

Bartjan de Hoop
Hester Gietema
Bram van Ginneken
Pieter Zanen
Gerard Groenewegen
Mathias Prokop

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

Received: 1 July 2008
Accepted: 6 October 2008
Published online: 19 November 2008
© The Author(s) 2008
This article is published with open access at
Springerlink.com

B. de Hoop (✉)
University Medical Center,
Heidelberglaan 100,
3508, GA, Utrecht, The Netherlands
e-mail: b.j.dehoop@umcutrecht.nl
Tel.: +31-887-553193

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

B. van Ginneken
Image Sciences Institute,
University Medical Center,
Utrecht, The Netherlands

P. Zanen
Department of Pulmonology,
University Medical Center,
Utrecht, The Netherlands

G. Groenewegen
Department of Oncology,
University Medical Center,
Utrecht, The Netherlands

Abstract We compared interexamination variability of CT lung nodule volumetry with six currently available semi-automated software packages to determine the minimum change needed to detect the growth of solid lung nodules. We had ethics committee approval. To simulate a follow-up examination with zero growth, we performed two low-dose unenhanced CT scans in 20 patients referred for pulmonary metastases. Between examinations, patients got off and on the table. Volumes of all pulmonary nodules were determined on both examinations using six nodule evaluation software packages. Variability (upper limit of the 95% confidence interval of the Bland-

Altman plot) was calculated for nodules for which segmentation was visually rated as adequate. We evaluated 214 nodules (mean diameter 10.9 mm, range 3.3 mm–30.0 mm). Software packages provided adequate segmentation in 71% to 86% of nodules ($p < 0.001$). In case of adequate segmentation, variability in volumetry between scans ranged from 16.4% to 22.3% for the various software packages. Variability with five to six software packages was significantly less for nodules ≥ 8 mm in diameter (range 12.9%–17.1%) than for nodules < 8 mm (range 18.5%–25.6%). Segmented volumes of each package were compared to each of the other packages. Systematic volume differences were detected in 11/15 comparisons. This hampers comparison of nodule volumes between software packages.

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

Introduction

Many small lung nodules are incidentally encountered with the use of thin collimations at multi-row detector computed tomography (MDCT). They are the major radiographic indicator of lung cancer and metastatic disease, despite the fact that most of them are caused by

benign conditions [1–3]. Accurate size and growth measurements are essential in the differential diagnosis of these nodules and also in an oncologic setting for observing the response to therapy.

Nodule volumes can be calculated using semi-automated nodule volumetry software. Reproducibility of volumes calculated by such software has been proven to exceed the

reproducibility of manual diameter measurements on which the RECIST criteria are based, but such nodule volumetry programs are still not yet widely implemented [1, 2]. Until recently, an obstacle for wide implementation of lung nodule volume measurements was the high rate of segmentation errors. This discourages the clinician to use such tools. Currently, new and more sophisticated segmentation software is being developed with the idea of alleviating this problem.

Volumetric measurements are usually based on segmentation of nodules on thin-section CT data sets and an algorithm that translates the segmented voxels into a nodule volume. For synthetic nodules, volumetry with various commercially available software packages proved to be very accurate [3]. In vivo accuracy, however, is less due to less sharply defined nodule borders, motion effects and complex geometry of adjacent structures. Earlier studies showed that a large degree of the interexamination variability of nodule volumetry can be explained by segmentation errors [4, 5], which were found to be common in irregularly shaped nodules. Currently, several software packages have been developed that claim to be able to adequately segment nodules irrespectively of size, shape and location. Variation in the results of volumetry may result in false-positive or false-negative conclusions with potential serious consequences for the patient. To avoid overinterpreting random changes in volumetric measurements, the diagnosis of real growth or regression typically requires that the difference in measured nodule size exceeds the upper limit of agreement. This upper limit of agreement, however, may be dependent on the software package used.

In this study, we simulated the situation of a baseline and follow-up chest exam on which a suspicious nodule was detected and subsequently evaluated by commercially available software for the presence of change. In order to determine the minimum amount of change needed to be 95% sure that the change was due to a real growth, we determined the interexamination variability with patients scanned twice in the same session and moved on and off the table. This variability was influenced by both patient factors (e.g., inspiration depth) and the software algorithm [4]. Since all software packages had to segment the same set of nodules, the approach we chose allows for determining a realistic number of the amount of change needed that incorporates all factors seen in clinical routine.

The objective of this study was to compare the interexamination variability of pulmonary nodule volumetry on repeat CTs using six currently available semi-automated software packages for CT. We used a dataset containing nodules of varying size, morphology and contact to pulmonary structures in order to define the upper limit of agreement for each software package and to determine whether these software packages could be used interchangeably.

Materials and methods

Patients and nodule selection

The study was approved by our institutional review board, and written informed consent was obtained from all patients after explanation about the risks, including the additional radiation dose. Twenty consecutive adult patients (15 men, 5 women, 40–84 years old, mean 57 years) with known pulmonary metastases were enrolled. All patients visited the oncology outpatient department and were referred for a chest CT for clinical indications. The presence of lung metastases had been previously shown on chest CT or chest radiography. Patients were referred for chest CT to monitor the effect of anticancer therapy. Primary tumors were melanoma (n=3), renal cell carcinoma (n=6), colorectal cancer (n=5), breast carcinoma (n=2), prostate cancer (n=1), seminoma (n=1), medullary thyroid cancer (n=1) and esophageal cancer (n=1).

All solid lesions with a minimum volume of 15 mm³ (corresponding to a diameter of about 3 mm) were included. Lung masses, defined as nodules exceeding 30 mm in diameter, were excluded from analysis. Nodules suspected of being metastases were included, as well as nodules that could potentially have a benign histology. Completely calcified nodules, however, were excluded. Only solid nodules were included since non-solid or partly solid nodules require a different segmentation approach, and not all of the evaluated segmentation software packages were developed for this task. A maximum of 50 nodules per patient were included.

To have an independent indication of nodule size, the maximum diameter of the nodule was measured using an electronic ruler. Nodules were categorized by size according to the Fleischner criteria [6]. Nodule shape was categorized based on 3D images into spherical, lobulated or irregular. A nodule was defined as spherical when it had an approximately constant radius and as lobular when it had a variable radius, but smooth outer margins. It was defined as irregular when the outer margins were not smooth. Attachment to the pleura or pulmonary vessels was noted.

Image acquisition

Two low-dose non-contrast-enhanced chest CTs were performed, followed by a contrast-enhanced standard-dose chest CT for clinical purposes. Between the two low-dose CT examinations, patients were asked to get off and on the table to simulate the conditions of a repeat CT examination for follow-up of a pulmonary nodule. Using this setup, growth or regression of the lung nodules could reliably be excluded.

All CT data were acquired on a 16-detector-row CT system (Mx8000 IDT, Philips Medical Systems, Cleveland, OH) using a spiral mode with 16×0.75-mm collimation.

The entire chest was examined. CT data were acquired in full inspiration. Exposure settings for the low-dose examinations were 30 mAs and 120 kVp or 140 kVp, depending on the patient's weight. The corresponding volume CT dose indices were 2.2 mGy and 3.5 mGy, respectively. Axial images were reconstructed at 1.0-mm thickness and 0.7-mm increment, using a moderately soft reconstruction kernel, the smallest field of view that included the outer rib margins and a 512×512 matrix.

Lung volume in both examinations was measured using the lung segmentation algorithm incorporated in the GE software.

Semi-automated volume measurements of pulmonary nodules

All nodule measurements were done by a single observer (2 years of experience in radiology with special interest in CT lung cancer evaluation). Nodules were identified using axial thin-slab maximum intensity projections (slab thickness 10 mm) displayed with window/center settings of 1,500/-500 HU. The same nodule was identified on the follow-up CT images using a printed screenshot.

The following segmentation algorithms were evaluated: Advantage ALA (GE, v7.4.63), Extended Brilliance Workspace (Philips, EBW v3.0), Lungcare I (Siemens, Somaris 5 VB 10A-W), Lungcare II (Siemens, Somaris 5 VE31H), OncoTreat (MEVIS, v1.6), and Vitrea (Vital images, v3.8.1, lung nodule evaluation add-on included). For the purpose of anonymization, the characters A to F were randomly assigned to the various packages.

In all algorithms, segmentation was initiated by clicking in the center of a nodule, starting a fully automated evaluation of the nodule. Next, all algorithms segment the nodule, calculate its volume and present the result. The segmented area was shown by the various software

packages by a thin line surrounding the area of the nodule or by a colored overlay. This segmentation was visually judged for accuracy. In order to minimize observer influence, only these automated results were used for comparisons in this study, except for when explicitly written otherwise.

We did a separate analysis for results obtained using manual correction of incomplete segmentations. Manual correction by the user was allowed to correct the segmentation by four of six packages. In case of a mismatch between nodule and segmentation, this feature was used to obtain the most precise segmentation feasible. The type of manual correction varied between the packages (Table 1). Two packages also allowed for a complete manual segmentation in case of failure; this feature was not used. Next, the segmentation was again visually judged for accuracy.

Evaluation of segmentation accuracy

In order to evaluate segmentation accuracy, all packages offered a volume-rendered display that could be turned and a thin-section image in at least one plane that could be scrolled back and forth. Two packages gave the possibility to evaluate the segmentation in other planes as well. The observer visually classified the segmentation accuracy into four categories: (1) 'excellent': excellent segmentation, the overlay completely matched the nodule; (2) 'satisfactory': although not perfect, the segmented volume is still representative of the nodule. The maximum mismatch between overlay and nodule is visually estimated not to exceed 20% in volume. (3) 'Poor': part of the nodule is segmented, but the segmented volume is not representative of the nodule (estimated mismatch >20%). (4) 'Failure': No segmentation or the result has no similarity with the lesion. An example of each classification can be found in

Table 1 Possibilities for manual correction used to correct the semi-automated nodule segmentation for the various software packages tested in this study

| | Advantage ALA | EBW | Lungcare I | Lungcare II | OncoTreat | Vitrea |
|---|---------------|-----|------------|-------------|-----------|--------|
| Manual adjustments to segmentation possible. | | • | • | • | • | • |
| Manually re-drawing (part of) the contour on one section possible, after which the software recalculates the segmentation. | | | | | • | • |
| User can choose to increase/decrease segmented volume, but user has no influence on how the segmentation will be changed. | | • | • | | | |
| User can adjust roundishness, which will effect the strength of an erosion/dilation step employed to separate lesions from adjacent pulmonary structures. | | | | | • | |

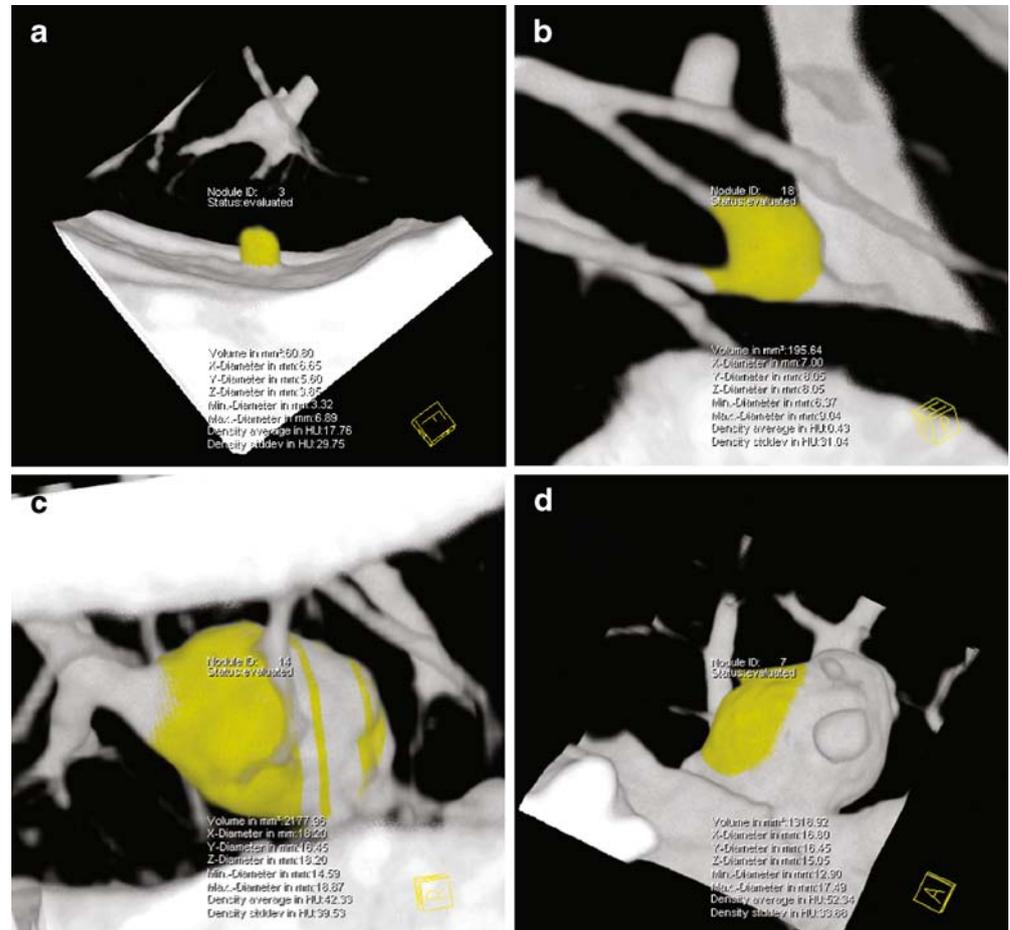
In our experience, manual correction using the 'increase/decrease nodule volume' or 'adjust roundishness' functions takes only seconds, while manually re-drawing a contour can take up to a few minutes. Note that these options may change with newer software versions. Systems are presented in alphabetic order. In order to anonymize the systems in the results section, a constant but different (randomly assigned) order (systems A to F) is used for the other tables and figures

Fig. 1. In order to exclude the influence of failed segmentations on the reproducibility of a software, nodules were grouped into ‘adequately’ (group 1 and 2) and ‘inadequately’ (group 3 and 4) segmented nodules. Inadequately segmented nodules were excluded from the calculations of inter-examination variability as these segmentations have no value and greatly influence volume measurement reproducibility, making meaningful comparisons impossible.

Reproducibility of visual assessment of segmentation

The intra- and interobserver reproducibility for the visual assessment of segmentation accuracy was tested. On one system, the observer performed the visual assessment twice with 1 week in between readings. A second observer, a CT technician with special training in evaluating and reporting cancer screening CTs with the use of volumetric software (>4,000 examinations in 3 years), repeated the visual assessment of the segmentation accuracy as well, on the same and on a second system. We identified the percentage of nodules in which the visual assessment of segmentation changed between adequate and inadequate.

Fig. 1 Examples of segmentation quality. a: Excellent segmentation b: Satisfactory segmentation. c: Poor segmentation. d: Failed segmentation



Statistical evaluation

All statistics were calculated using Microsoft Excel XP (Microsoft, Redmond, Wash.) and the SPSS statistical software package version 15 (SPSS, Chicago, Ill.).

To assess the effects of inspiration level, we calculated the Pearson correlation coefficient between the relative difference and the ratio of lung volumes (first/second examination).

In order to compare the number of adequately segmented nodules per system, a binominal test was applied, using the percentage of adequately segmented nodules of the best system as test proportion.

Differences in volume (ΔV) were calculated by subtracting the volume measured on the first scan (V_1) from the volume measured on the second scan (V_2). This difference was then normalized with respect to mean nodule volume to assess relative differences:

$$\Delta V_{rel} = 100\% \cdot \frac{V_2 - V_1}{(V_1 + V_2)/2}.$$

The histogram of relative differences showed a normal distribution for all packages (tested with the Kolmogorov-

Smirnov test). Because the same nodule was measured twice on successive chest CTs, a mean relative difference close to 0 can be expected. In fact, none of the packages had a mean relative difference higher than 1.1%. We therefore decided to use only the upper limit of agreement of the 95% CI of the relative differences as assessed according to the method proposed by Bland and Altman [7] as the measure of interexamination variability. An increase in nodule volume above this upper limit of agreement can, with 95% confidence, be attributed to real growth.

To compare the various software packages with respect to interexamination variability, we used an F-test on a subgroup of nodules that were adequately segmented on both scans by all packages.

We also tested whether there was a significant difference in interexamination variability between excellently and satisfactorily segmented nodules. For each software package separately, an F-test was used to compare interexamination variability for all those nodules that were classified as excellently or satisfactorily segmented with this specific software.

An F-test was also used to test for differences in interexamination variability before and after manual correction by the user.

Influence of nodule diameter on interexamination variability was tested using one-way ANOVA.

In order to detect systematic differences in measured volumes between packages, we performed a mixed model variance analysis of nodule volumes on a subset of nodules that were adequately segmented by all programs.

Results

Nodule characteristics

A total of 214 solid pulmonary nodules fulfilled the inclusion criteria. Mean diameter, measured with the electronic ruler, was 10.9 mm (range 3.3 mm–30.0 mm). The number of eligible nodules per patient ranged from 0 (no visible metastases visible after therapy) in nine patients to 50 nodules in one patient, with a median of 19 in the patients with nodules. The database included 91 round, 39 lobular and 84 irregular nodules. Seventy-six had no contact to pulmonary structures, 60 were pleura attached, and 78 nodules were attached to a vessel or to both vessel and pleura.

Influence of inspirational level on interexamination variability

Ratio of lung volume between scan 1 and 2 ranged from 81% to 126%, with a mean of 102%, weighted to the number of nodules per patient. Inspirational level did not significantly correlate to the relative difference for any of

the software packages. Mean correlation coefficient for all software packages was -0.11 ($p=0.30$). Figure 2 shows the software packages with the weakest and strongest correlation to inspirational level.

Reproducibility of visual assessment of segmentation accuracy

Both intra- and interobserver reproducibility of the visual assessment of segmentation accuracy was high. Only for 2 of 214 nodules (1%), the visual assessment of segmentation changed between adequate and inadequate when the same observer repeated the measurements. The second observer changed the accuracy score for 3/214 (1.4%) nodules on the first and 6/214 (2.8%) nodules on the second system.

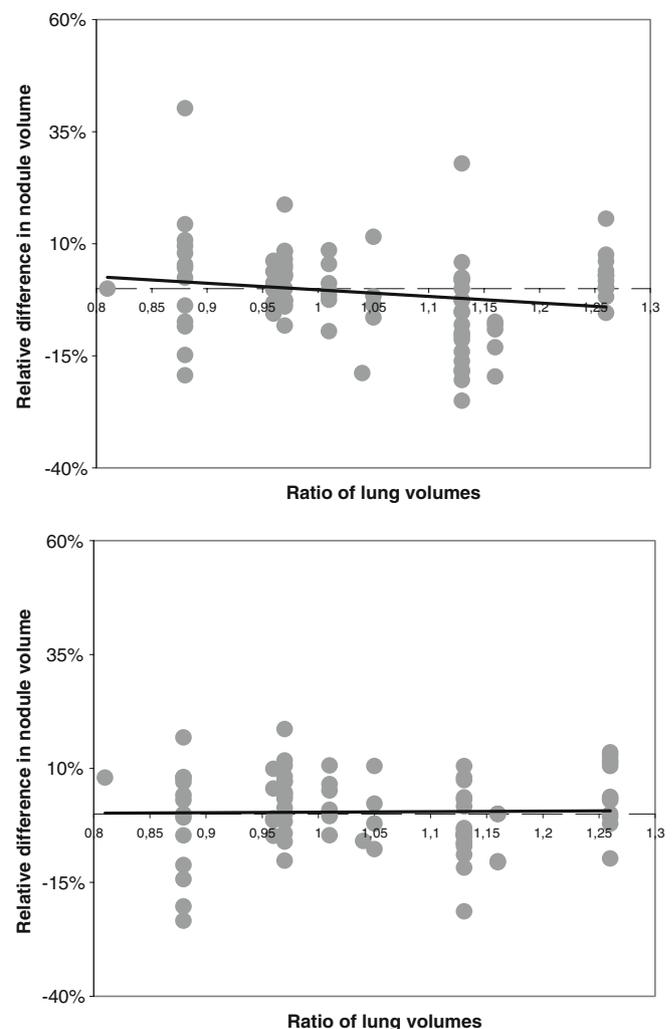


Fig. 2 Correlation between ratios of lung volumes (lung volume on second scan divided by lung volume on first scan) and differences in measured nodule size. The graphs show the software package with the strongest correlation (a: software A, $r=-0.18$, $p=0.10$) and the weakest correlation (b: software F, $r=0.02$, $p=0.87$)

Segmentation accuracy

The number of nodules that were segmented adequately by each individual package varied significantly from between 71% and 86% before manual correction and between 71% and 98% after correction depending on the software package used ($p < 0.001$) (Fig. 3).

Interexamination variability

Eighty-nine (42%) of all nodules were adequately segmented by all packages. In this dataset, interexamination variability of software packages B and D was significantly lower than that of package C, E and F. Extent of variability of package C also differed significantly from package A (Table 2).

Significant differences were seen in interexamination variability between excellently and satisfactorily segmented nodules for three of the six software packages (Table 2).

The overall measurement variability of the software packages before and after manual correction is given in Fig. 4. For each individual software package, Fig. 4 shows the variability for all nodules adequately segmented by this particular software package. Note that the numbers differ from those in Table 2 because for each package all adequately segmented nodules were included and not only the subset of 89 nodules that were adequately segmented by all software packages. Compared to the automated results, the upper limit of agreement did not change significantly after manual correction for any of the software packages.

Nodule diameter significantly influenced the extent of interexamination variability in five out of six software packages. This influence is shown in Table 2. Due to a low number of nodules (<40) in the groups smaller than <8 mm, these categories were combined.

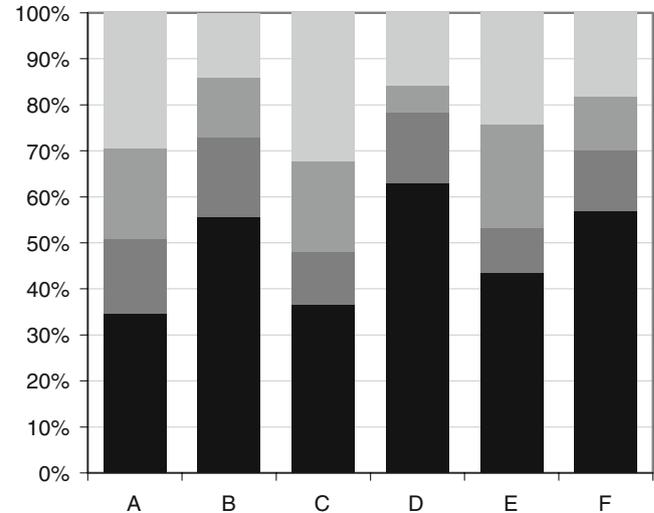
Systematic volume differences

In the dataset of nodules that were adequately segmented by all packages, mean nodule volume per software packages is given in Table 2. The mixed model variance analysis showed significant systematic differences in mean volumes in 11 of the 15 ($6 \times 5/2$) possible pairs of software packages.

Discussion

In this study we show substantial differences in segmentation performance among six currently available pulmonary nodule segmentation software packages in a dataset of nodules with a variety in size, morphology and contact to pulmonary structures. The best package segmented 86% of all nodules in both the first and the second examination

a Segmentation accuracy (without manual adjustments)



b Segmentation accuracy (after manual adjustments)

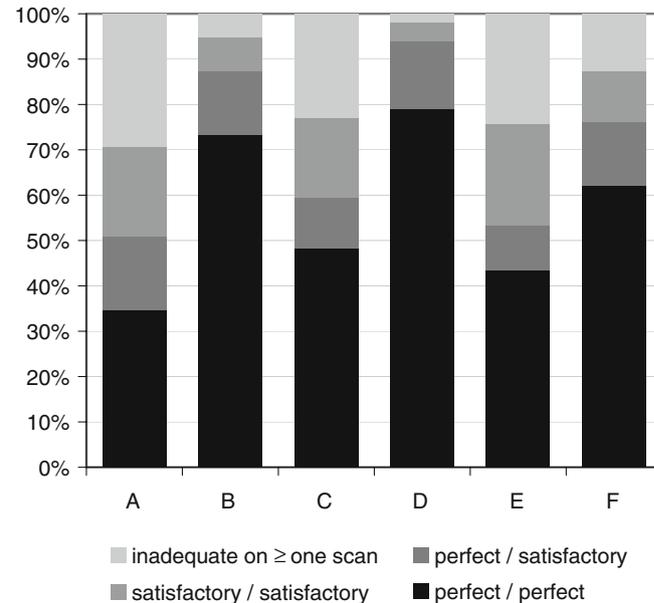


Fig. 3 Comparison of visual rating of segmentation accuracy for the various software packages. The graph displays the percentage of nodules for which segmentation was rated “excellent” and/or “satisfactory” on scan 1 and scan 2. Nodules for which segmentation was rated “poor” or “failure” on at least one of the scans are summarized as “inadequately” segmented nodules. a: Displays the raw results without manual adjustment of nodule contours, while (b) displays results after manual correction of the segmentation. Note that systems A and E did not allow for manual correction

with excellent or satisfactory accuracy. All software packages showed similar interexamination variability, but there were significant differences in absolute nodule volumes between software packages. Manual correction substantially improved the number of accurate segmentations without significantly affecting reproducibility.

Table 2 Comparison of interexamination variability

| Software package | A | B | C | D | E | F |
|-------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| a. Common dataset | 17.0% | 13.1% | 20.8% | 13.4% | 20.5% | 19.6% |
| b. Individual datasets | | | | | | |
| Excellent | 15.9% | 15.4% | 21.0% | 13.5% | 20.4% | 19.5% |
| Satisfactory | 21.9% | 18.0% | 22.9% | 27.6% | 19.5% | 19.7% |
| c. Influence of size | | | | | | |
| <8 mm | 21.2% | 18.5% | 24.9% | 19.5% | 25.6% | 24.5% |
| ≥8 mm | 17.1% | 14.5% | 16.8% | 16.1% | 12.9% | 13.8% |
| d. Median nodule volume | 80 mm ³ | 78 mm ³ | 59 mm ³ | 85 mm ³ | 57 mm ³ | 95 mm ³ |

(a) Variability in the common dataset of 89 nodules that were adequately segmented by all software packages. (b) Variability for the individual datasets of all nodules that were excellently and satisfactorily segmented by each particular software package. A nodule that was rated as excellently segmented in one CT examination and satisfactorily in the other was classified as 'satisfactory.' Significant differences are printed bold. (c) Variability for different sizes of nodules, also for the individual datasets. Note that results for different software packages are therefore not comparable. Significant differences are printed in bold. (d) Median nodule volume in the common dataset of 89 nodules that were sufficiently segmented by all software packages. Despite the fact that the sets contain identical nodules, there are substantial differences in nodule volumes measured by the various programs

High segmentation accuracy is a prerequisite for adequate performance of nodule volumetry software. It is obvious that segmentations that include surrounding structures or do not include part of a nodule may lead to wrong management decisions. We found substantial variations in segmentation accuracy between software packages.

We also found that the differences in variability were comparatively small when excellent and satisfactory (small segmentation errors included) segmentations were compared. While there were significant differences between excellent and satisfactory segmentations in some packages, these differences were comparatively small and in the order of the differences between software packages.

In a comparison using the common dataset of 89 nodules that were adequately segmented by all software packages, differences between software packages were significant for a minority of packages, but absolute differences were small even between these packages. Overall, no system had a variability of more than 22.3% for the automated volumetry. In case of an adequate segmentation, an increase in measured nodule volume of more than 22.3% can, with 95% confidence, be attributed to real growth for all software packages tested. To put this variability into perspective, a volume difference of 22.3% in a nodule with a diameter of 5 mm equals a diameter change of only ~0.3 mm. The RECIST criteria for progressive disease are a 20% increase in the sum of the longest diameter of target lesions, equivalent to a 73% volume increase. The interexamination variability of all software packages was well below this threshold, which creates the opportunity to detect progressive disease earlier.

There was, however, a significant difference in absolute nodule volumes among software packages. This can lead to variations in management decisions: Software X and Y will measure higher volume (and thus, effective diameter) than software Z, for example, and will therefore induce more

aggressive management decisions if recommendations by the Fleischner society are followed. [6]. Consequently, the nodule shown in Fig. 5 would be treated differently depending on the segmentation algorithm used. As shown in Table 2, the size changes required to detect significant growth may be substantially greater when different software packages are used for baseline and follow-up evaluation instead of the same package for both examinations. In addition, there is a substantial bias, which means that the systematic differences in mean volumes can lead to a situation in which a growing nodule will appear to have shrunk (e.g., first measurement with software F, second measurement with software E) or a stable nodule will appear to have grown (e.g., first measurement with software E, second measurement with software F).

Diameter size influenced interexamination variability in five of the six software packages. Variability decreased as nodule size increased because a small segmentation error has a larger influence in a small nodule compared to a larger nodule. It is therefore important to use a larger confidence interval for small nodules, depending on the software used.

Semi-automated volumetric measurements have been tested before. Wormanns et al. applied a similar study design and found a 95% CI for the relative difference of -20.6% to 21.9% using an older version of Lungcare (Siemens, Somaris/5 VB 10A-W) [5], a version that was also tested in this study. After exclusion of all nodules that were inadequately segmented by this particular package, our results show comparable variability for this particular package. Kostis et al. used a segmentation algorithm developed in an academic environment to study variability of volume measurements [8]. Although all juxtapleural nodules were excluded, they demonstrated a standard deviation of the relative differences of 16.1%, which would amount to an interexamination variability of 31.6%. The

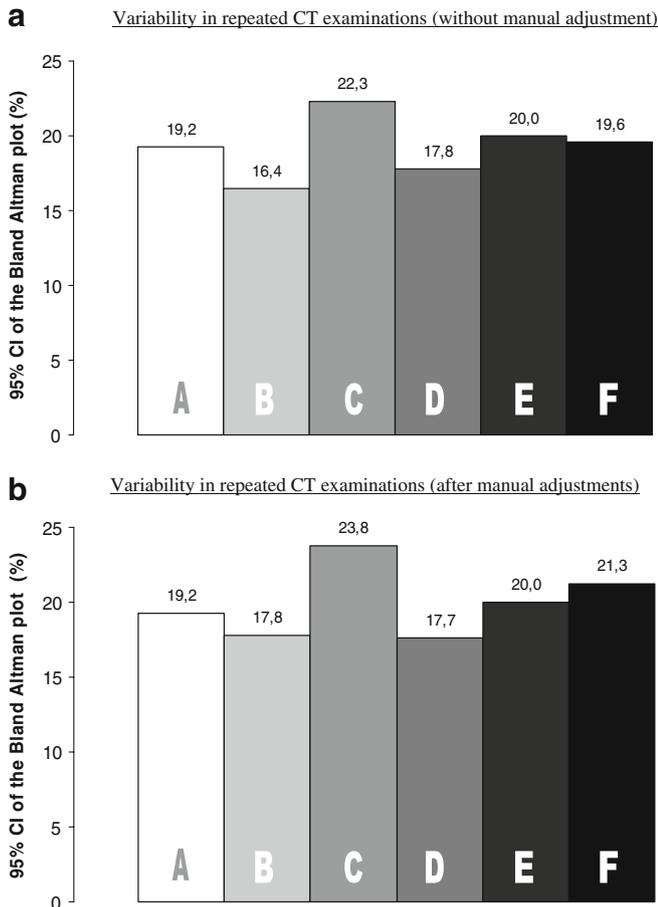


Fig. 4 Comparison of overall variability in repeated CT examinations for the various software packages, without (a) and after (b) manual adjustments. For each package only those nodules were considered for which segmentation was visually rated as adequate (excellent or satisfactory on both scans). The percentage of nodules included in these calculations therefore varied between (a) and (b), and per software package (see Fig. 3). Overall interexamination variability is expressed as the upper limit of the 95% CI for the relative differences between scan 1 and 2. **a** Variability in repeated CT examinations (without manual adjustment). **b** Variability in repeated CT examinations (after manual adjustments)

Advance Lung Analysis (ALA) package from GE Healthcare was tested in a similar way, but had a smaller database than the one used in this study by Goodman et al. [9]. They reported a variability of 25.6% for this particular package.

Significant differences in variability existed between excellently and satisfactorily segmented nodules. Still, we chose to combine these categories for several reasons. Firstly, although significant, the differences between the two categories are relatively small. Secondly, we wanted to mimic daily clinical practice in which a small segmentation error can be missed easily, and small errors are still common (Fig. 3). Therefore, including ‘satisfactorily’ segmented nodules will give a more realistic representation of the variability between two sequential CT scans.

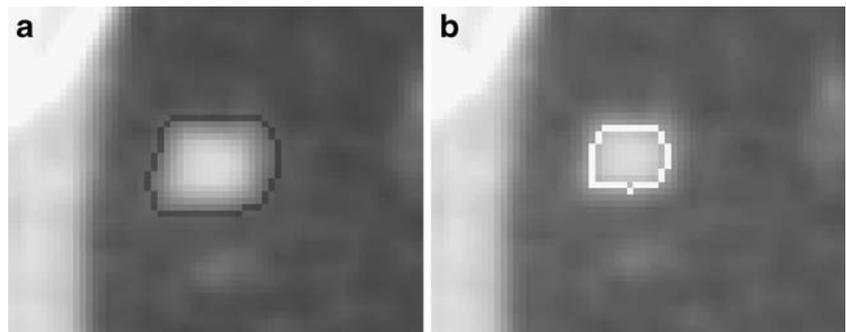
Inspiration depth did not significantly influence variability of volume measurements on sequential CT examinations. Nevertheless, a trend was visible towards smaller nodule size with increasing lung volume. This effect is small and appears to be substantially less important than other factors. It only had a marginal influence on the interexamination variability, as the mean ratio of lung volume for all patients was close to one. This is also illustrated by the mean relative difference in nodule volume between the two examinations, which was no more than 1.1% for any of the software packages.

Our study has limitations. Only one observer did all the measurements. However, all evaluated software packages required only a single click on the nodule to start a fully automated evaluation of the nodules, minimizing the impact of a human observer. The observer further only influenced the results by rating the segmentations for accuracy. The reproducibility tests in this study demonstrated that this rating process was very reproducible by the same and by a second observer.

The purpose of this study was to compare interexamination variability. Segmentation accuracy was only visually assessed. Accuracy of pulmonary nodule segmentation software has been tested using synthetic nodules and has been shown to be accurate, although systematic biases were frequent [3, 10–14]. However, as long as a software package is consistent in under- or overestimation of a volume, this will only marginally influence measured growth rates.

Our results are only applicable to images with the same slice thickness as used in this study. Recent literature suggests

Fig. 5 Systematic volume differences between software packages. Although both packages draw an accurate outline around the nodule, (a) shows a volume of 82 mm³ (diameter ~5.4 mm), while in (b) the calculated volume was only 32 mm³ (diameter ~3.9 mm)



that segmentation performance and accuracy are influenced by this parameter [15]. Whether it also influences inter-examination variability has not been tested. We chose to use 1.0-mm slice thickness since accuracy is best in thin slices. The reconstruction kernel has also been shown to influence accuracy, but these influences were small and clinically irrelevant [10]. Low-dose imaging was chosen in order to minimize the radiation burden for the participants. With thin slices, low dose imaging does not affect the accuracy of volumetry compared to the standard dose [15, 16].

The segmentations of a nodule in examination 1 and 2 were performed separately, and the operator was blinded for the results of the assessments. This was done in order to acquire an objective volume measurement for each nodule. This did not affect the fully automated measurements, but this approach could have influenced the variability after manual correction because it may have concealed obvious segmentation differences. In clinical practice, a clinician will frequently view both segmentations together so that obvious differences will be noted more rapidly. Therefore, variability might be less in a clinical setting in which parallel viewing of two sequential CT examinations is used and manual correction, when available, is used to minimize segmentation differences. In this study, we mainly included metastases, so the results may have limitations when transferred to a nodule suspected of lung cancer. However, the shape of the nodules was very variable, and they by no means only included round or lobulated lesions: a substantial proportion of nodules (39%) had irregular contours.

We had only one observer that performed the scoring, which meant that our results include effects of intra-observer variability, but not inter-observer variability. However, in clinical practice nodule measurements performed for follow-up of lesions will frequently be performed by the same person. In addition, inter-observer variability has been demonstrated by Wormanns et al. to be almost negligible compared to the variability caused by other factors [5].

In conclusion, substantial variations in segmentation performance exist among current lung nodule software packages. However, in case of a visually adequate segmentation, the change in nodule volume required to be 95% sure that there is real growth is similar, ranging 16% to 22%, for all software packages. It varies, however, with nodule size and to a much lesser degree to the difference in inspirational level. Systematic differences in volume measurements between packages hamper comparison of measurement results from different software packages for evaluation of a single nodule and can influence treatment decisions. Treatment decisions based on nodule volumetry need therefore to be based on identical software for volumetry on baseline and follow-up CT examinations.

Acknowledgement We thank Wibeke van Leeuwen and Caroline van Bavel for their support in collecting the data.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Marten K, Engelke C (2007) Computer-aided detection and automated CT volumetry of pulmonary nodules. *Eur Radiol* 17:888–901
- Therasse P, Arbutk SG, Eisenhauer EA et al (2000) New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 92:205–216
- Yankelevitz DF, Reeves AP, Kostis WJ et al (2000) Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. *Radiology* 217:251–256
- Gietema HA, Schaefer-Prokop CM, Mali WPTM et al (2007) Pulmonary nodules: Interscan variability of semi-automated volume measurements with multisection CT influence of inspiration level, nodule size, and segmentation performance. *Radiology* 245:888–894
- Wormanns D, Kohl G, Klotz E et al (2004) Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. *Eur Radiol* 14:86–92
- MacMahon H, Austin JHM, Gamsu G et al (2005) Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner society. *Radiology* 237:395–400
- Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1:307–310
- Kostis WJ, Yankelevitz DF, Reeves AP et al (2004) Small pulmonary nodules: reproducibility of three-dimensional volumetric measurement and estimation of time to follow-up CT. *Radiology* 231:446–452
- Goodman LR, Gulsun M, Washington L et al (2006) Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. *Am J Roentgenol* 186:989–994
- Ravenel JG, Leue WM, Nietert PJ et al (2008) Pulmonary nodule volume: effects of reconstruction parameters on automated measurements—a phantom study. *Radiology* 247:400–408
- Das M, Ley-Zaporozhan J, Gietema HA et al (2007) Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners. *Eur Radiol* 17:1979–1984
- Das M, Muhlenbruch G, Katoh M et al (2007) Automated volumetry of solid pulmonary nodules in a phantom: accuracy across different CT scanner technologies. *Invest Radiol* 42:297–302
- Ko JP, Rusinek H, Jacobs EL et al (2003) Small pulmonary nodules: volume measurement at chest CT—phantom study. *Radiology* 228:864–870
- Yankelevitz DF, Gupta R, Zhao B, Henschke CI (1999) Small pulmonary nodules: evaluation with repeat CT—Preliminary experience. *Radiology* 212:561–566
- Goo JM, Tongdee T, Tongdee R et al (2005) Volumetric measurement of synthetic lung nodules with multi-detector row CT: effect of various image reconstruction parameters and segmentation thresholds on measurement accuracy. *Radiology* 235:850–856
- Larici AR, Storto ML, Torge M et al (2008) Automated volumetry of pulmonary nodules on multidetector CT: influence of slice thickness, reconstruction algorithm and tube current. Preliminary results. *Radiol Med (Torino)* 113:29–42