correlation coefficient (CC) and R-square coefficient from the region test are presented in Table 9. Both CC and R-square coefficient for all relationships show an encouraging high number that suggests high degree of similarity across all the observers. One observation that can be noticed is that Obs-2 and Obs-3 relationship gave the highest CC value and R-squared value for both right and left lung suggesting Obs-2 and Obs-3 are the most similar to each other.

Bland-Altman plots

The graphical comparison of the lung area traced by the three observers is done using Bland Altman (BA) plots for right and left lung as shown in Fig. 4. From the BA plots, it can be seen that the difference between the manual tracing of Obs-2 and Obs-3 is in a higher

Table 9 CC and R-square coefficient between observers

Obs	Right lung		Left lung	
	CC	R-square	CC	R-square
Obs-1 vs. Obs-2	0.9939	0.9878	0.9928	0.9857
Obs-1 vs. Obs-3	0.9941	0.9882	0.9928	0.9856
Obs-2 vs. Obs-3	0.9996	0.9992	0.9992	0.9985

agreement compared to difference between Obs-1 and Obs-3, and difference between Obs-1 and Obs-3. This is due to the smaller two SD ranges of the difference between Obs-2 and Obs-3. This in turn shows that the variability between Obs-2 and Obs-3 is relatively lower than the variability between Obs-1 and Obs-2 and variability between Obs-1 and Obs-3. The low variability means that Obs-2 and Obs-3 managed to produce more similar manual tracings over the different slices of HRCT. A similar behavior was observed for the left lung.

T-test, Mann-Whiney and chi-squared test between three observers & its interpretation

Next, a two tailed T-test and Mann–Whitney test is also done to show the variability of the observers when compared to one another for both right and left lungs. Since there are three observers there are three categories of relationship as shown in Table 10. Both tests reveal that Obs-2 and Obs-3 have high similarity with 0.98 and 0.99 for the T-test and Mann–Whitney test respectively for the right lung in Table 10. For the left lung the values also reflect that of the right lung with 0.99 for both T-test and Mann–Whitney test in Table 10. These numbers from the T-test and Mann–Whitney test indicates the closeness between the observers. For the cases of Obs-1 and Obs-2, the value of T-test and Mann–Whitney are similar to each other at 0.57 and 0.50 respectively for the right lung and 0.54 and 0.59 respectively for left lung. The high *P*-values reject the null-hypothesis that the observers are completely different.

Besides the T-test and Mann–Whitney test, a Chi-Squared test was performed to evaluate the three observers and the results are displayed in Table 10. The results from Chi-Squared test did not show in detail the difference between observers as shown in T-test and Mann–Whitney Test, however the results showed that there was high level of similarity between observers.

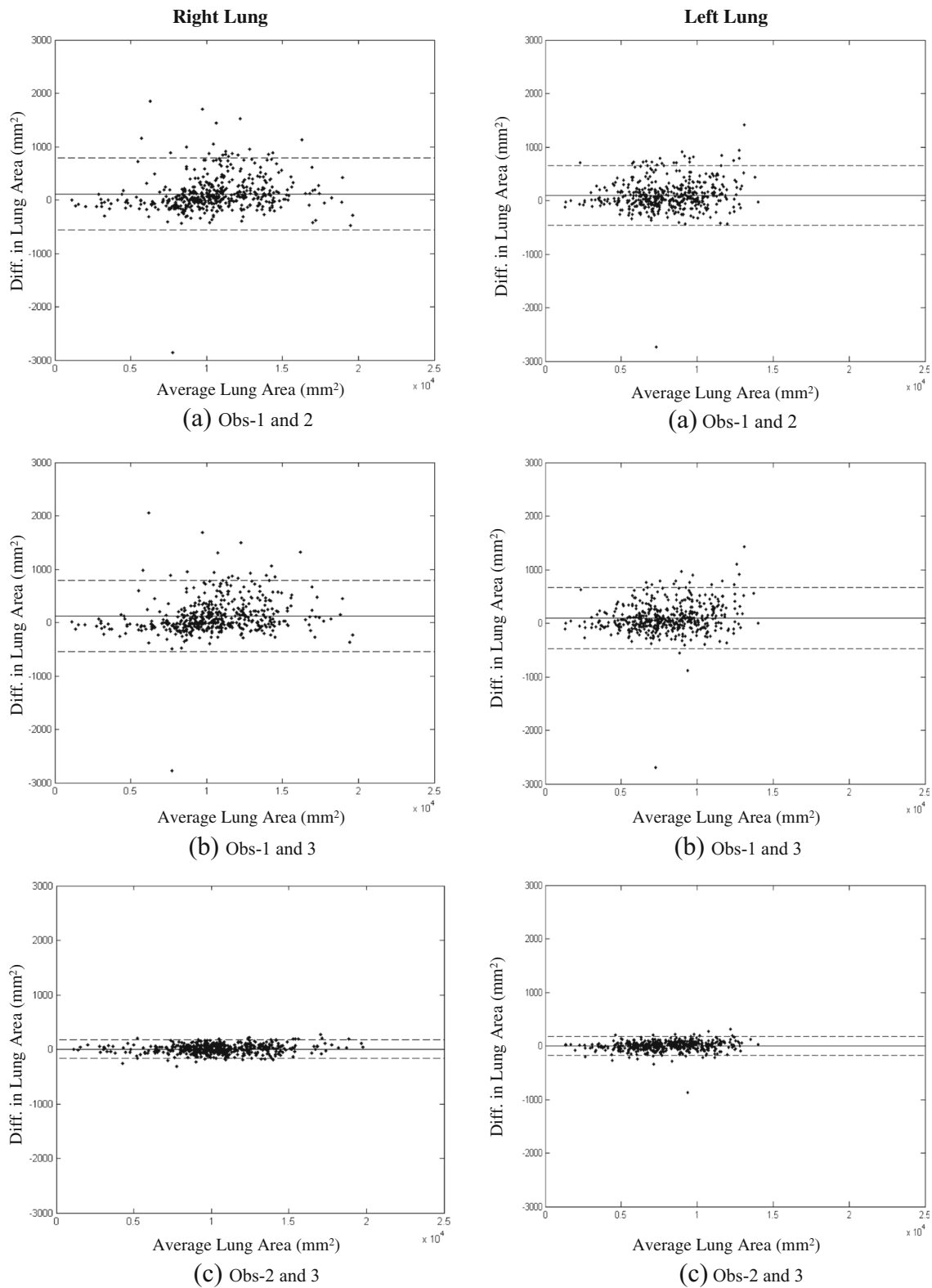


Fig. 4 BA plots of different observers for *right lung* and *left lung*

This can be seen from the *P*-values of 0.239 which is more than 0.05 across all relationships between Obs-1, Obs-2 and Obs-3. The *P*-value coincides with a

Contingency Coefficient of 0.99 from the Chi-Squared test which also suggest high degree of association between the observers.

Table 10 T-test and Mann–Whitney test of inter-observer variability on both lungs

Observer	Right lung			Left lung		
	T-test P-value	Mann–Whitney test P-value	Chi-squared test P-value	T-test P-value	Mann–Whitney Test P-value	Chi-squared test P-value
Obs-1 vs. Obs-2	0.57	0.5	0.239	0.54	0.59	0.239
Obs-1 vs. Obs-3	0.55	0.48	0.239	0.55	0.59	0.239
Obs-2 vs. Obs-3	0.98	0.99	0.239	0.99	0.99	0.239

Inter-observer variability between diseased and control lungs

Area quantification of diseased vs. controls using three observers

The observations of the ground truth from all the observers were then compared to the ALDS segmentation visually in the form of overlays for normal cases in Fig. 5 and for abnormal cases in Fig. 6. The green boundaries represent the segmentation and the red boundaries represent the ground truth from different observers. For normal cases the similarity between the green borders and red borders are almost in-distinguishable, indicating high similarity. This is because normal lungs are easier to segment due to clearer borders between lungs and body. For abnormal cases the comparison is done as in Fig. 6. For example for level 4 for left lung shown in Fig. 6, it can be seen that the green border is further away from the red boundary and the green region is smaller than the red region. The abnormal lungs can have complicated and more vague borders than the lungs and body region.

In Tables 11 and 12, the area of each level of segment is compared to that of the three observers for right lung and left lung. The lung area of the right lung is typically larger than

that of the left lung because of the position of the heart [7]. The diseased lung area theoretically should be smaller than that of normal lungs [34]. However in certain levels the diseased lung could actually be a similar size with the ground truth signifying accurate segmentation. The disease may not be prevalent in certain levels, which would explain the good segmentation in those levels. In Tables 11 and 12, it can be seen that for certain levels the area of the abnormal lungs is smaller than that of the normal lungs in Level 4 (L4) and Level 5 (L5) for Obs-2 and Obs-3 for the left lung. The variation between observers can be seen for the normal lung in Table 11 where Obs-1 has a higher mean area compared to Obs-2 and Obs-3.

Regression plots of diseased vs. controls using three observers

For all three observers, the regression plots indicate the abnormal right lung and left lung labeled ‘x’ and normal right lung and left lung area labeled ‘o’ in Fig. 7. The trend line represents the ideal case where the segmented lung area is identical to that of ground truth lung area indicating accurate segmentation. Deviation from the trend line would indicate poor segmentation. It is encouraging that most of the points are close to the trend line signifying accurate segmentation. However from Fig. 7 it is seen that most of the large deviations from the trend line are points

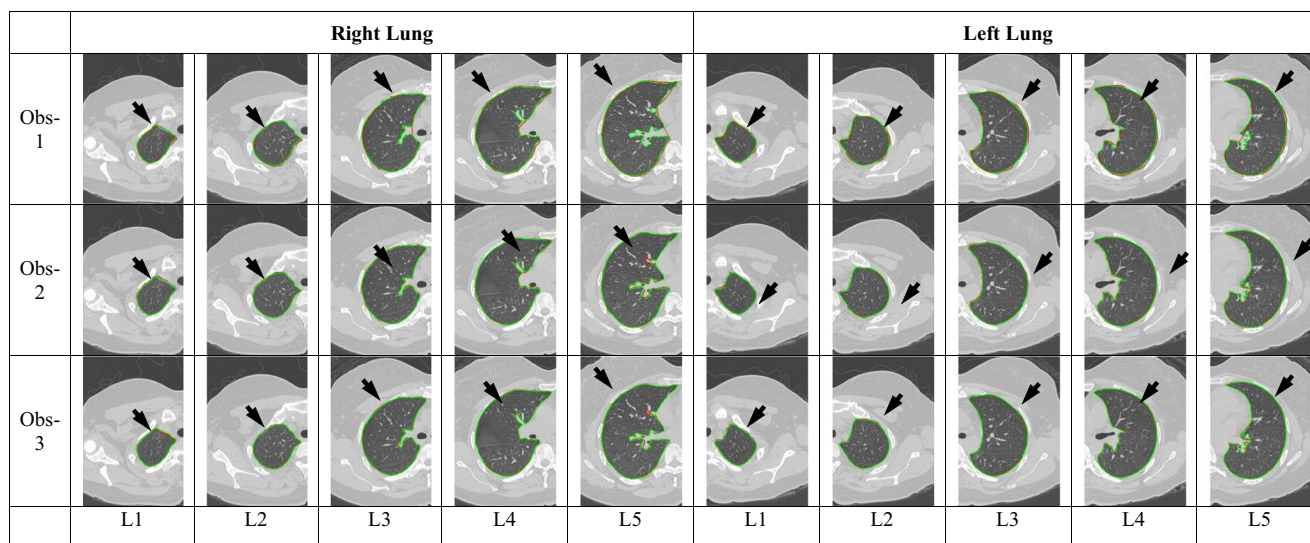


Fig. 5 Overlay of segmentation (green borders) and three observers (red borders) for normal right lung and left lung

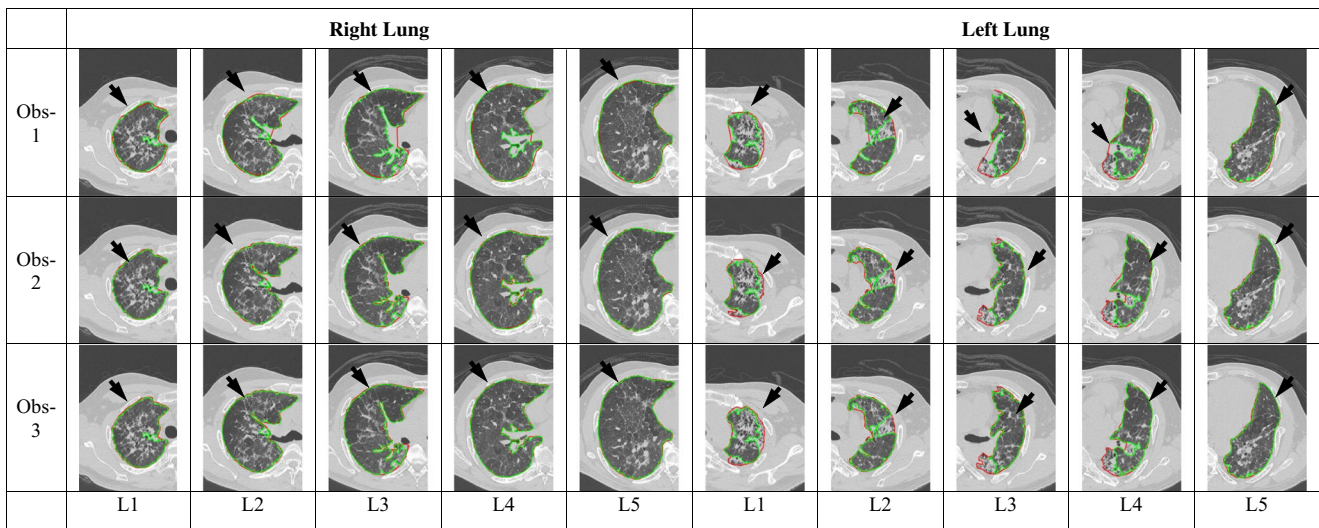


Fig. 6 Overlay of segmentation (green) and three observers (red) borders for abnormal right lung and left lung

labeled ‘x’ which are abnormal lungs. This feature is present in analysis of all three observers. The abnormal lungs can be detected based on the large lung area difference between segmentation and ground truth for all three observers. The same feature is seen for the left lung. It is encouraging that all three observers exhibit the feature of abnormal lungs deviate from the trend line.

Graphical representation highlighting the differences between normal and abnormal lungs using Bland-Altman (BA) plots for all three observers are shown in Figs. 8 and 9. It is noticeable that the two standard deviation (2SD) ranges shown as the two dotted lines are larger in abnormal lungs as compared to that of the normal lungs for both left lung and right lung. This supports the feature that abnormal lungs are detected based on low similarity to that of the ground truth for all three observers. This difference can be seen in both left lung and right lung.

Discussion

The purpose of this study was to investigate the inter-observer variability in the analysis of lung segmentation and

segmentation performance. From the results section, we see visually the high performance of the system for segmentation in the places where the manual borders of all three observers can be totally overlapped against the borders of the segmentation (green) (see Fig. 5) and high performance numbers summarized in Tables 6 and 7. The general statistics from Tables 2 and 3 shows that Obs-1 is slightly different than Obs-2 and Obs-3. The Bland-Altman plots in Fig. 4 show that Obs-1 has a higher difference when compared with Obs-2 and Obs-3. Obs-2 when compared with Obs-3 has a smaller difference shown by the smaller 2SD ranges.

In terms of outliers, Obs-1 vs. Obs-2 and Obs-1 vs. Obs-3 comparisons has more spread out outliers as compared to Obs-2 vs. Obs-3 comparison as seen for the right lung and left lung in Fig. 4. When the outliers are removed the 2SD ranges decreased for all comparisons suggesting that the observers are more in agreement. The decrease in 2SD ranges signify that the Observers show less variability when the outliers are removed. These outliers are actually samples of manual tracing that observers actually differ in. This difference can be due to of difference of opinion or an error from tracing. This is possible considering that manual tracing is not definitive and

Table 11 Normal and abnormal lung area from ALDS and ground truths from observers (Obs) for right lung

Level	Right lung							
	Normal lung average area (×1000 mm ²)				Abnormal lung average area (×1000 mm ²)			
	ALDS	Obs-1	Obs-2	Obs-3	ALDS	Obs-1	Obs-2	Obs-3
L1	7.152	8.0834	7.145	7.138	8.131	8.039	8.041	8.084
L2	8.610	9.706	8.557	8.536	9.766	9.762	9.773	9.706
L3	10.406	10.674	10.328	10.281	10.649	10.853	10.626	10.674
L4	11.566	11.641	11.526	11.466	11.522	11.972	11.649	11.641
L5	12.575	12.336	12.505	12.464	12.159	12.403	12.347	12.336
Mean	10.099	10.506	10.048	10.013	10.462	10.625	10.505	10.506

Table 12 Normal and abnormal lung area from ALDS and ground truths from observers (Obs) for left lung

Level	Left lung							
	Normal lung average area ($\times 1000 \text{ mm}^2$)				Abnormal lung average area ($\times 1000 \text{ mm}^2$)			
	ALDS	Obs-1	Obs-2	Obs-3	ALDS	Obs-1	Obs-2	Obs-3
L1	6.452	6.419	6.445	6.422	7.199	7.204	7.192	7.251
L2	7.887	7.862	7.832	7.819	8.174	8.329	8.272	8.231
L3	9.333	9.219	9.272	9.228	8.433	8.698	8.440	8.486
L4	9.655	9.766	9.599	9.545	7.768	8.054	7.894	7.869
L5	9.273	9.288	9.228	9.204	7.214	7.411	7.371	7.374
Mean	8.546	8.537	8.501	8.469	7.765	7.949	7.842	7.850

the process of tracing is a long and tedious process that may give rise to fatigue as well [4]. From the comparisons 2SD ranges in Fig. 4, it can be suggested that Obs-2 vs. Obs-3 have the highest agreement denoted by the smaller 2SD ranges. Thus Obs-2 and Obs-3 relatively similar and have less variability as compared to Obs-1.

The effect of different observers on the system is shown by the difference in performance of segmentation as in Tables 6 and 7. For the right lung the Dice Similarity Coefficient (DSC) values were 97.25, 98.58 and 98.53 % for Obs-1, Obs-2 and Obs-3 respectively. In terms of Jaccard Index for the right lung, the three observers yielded 94.69, 97.24 and 97.15 % for Obs-1, Obs-2 and Obs-3 respectively. For the left lung, the DSC values were 96.70, 98.21, and 98.26 % for Obs-1, Obs-2 and Obs-3 respectively. The Jaccard's Index values for the left lung were 92.75, 96.52 and 96.62 % for Obs-1, Obs-2 and Obs-3 respectively. These values showed that the ALDS segmentation accuracy was still high for three observers with small difference. This shows the ALDS system was able to segment the lung with an acceptable accuracy for all three observers being compared with their manual delineations. This is of significance especially in to validate that the high performance of the ALDS segmentation was able to be repeated by another observer that is Obs-2 and Obs-3 or in Obs-1's case was able to reach a comparable performance to all other observers.

The *P*-values in Table 10 suggest and support the notion that Obs-2 and Obs-3 are very similar with the high *P*-value of up to 0.98 and 0.99 for right and left lung respectively. This suggests that the level of tracing can be repeated by another person and is very encouraging. When comparing Obs-1 vs. Obs-2 and Obs-1 vs. Obs-3 both yield values that are satisfactory and suggest that there is variability between observers. The *P*-values for both T-test and Mann–Whitney test increased significantly when the outliers from the Bland–Altman plot are removed for Obs-1 vs. Obs-2 and Obs-1 vs. Obs-3 comparisons. The rise in *P*-values indicates the ability of the tracings to be more actually be more similar with the omission of outliers. Thus the observers have the ability to have less variability which shows that the level of manual

tracing can be repeated. The Chi-Squared tests show that the *P*-values for all relationships are higher than 0.05, thus the observers are not independent from one another. We also show the correlation coefficient (CC) and R-squared coefficient, showing the similarity between the observers. The relationship between Obs-2 and Obs-3 showed the highest CC of 0.999 which is slightly higher than 0.994 for both Obs-1 vs. Obs-2 and Obs-1 vs. Obs-3, relationships respectively, as seen in Table 9.

Besides showing segmentation performance, introducing different observers could also affect the ability of the ALDS segmentation system to determine the abnormal lungs. From the regression plots in Fig. 6, all three observers produced visually very similar plots. The feature of abnormal lungs causing the most deviation from the ground truth is evident in all three observers' delineations for both left and right lungs. Bland–Altman (BA) plots for each observer for both abnormal lung and normal lung were shown in Figs. 8 and 9. For both left lung and right lung, the abnormal plots were visually similar. For normal lungs, Obs-1 was noticeable different than Obs-2 and Obs-3 for both left lung and right lung. The two standard deviation (2SD) ranges represented by dotted lines had a bigger range in Obs-1. This is the variability introduced when a different observer is enlisted. Although there was a difference between Obs-1 compared to Obs-2 and Obs-3, the 2SD ranges of normal lungs are smaller than abnormal lungs for all observers. This can be said to support the feature of difference between normal and abnormal lungs.

There are few possibilities of variability for this case of comparison. The first is the number of points done by an observer. The more the points used could give a more accurate and detailed tracing and vice versa. On average the points traced by Obs-2 and Obs-3 is higher than that of Obs-1. Figure 10 shows an example of tracing. Obs-1 traced 49 points, Obs-2 traced 102 points and Obs-3 traced 92 points. Due to the lesser number of points by Obs-1 details such as the detail of tracing as indicated by the white arrows in Fig. 10. From here it can be said since the lesser number of points decreases the detail of the tracing, it cases the similarity

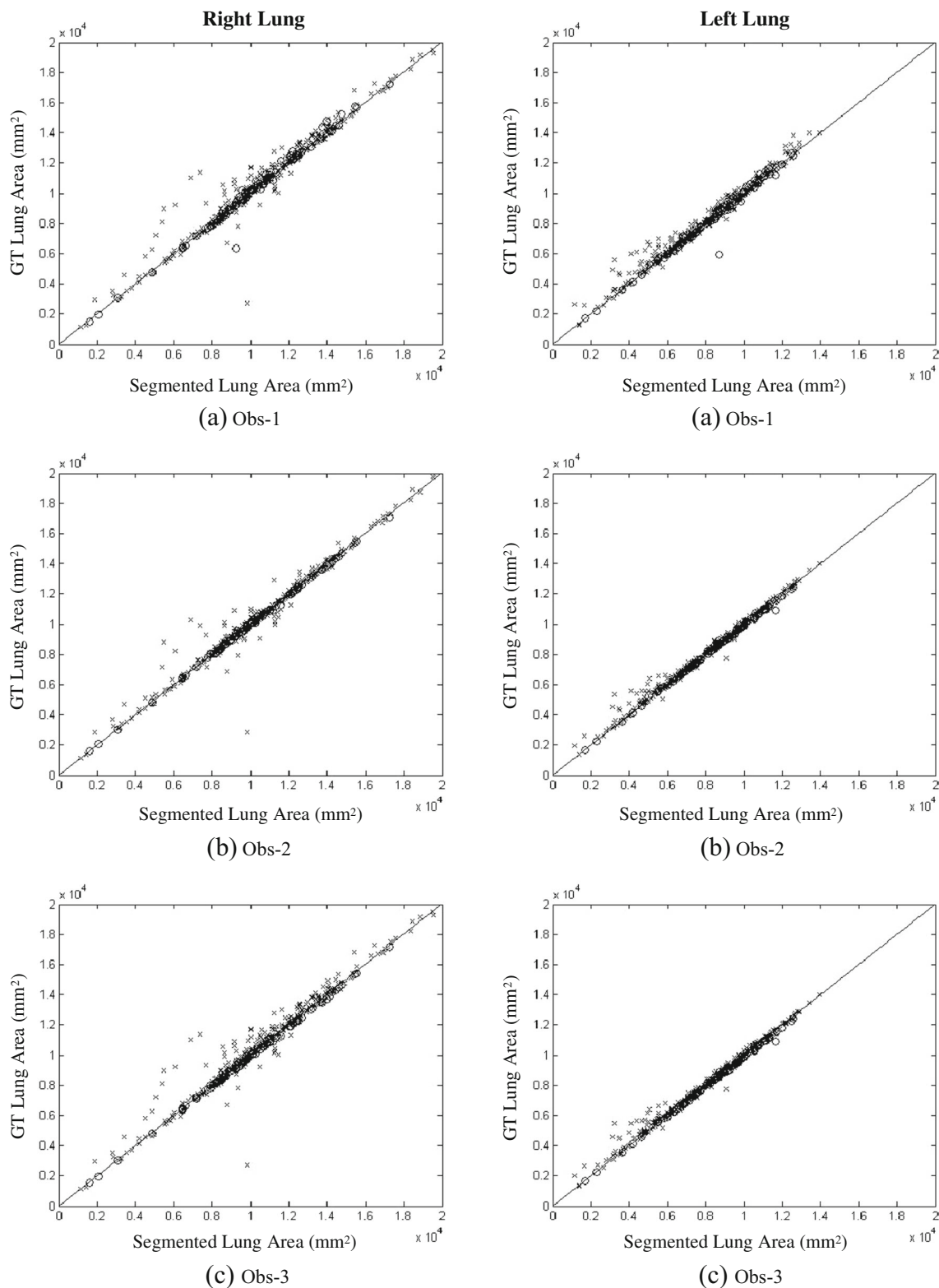


Fig. 7 Regression plot of normal labelled 'o' and abnormal lung 'x' for *right lung* and *left lung*

coefficient to vary as well. From a randomly selected 10 images, the average points plotted were 43, 89 and 70 for Obs-1, Obs-2 and Obs-3 respectively. Obs-1 has a significantly lesser points than Obs-2 and Obs-3 for these 10 images. Overall

Obs-1 also has the least Dice Similarity Coefficient (DSC) of all three observers at 97.25 % compared to that of 98.58 and 98.53 % for Obs-2 and Obs-3 respectively. Thus the lesser amount of points traced that result in a lesser detailed tracing

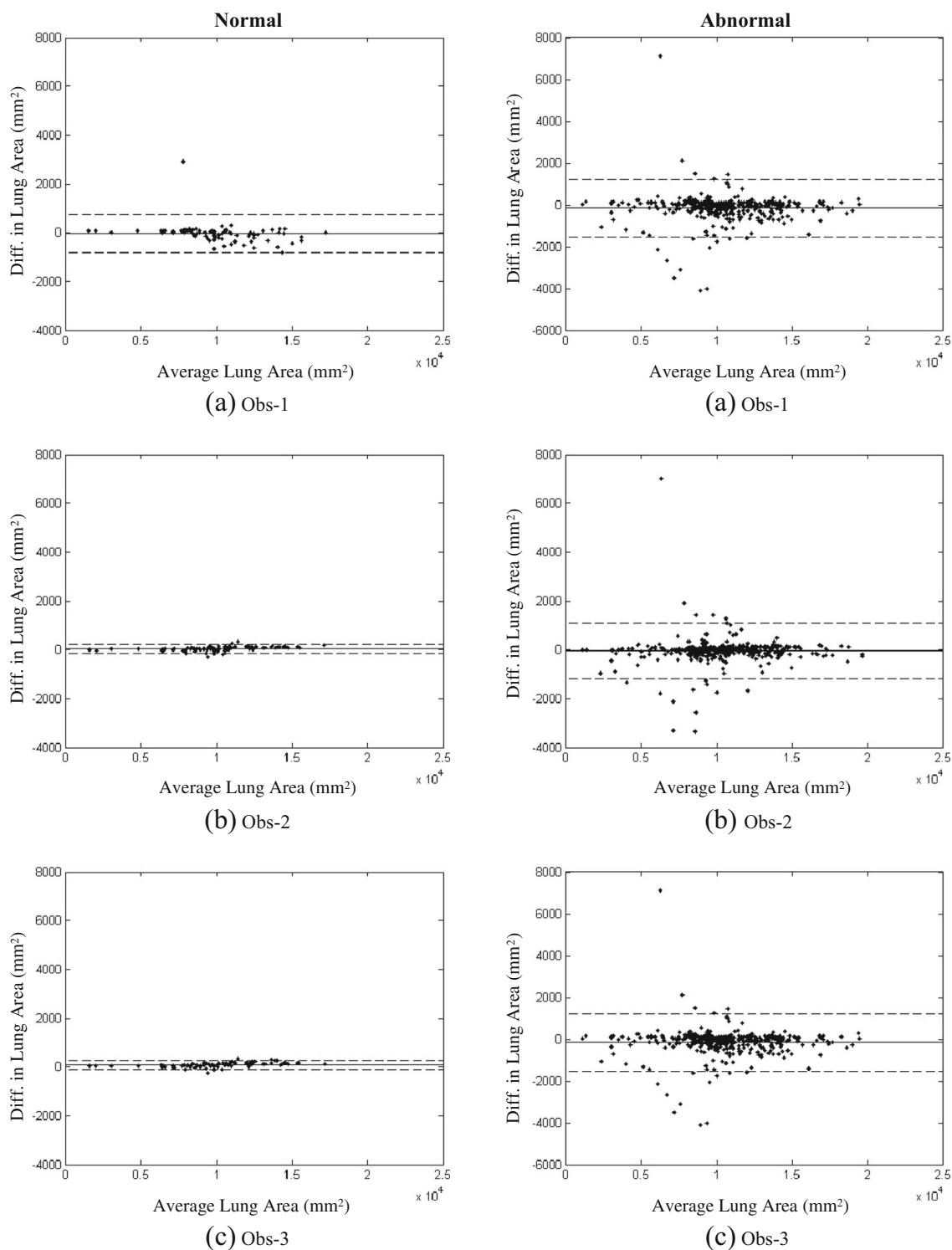


Fig. 8 BA plot of normal lung and abnormal lung by three observers and segmentation for *right lung*

could lower the system’s performance, however the results still show an encourage correlation between the system’s output and all three observer’s tracing. However since the Observer Deterioration Factor was less than 10 %, the tracings of the Obsever-1 is acceptable, however can be improved under rigorous training.

Similar studies done for inter-observer variability such as a study by Hu et al. proposed a segmentation algorithm tested on 8 eight normal patients that was also compared to two observers where the distance between the segmentation and ground truth was calculated based on pixels was presented in the form of mean, max and root mean square [17]. Comparatively, our

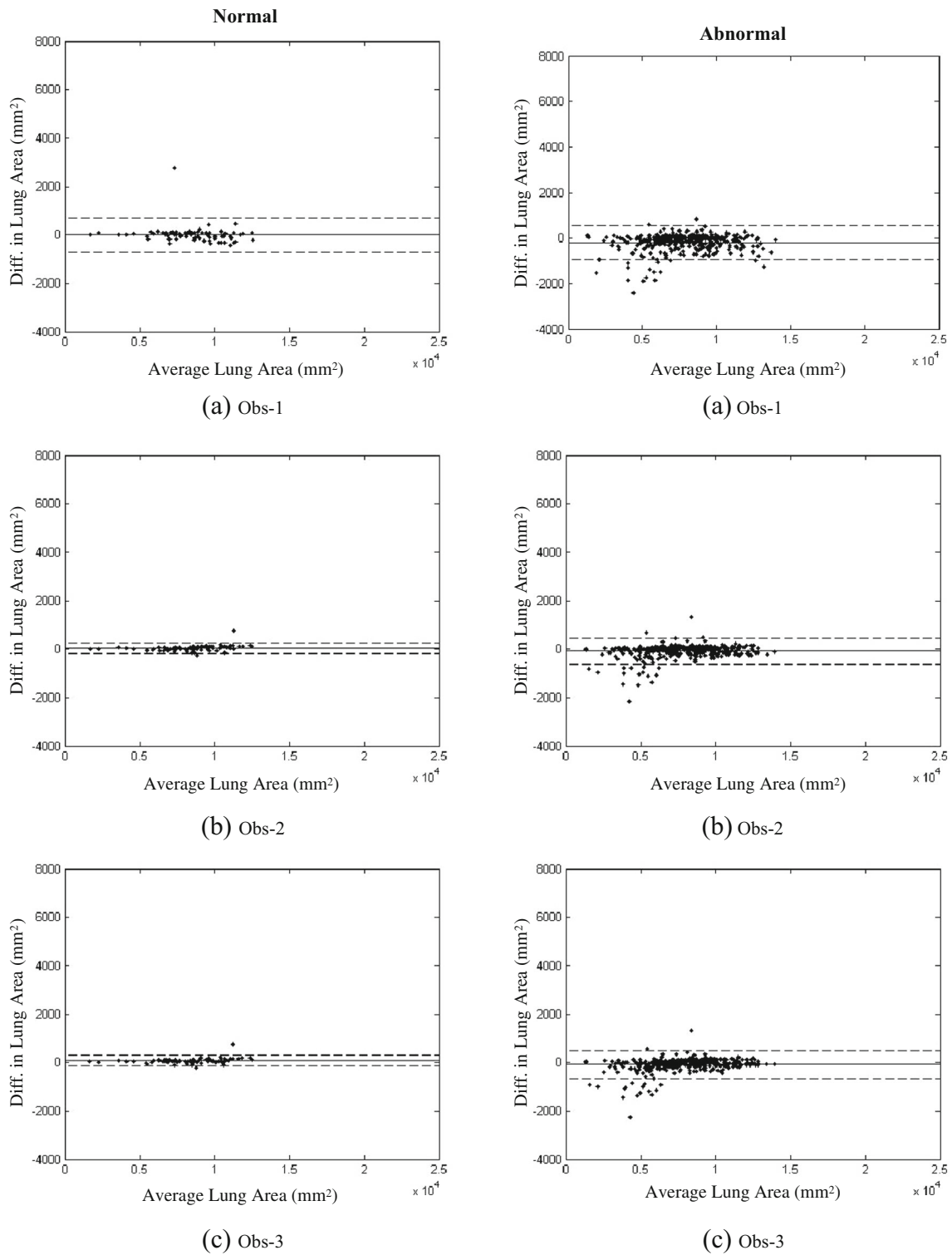


Fig. 9 BA plot of normal lung and abnormal lung by three observers and segmentation for *left lung*

study has three observers and 96 patients consisting of 15 normal and 81 abnormal patients. Our study showed visually as well as numerically using various methods.

Nery et al. did an inter-observer analysis compared with a lung segmentation based on watershed [28]. The study

compared the performance of segmentation based on two observers who were physicians. The difference between the segmentation and observers' delineation was calculated with Dice Similarity Coefficient (DSC) and Pratt's figure of merit. Observer analysis was done on two images. Comparison was

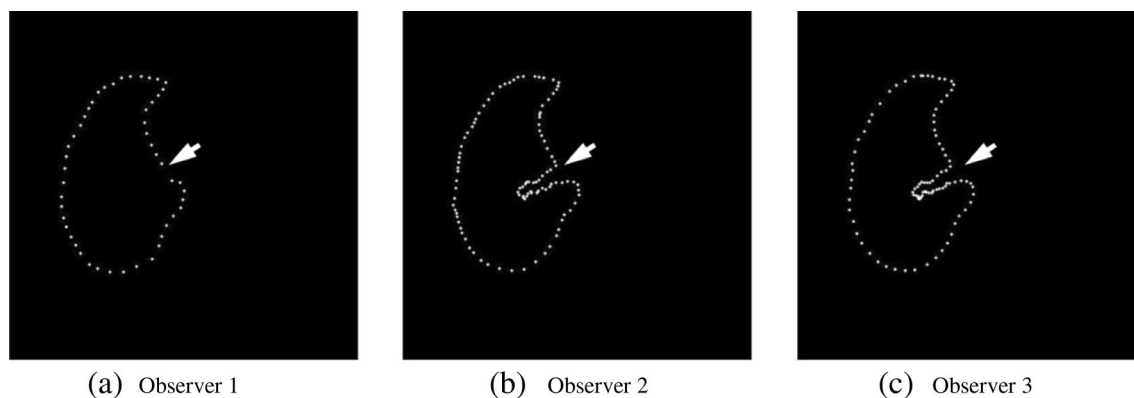


Fig. 10 Sample of observers' tracing *right lung*

done between observers and also observer to the segmentation algorithm for two patients. This comparison is considered to be limited because of the low number of patients. In our study we have three observers compared to each other and the segmentation across 96 patients' images. Another work by Nery et al. evaluated the inter-observer analysis based on Pratt's figure of merit and mean error compared to also a lung segmentation based on watershed [29]. The work also evaluated the Bland Altman plots of the difference of area and pixels of the lung delineations from the observers and segmentation method. However the work was also based on three observers alone on 41 CT slices which are relatively small in terms of data.

Santos et al. did an inter-observer analysis compared with a lung segmentation method based on gray-level thresholding [39]. The study did the analysis on 30 random selected images from eight patients. The observers were six radiologists. Although the number of observers was high, the number of randomly selected image to be evaluated is very low. Santos et al. compared the observers' tracings based on Pratt's figure of merit, mean distance and maximum distance. Kuhnigk et al. also evaluated the inter-observer variability compared with a lung segmentation method based on watershed method [22]. The study compared results from five observers on 24 patients. This again is relatively low in terms of number of patients compared to our study. Kuhnigk et al. showed an incomplete difference between observers by showing the volume and mean distance for the right lung only.

The strength of this paper is that it was able to demonstrate the variability and agreement between observers visually through various plots as well as numerically through various methods. Besides that the study also evaluates three observers for a relatively large database of 96 patients. The relative high number of patients utilized offers diversity of data which consists of patients with normal lungs, ILD lungs and Non-ILD lungs. This allows validation to be done for the segmentation accuracy from all three observers. This paper also investigates the characteristic of the segmentation system that requires a human interaction thus having a diverse data would give a more holistic presentation of segmentation with different observers.

While this study offers several diagnostic advantages, there are certain aspects of this study which can be considered as a limitation and extensions. They are: (i) the role of intra-observer variability analysis [38]. This requires ground truth tracings of the lungs borders by taking same observer at different times. We intend to consider this in our future research. Note, that since the two analyses (inter- and intra-) are independent to each other, the current inter-observer variability analysis holds valid in such a scenario. (ii) Since this work was done by taking five slices independently for each volume, one of the extensions of this work would be to use the entire lung volume in 3D by considering the spatial information of the neighborhood slices [11, 47]. Under the 3D model framework, one can again attempt the inter- and intra-observer variability analysis paradigm. (iii) Lastly, an extension would be to adapt Suri's strategy for stratification of lung cancer stages [2, 3] using the 2D/3D tissue characterization approaches [13, 24, 44, 45] using machine learning paradigm [41].

Conclusion

The study performed the inter-observer variability analysis of the manually traced lung borders by three observers which was also compared against the automated delineation system. The study presented the following statistical tests: (i) test for normality using D'Agostino-Pearson test; (ii) ANOVA test for studying the similarity between observers; (iii) significant difference tests using T, Mann-Whitney and Chi-Squared tests. We showed that all the three set of tests were successful, which includes normality and ANOVA. The T-test, Mann-Whitney test and Chi-Squared test showed that there is no significant difference for all three observers. The regression test showed high degree of correlation between all observers. The performance indices DSC, JI between observers and automated system for the right lung were 97.25 and 94.65 %, respectively for Obs-1, 98.58 and 97.24 % respectively for Obs-2 and, 98.53 and 97.15 % respectively for Obs-3. For the left lung the performance indices DSC and JI were 96.70

and 93.66 % respectively for Obs-1, 98.21 and 96.52 % respectively for Obs-2, and 98.26 and 96.62 % respectively for Obs-3. Mean HD for Obs-2 and Obs-3 are less than 10 mm while Obs-1 is less than 20 mm, which is consistent with the experience and assumptions of the three observers. Although, Observer-1 has lesser experience compared to Observer-2 and Observer-3, the Observer Deterioration Factor (ODF) shows that Observer-1 has less than 10 % difference compared to the other two, which is under acceptable range as per our analysis.

Acknowledgments This research was supported Research University Grant: GUP QK130000.2540.06H35, Universiti Teknologi Malaysia and Ministry of Higher Education Malaysia. We would like to thank Dr. Amir Zeki, Pulmonary and Critical Care Medicine Division, UC Davis School of Medicine, California, Sacramento, USA, as well as all the clinicians and radiologists who contributed and made this study a success. We are grateful to AtheroPoint™ LLC, Roseville, CA, USA for gracefully letting us use ImgTracer™ 1.0 software for tracing the manual borders of the lung.

References

- Abdullah, N., Mesurole, B., El-Khoury, M., and Kao, E., Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. *Radiology* 252:665–672, 2009.
- Acharya, U. R., Sree, S. V., Krishnan, M. M. R., Krishnananda, N., Ranjan, S., Umesh, P., and Suri, J. S., Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. *Comput. Methods Programs Biomed.* 112:624–632, 2013.
- Acharya, U. R., Sree, S. V., Krishnan, M. M. R., Molinari, F., Saba, L., Ho, S. Y. S., Ahuja, A. T., Ho, S. C., Nicolaides, A., and Suri, J. S., Atherosclerotic risk stratification strategy for carotid arteries using texture-based features. *Ultrasound Med. Biol.* 38:899–915, 2012.
- Alberola-López, C., Martín-Fernández, M., and Ruiz-Alzola, J., Comments on: a methodology for evaluation of boundary detection algorithms on medical images. *IEEE Trans. Med. Imaging* 23:658–660, 2004.
- Centers for Disease Control and Prevention, Deaths: final data for 2005. *Natl. Cent. Heal. Stat. Reports.* 56. 2008.
- Centers for Disease Control and Prevention, Deaths: final data for 2004. *Natl. Vital Stat. Reports.* 55. 2007.
- Churg, A., *Thurbeck's pathology of the lung*. Thieme New York, NY, 2005.
- Dehmeshki, J., Amin, H., and Valdivieso, M., X. Ye Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. *IEEE Trans. Med. Imaging* 27:467–480, 2008.
- Devan, L., Santosham, R., and Hariharan, R., Automated texture-based characterization of fibrosis and carcinoma using low-dose lung CT images. *Int. J. Imaging. Syst. Technol.* 24:39–44, 2014.
- Doi, K., Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput. Med. Imaging Graph.* 31:198–211, 2007.
- El-Baz, A., Suri, J. S., Lung imaging and computer aided diagnosis. CRC Press. 2011.
- Erasmus, J. J., Gladish, G. W., Broemeling, L., Sabloff, B. S., Truong, M. T., Herbst, R. S., and Munden, R. F., Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J. Clin. Oncol.* 21:2574–2582, 2003.
- Gupta, A., Kesavabhotla, K., Baradaran, H., Kamel, H., Pandya, A., Giambrone, A. E., Wright, D., Pain, K. J., Mtui, E. E., Suri, J. S., and Sanelli, P. C., Plaque echolucency and stroke risk in asymptomatic carotid stenosis a systematic review and meta-analysis. *Stroke* 46:91–97, 2015.
- Farag, A., Suri, J. S., Deformable models: theory and biomaterial applications. Springer Sci. Bus. Media. 2007.
- Hollander, M., Wolfe, D. A., Nonparametric statistical methods, John Wiley & Sons. 1999.
- Hopper, K. D., Kasales, C. J., Van Slyke, M. A., Schwartz, T. A., TenHave, T. R., Jozefiak, J. A., and Van Slyke, M. A., Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR Am. J. Roentgenol.* 167:851–854, 1996.
- Hu, S., Hoffman, E. A., and Reinhardt, J. M., Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Trans. Med. Imaging* 20:490–498, 2001.
- Huttenlocher, P., Klanderma, G. A., and Rucklidge, W. J., Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 15:850–863, 1993. doi:10.1109/34.232073.
- Jackson, S., Research methods and statistics: a critical thinking approach. Cengage Learn. 2011.
- Ko, J. P., Rusinek, H., Jacobs, E. L., Babb, J. S., Betke, M., Naidich, D. P., McGuinness, G., and Naidich, D. P., Small pulmonary nodules: volume measurement at chest CT—phantom study 1. *Radiology* 228:864–870, 2003.
- Korfatis, P., Kalogeropoulou, C., Karahaliou, A., Kazantzi, A., Skiadopoulos, S., and Costaridou, L., Texture classification-based segmentation of lung affected by interstitial pneumonia in high-resolution CT. *Med. Phys.* 35:5290–5302, 2008.
- Kuhnigk, J. J. M., Hahn, H., Hindennach, M., Dicken, V., Krass, S., and Peitgen, H. O., Lung lobe segmentation by anatomy-guided 3D watershed transform. *Med. Imaging* 2003:1482–1490, 2003.
- Martin Bland, J., and Altman, D., Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327:307–310, 1986.
- Molinari, F., Mantovani, A., Deandrea, M., Limone, P., Garberoglio, R., and Suri, J. S., Characterization of single thyroid nodules by contrast-enhanced 3-D ultrasound. *Ultrasound Med. Biol.* 36:1616–1625, 2010.
- Molinari, F., Meiburger, K. M., Zeng, G., Nicolaides, A., and Suri, J. S., CAUDLES-EF: carotid automated ultrasound double line extraction system using edge flow. *Ultrasound Imaging.* 129–162. 2012.
- Nagaraj, S., Rao, G. N., and Koteswararao, K., The role of pattern recognition in computer-aided diagnosis and computer-aided detection in medical imaging: a clinical validation. *Int. J. Comput. Appl.* 8:18–22, 2010.
- Nandy, K., Interactive segmentation and tracking in optical microscopic images. *Cytometry Part A* 81:357–359, 2012.
- Nery, F., Silva, J. S., Ferreira, N. C., Caramelo, F. J., and Faustino, R., Automated identification of the lung contours in positron emission tomography. *J. Instrum.* 8:C03018, 2013.
- Nery, F., Silvestre, J., Ferreira, N. C., Caramelo, F., Silva, J. S., and Faustino, R., An algorithm for the pulmonary border extraction in PET Images. *Procedia Technol.* 5:876–884, 2012.
- Noor, N. M., Than, J. C. M., Rijal, O. M., Kassim, R. M., Yunus, A., Zeki, A. A., Anzidei, M., Saba, L., and Suri, J. S., Automatic lung segmentation using control feedback system: morphology and texture paradigm. *J. Med. Syst.* 39:1–18, 2015.
- O'Dwyer, D. N., Armstrong, M. E., Cooke, G., Dodd, J. D., Veale, D. J., and Donnelly, S. C., Rheumatoid arthritis (RA) associated interstitial lung disease (ILD). *Eur. J. Intern. Med.* 24:597–603, 2013.

32. Osareh, A., and Shadgar, B., A segmentation method of lung cavities using region aided geometric snakes. *J. Med. Syst.* 34:419–433, 2010.
33. Otsu, N., A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man. Cybern.* 9:62–66, 1975.
34. Peroš-Golubičić, T., Sharma, O., Clinical atlas of interstitial lung disease. Springer. 2006.
35. Pope, A., Reproducibility: intraobserver and interobserver variability, *Biostat. Radiol.* Springer. 125–140. 2009.
36. van Rikxoort, E. M., and van Ginneken, B., Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review. *Phys. Med. Biol.* 58:R187, 2013.
37. van Rikxoort, E. M., de Hoop, B., van de Vorst, S., Prokop, M., and van Ginneken, B., Automatic segmentation of pulmonary segments from volumetric chest CT scans. *IEEE Trans. Med. Imaging* 28: 621–30, 2009.
38. Saba, L., Molinari, F., Meiburger, K. M., Acharya, U. R., Nicolaides, A., and Suri, J. S., Inter-and intra-observer variability analysis of completely automated cIMT measurement software (AtheroEdge™) and its benchmarking against commercial ultrasound scanner and expert Readers. *Comput. Biol. Med.* 43:1261–1272, 2013.
39. Santos, B. S., Ferreira, C., Sousa Santos, B., Silva, J. S., Silva, A., and Teixeira, L., Quantitative evaluation of a pulmonary contour segmentation algorithm in x-ray computed tomography images. *Acad. Radiol.* 11:868–878, 2004.
40. Schwarz, M. I., Matthay, R. A., Sahn, S. A., Stanford, R. M., Larry, B., and Scheinhorn, D. J., Interstitial lung disease in polymyositis and dermatomyositis: analysis of six cases and review of the literature. *Medicine* 55:89–104, 1976.
41. Sharma, A. M., Gupta, A., Kumar, P. K., Rajan, J., Saba, L., Nobutaka, I., Laird, J. R., Nicolades, A., and Suri, J. S., A review on carotid ultrasound atherosclerotic tissue characterization and stroke risk stratification in machine learning framework. *Curr. Atheroscler. Rep.* 17:1–13, 2015.
42. Sharman, P., and Wood-Baker, R., Interstitial lung disease due to fumes from heat-cutting polymer rope. *Occup. Med.* 63:451–453, 2013.
43. Sheehan, F. H., Stewart, D. K., Dodge, H. T., Mitten, S., Bolson, E. L., and Brown, B. G., Variability in the measurement of regional left ventricular wall motion from contrast angiograms. *Circulation* 68: 550–559, 1983.
44. Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., and Suri, J. S., Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Syst. Appl.* 42:6184–6195, 2015.
45. Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., and Suri, J. S., Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: a first comparative study of its kind. *Comput. Methods Programs Biomed.* 126:98–108, 2016.
46. Singh, S., Maxwell, J., Baker, J. A., Nicholas, J. L., and Lo, J. Y., Computer-aided classification of breast masses: performance and interobserver variability of expert radiologists versus residents 1. *Radiology* 258:73–80, 2011.
47. Trivedi, R., Saba, L., Suri, J. S., 3D imaging technologies in atherosclerosis. Springer. 2015.
48. Wang, A., and Yan, H., Delineating low-count defective-contour SPECT lung scans for PE diagnosis using adaptive dual exponential thresholding and active contours. *Int. J. Imaging Syst. Technol.* 20: 149–154, 2010.
49. Watadani, T., Sakai, F., Johkoh, T., and Noma, S., Interobserver variability in the CT assessment of honeycombing in the lungs. *Radiology* 266:936–944, 2013.
50. World Health Organization, World health statistics 2011. Geneva WHO. 2011.
51. Wormanns, D., Diederich, S., and Lentschig, M., Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. *Eur. Radiol.* 10:710–713, 2000.