No Benefit for Consensus Double Reading at Baseline Screening for Lung Cancer with the Use of Semiautomated Volumetry Software

Wang, Y.; van Klaveren, R.J.; de Bock, G.H.; Zhao, Y.; Vernhout, R.; Leusveld, A.; Scholten, E.; Verschakelen, J.; Mali, W.; de Koning, H.

Published in:
Radiology

DOI:
10.1148/radiol.11102289

No Benefit for Consensus Double Reading at Baseline Screening for Lung Cancer with the Use of Semiautomated Volumetry Software[1]

Ying Wang, MD[2]
Rob J. van Klaveren, MD, PhD
Geertruida H. de Bock, MD, PhD
Yingru Zhao, MD
René Vernhout, MD
Anne Leusveld, MD
Ernst Scholten, MD
Johny Verschakelen, MD, PhD
Willem Mali, MD, PhD
Harry de Koning, MD, PhD
Matthijs Oudkerk, MD, PhD

Purpose: To retrospectively evaluate the performance of consensus double reading compared with single reading at baseline screening of a lung cancer computed tomography (CT) screening trial.

Materials and Methods: The study was approved by the Dutch Minister of Health and ethical committees. Written informed consent was obtained from all participants. The benefit of consensus double reading was expressed by the percentage change in cancer detection rate, recall rate, number of additional nodules detected, and change in sensitivity and specificity in 7557 participants. The reference standard was a retrospective analysis of the serial CT scans performed in participants diagnosed with lung cancer during a 2-year period after baseline. Semiautomated volumetric software was used for nodule evaluation. McNemar tests were performed to test statistical significance. In addition, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated and 95% confidence intervals (CIs) constructed.

Results: Seventy-four cases of lung cancer were qualified as detectable at baseline. Compared with single reading, consensus double reading did not increase the cancer detection rate (2.7%; 95% CI: −1.0%, 6.4%; $P = .50$) or change the recall rate (20.6% vs 20.8%, $P = .28$), but led to the detection of 19.0% (1635 of 8623; 95% CI: 18.0%, 19.9%, $P < .01$) more nodules. The sensitivity, specificity, PPV, and NPV were 95.9% (71 of 74), 80.2% (6001 of 7483), 4.6% (71 of 1553) and 99.9% (6001 of 6004) for single reading and 98.6% (73 of 74), 80.0% (1497 of 7483), 4.6% (73 of 1570), and 99.9% (5986 of 5987) for consensus double reading, respectively.

Conclusion: There is no statistically significant benefit for consensus double reading at baseline screening for lung cancer with the use of a nodule management strategy based solely on semiautomated volumetry.

[1] From the Departments of Radiology (Y.W., Y.Z., A.L., M.O.) and Epidemiology (G.H.d.B.), University Medical Center Groningen, Hanzeplein 1, 9700 RB Groningen, the Netherlands; Departments of Pulmonology (R.J.v.K., R.V.) and Public Health (H.d.K.), Erasmus Medical Center, Rotterdam, the Netherlands; Department of Radiology, Kennemer Gasthuis, Haarlem, the Netherlands (E.S.); Department of Radiology, University Hospital Gasthuisberg, Leuven, Belgium (J.V.); and Department of Radiology, University of Utrecht Medical Center, Utrecht, the Netherlands (W.M.). Received November 19, 2010; revision requested January 10, 2011; revision received May 23; accepted June 21; final version accepted August 23. **Address correspondence to** M.O. (e-mail: *m.oudkerk@rad.umcg.nl*).

[2] **Current address:** Department of Radiology, Tianjin Medical University General Hospital, Tianjin, China.

Achieving a maximum diagnostic yield in cancer screening programs is dependent not only on the image quality but also on the appropriate reading of the images. Studies in breast cancer screening have shown that radiologists can sometimes differ substantially in the interpretations of mammograms and their recommendations for further management (1–3). Efforts to improve accuracy and to reduce variability in the interpretation may potentially increase the effectiveness of a screening program.

Several studies have investigated the benefit of double reading in breast cancer screening and have shown that double reading increased the cancer detection rate by 6%–15% compared with single reading (4–12). Taking the costs of double reading into account, double reading also appeared to be more cost effective than a single reading policy (13,14). Although double reading is recommended for breast cancer screening today, there is inconsistency in the data reported so far, as some investigators found an increase in the cancer detection rate of only 2%–5% after double reading (15–17).

Interobserver variability in reader performance for the detection and characterization of pulmonary nodules has been found to be relatively high with chest computed tomographic (CT) imaging both in the clinical setting and in lung cancer screening (18–21). So far, the effect of double reading on the cancer detection in lung cancer CT screening has, to our knowledge, not been investigated. The purpose of this study was, therefore, to retrospectively evaluate the performance of consensus double reading compared with single reading during the baseline period of a low-dose CT lung cancer screening trial.

Advances in Knowledge

- No difference was found in the recall rate between single and consensus double reading (20.6% versus 20.8%; $P = .28$) at baseline lung cancer screening with the use of a nodule management strategy based on semiautomated volumetry.
- Consensus double reading led to the detection of 19.0% (1635 of 8623; $P < .0001$) more pulmonary nodules.

Materials and Methods

Study Group

The study was approved by the Dutch Minister of Health and the ethical committees of all four participating hospitals. Written informed consent was obtained from all participants. The original approval and informed consent for the screening study included the ability to use data for future research, including the current prospective "side study." This study is a side study of the Dutch-Belgian multicenter randomized controlled low-dose CT lung cancer screening trial (the NELSON trial [Nedelands-Leuvens Longkanker Screenings Onderzoek]) (23). Although we received four Leonardo workstations and Lung Care software from Siemens Germany for the NELSON trial, we had full control over the data and the results submitted for publication.

Participants had to be current or former smokers, with a history of more than 15 cigarettes per day for over 25 years or more than 10 cigarettes per day for over 30 years, and were between 50 and 75 years of age. Between April 2004 and December 2006, 7557 participants were included. Their mean age was 59 years ± 6 (standard deviation). Of these participants, 84% (6310 of 7557) were male (mean age, 59 years ± 6) and 16% (1247 of 7557) were female (mean age, 58 years ± 6). All 7553 participants who underwent baseline chest CT scanning were analyzed in this side study.

Implication for Patient Care

- With the use of a nodule management strategy based solely on semiautomated volumetry, there is no substantial benefit for consensus double reading at baseline screening for lung cancer.

Imaging Methods

All scans were performed with 16–detector row helical CT scanners (Sensation 16, Siemens Medical Systems, Forchheim, Germany; Mx8000 IDT or Brilliance 16P, Philips Medical Systems, Cleveland, Ohio) with the following parameters: 0.5 second tube rotation, 0.75 mm single section collimation, and 15 or 18 mm table feed per rotation (pitch = 1.3–1.5). A caudo-cranial scan direction without contrast material was used. Scans were performed from the level of the lung bases (posterior recesses) to the lung apex with the help of a scout view. Depending on the body weight (<50, 50–80, and >80 kg), the kilovolts-peak settings were 80–90, 120, and 140 kVp, respectively. The milliampere-second values were 20–30 mAs and were adjusted accordingly dependent on the machine used. This corresponds to an effective radiation dose of less than 1.6 mSv. We reconstructed axial images with 1.0 mm thickness at 0.7 mm increments. A standard soft-tissue reconstruction algorithm was used for reconstruction (Siemens: B30 filter; Philips: B filter). All images were reconstructed with a field of view large enough to cover the entire lung parenchyma.

Published online before print
10.1148/radiol.11102289 **Content codes:** CH CT

Radiology 2012; 262:320–326

Abbreviations:
CI = confidence interval
NPV = negative predictive value
PPV = positive predictive value
VDT = volume doubling time

Author contributions:
Guarantors of integrity of entire study, Y.W., R.J.v.K., M.O.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, Y.W., R.J.v.K., A.L.; clinical studies, Y.W., R.J.v.K., G.H.d.B., Y.Z., A.L., E.S., J.V.; experimental studies, Y.W., J.V., M.O.; statistical analysis, Y.W., R.J.v.K., G.H.d.B., R.V., M.O.; and manuscript editing, Y.W., R.J.v.K., G.H.d.B., R.V., A.L., W.M., M.O.

Potential conflicts of interest are listed at the end of this article.

Nodule Management Protocol

In our NELSON nodule management strategy, the probability that a nodule is lung cancer at baseline screening is based on the measured volume (size) of detected nodules (24,25). Based on semiautomated volumetry, the outcome of the screening test was positive if any noncalcified nodule on a CT scan had a solid component larger than 500 mm^3 (>9.8 mm in diameter) and indeterminate if the volume of the largest solid nodule or the solid component of a partial-solid nodule was between 50 and 500 mm^3 (4.6–9.8 mm in diameter) or more than 8 mm for nonsolid nodules. Otherwise, the test was negative and the participants were asked to undergo an annual follow-up. In participants with an indeterminate test result, a 3-month recall CT scan was performed to assess whether the nodule had grown. In case of growth and a volume doubling time (VDT) of less than 400 days, the outcome of the screening test for this group was positive; otherwise it was negative. Growth was defined as a percentage volume change of 25% or more between the first and second scan. All positive findings were referred to the chest physician of choice via the general practitioner, usually the chest physician associated with the screening center. For work-up and diagnosis, a uniform procedure was followed (23). If lung cancer was diagnosed, the participant was treated appropriately and the case was classified as test true-positive; otherwise the participant was scheduled for the second round CT scan. If baseline-detectable nodules were not diagnosed as cancers in a 2-year period, the original image was considered true-negative.

Reading Procedure

At all screening sites, digital workstations (Leonardo; Siemens Medical Solutions, Erlangen, Germany) were used for image interpretation with U.S. Food and Drug Administration–approved commercially available software for semiautomated volume measurements (LungCare, version Somaris/5: VA70C-W; Siemens Medical Solutions) (26,27). Pulmonary nodules were identified on axial thin-slab maximum intensity projections with cine mode. The section thickness of maximum intensity projections was set as 6 or 8 mm. Images were initially displayed with a window level of −500 HU and window width of 1500 HU, but the readers were free to alter these settings at their discretion.

First, the CT images were read at the four screening sites by one of the 13 local readers. Readers had 0 to more than 20 years (median, 6 years) of experience in the interpretation of thoracic CT images. The interpretations encompassed the evaluation of the nodule features (location, size, morphology) and the provision of the corresponding screening test result based on the highest nodule category. These data were uploaded into the online management system and were considered the results of single reading. Subsequently, the CT images were sent to the central site for second reading by one of the two central radiologists, both of whom had 6 years of experience in the interpretation of thoracic CT images. When second readers interpreted the CT images, they were not blinded to the results of first reading and could alter their interpretations. After their interpretations, the second readers evaluated the agreement between the first and second reading with regard to the screening test result. In case of a discrepancy, the second readers informed the first readers and they reevaluated the CT image to reach consensus. If no consensus was reached, arbitration from an expert radiologist with more than 20 years of experience was performed, and the interpretations of the second reader were changed according to the opinion of the consensus panel, which was regarded as the final result of consensus double reading.

Reference Standard

The reference standard was formed through retrospective analysis of the serial CT scans performed in 130 participants in whom lung cancer was diagnosed during a 2-year period after baseline screening. From the 130 lung cancer cases detected during 2 years of follow-up, 56 were excluded because they were not detectable at baseline screening (Figure). This included both the screening-detected and the interval lung cancer cases. Each lung cancer case was retrospectively matched to one particular pulmonary nodule detected at the baseline CT scan if possible. Cancer cases originating from baseline nodules larger than 500 mm^3 or baseline nodules with fast growth (VDT, <400 days) at 3-month follow-up were included in this study as ground truth because they could have been diagnosed during baseline screening according to our protocol (23). Exclusion of the other lung cancers cases from our evaluation did not influence the results of our study on the value of consensus double reading because they could not be diagnosed at either reading. An interval lung cancer was defined as a lung cancer detected during the 1-year period following a negative baseline scan or second round test result. These interval lung cancers were identified through linkage of the participants with the national pathology database and by active information collection from appointments, general practitioners, letters, and phone calls.

Statistical Analysis

At baseline screening of the NELSON lung cancer screening trial, the performance of consensus double reading was compared with single reading to investigate whether there exists an added value for consensus double reading. The primary outcome measure was the cancer detection rate and the proportion of early stage (stages I and II) disease detected. The secondary outcome measure was the recall rate, which was based on the highest nodule category detected. In addition to the overall recall difference, we also evaluated this in cancer and noncancer cases separately. Furthermore, the percentage of additional pulmonary nodules detected by consensus double reading was calculated as the number of nodules detected by both readers minus those detected by the first reader only, divided by the total number of nodules detected by both readers. Finally, the diagnostic accuracy of single and consensus double reading were expressed as sensitivity,

specificity, positive predictive value (PPV), and negative predictive value (NPV), with their 95% confidence intervals (CIs). A 3-month recall CT scan and referral to chest physicians were both regarded as a positive finding in this analysis. The McNemar test was used to compare the differences, taking the matched nature of the data into account (28). $P \leq .05$ was considered to indicate a statistically significant difference. All analyses were performed with statistical software (SPSS, version 16.0; SPSS, Chicago, Ill).

Results

Among 7557 participants who underwent baseline screening, lung cancer was diagnosed in 130 after 2 years of follow-up after the baseline CT scan. After retrospectively review, 74 participants with lung cancer could be included in this study: Diagnosis was made in 53 participants after a positive finding at baseline screening, in 17 after a positive test result following a 3-month recall CT scan, and in three at 1 year after baseline screening as a result of a negative work-up following a positive baseline test result, and one participant had an interval lung cancer (Figure).

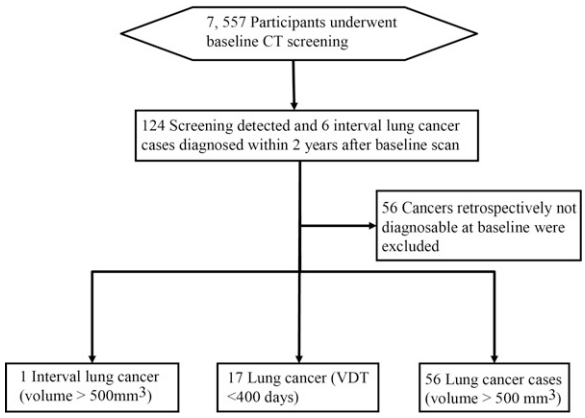
Single reading enabled the detection of 95.9% (71 of 74) and consensus double reading 98.6% (73 of 74) of all cancer cases. The mean difference was 2.7% (95% CI: −1.0%, 6.4%; $P = .50$) (Table 1). The agreement between two readings was 97.3% (72 of 74; 95% CI: 90.1%, 99.8%). Second reading provided 0.3 (two of 7553) additional screening-detected cancer cases per 1000 participants. In total, 46 of 74 lung cancer cases were detected at an early stage. Consensus double reading led to the detection of one additional early stage case. The early stage detection rate increased from 60.8% (45 of 74) to 62.2% (46 of 74), with a mean difference of 1.3% (one of 74; 95% CI: −1.3%, 4%; $P > .99$).

The overall recall rates are 20.6% (1553 of 7553) and 20.8% (1470 of 7553) for single and consensus double reading, respectively. The mean difference was 0.2% (83 of 7553; 95% CI: 0%–0.2%, 0.6%; $P = .28$). The final screening results changed in 3.1% (233 of 7553) of all participants. The agreement between two readings was 96.9% (7320 of 7553; 95% CI: 96.5%, 97.3%).

Of the total 8623 pulmonary nodules detected at baseline screening, 19.0% (1635 of 8623; 95% CI: 18.0%, 19.9%; $P < .01$) were additionally detected by means of consensus double reading (Table 2). These nodules included both true- and false-positive findings.

Among the 74 cancer cases, there were four cases with disagreement: Consensus double reading upgraded the results of the screening test in three (4.2%) cases and downgraded the outcome in one (1.4%) case; in two cases this was caused by the additional detection of the cancer lesion by consensus double reading, and in the other two cases this was caused by disagreement on the measured volume of malignant nodule. In 7483 noncancer cases, consensus double reading upgraded the results of the screening test in 121 cases and downgraded the result in 108. The mean difference was 0.2% (13 of 7483; 95% CI: −0.2%, 0.6%; $P = .43$) (Table 3).

For single reading, the sensitivity, specificity, PPV, and NPV were, respectively, 95.9% (71 of 74; 95% CI: 87.8%, 98.9%), 80.2% (6001 of 7483; 95% CI: 79.3%, 81.1%), 4.6% (71 of 1553; 95% CI: 3.6%, 5.8%), 99.9% (6001 of 6004; 95% CI: 99.8%, 100%). For consensus double reading, the sensitivity, specificity, PPV, and NPV were, respectively, 98.6% (73 of 74; 95% CI: 91.7%, 99.9%), 80.0% (1497 of 7483; 95% CI: 79.1%, 80.9%), 4.6% (73 of 1570; 95% CI: 3.7%, 5.8%), and 99.9% (5986 of 5987, 95% CI: 99.9%, 100%) (Table 4).



Flowchart for the selection of cancer cases.

Table 1

Change in Lung Cancer Detection Rate by Single and Consensus Double Reading at Baseline CT Screening

	Consensus Double Reading		
Single Reading	Negative Finding	Positive Finding	Total
Negative finding	1	2	3 (4.1)
Positive finding	0	71	71 (95.9)
Total	1 (1.4)	73 (98.6)	74

Note.—Data are numbers of participants, and numbers in parentheses are percentages. The difference in lung cancer detection rate was 2.7% (two of 74; 95% CI: −1.0%, 6.4%).

Discussion

At baseline screening of the NELSON lung cancer screening trial, we observed

Table 2

Additional Pulmonary Nodules Detected by Consensus Double Reading at Baseline CT Screening

Nodule Category/Volume	Single Reading	Additional Nodules Detected by Double Reading	Total
Benign nodules and nodules <50 mm^3 (<4.6 mm)	5090 (72.7)	1166 (71.3)	6256 (72.6)
50–500 mm^3 (4.6–9.8 mm)	1796 (25.7)	440 (26.9)	2236 (25.9)
>500 mm^3 (>9.8 mm)	102 (1.5)	29 (1.8)	131 (1.5)
Total	6988	1635	8623

Note.—Data are numbers of nodules, and numbers in parentheses are percentages.

Table 3

Baseline CT Screening Results of Single and Consensus Double Reading in 7557 Participants

	Consensus Double Reading			
Single Reading	Annual Follow-up	3-month Recall CT scan	Referral to Chest Physician	Total
Lung cancer cases				
Annual follow-up	1 (1.4)	1 (1.4)	1 (1.4)	3 (4.1)
3-month recall CT	0	15 (20.3)	1 (1.4)	16 (21.6)
Referral to chest physician	0	1 (1.4)	54 (73.0)	55 (70.4)
Total	1 (1.4)	17 (23.0)	56 (75.7)	74 (100)
Non–lung cancer cases				
Annual follow-up	5887 (78.7)	114 (1.5)	0	6001 (80.2)
3-month recall CT	97 (1.3)	1311 (17.5)	7 (0.1)	1415 (19.0)
Referral to chest physician	2 (0.01)	9 (0.1)	56 (0.7)	67 (0.9)
Total	5986 (80.0)	1434 (19.1)	63 (0.8)	7483 (100)

Note.—Data in parentheses are percentages.

that consensus double reading led, compared with single reading, to the detection of 2.7% (two of 74) more subjects with lung cancer, 1.3% (one of 74) more cases of early stage disease, an 0.2% (83 of 7553) increase in the recall rate, 19% (1635 of 8623) more nodules and a 2.7% (two of 74) increase in sensitivity and a 0.2% (15 of 7483) decrease in specificity.

An important parameter for the effectiveness of double reading in a cancer screening program is the degree by which the cancer detection rate increases (29). A non–statistically significant increase in cancer detection rate (2.7% [two of 74]) and early stage cancer detection rate (1.3% [one of 74]) by consensus double reading was observed in our study; however, insufficient power led to our conclusion that there is no evidence that the performances of the two readings were different. A possible explanation for the lack of the statistical significance is that the power of our study was not strong enough due to the relatively small number of cancer cases. We believe, however, that even with a larger number of cases the added value of consensus double reading in lung cancer screening would still be limited, because the performance of single reading in our screening trial was quite good and left little space for improvement for the double reading. It could also be hypothesized that during consensus double reading, first readers might not perform as well because they are aware of the fact that there will be another interpretation, leading to carelessness and eventually to a lower accuracy. Similarly, second readers may have become careless and simply agreed with the first reader's interpretation, since they were not blinded to the first reader's conclusions, leading to a lack of change in the cancer detection rate. However, this phenomenon was not observed in our study since only one and three cancers were missed at consensus double reading and single reading, respectively. This can be explained by the fact that multidetector CT performed at low-dose level provides high spatial resolution and attenuation and the use of maximum intensity projections reconstruction algorithms with cine mode further facilitates the identification of abnormalities (22). Second, our nodule management strategy is an objective, software-driven approach (23,25), in which the recall is determined only by the volume and the VDT of the nodules detected without further subjective interpretation. As reported before, semiautomated software volume measurement of pulmonary nodules is highly repeatable (30). This also explains why the 2.7% increase in our study is much lower than the 6%–15% observed for breast cancer screening (4–12). In breast cancer screening, lesion identification is relatively difficult due to the small difference in density between the lesion and the surrounding normal breast tissue at mammography. As a result, classification of the detected nodules is rather subjective according to the Breast Imaging Reporting and Data System (31).

Another important parameter for the value of double reading is the recall rate in noncancer cases. In noncancer cases, consensus double reading upgraded the outcome of the screening test in 121 (1.6%) cases and downgraded the outcome in 108 (1.4%). This led to a non–statistically significant 0.2% (13 of 7483) increase in the recall rate, equivalent to a 0.2% reduction in the specificity. Our interpretation is that consensus double reading enhanced adherence to the NELSON nodule management strategy without changing the

Table 4

Diagnostic Accuracy of Single and Consensus Double Reading during the Baseline Period of Low-Dose CT Lung Cancer Screening Trial

A: Accuracy of Single Reading			
	Lung Cancer		
Finding	Yes	No	Total
Positive	71	1482	1553
Negative	3	6001	6004
Total	74	7483	7557
B: Accuracy of Consensus Double Reading			
Finding	Lung Cancer		
	Yes	No	Total
Positive	73	1497	1570
Negative	1	5986	5987
Total	74	7483	7557

Note.—Data are numbers of participants.

recall rate. This implies that there is no difference in the specificity of the two reading strategies, which is in accordance with findings in breast cancer screening, in which no change in specificity was observed as a result of consensus double reading (6,9).

Furthermore, we observed that 19.0% (1635 of 8623) more pulmonary nodules were detected by means of consensus double reading. This observation is in line with previous studies. Gruden et al (22) explored the interreader variability in a lung cancer CT screening project and showed that the difference between readers could have occurred in lesion detection, characterization of a lesion as a nodule or nonnodule, or lesion measurement. The interobserver agreement was moderate to substantial, and potential for considerable improvement existed. Similar results have been reported in clinical settings with a relatively high interobserver variability for the detection and characterization of pulmonary nodules (16–18). Theoretically, the more nodules identified, the higher the probability to detect cancer. However, the 19.0% increase in the nodule detection rate observed in our study only lead to a 3.1% (233 of 7553) change in the outcome of the screening test (positive, indeterminate, or negative) and a 2.7% (two of 74) increase in the cancer detection rate. The reason for this discordance could be explained by the fact that first readers pay more attention to the larger, suspicious nodules and potentially neglect the smaller ones. In our nodule management protocol, the test result was based on the highest nodule category. Therefore, the detection of additional nodules in the vast majority of the participants did not change the test result and cancer detection.

A limitation of our study was that the power of our analysis on the value of consensus double reading is not strong enough because the study was powered to assess whether CT screening for lung cancer will lead to a decrease in lung cancer mortality. For that, we further performed an ad hoc power analysis based on the data derived from current study. Given a power of 90%, a significance level of 5%, an expected proportion of upgrading recall for cancer cases of 4.1% (three of 74), and a downgrading proportion of 1.4% (one of 74), a sample of 789 double readings of baseline cancers should be needed to make the contribution of a second reading a statistically significant (two-sided McNemar test). On the basis of the 0.9% cancer detection rate in our study, 80531 participants need to be enrolled; a study of that size will never be conducted. As this ad hoc analysis is only based on four discordant cases, the number of participants needed to demonstrate a statistically significant contribution from a second reading will vary widely, but will at least be much larger than the 7557 participants used in this study. Furthermore, even if this significance could be achieved, the human cost of one extra detected early stage–lung cancer corresponds to 7557 second readings, which is equivalent to 253 working days for a radiologist with a throughput of 30 readings per day. Therefore, after weighing the advantages and the cost of consensus doubling reading, we do not recommend consensus double reading in lung cancer screening with the use of our nodule management strategy based on semiautomated volumetry.

A further limitation of our study was that we only included baseline-diagnosable lung cancer, which is either a lesion larger than 500 mm^3 or a fast-growing nodule in our study with a VDT of less than 400 days at a 3-month recall CT scan. Lung cancer cases originating from new lesions at later rounds or malignancies that have been diagnosed in nodules with a VDT of more than 400 days were excluded. Therefore, the result of this study was only applicable for baseline screening with the use of the NELSON nodule management strategy. With the use of different nodule management strategies, interobserver variability and the range of baseline-diagnosable lung cancer cases might be different and lead to a different conclusions with regard to the value of second reading in CT screening for lung cancer.

In conclusion, there is no statistically significant benefit for consensus double reading at baseline screening for lung cancer with the use of nodule management strategy based solely on semiautomated volumetry.

Disclosures of Potential Conflicts of Interest: **Y.W.** No potential conflicts of interest to disclose. **R.J.v.K.** No potential conflicts of interest to disclose. **G.H.d.B.** No potential conflicts of interest to disclose. **Y.Z.** No potential conflicts of interest to disclose. **R.V.** No potential conflicts of interest to disclose. **A.L.** No potential conflicts of interest to disclose. **E.S.** No potential conflicts of interest to disclose. **J.V.** No potential conflicts of interest to disclose. **W.M.** No potential conflicts of interest to disclose. **H.d.K.** No potential conflicts of interest to disclose. **M.O.** No potential conflicts of interest to disclose.

References

1. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331(22):1493–1499.
2. Taplin SH, Rutter CM, Elmore JG, Seger D, White D, Brenner RJ. Accuracy of screening mammography using single versus independent double interpretation. AJR Am J Roentgenol 2000;174(5):1257–1262.
3. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224(3):861–869.
4. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. Clin Radiol 1994;49(4):248–251.
5. Ciatto S, Ambrogetti D, Bonardi R, et al. Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. J Med Screen 2005;12(2):103–106.
6. Warren RM, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Br J Radiol 1995;68(813):958–962.
7. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191(1):241–244.
8. Deans HE, Everington D, Cordiner C, Kirkpatrick AE, Lindsay E. Scottish experience of double reading in the National Breast Screening Programme. Breast 1998;7(2):75–79.
9. Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films—one radiologist or two? Clin Radiol 1993;48(6):414–421.
10. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol 2003;180(5):1461–1467.
11. Shaw CM, Flanagan FL, Fenlon HM, McNicholas MM. Consensus review of discordant findings maximizes cancer detection rate in double-reader screening mammography: Irish National Breast Screening Program experience. Radiology 2009;250(2):354–362.
12. Hofvind S, Geller BM, Rosenberg RD, Skaane P. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology 2009;253(3):652–660.
13. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ 1996;312(7034):809–812.
14. Leivo T, Salminen T, Sintonen H, et al. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat 1999;54(3):261–267.
15. Ciatto S, Del Turco MR, Morrone D, et al. Independent double reading of screening mammograms. J Med Screen 1995;2(2):99–101.
16. Denton ER, Field S. Just how valuable is double reporting in screening mammography? Clin Radiol 1997;52(6):466–468.
17. Matson M, Hibbert J, Field S. Double reporting of screening mammograms. Clin Radiol 1997;52(7):567.
18. Wormanns D, Diederich S, Lentschig MG, Winter F, Heindel W. Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. Eur Radiol 2000;10(5):710–713.
19. Burns J, Haramati LB, Whitney K, Zelefsky MN. Consistency of reporting basic characteristics of lung nodules and masses on computed tomography. Acad Radiol 2004;11(2):233–237.
20. Leader JK, Warfel TE, Fuhrman CR, et al. Pulmonary nodule detection with low-dose CT of the lung: agreement among radiologists. AJR Am J Roentgenol 2005;185(4):973–978.
21. Gierada DS, Pilgram TK, Ford M, et al. Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. Radiology 2008;246(1):265–272.
22. Gruden JF, Ouanounou S, Tigges S, Norris SD, Klausner TS. Incremental benefit of maximum-intensity-projection images on observer detection of small pulmonary nodules revealed by multidetector CT. AJR Am J Roentgenol 2002;179(1):149–157.
23. Xu DM, Gietema H, de Koning H, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 2006;54(2):177–184.
24. Henschke CI, Yankelevitz DF, Naidich DP, et al. CT screening for lung cancer: suspiciousness of nodules according to size on baseline scans. Radiology 2004;231(1):164–168.
25. van Klaveren RJ, Oudkerk M, Prokop M, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009;361(23):2221–2229.
26. Bolte H, Riedel C, Müller-Hülsbeck S, et al. Precision of computer-aided volumetry of artificial small solid pulmonary nodules in ex vivo porcine lungs. Br J Radiol 2007;80(954):414–421.
27. Wormanns D, Kohl G, Klotz E, et al. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. Eur Radiol 2004;14(1):86–92.
28. McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–157.
29. Williams LJ, Hartswood M, Prescott RJ. Methodological issues in mammography double reading studies. J Med Screen 1998;5(4):202–206.
30. Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, et al. Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. Radiology 2008;248(2):625–631.
31. American College of Radiology. Breast imaging reporting and data system (BI-RADS). 3rd ed. Reston, Va: American College of Radiology, 1998.