Eur Radiol (2010) 20: 1878–1885
DOI 10.1007/s00330-010-1749-z

H. Ashraf
B. de Hoop
S. B. Shaker
A. Dirksen
K. S. Bach
H. Hansen
M. Prokop
J. H. Pedersen

Lung nodule volumetry: segmentation algorithms within the same software package cannot be used interchangeably

Received: 6 November 2009
Revised: 9 January 2010
Accepted: 21 January 2010
Published online: 20 March 2010

H. Ashraf (✉) · K. S. Bach · H. Hansen
Department of Radiology, Gentofte Hospital, Copenhagen University, Niels Andersens vej 65, 2900 Hellerup, Denmark
e-mail: Haseem@dadlnet.dk

B. de Hoop · M. Prokop
Department of Radiology, University Medical Centre Utrecht, Utrecht, The Netherlands

S. B. Shaker · A. Dirksen
Department of Respiratory Medicine, Gentofte Hospital, Copenhagen University, Hellerup, Denmark

M. Prokop
Department of Radiology, Radboud University Nijmegen, Nijmegen, The Netherlands

J. H. Pedersen
Department of Cardiothoracic Surgery RT, Rigshospitalet, Copenhagen University, Copenhagen, Denmark

Abstract *Objective:* We examined the reproducibility of lung nodule volumetry software that offers three different volumetry algorithms. *Methods:* In a lung cancer screening trial, 188 baseline nodules >5 mm were identified. Including follow-ups, these nodules formed a study-set of 545 nodules. Nodules were independently double read by two readers using commercially available volumetry software. The software offers readers three different analysing algorithms. We compared the inter-observer variability of nodule volumetry when the readers used the same and different algorithms. *Results:* Both readers were able to correctly segment and measure 72% of nodules. In 80% of these cases, the readers chose the same algorithm. When readers used the same algorithm, exactly the same volume was measured in 50% of readings and a difference of >25% was observed in 4%. When the readers used different algorithms, 83% of measurements showed a difference of >25%. *Conclusion:* Modern volumetric software failed to correctly segment a high number of screen detected nodules. While choosing a different algorithm can yield better segmentation of a lung nodule, reproducibility of volumetric measurements deteriorates substantially when different algorithms were used. It is crucial even in the same software package to choose identical parameters for follow-up.

Keywords Pulmonary nodules · Volumetry · Segmentation · Reproducibility · Computed tomography

Introduction

Since the introduction of computed tomography (CT), the technique has improved significantly, and many more nodules are detected with modern techniques. Thinner slices and faster rotation time allow a rapid and detailed evaluation of the lung [1]. Furthermore, low-dose CT techniques have reduced the radiation exposure and made use of repeat imaging more acceptable from an ethical point of view [2, 3].

Assessment of growth is a key issue in the diagnostic workup of lung nodules found on CT [4]. Rapid growth of lung nodules is associated with malignant lung disease and repeat imaging is essential [5]. Previously, assessment of lung nodules was performed manually, by measuring the nodule in three dimensions (x, y and z) [6]. Recently pulmonary nodule evaluation software has been launched that allows for semi-automated volumetric measurements and is increasingly being used for the diagnostic workup of lung nodules [7, 8].

In lung cancer screening trials with low-dose CT, nodule volumetry is increasingly used for follow-up of indeterminate nodules in order to detect growth and thus, identify suspected malignant lesions [9]. Nodule volumetry software

is available from various vendors but has been shown to vary with respect to absolute measured volume as well as reproducibility of volumetric measurements [10].

Correct segmentation of a pulmonary nodule is the prerequisite for accurate volumetry. In this study, we examined one particular volumetry software package [11] that approaches the issue of nodule segmentation by providing three distinct segmentation options, which include a generic segmentation (*All sizes*) and two segmentation options that are specifically aimed at small nodules (*Small size*) and non-solid nodules (*Subsolid*). We examined inter-observer variability in a lung cancer screening setting under the condition that observers would start with one specific algorithm and then chose to step up to the next algorithms should nodule segmentation with the first one fail. We compared inter-observer variability if both observers chose the same algorithm and if they chose different algorithms. For each approach the percentage of nodules in which differences in measured volumes exceeded 25% was recorded.

Materials and methods

Patients

The study population was selected from the Danish Lung Cancer Screening Trial (DLCST). The DLCST is a 5-year trial investigating the effect of annual screening with low-dose CT on lung cancer mortality. Participants were current or former smokers aged between 50 and 70 years at inclusion with a smoking history of more than 20 pack years [12].

The CT images were screened by two radiologists (K.S.B. and H.H.) and all non-calcified nodules with a diameter over 5 mm (manual measurement) were included in this study. All screen-detected nodules were tabulated along with information regarding the lung segment in which the nodule was found. In the event of disagreement between the radiologists consensus was obtained and registered.

Depending on the radiological degree of suspicion, the nodules were either surgically resected or underwent repeat imaging after 3 months to evaluate growth. Included in this study were nodules >5 mm detected at baseline screening starting November 2004, and their follow-up images up to April 2008.

Methods

All imaging was performed on multidetector (MD) CT (16-row, MX 8000 IDT, Philips Medical Systems, Cleveland, Ohio, USA). Imaging was performed supine at full inspiration in the caudo-cranial direction including the entire lungs. A low-dose technique with 140 kV and 40 mAs was used. Imaging was performed with spiral data acquisition with the following acquisition parameters: section collimation 16×0.75 mm, pitch 1.5 and rotation time 0.5 s. Images were reconstructed with 3-mm slice thickness at 1.5-mm increments using a soft algorithm (Kernel A) [12].

The reproducibility readings of the present study were done by two trained observers (1st reader, H.A., and 2nd reader, B.d.H.) with more than 2 years' experience in evaluating lung screening imaging with semi-automated nodule volumetry software [11] (Syngo LungCARE CT, Siemens Medical Solutions, Erlangen, Germany). The observers were participating in different screening trials, the Danish DLSCT (H.A.) and the Dutch-Belgian NELSON (B.d.H.) trials. To ensure that nodules were correctly matched, a CT slice on which the nodule was clearly marked was available for both readers. Otherwise each reader was blinded to the readings of the other reader.

The analysis procedure for solid nodules consisted of a step-up approach in which first the *Small size* algorithm was tried, and in the event of failure of proper segmentation the *All sizes* algorithm was tried. This evaluation was performed independently by the two readers. In particular, the following steps were taken: after positioning a seed point in the nodule, the software produced a visual three-dimensional (3D) presentation of the detected nodule highlighting the voxels of the nodule for which the volume was calculated (Fig. 1). If the segmentation was visually judged to include the whole nodule and no surrounding structures such as vessels and pleura, the segmentation was considered successful. If this visual validation of the nodule showed incorrect segmentation, the reader tried to segment the nodule three times with the same algorithm before concluding that the nodule could not be correctly segmented by this algorithm. In the case of part-solid nodules, only the solid part was analysed either with the *Small size* or the *All sizes* algorithm. The *Subsolid* algorithm was applied in the case of pure non-solid ground glass opacity.

Bland-Altman plots were used to compare volumetric results for those nodules in which the readers had used the same algorithm and for those nodules in which the readers had used different algorithms. Results were analysed using R statistical software version 2.7.1, and a significance level of 0.05 was applied. The differences between readers were normally distributed. An *F*-test was used to compare the variances achieved with the various algorithms.

Results

At baseline screening, 188 nodules were found in 161 participants. Including repeat imaging for follow-up of these nodules, 545 nodules on 488 CTs could be included in this study. In 154 of the 545 nodules (28%), one (10%) or both (18%) readers were unable to correctly segment the nodule using all available segmentation algorithms. In the

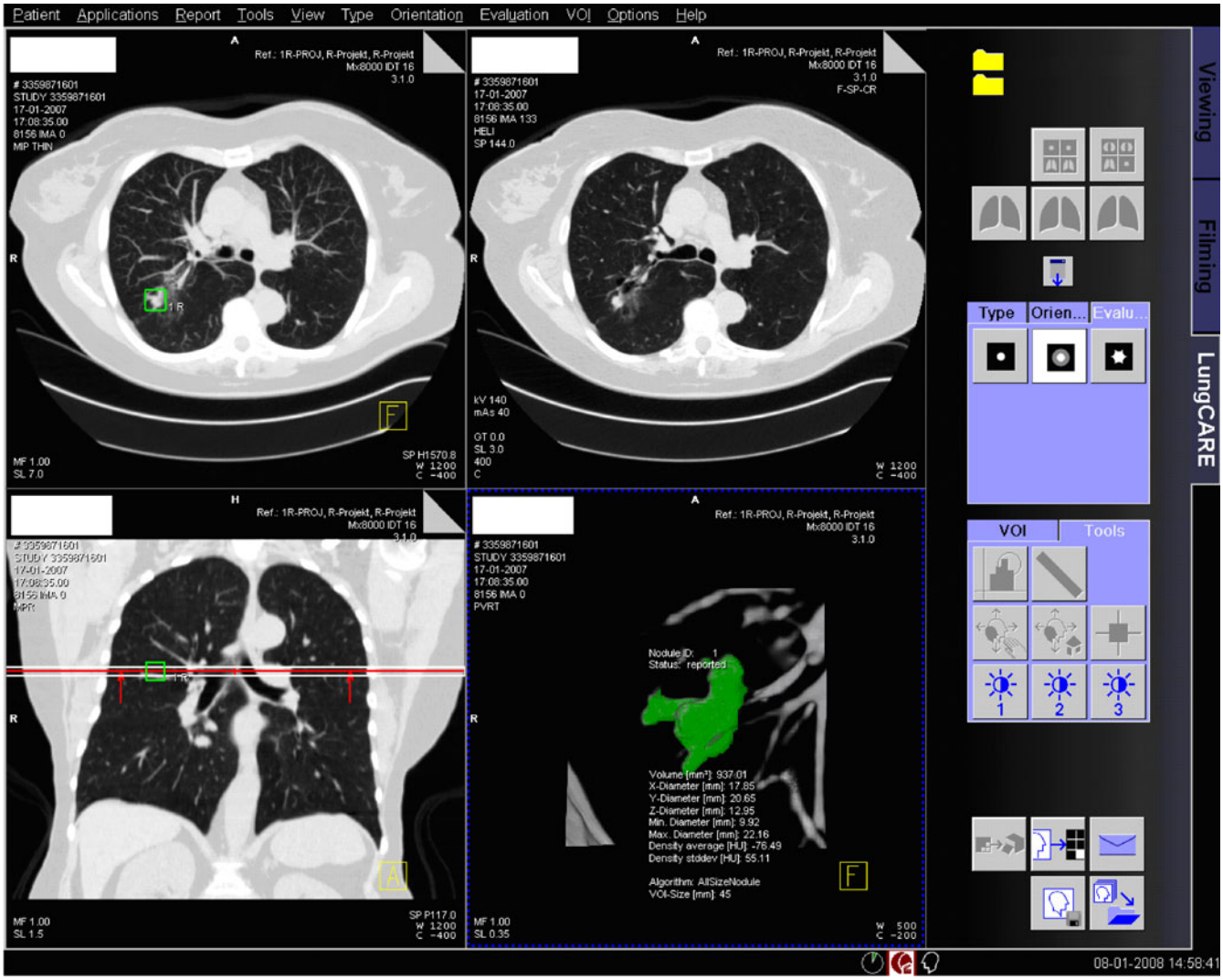

Fig. 1 Screenshot of Siemens LungCARE software. The *right lower window* displays a visual 3D presentation of the nodule

remaining 391 cases in which both readers found at least one algorithm that correctly segmented the nodule, they chose the same algorithm in 311 cases (80%) (Table 1).

When the two readers chose the same algorithm, they found exactly the same volume in 50% of cases. In 4% of cases, the difference in volume was larger than 25%. The percentage variation in volume measurements (percent of minimal reading) between readers was significantly smaller for the *Subsolid* algorithm compared with the *All sizes* algorithm (*F*-test, $p<0.001$), which again had less variation than the *Small size* algorithm (*F*-test, $p<0.001$) (Fig. 2). However, when measuring the variation in absolute terms (i.e. mm^3), the *Small size* algorithm showed least variability and *All sizes* algorithms had the highest variability ($p<0.01$, data not shown).

When the readers chose different algorithms, the volume determined by the *Subsolid* algorithm was always larger than that obtained with the *All sizes* algorithm ($p<0.001$), which again was always larger than that of the *Small size* algorithm (Fig. 3) ($p<0.001$). *All sizes* measurements were on average 89% (95% CI: 60–118%) larger than *Small size* measurement of the same nodule, and in 80% *All sizes* readings were more than 25% larger than *Small size* readings. *Subsolid* measurements were always more than 25% larger than readings using one of the algorithms for solid nodules, i.e. *Small size* or *All sizes*. On average volumetric results obtained with the *Subsolid* algorithm were 1,428% (95% CI: 508–2,347%) larger than those from algorithms for solid nodules (Table 1, Fig. 4a-c).

Discussion

In this study we examined one particular volumetry software package (LungCARE CT version VE25A, Siemens Medical Solutions, Erlangen, Germany) that offers several

Table 1 Algorithm applied by two readers, nodule characteristics and differences between readings

	Algorithm of 1st reader	Algorithm of 2nd reader	No. of nodule readings, (%)	Volume of nodules, Mean (range) mm^3	Difference between readers (1st – 2nd)			
					Mean (SD) Percent of minimal reading	Range	No difference in volume	Difference in volume >25%
Same algorithms	Small size	Small size	252 (65)	160 (7–2,088)	−1 (11)	−65 to 48	58%	4%
	All sizes	All sizes	36 (9)	1,056 (146–3,486)	1 (6)	−14 to 31	22%	3%
	Subsolid	Subsolid	23 (6)	392 (51–1,296)	−1 (3)	−5 to 8	4%	0%
	Total		311 (80)	281 (7–3,486)	0 (10)	−65 to 48	50%	4%
Different algorithms	Small size	All sizes	35 (9)	255 (6–995)	−100 (154)	−890 to −6	0%	77%
	All sizes	Small size	34(8)	304 (9–1,780)	77 (74)	14–406	0%	82%
	Small/All sizes	Subsolid	11 (3)	249 (5–2,107)	−1,428 (1,368)	−4,379 to −196	0%	100%
	Subsolid	Small/All sizes	0 (0)	–	–	–	–	–
	Total		80 (20)	275 (5-2,107)	−207 (705)	−4,379 to 406	0%	83%

algorithms for the analysis of nodules depending on the morphology of the nodule. Former versions of the software without this option have been tested [13–15]. Although several options may broaden the utility of the software, full understanding of these new features and a high reproducibility of measurements is a key issue before software can be used in clinical decision-making. We found volumetric measurements of screen-detected nodules were reasonably reproducible when readers used the same algorithm. However, it is not a good idea to use a step-up approach with different segmentation algorithms in order to try to optimise nodule segmentation. Even if the vendor offers different algorithms within the same software package, you should always stick to the same algorithm and record it in order to avoid massive measurement errors.

In the NELSON lung cancer screening trial, where nodules were segmented by volumetry software, volume measurements were identical in a very high percentage (89%) [13]. However, as in other recently published studies [14, 15], this study excluded subsolid, semi-solid, pleura-based and vessel-connected nodules. Only nodules surrounded by lung tissue (intraparenchymal nodules) were

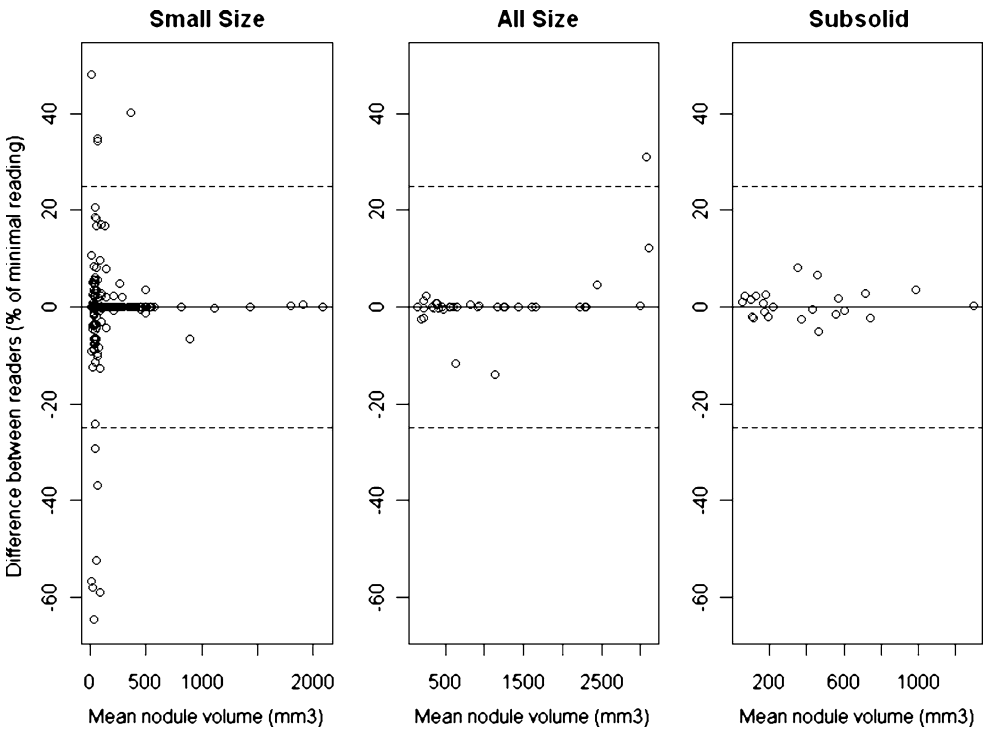

Fig. 2 Bland-Altman plots when the same algorithm was chosen by both readers. *Small size*=252 nodules, *All sizes*=36 nodules and *Subsolid*=23 nodules. *Dotted lines* indicate 25% variation. Because of the relationship between nodule size and variability (and to simplify), lines corresponding to the 95% confidence interval (CI) were omitted

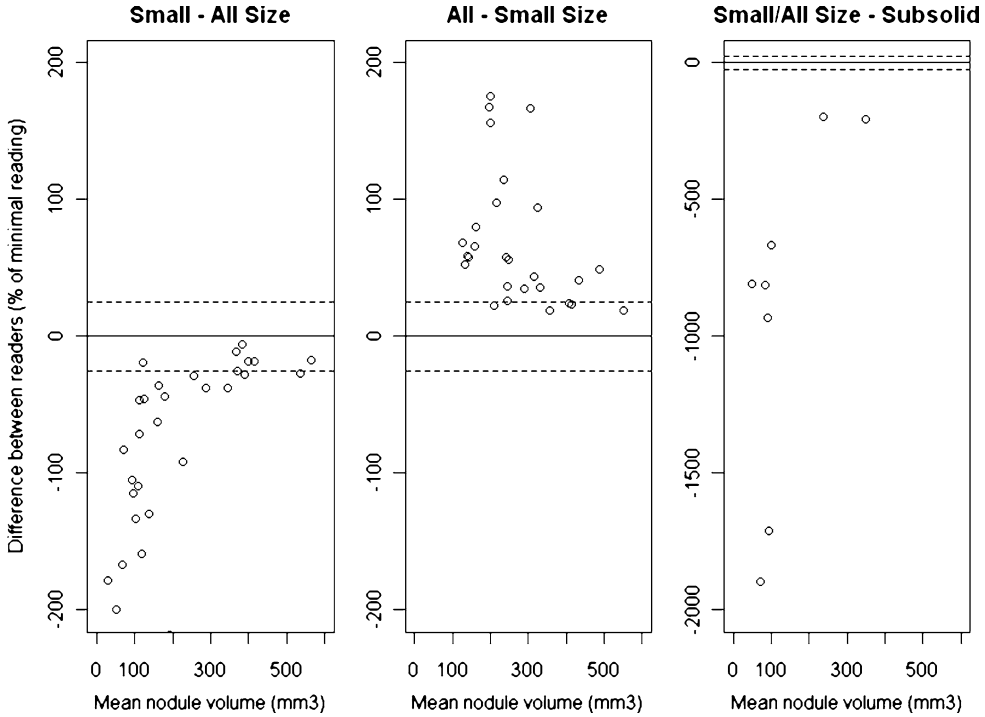

Fig. 3 Bland-Altman plots when the readers chose different algorithms. *Small size/All sizes*= 35 nodules, *All sizes/Small size*=34 nodules and *All sizes/ Small size – Subsolid*=11 nodules. *Dotted lines* indicate 25% variation. Because of the relationship between nodule size and variability (and to simplify), lines corresponding to the 95% CI were omitted

included, and they are known to be more reproducible when evaluated by pulmonary nodule evaluation software [15]. In a study [15] by the NELSON group of 4,225 nodules in 2,239 participants, they found complete agreement in volume in 86% and a disagreement ≥25% in 2% of nodules only. However, this study included solid 15- to 500-mm^3 nodules only, and if readers manually modified the volume in the prospective lung cancer screening study, the nodules were excluded as well. In the present study no nodules were excluded; we also included semi-solid and ground-glass lesions which, in our opinion, will provide a more representative result, as in everyday clinical practice all sorts of nodules must be assessed.

In 28% of the readings, at least one of the readers could not determine the volume (n=154). The relatively high number of nodules that could not be measured emphasises the necessity of visual validation of nodule segmentation

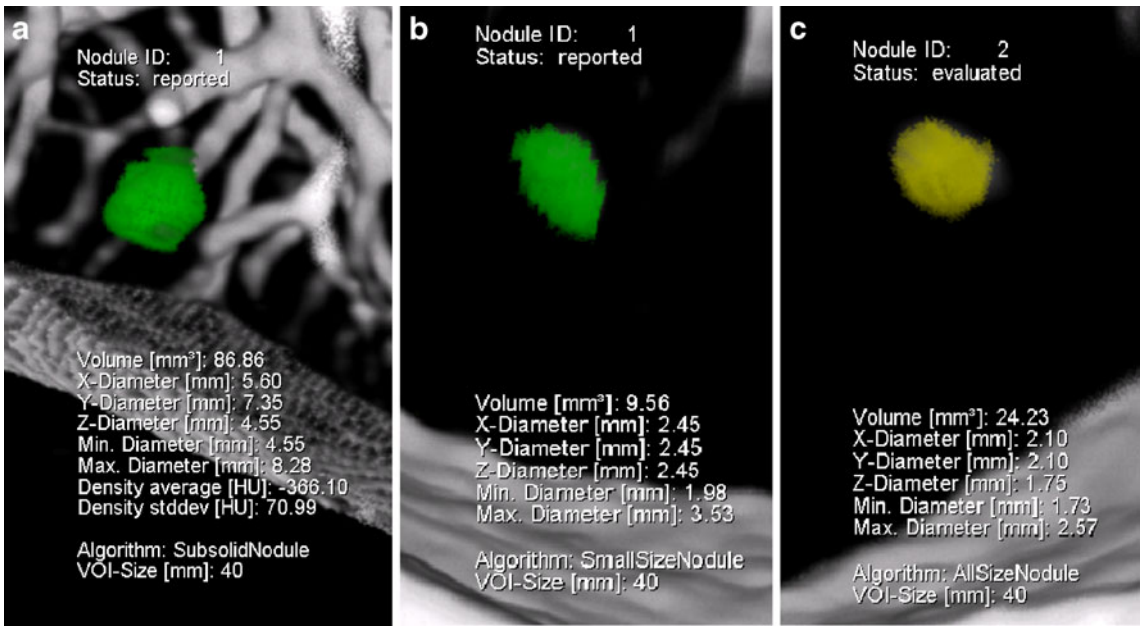

Fig. 4 The same nodule analysed at the same time with the *Subsolid* (**a**), *Small size* (**b**) and *All sizes* (**c**) algorithms

with skilled human interference in the reading process and demonstrates why fully automated detection of lung nodules is still unrealistic with the currently tested state-of-art software. A semi-automatic approach with a manual selection of nodules and supervision of the nodule rendering procedure is necessary to ensure accuracy of the volumetric measurement.

When comparing the difference between measurement in percent of the minimal volume the inter-observer variability was highest for *Small size* measurements and least using the *Subsolid* algorithm ($p<0.01$) (Fig. 2). However, when comparing volume differences in absolute values *Small size* measurements showed the least variability and *All sizes* the highest ($p<0.01$) (data not shown). This is because the size of the nodule has great influence on the variation coefficient, which is visualised in Fig. 2 where variability increases with decreasing volume. This effect is also seen when different algorithms are applied (Fig. 3); the difference between the readers tends to be smaller with increasing nodule size. Based on our data, it could not be decided which algorithm was most reproducible, because the algorithm was not randomly chosen. Furthermore, the variability of the volume measurement depended on several factors, such as nodule size, morphological characteristics of the nodule, and whether the variability was calculated in relative or in absolute terms.

One explanation for the observed volume disagreement between two readers is related to the fact that semiautomatic volumetric measurements may vary according to the positioning of the seed point by the observer, which is a non-automated part of the procedure. When a spherical 3D template gradually expands from the seed point, different starting positions within the nodules may lead to different volumetric results. Obviously, the chance of picking the same seed point is inversely related to the size of the nodule, and this may explain the high percentage of identical readings when using the *Small size* algorithm (Table 1).

In 20% of the correctly segmented nodules ($n=80$), the readers used different algorithms to analyse the same nodule, and this had a significant influence on the volume measured. If the same nodule was measured with two different algorithms, the *All sizes* measurements were always larger than the measurements in which the *Small size* algorithm was used, and usually (80%) the difference in volume exceeded 25%. Furthermore, measurements performed with the *Subsolid* algorithm were always (Fig. 4a) larger than the *Small size/All sizes* measurements (Fig. 4b, c). This shows that the *Subsolid* algorithm detects a larger volume compared with the solid algorithms, as part-solid nodules usually have a solid core surrounded by a larger subsolid sphere. The difference between *All sizes* and *Small size* measurements is less obvious and tended to be smaller with increasing nodule size (Fig. 3).

Some previous studies have indicated that a minimal growth of 25% is required to avoid confusion with random measurement variability [14–16]. In the NELSON and DLCST trials, the limit of 25% growth is implemented in the study protocol. Growth <25% is considered insignificant and no special follow-up is required, only nodules that grow >25% are referred for additional diagnostic workup [9, 12]. Therefore, when using pulmonary nodule evaluation software for repeated nodule measurements the variability should be <25%. Otherwise, variability >25% in repeated measurements may result in false-positive growth estimation.

The NELSON group has reported on the variability of volume analysis using pulmonary nodule evaluation software, and in one of their first studies analysing 430 nodules [13] they found for nodules with a discrepancy between two readings that 95% of the variability was between −22% and 29%. However, for most nodules (89%) there was no difference between readings. The variability was above 25%, i.e. false positive growth, in only 1.2% of nodules. In a later study of 218 nodules also from the NELSON group [14], variability was found to be dependent on nodule size as the variability lowered with increasing size, which was consistent with our findings (Fig. 3). In this study [14] significant growth was defined as being growth beyond the 95% CI of the variability, which was estimated to be −21% to 24%, which was comparable to the first study [13]. Furthermore, the NELSON group investigated the influence of nodule morphology and concluded that for irregular nodules the cut-off point for significant growth should be at 30% relative growth and only 15% for spherical nodules. These findings were later confirmed in a large study from NELSON [15] consisting of 2,367 nodules which also included attached nodules. The odds ratio for irregular nodules having variability above 15% was 9.1 (95% CI: 6.1–15.1) compared with spherical nodules. In all the NELSON studies [13–15] the same software package (LungCARE) was applied, and a comparison of the performances of six different software packages showed that none of the systems had a variability for adequately segmented nodules of more than 22.3% [10]. Overall, the studies above comply well with the 25% definition of significant growth used in most lung cancer screening trials [9, 12, 16]. However, when analysing subgroups of nodules dependent on size and morphology, the cut-off of 25% may be challenged, as smaller size and irregular nodules may require a higher cut-off. A recent study [17] further challenged the 25% definition of significant growth by suggesting that even 30% observed growth may not prove real growth. In our study, 4% had variability over 25% when the same algorithm was applied, and use of different algorithms resulted in 83% variability above 25% (Table 1). As mentioned previously, we also observed larger variability for smaller nodules, indicating the inappropriateness of applying the same threshold for growth to all nodules. To avoid false-positive growth, which is essential in lung cancer screening programs, the size and morphology of the nodule should be taken into

account and a sensible and customised approach to the definition of significant growth should be applied.

This study has limitations. Although the software is widely used, all results reported are valid only for the particular software release we used (LungCARE VE025A). Furthermore, we used a 3-mm slice thickness, while 1-mm slices are preferred because segmentation accuracy and reproducibility has proven to be superior with the use of the thin slices [18]. However, in the DLCST the radiologist used 3-mm slices when screening for nodules [12], and therefore this slice thickness was chosen in the present study.

We consider it to be a strength of the study that all nodules were included, and not just nodules surrounded by lung tissue (intraparenchymal nodules). Also the fact that the double readings were performed in two institutions in different countries strengthens the external validity of the study. The readers were completely blinded in the sense that they had no information regarding the choice of algorithm or the results of the other reader.

The development of new and improved pulmonary nodule evaluation software is a promising tool for the diagnostic workup of indeterminate lung nodules. In this context, the reproducibility of the volumetric measurements is a key issue before the results can be used in everyday clinical decision making. New versions of software are not always comparable with former versions. Siemens syngo LungCARE CT version VE25A used in the DLCST allowed the choice of various algorithms for the analysis of lung nodules. Former versions of the software used in the NELSON study did not have this option. The use of different algorithms presents a challenge when pooling data from several trials with the aim of gaining more statistical power, as volumetric measurements may not be directly comparable. A complete and independent understanding and validation of the different software packages requires full access to technical details behind the algorithm, and this is usually incompatible with the policy of software companies. However, a close cooperation between clinicians and software companies is desirable to ensure continued development of high-quality software.

Conclusion

We found volumetric measurements to be reproducible using Siemens *syngo* LungCARE CT version VE25A, when using the same nodule-analysing algorithm. Modern volumetric software failed to correctly segment a high number of manually detected screen nodules (28%). Provided the same software algorithm was used, 96% of the volumetric measurements showed a variability of less than 25%. However, segmentation algorithms within the same software package cannot be used interchangeably, and using the same analysing algorithm is essential for correct longitudinal assessment of lung nodules.

References

1. Verschakelen JA, Bogaert J, de Wever W (2002) Computed tomography in staging for lung cancer. Eur Respir J 19:40–48
2. Henschke CI, McCauley DI, Yankelevitz DF et al (1999) Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 354:99–105
3. Swensen SJ, Jett JR, Hartman TE et al (2003) Lung cancer screening with CT: Mayo Clinic experience. Radiology 226:756–761
4. Xu DM, van der Zaag-Loonen HJ, Oudkerk M et al (2009) Smooth or attached solid indeterminate nodules detected at baseline CT screening in the NELSON study: cancer risk during 1 year of follow-up. Radiology 250:264–272
5. Hasegawa M, Sone S, Takashima S et al (2000) Growth rate of small lung cancers detected on mass CT screening. Br J Radiol 73:1252–1259
6. Pauls S, Kürschner C, Dharaiya E et al (2008) Comparison of manual and automated size measurements of lung metastases on MDCT images: potential influence on therapeutic decisions. Eur J Radiol 66:19–26
7. Kostis WJ, Reeves AP, Yankelevitz DF et al (2003) Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. IEEE Trans Med Imaging 22:1259–1274
8. Yankelevitz DF, Anthony PR, William JK, MS et al (2000) Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. Radiology 217:251–256
9. Xu D, Gietema H, de Koning H et al (2006) Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 54:177–184
10. De Hoop B, Gietema HA, van Ginneken B et al (2009) A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated CT examinations. Eur Radiol 19:800–808
11. Siemens LungCARE CT and syngo LungCAD Homepage. Available via Siemens. http://www.medical.siemens.com/webapp/wcs/stores/servlet/ProductDisplay~q_catalogId~e_-11~a_catTree~e_100010,1007660,12752,1008405,1008410~a_langId~e_-11~a_productId~e_11611~a_storeId~e_10001.htm. Accessed 1 March 2009
12. Pedersen JH, Ashraf H, Dirksen A et al (2009) The Danish randomized lung cancer CT screening trial—overall design and results of the prevalence round. J Thorac Oncol 4:608–614

13. Gietema HA, Wang Y, Xu D et al (2006) Pulmonary nodules detected at lung cancer screening: interobserver variability of semiautomated volume measurements. Radiology 241:251–256
14. Gietema HA, Schaefer-Prokop CM, Mali WPTM et al (2007) Pulmonary nodules: interscan variability of semiautomated volume measurements with multisection CT—influence of inspiration level, nodule size, and segmentation performance. Radiology 245:888–894
15. Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ et al (2008) Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. Radiology 248:625–631
16. Marchiano A, Calabro E, Civelli E et al (2009) Pulmonary nodules: volume repeatability at multidetector CT lung cancer screening. Radiology 251:919–925
17. Rampinelli C, Fiori DE, Raimondi S et al (2009) In vivo repeatability of automated volume calculations of small pulmonary nodules with CT. AJR Am J Roentgenol 192:1657–1661
18. Gurung J, Maataoui A, Khan M et al (2006) Automated detection of lung nodules in multidetector CT: influence of different reconstruction protocols on performance of a software prototype. Rofo 178:71–77