Medical Image Analysis 14 (2010) 707–722

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study

Bram van Ginneken [a,b,*], Samuel G. Armato III [c], Bartjan de Hoop [d], Saskia van Amelsvoort-van de Vorst [d], Thomas Duindam [a], Meindert Niemeijer [a], Keelin Murphy [a], Arnold Schilham [a], Alessandra Retico [e], Maria Evelina Fantacci [e,f], Niccolò Camarlinghi [e,f], Francesco Bagagli [e,f], Ilaria Gori [e,g], Takeshi Hara [h], Hiroshi Fujita [h], Gianfranco Gargano [i,j], Roberto Bellotti [i,j], Sabina Tangaro [j], Lourdes Bolaños [k,l], Francesco De Carlo [j], Piergiorgio Cerello [k], Sorin Cristian Cheran [k], Ernesto Lopez Torres [l], Mathias Prokop [d,b]

[a] *Image Sciences Institute, University Medical Center Utrecht, The Netherlands*
[b] *Department of Radiology, Radboud University Nijmegen Medical Centre, The Netherlands*
[c] *Department of Radiology, University of Chicago, USA*
[d] *Department of Radiology, University Medical Center Utrecht, The Netherlands*
[e] *Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Pisa, Italy*
[f] *Dipartimento di Fisica dell'Università di Pisa, Pisa, Italy*
[g] *Bracco Imaging S.p.A., Milano, Italy*
[h] *Department of Intelligent Image Information, Gifu University Graduate School of Medicine, Gifu, Japan*
[i] *Dipartimento Interateneo 'M. Merlin' dell'Univerisità degli Studi di Bari, Italy*
[j] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy*
[k] *Istituto Nazionale di Fisica Nucleare, Sezione di Torino, Italy*
[l] *Caeden, Cuba*

ARTICLE INFO

Article history:
Received 14 August 2009
Received in revised form 14 May 2010
Accepted 25 May 2010
Available online 4 June 2010

Keywords:
Computer-aided detection
Computed tomography
Lung nodules
Lung cancer

ABSTRACT

Numerous publications and commercial systems are available that deal with automatic detection of pulmonary nodules in thoracic computed tomography scans, but a comparative study where many systems are applied to the same data set has not yet been performed. This paper introduces ANODE09 (http://anode09.isi.uu.nl), a database of 55 scans from a lung cancer screening program and a web-based framework for objective evaluation of nodule detection algorithms. Any team can upload results to facilitate benchmarking. The performance of six algorithms for which results are available are compared; five from academic groups and one commercially available system. A method to combine the output of multiple systems is proposed. Results show a substantial performance difference between algorithms, and demonstrate that combining the output of algorithms leads to marked performance improvements.

1. Introduction

Computer-aided detection (CAD) has become one of the most active research areas within medical image analysis. The detection of pulmonary nodules from volumetric computed tomography (CT) scans is one of the most studied CAD applications (Sluimer et al., 2006). There are several reasons for the interest in this task. First, lung cancer is the most deadly cancer and early detection may be the most promising strategy to reduce lung cancer mortality. With CT, small lung nodules can be identified. If these nodules are malignant, they usually represent early stage lung cancer and with surgical intervention there is a high chance of long-term survival of the patient (MacMahon et al., 2005). Second, the advent of multi-detector row CT scanners with fast gantry rotation times has made it possible to scan the entire chest in a few seconds, well within a single breath-hold. This generates high quality scans with isotropic voxels around 0.35 mm^3, that can be obtained with a low-dose and high patient throughput. As a result, there are currently many ongoing trials that investigate the efficacy and effectiveness of lung cancer screening with low-dose CT among high risk individuals (Infante et al., 2009; van den Bergh et al., 2008; Gohagan et al., 2004; Henschke, 2007). In addition to screening, chest CT exams are being used more often for a wide range of diagnostic tasks. It is always important to report findings of nodules in these scans, and this can be a cumbersome, time-consuming task because the scans contain 300–500 slices. It appears best to use dedicated visualization settings (sliding maximum intensity projections of

* Corresponding author at: Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, The Netherlands. Tel.: + 31 30 250 4635; fax: +31 30 251 3399.
E-mail address: bram@isi.uu.nl (B. van Ginneken).

1361-8415/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2010.05.005

around 10 mm) for optimal detection performance (Gruden et al., 2002), but such settings may not be optimal for the detection of other abnormalities. CAD of nodules may therefore become a practical necessity for time-efficient interpretation of chest CT scans.

Although at first glance the detection of nodules may seem a fairly straightforward task, it turns out that nodules come in many forms. There are nodules that are easy to detect, for example a round, well-marginated, solid nodule of 4–10 mm in diameter, located in the periphery of the lung. But much smaller nodules are also visible on CT, and there are nodules with complex vascular attachments located in regions with large vessels, and part-solid and non-solid nodules with densities only slightly above those of the surrounding lung parenchyma. If a scan contains abnormalities, the lungs may contain many lesions that are somewhat nodular in appearance, but unrelated to lung cancer. It is therefore difficult, if not meaningless, to compare the performance of two nodule CAD systems that have been tested on different databases. Another reason why it is hard to compare results obtained on different databases is that many studies, especially older ones, have used scans with thick sections, in the range of 2.5–10 mm. For the detection of nodules, using scans with such thick sections is not recommended because they introduce a partial-volume effect for smaller nodules and prevent the use of visualization techniques such as sliding maximum intensity projections and volume rendering which improve nodule detectability (Peloschek et al., 2007).

A large number of systems for nodule detection have been proposed in the literature (Li et al., 2008; Arimura et al., 2004; Armato et al., 2001, 2002; Bae et al., 2005; Bellotti et al., 2007; Brown et al., 2003; Dehmeshki et al., 2007; Enquobahrie et al., 2007; Farag et al., 2004; Ge et al., 2005; Ko and Betke, 2001; Kung et al., 2004; Lee et al., 2001; Matsumoto et al., 2006; McCulloch et al., 2004; Mendonça et al., 2007; Murphy et al., 2007; Novak et al., 2004; Osman et al., 2007; Paik et al., 2004; Retico et al., 2008; Suzuki et al., 2003; Wei et al., 2002; Wiemker et al., 2005, 2002; Ye et al., 2007; Zhang et al., 2007; Zhao et al., 2003; Fotin et al., 2009). In addition, several commercial systems for nodule detection are available and many workstations that radiologists routinely use to interpret CT scans provide on-board nodule detection capabilities. The reported performance of systems varies tremendously. In a recent literature survey (Li, 2007), a comparison of nine systems yielded sensitivities from 70% to 90% with a range of 0.5 to 15 false positive detections per scan. Even when the same CAD system is evaluated, results can vary substantially. A study from 2005 (Lee et al., 2005) measured performance of the ImageChecker CT LN-1000, developed by R2 Technology (Sunnyvale, CA). This technology has been acquired by Mevis (Pewaukee, WI) and recently released as Visia CT Lung. The system was applied to 70 scans with 78 nodules. CAD detected 47 (60%) of these and produced 1.56 false-positive nodules per scan. In another study (Das et al., 2006), the ImageChecker CT (no version number was given) obtained 73% sensitivity and six false positives (FPs) per scan. Finally, in a recent study (Godoy et al., 2008) the results for ImageChecker CT V2.0, in a study partly funded by R2 Technology, achieved a sensitivity of 87.7% for lung cancer nodules with a diameter of 4 mm and larger with either solid or semi-solid morphology, at a false positive rate of 0.9 per scan.

A major step forward to more objective measurement of CAD performance is the creation of a publicly available database by the Lung Image Database Consortium (LIDC) (Armato et al., 2004). Annotated chest CT scans are available on-line at http://imaging.nci.nih.gov/ncia/. So far, this collection contains 400 scans. One of the LIDC groups also made data available at http://www.via.cornell.edu/databases/lungdb.html. Because the data and annotations are freely available, companies and research groups may report their results on different subsets of the databases and will almost certainly perform evaluation in different ways, making the results again difficult to compare.

The purpose of this paper is to present a new database of state-of-the-art CT scans from a lung cancer screening trial, and a framework for the evaluation of CAD algorithms applied to this data set. To alleviate the problem that observers tend to show substantial disagreement on what constitutes a nodule (Armato et al., 2009), we introduce the concept of relevant and irrelevant findings. Irrelevant findings are nodules that are unlikely to be cancer, such as calcified nodules or very small nodules. These irrelevant findings have been marked in the database and if a CAD system detects such a lesion, the output of the system is ignored (i.e., not counted as either a true positive or a false positive). Evaluation is performed using free-response receiver operating characteristic (FROC) analysis and the results are computed automatically after a list of the coordinates of findings, along with a degree of suspicion generated by the CAD system is submitted to the ANODE09 web site (http://anode09.isi.uu.nl). This ensures that every system is evaluated in exactly the same way, using the same software, and that the results are directly comparable. The only factors affecting differences in results then would be the CAD system, not the data or the details of the evaluation procedure. This paper describes the database and the evaluation procedure in detail in Sections 2 and 3. In Section 4 six systems whose findings have been submitted are described and their results are given in Section 5. They include recent and older CAD systems developed by academic groups and one commercial system.

The other major contribution of this paper is a generic method to combine the output of multiple CAD systems, outlined in Section 4.7. This is perhaps an even more compelling reason to have organized the ANODE09 study. There is in fact no reason to assume that a single CAD scheme would be optimal for nodule detection. It is more likely that different methods have complementary strengths, and the availability of multiple system's outputs on a single database allows us to test this in practice. It will be shown that combining CAD systems can substantially improve the overall performance. The implications and limitations of this study are discussed in Section 6 and we draw conclusions in Section 7.

2. Data

The ANODE09 data set consists of 55 anonymized CT scans. Five scans are examples and are made available with radiologist annotations. These scans are not used in the evaluation of algorithms and can be used for training CAD algorithms or optimizing their internal settings, if desired. The remaining 50 scans are for testing. The reference annotations for those 50 scans are not publicly available.

All data has been provided by the University Medical Center Utrecht and originates from the NELSON study, the largest CT lung cancer screening trial in Europe. Current and former heavy smokers, mainly men, aged 50–75 years are included in this study. Scans were acquired on a 16 or 64-slice CT scanner (Philips Medical Systems, Cleveland, OH) using a spiral mode with 16×0.75 mm or 64×0.75 mm collimation. The entire chest was scanned in 4–10 s using a caudo-cranial scan direction to minimize breathing artifacts. Scans were performed in full inspiration, without spirometric control. Exposure settings were low-dose: 30 mAs and 120 kVp (volume CT dose index, $CTDI_{vol} = 2.2$ mGy) for patients weighing less than 80 kg, and 30 mAs at 140 kVp for those weighing over 80 kg ($CTDI_{vol} = 3.5$ mGy). Axial images with a 512×512 matrix were reconstructed at 1.0 mm thickness and 0.7 mm increment, using a moderately soft reconstruction kernel (Philips B) and the smallest field of view that included the outer rib margins at the widest dimension of the thorax. As a result of this scanning procedure, where the field of view is adjusted to patient size, the resolution in the x and y-direction varied from 0.59 to 0.83 mm with an

average of 0.71. The data is therefore nearly isotropic. More information about the acquisition process and the screening study from which the data originates is available elsewhere (Xu et al., 2006; van Iersel et al., 2006).

The large majority of the ANODE09 scans were randomly selected from the entire Utrecht database of the NELSON screening program. A small number of scans were randomly picked from the 1% of scans in the entire database which contained the largest number of annotations. Scans that contained evident interstitial lung disease, which can lead to the presence of hundreds of usually small nodular findings, were excluded. The reason for adding some scans with more than the average number of findings is that we aimed to have a reasonable number of nodules in a test set that was not too large, to make web-based distribution of the data feasible.

It should be noted that the ANODE09 data set contains relatively few larger nodules, especially compared to other databases on which results for nodule CAD systems have been reported in the literature. We did not, as was done for example in Fotin et al. (2009), specifically add cases with larger nodules. The ANODE09 set can be considered representative of findings among asymptomatic heavy smokers.

3. Annotation and evaluation

3.1. Annotation process and irrelevant findings

In the NELSON study, nodules – defined as a round opacity, at least moderately well marginated and no greater than 3 cm in maximum diameter (Austin et al., 1996) – were divided into four groups (Xu et al., 2006). Class 1 contained nodules with fat, benign calcifications or other benign characteristics. The other groups contained nodules without benign characteristics. Class 2 nodules had a volume below 50 mm^3. All volume measurements were done in 3D on Siemens workstations using the Syngo Lungcare software package (Somaris/5 VB 10A-W). If 3D segmentation failed, a diameter was drawn on an axial section. Class 3 contained solid, part-solid or non-solid nodules with a volume between 50 and 500 mm^3. The equivalent diameters[1] are 4.57 mm and 9.84 mm, respectively. Larger nodules fell into class 4 and participants with such a nodule were referred to a pulmonologist for work-up and diagnosis. Participants with a class 3 nodule were invited for a 3 month repeat scan. Finding a nodule in class 2 did not change the follow-up, and there was no lower size limit for class 2 nodules. Therefore not all of such small nodules may have been annotated. Scans were read by an experienced observer and by a second experienced observer in an unblinded fashion.

From our experience in the NELSON study we have learnt that it is not easy to distinguish nodules from findings that mimic a nodule. This is supported by the literature, (e.g. Lee et al., 2005). Most of these findings are scars, but other examples are vessels with a local outpouching and pleural plaques. The LIDC study (Armato et al., 2004; Armato et al., 2007; Armato et al., 2009) has made explicit the variation among radiologists in the identification of lung nodules. In the LIDC study four observers indicated nodules in 90 chest CT scans in a two step process, first blinded, next unblinded, so that they could see the results of the three other readers. It was found (Ochs et al., 2007) that for nodules $\geqslant$3 mm, there were 174 nodules where at least 1 of 4 observers said it was a nodule, for 146 of those at least 2 of 4 observers agreed, for 121 at least 3 agreed and for 90 all four agreed. These results indicate that there is a large number of nodules for which human expert observers agree, but an approximately equally large group of findings about which there is no consensus among observers. If a CAD system placed a marker on such a nodule, should it be considered a true positive or a false positive?

To partly circumvent this problem, we introduced a second category of findings in the ANODE09 study. We call this category 'irrelevant findings', as opposed to 'relevant' or actionable findings, i.e. the nodules that a CAD system definitively should detect. Any CAD marks in regions around irrelevant findings are ignored in the evaluation, as explained in Section 3.2. There are three types of irrelevant findings: findings that mimic a nodule but that an expert observer believes not to be a nodule, nodules with benign characteristics (class 1 in the NELSON protocol), and nodules that are too small to be relevant.

Almost all very small lung nodules are benign and are normal pulmonary lymph nodes or small granulomas (Henschke et al., 2004). Here one needs to use a threshold for volume or effective diameter. We decided to use the threshold of 4 mm effective diameter, because it is the one currently recommended by the Fleischner society (MacMahon et al., 2005) and many CAD systems use this threshold as well. This is a slightly smaller size than what is used in the NELSON study. In some scans with many nodules, some nodules were also listed as irrelevant (and thus ignored in the evaluation) although they did meet all the requirements of relevant nodules. This was done to prevent the results of a CAD algorithm on a few scans dominating the assessment of its performance.

The rationale for introducing irrelevant findings is that it is unfair or at least debatable to call a mark on such a finding a false positive. Accurate segmentation of nodules is an extraordinarily difficult task (de Hoop et al., 2009) and therefore in ANODE09 a mark on a nodule slightly below 4 mm in diameter according to our segmentation procedure will not count as an error. Similarly, a mark on a calcified nodule may be appreciated by some radiologists and should not count as an error. As it is difficult to distinguish scarring and other abnormalities from nodules that may represent lung cancer, it would be unfair to count a mark on such a lesion as wrong as an obvious false positive that is placed, for example, on a vessel bifurcation.

To implement this, two observers annotated in a blinded fashion all 55 ANODE09 scans using the NELSON annotations as a basis. The majority of relevant findings were already contained in the NELSON annotations. Findings that were not in the NELSON annotations were added, and all findings were labeled as relevant or irrelevant. One observer was a very experienced reader from the NELSON trial, the other one was a radiologist in training. A third observer, an experienced radiologist, resolved cases where the two observers disagreed. All findings were segmented with an in house implementation of an algorithm comparable to (Kostis et al., 2003), where the parameters were adjusted interactively by a human operator until a satisfactory segmentation was obtained. Findings below 4 mm were listed as irrelevant. There was no lower size limit specified, but in practice the smallest annotated irrelevant nodules have a diameter around 2 mm. In the 50 test scans of the ANODE09 set we recorded 207 relevant and 433 irrelevant findings. In the five example scans 39 relevant nodules and 31 irrelevant findings were annotated.

For each annotation the scan name, x, y, z coordinates of the point clicked by the observer and diameter were stored. In addition, for each relevant finding it was recorded if it was in contact with the pleura (29%), a fissure (17%), or a vessel (42%). This was done based on visual assessment by one observer. It is especially difficult to judge if a nodule is in contact with vasculature. Probably all nodules are in contact with very small vessels close to or below the resolution of a CT scan, so it is hard to draw the distinction. This issue is not so critical; the categorization was only made to al-

[1] The effective or equivalent diameter of a nodule is the diameter of a sphere with the same volume as a 3D segmentation of that nodule. Throughout this paper we give the size of nodules in mm and these lengths always mean effective diameter and are usually derived from a 3D segmentation.

low us to define different groups of nodules and report performance of methods for different subsets: pleural nodules, peri-fissural nodules, vascular nodules and isolated nodules. Note that a nodule can belong to more than one category of the first three. A nodule is isolated if it is not in contact with the pleura, a fissure or a vessel. This was the case for 20% of all nodules. Nodules were also divided into small and large nodules. The cut-off point was chosen to be 5 mm. At this point, 45% of nodules were large. Although the difference between a 4 or 5 mm nodule may seem small, note that it corresponds to almost a doubling in volume. Few nodules were above 7 mm (10%) and very few above 9 mm (2%). The densities of nodules varied, from calcified (irrelevant findings) to solid, to part-solid and non-solid. Part-solid and non-solid nodules were not included as separate categories as these were relatively rare among the relevant findings. Examples of different types of nodules and irrelevant findings are given in Fig. 1.

3.2. Evaluation: hit criterion

The results of CAD systems that have processed the test scans must be submitted on-line in the form of a text file with a set S of findings, specified by a scan name (test01 to test50), a 3D position (x, y, and z coordinate) and a degree of suspicion p. In order to limit the amount of computational processing required for the evaluation, only the 2000 most suspicious findings are analyzed. In the evaluation procedure it is determined for each finding if its distance to any nodule (relevant finding) in the scan is less than 1.5 times the radius of that nodule. If so, this signifies a hit. The factor 1.5 is used to make sure a 'near hit' is allowed, and to compensate for the fact that nodules are not perfectly spherical while distances between center points are used in the computations. We experimented with higher and lower values for this factor but found the overall results to be very stable for a wide range of values.

If a hit on a relevant finding is produced, the finding will count as a true positive (TP) and increase the overall average sensitivity of the algorithm. The relevant finding is then removed from the reference set so that it can 'hit' only once. If no hit is produced, it is determined if the distance of the finding to any irrelevant finding in the scan is less than 1.5 times the radius of that finding. If so, the finding does not count as true positive, nor as false positive; it will simply be discarded. Otherwise, the finding will be considered a false positive (FP).

3.3. FROC analysis

Results are evaluated with free-response receiver operating characteristic (FROC) analysis (see Operating Characteristic Analysis in Medical Imaging, 2008, Chapter 5). This means that the sensitivity (the fraction of true nodules in all test scans detected by the system, given by TP/n where n is the total number of relevant findings in all scans, so n = 207 in this study) is plotted as a function of

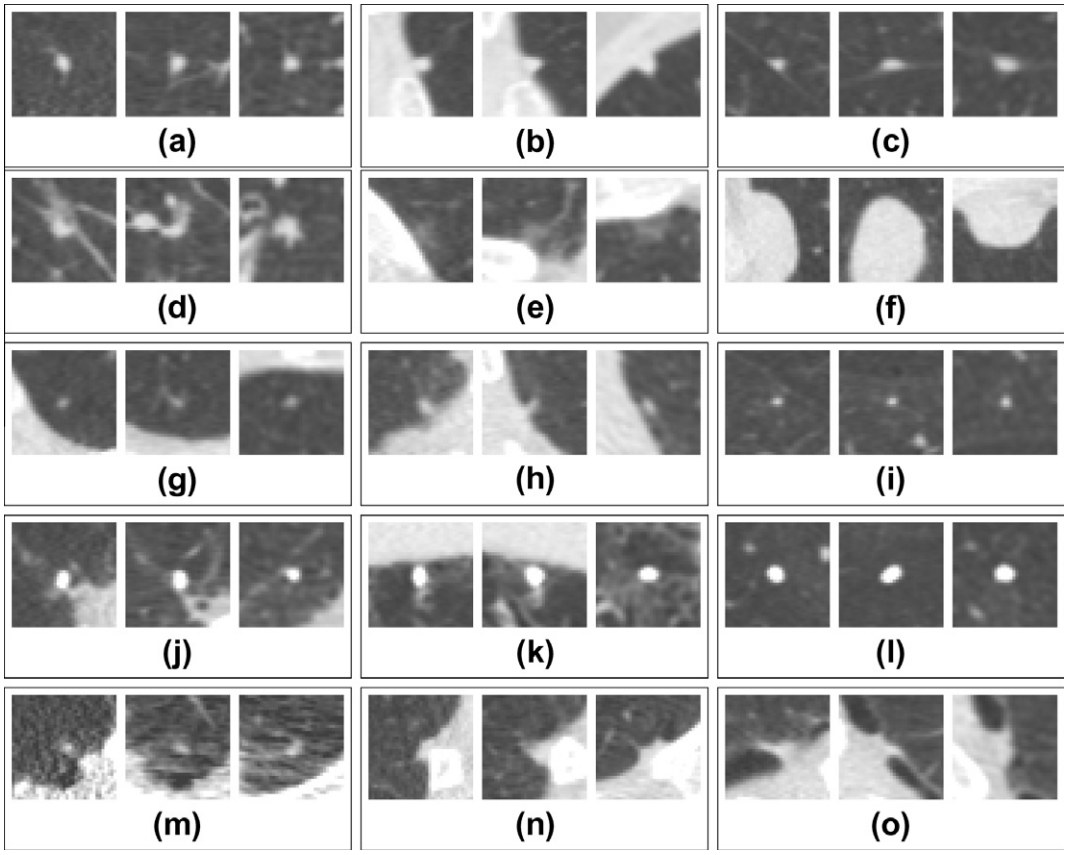

Fig. 1. Examples of relevant and irrelevant findings. In every box a nodule is displayed in a sagittal, coronal and axial view, 35 voxels (approximately 25 mm) around the center point using a lung window (center −600 HU, width 1600 HU). The top row shows three small nodules, (a) an isolated nodule of 4.4 mm; (b) a pleural nodule of 4.2 mm and (c) a peri-fissural nodule of 4.8 mm (the thin line visible on each view is the fissure). The second row shows three large nodules, (d) a nodule of 5.9 mm with vascular attachments; (e) a ground-glass nodule of 5.4 mm (a relatively rare finding and therefore not used as a separate category in this study) and (f) a large pleural nodule (18.4 mm). The third row shows nodules that were too small (below 4 mm) to be relevant. Nodules measure (g) 3.2 mm, (h) 3.5 mm, and (i) 2.3 mm, respectively. The fourth row (j–l) shows three examples of calcified nodules. Calcification is a benign characteristic and therefore these were considered irrelevant findings. Even with the used window level it is evident that the nodules are too bright (dense). The last row shows several lesions that were not considered nodules, but (m) apical scarring, (n) pleural thickening and a (o) a nodular abnormality next to an emphysematous bulla, unrelated to lung cancer.

the average number of false positive markers per scan (given by FP/m where m is the total number of scans, so $m = 50$ in this study). To obtain a point on the FROC curve, only those findings of a CAD system whose degree of suspicion $p \geqslant t$, where t is a threshold, are selected, and the number of false positives FP and true positives TP is determined, according to the procedure outlined in Section 3.2. Each unique value p in S defines a unique point on the FROC curve, using that p value as the threshold t. Between these points, straight lines are drawn to produce the FROC curve. The point with the lowest false positive rate is connected to (0,0). From the point with the highest false positive rate, the FROC curve is extended by a straight horizontal line.

3.4. Scoring system

To extract a single score from the FROC curve, we measure the sensitivity at seven predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan. Note that since we connect points on the FROC with straight lines as outlined above, we can always exactly compute these sensitivities from the curve, even if there is no threshold t that precisely produces these false positive rates. These seven sensitivities are averaged to obtain an overall score of a system. Clearly a perfect system will have a score of 1 and the lowest possible score is 0. Most CAD systems in clinical use today have their internal threshold set to operate somewhere between 1 and 4 false positives per scan on average (most systems do not allow the user to vary the threshold). To make the task more challenging, we included lower false positive rates than those used in clinical practice in our evaluation. This determines if a system can also identify a significant percentage of nodules with very few false alarms, as might be needed for CAD algorithms that operate autonomously.

From the previous exposition, it should be clear that to obtain a good score, systems should include enough findings in their results to reach the point of 8 FPs per scan. It is also recommended to include enough distinct values for the degree of suspicion p to produce a decent number of unique points on the FROC curve. In the extreme case that all findings are assigned the same p value, there will be only one point on the curve defined, and a straight line will be drawn from (0,0) to this point, and a horizontal line will extend from that point to the right.

4. Methods

In this section a brief description is provided of six methods that have been applied to the ANODE09 data set. These methods are listed as A–F in the remainder of this paper. Two more methods have been submitted (Schneider et al., 2009; Dolejsí and J. Kybic, 2009) but their performance was much lower than that of the other systems and therefore they have not been included in this analysis. For each method the main steps of the algorithm are given. It is also listed what training data was used. If available, typical performance of the system on previously used evaluation data is provided.

This section also presents a general method to combine the output of multiple CAD systems.

4.1. Method A: Fujitalab

This method was developed at Gifu University, Japan. A key original element in this detection approach is the analysis of nodule patterns with second-order local autocorrelation features in 3D space and multi-regression analysis. The second-order local autocorrelation features were expressed as a feature vector calculated from the voxel values in a $3 \times 3 \times 3$ region. From a region of this size 235 combination patterns can be obtained, excluding combinations which can be obtained by parallel movement of the center of the region. For each combination, the voxel values were multiplied, and the result was expressed as a component of the feature vector.

Using multi-regression analysis, the weighting factor for these 235 elements and a constant value were determined to indicate the training values. The training value was defined as the likelihood of nodules. A nodular shadow gave a 3D Gaussian distribution for the training output; on the other hand, a normal shadow gave a zero output.

The complete scheme involved the following steps: Segmentation of lung region; 3D matched filtering using 3D Fourier Transforms; 3D gradient concentration filtering; identification of initial candidates of nodules; false-positive reduction; analysis of the nodule images from the 235 patterns using the multi-regression analysis; calculation of mutual correlation between the training pattern and the estimated image; elimination of false positives using a rule-based approach; and calculation of the final detection results.

The lung region was segmented with gray-scale thresholding and 3D component labeling. The gradient concentration filter was designed to enhance rounded convex regions by measuring the degree of convergence of the gradient vectors around a point of interest. However, this method of using gradient concentration filters for 3D image processing is time-consuming and the segmentation results are not very satisfactory. Therefore, an improved gradient concentration filter that limits the region in which the degree of convergence is calculated was used in this study. This limited region was considered to be the one that possibility includes nodules such as rounded convex regions. The calculation time could be shortened by limiting the calculation to a given region. Additionally, good segmentation results were obtained in this case. As for the nodules, the output value of the degree-of-concentration showed a high value compared with a blood vessel region. Then, pixels with a high output value of the degree-of-concentration filter were used as starting points for a region growing technique and in this way candidate regions were obtained.

Image features, i.e. size, degree of sphericity, aspect ratio, mean value of the degree of convergence, and the maximum value of the degree of convergence, were used for elimination of false positives.

Next, the autocorrelation features and multi-regression analysis was applied to the remaining candidates. The output of multi-regression is expected to be a continuous value; hence, the comparison between the training patterns obtained in multi-regression and the output also emphasized the nodular shadows. False-positives were therefore further eliminated by using the correlation value and the volume of the remaining candidates.

For training, the five example cases from the ANODE09 study were used. The CAD system was implemented in C/C++. The average processing time per case is 10 min on a PC with 16GB memory and a 2.0 GHz Opteron dual processor.

4.2. Method B: region growing volume plateau

This method was developed at the University of Bari, Italy, in the MAGIC-5 research project (Bellotti et al., 2007). The method has been published (Bellotti et al., 2007) and was slightly modified for this study.

The system consisted of three steps: (1) the lung parenchymal volume was segmented in the whole CT volume; (2) a region growing algorithm was iteratively applied to the segmented volume to detect candidate nodules; (3) a double-threshold and a neural network were applied to reduce false positives and classify the findings.

The lung parenchymal segmentation started with a simple-threshold 3D region growing applied to the CT volume. The result is a binary mask of the respiratory system, containing the trachea, the bronchi, and the lungs. The next step was the segmentation of the external airways (trachea and bronchi) by a 3D region growing with wave-front simulation and suitable stop conditions, allowing a proper handling of the hilar region. Particular attention was given to detecting and solving the problem of the apparent fusion between the lungs, caused by partial-volume effects. 3D morphology operations ensured the accurate inclusion of all the nodules (internal, pleural, and vascular) in the segmented volume.

The second step detected candidate nodules inside the segmented volume. This functionality was implemented by a region growing algorithm with an inclusion rule given by the logical AND of two rules: a voxel was included in the region if its density averaged with its first order neighbors was larger than a threshold t_1, and a voxel was included in the region if its density was larger than a threshold t_2.

The threshold t_1 was dynamically defined for each nodule candidate. Starting from an initial value, t_1 was decreased to obtain a curve providing the volume as a function of the threshold. In general, this curve shows a decrease followed by a plateau due to difference in density between the background and the nodule candidate. From this curve it is possible to infer the best t_1 value as the smallest in the range of the plateau.

The t_2 threshold and the starting value of t_1 were chosen in order to maximize the detection rate (the fraction of selected nodules with respect to the total number of nodules diagnosed by the radiologist). The seed points were searched automatically as follows: the segmented volume is scanned until a voxel matching the inclusion rules (with thresholds t_2 and t_1) was found; this voxel was used as seed point and the growth was started. Once the region was completely grown, it was removed from the scan and stored for further analysis. Then the search for new seed points was restarted. This process was iterated until no more seed points matching the inclusion rules were found.

For each candidate nodule the following features were calculated: sphericity, ellipticity, maximum intensity, intensity standard deviation, Shannon entropy, volume, maximum radius. Almost all the FPs findings refer to candidates with too few voxels or to non-spherical candidates and could be easily ruled out by a simple-threshold on the volume and the sphericity. A further reduction of false detections was obtained by means of a classification step carried out by a supervised two-layered feed-forward neural network, trained with a gradient descent learning rule and with a sigmoid transfer function. The output of the neural network was used as degree of suspicion for each candidate.

Initially, results based on the training data described in Bellotti et al. (2007) were submitted to the ANODE09 organizers. This is the training data also used by methods C and D. Later it was found that training the system with the example scans of the ANODE09 data set produced slightly better results on those example sets (tested through cross-validation) and therefore these example scans were used in the final submission presented in this paper.

The method runs in an average of 15 min per scan on an Intel Xeon Dual Quad Core 2.6 GHz CPU" with 16 GB of RAM, using only a single core. The implementation is in C++ using Root-CERN Data Analysis Framework, an open-source C++ framework developed at CERN for high energy physics data analysis.

4.3. Method C: Channeler Ant model

This method has been developed by researchers from the University of Torino, Italy and CEADEN in Cuba within the framework of the MAGIC-5 research project (Bellotti et al., 2007). The system is aimed at segmentation of generic 3D objects of unknown shape and can therefore be adapted to the automated search for lung nodules in low-dose CT scans.

The training data consisted of a set of low-dose lung CT scans collected by the Pisa Center of the ITALUNG-CT trial, the first Italian randomized controlled trial for the screening of lung cancer (Pegna et al., 2009). The CT scans were acquired with a 4-slice spiral CT scanner (Siemens volume zoom) according to a low-dose protocol (tube voltage: 140 kV, tube current: 20 mA, mean equivalent dose 0.6 mSv), with 1.25 mm slice collimation. Slices were reconstructed at 1 mm thickness, using a medium sharp reconstruction kernel (Siemens B50f). The number of slices per scan was approximately 300, each slice being a 512 by 512 pixel matrix, with pixel sizes ranging from 0.53 to 0.74 mm. The scans were annotated by experienced radiologists with a dedicated annotation and visualization tool (Gori et al., 2007b).

The method started with lung segmentation. The lung parenchyma was identified by means of a 3D region growing method and a wave-front algorithm for the definition of the lung surface on the inner side.

The Channeler Ant model (Cerello et al., 2008) was used as a segmentation method for the vessel tree and the nodules candidates. Ant colonies were released on selected position of a 3D matrix, i.e. the anthill. Each ant behaves according to a predefined set of rules (Cerello et al., 2008) and releases a quantity of pheromone while moving in the 3D environment defined by the lung volume. When the colony was extinct and no more voxels matched the required conditions to become anthills, the information provided by the pheromone map was analyzed. Ants explore (i.e. live in) a 3D environment described in terms of positions and intensities of voxels. Their life cycle is a sequence of atomic time steps, during which ants move from one voxel to one of its 26 neighbors. The behavior of ants was defined by a set of rules that specify how they move in the environment, how much pheromone they release before moving to another location, when they reproduce or die. The environment is defined by the voxel image intensities, which can be thought of as the amount of available food for the colony: therefore, voxel intensities should be progressively consumed when the number of visits increases. This mechanism, required to make the colony evolve and explore the environment, was implemented in a complementary way: whenever the limit to the maximum number of visits in a voxel was reached, the voxel was no more available as a destination.

The ant colony started evolving from a voxel at the root of the vessel tree. When all the ants in the colony have died, the process stopped and the segmented object was removed from the original image and its coordinates were added to a list. In the remaining image, a voxel with intensity greater than a predefined threshold became the new anthill and a new ant colony was deployed. If the number of voxels of an object was large with respect to the maximum expected size of a nodule, as it happens with the bronchial and vascular trees, the object was processed and smaller connected objects are looked for. The process finished when all the voxels inside the matrix with intensity above the threshold had been analyzed. From the segmented objects five features were extracted: number of voxels, maximum intensity, average intensity, standard deviation of intensity and sphericity. A feed-forward artificial neural network was implemented in order to classify the segmented objects.

A limitation of the method is that nodules with diameter smaller than 3 mm attached to the vascular tree cannot be detected. When the system was applied to the training data set, using cross-validation, a sensitivity of 46% and 64% was obtained at an average of 2 and 6 false positives per scan, respectively.

The Channeler Ants run in an average of 550 s per CT scan on an iMac with a 2.4 GHz Intel Core 2 Duo processor and 2 GB RAM. The software is implemented in C++ as an extension of the Root func-

tionality and shares its code repository with the previously described method.

4.4. Method D: Voxel-based neural approach

This method was jointly developed by researchers from INFN and the University of Pisa, Italy, and a researcher from Bracco Imaging S.p.A. within the MAGIC-5 research project (Bellotti et al., 2007). The method is described in Retico et al. (2009), Gori et al. (2009). A subset of the ITALUNG-CT data set (Pegna et al., 2009) that was also used in method C was available to train and validate the system.

First, lung nodules were partitioned in two main classes, depending on their location in the lung. A nodule was labeled either as internal if fully contained in the lung parenchyma or as juxtapleural if connected to the pleura. The internal and juxtapleural nodule classes surely included the ANODE09 categories of isolated and pleural nodules, respectively. Nodules belonging to the other ANODE09 categories (peri-fissural and vascular) could either fall into the internal or into the juxtapleural nodule class.

The system dealt differently with internal and juxtapleural nodules, by means of two dedicated procedures: CAD_I for internal and CAD_{JP} for juxtapleural nodules. Both are three-step procedures (Buscema, 2004; Gori and Mattiuzzi, 2008; Gori et al., 2007; Retico et al., 2008; Retico et al., 2009; Gori et al., 2009):

1. Lung segmentation: an approach based on thresholding, region growing and morphological operators is implemented, once the scans have been isotropically resampled. In order to outline the shape of the pleura irregularities (including juxtapleural nodules), the lung boundaries were not smoothed. The identified lung mask, including vessels and airway walls, was used for CAD_I, whereas its boundary was used for CAD_{JP}.
2. Candidate nodule selection:
   - CAD_I: internal nodules were modeled as spherical objects with a Gaussian profile, following the approach proposed in (Li et al., 2003); the 3D matrix of data was filtered with a multi-scale filter function built to discriminate between spherical objects and objects with planar or elongated shapes. The local maxima of the 3D filtered matrix were the internal candidate nodule locations. A large number of false positives were included at this stage, above all crossings between blood vessels.
   - CAD_{JP}: to identify juxtapleural candidate nodules, pleura surface normals were constructed and each voxel was assigned a score proportional to the number of normals intersecting in it. To deal with noise, cylinders with Gaussian profile were considered instead of segments (Paik et al., 2004). The local maxima of the 3D score matrix were the juxtapleural candidate nodule locations. A large number of FPs was found, mostly due to irregularities in the pleura surface (e.g. apical scars, pleural thickening and plaques) and movement artifacts.
3. FP reduction: an original procedure, the Voxel-based neural approach (Gori et al., 2007; Retico et al., 2008; Retico et al., 2009; Gori et al., 2009), was developed to reduce the number of FPs in the lists of internal and juxtapleural candidate nodules. First, a region of interest (ROI) including voxels belonging to the candidate nodule was defined from each location provided by the previous step. For internal candidate nodules, a simple procedure based on relative thresholding was implemented, while for juxtapleural candidate nodules a morphological opening-based algorithm was used. The basic idea is to associate to each voxel of a ROI a feature vector constituted by the intensity values of its 3D neighbors and the eigenvalues of the gradient matrix and of the Hessian matrix. Feature vectors were then classified by a three-layer feed-forward neural network which is trained to assign each voxel either to the nodule or normal tissue target class. A ROI was assigned a degree of suspicion p, defined as the percentage of voxels tagged as nodule by the neural classifier.

The final list of findings was simply obtained by merging the output lists of findings generated by CAD_I and CAD_{JP}.

The training data used for CAD_I consisted of 30 internal nodules contained in 15 CT scans, whereas 28 nodules belonging to 14 CT scans were used for CAD_{JP}. The diameters of these nodules were in the 4–12 mm range; in particular, the 65% of them was in the 4–6 mm range. Calcified solid nodules were not considered. As only a very small number of part-solid or non-solid nodules were annotated in the available data set, they were not included.

System performance was evaluated on a prediction set of thirty other scans extracted from the ITALUNG-CT data set, containing 35 internal and 32 juxtapleural nodules. A sensitivity of 78% and 70% was measured at 8 and 4 false positive detection per scan, respectively (Gori et al., 2009).

The algorithm for detecting internal nodules runs in an average of 12 min and the algorithm for juxtapleural nodules requires 15 min per scan, both running on a Dual Opteron 2.5 GHz machine with 4 GB of RAM, using a single core. The systems were developed in C++ using the Insight Toolkit (http://www.itk.org/), an open-source framework for segmentation and registration of medical images, while the neural networks were implemented using FANN (Fast Artificial Neural Network) (http://leenissen.dk/fann/), an open source neural network library.

4.5. Method E: ISI-CAD

This method was developed at the University Medical Center Utrecht, the Netherlands, by the group who organized the ANODE09 study. The method is described in detail in Murphy et al. (2007), Murphy et al. (2009).

First the scan was subsampled to isotropic resolution and axial slices of 256 by 256. The lungs were segmented by region growing and post-processing, including morphological smoothing of the lung boundaries (van Rikxoort et al., 2009). To extract nodule candidates, the shape index and curvedness were computed at a fixed scale of 1 voxel. Voxels for which these values are within preset ranges are clustered into a candidate structure. This procedure yielded on average around 700 structures per scan.

False-positive candidates are removed by means of a two-step approach using k-nearest neighbor classification (kNN). The kNN classifiers are trained using features of the image intensity gradients and gray-values in addition to further measures of shape index and curvedness profiles in the candidate regions. The initial classification step uses a small number of relatively simple features to quickly reduce the most obviously incorrect candidates. These are not further processed. After this first stage around 80 candidates per scan remain. The second classifier employs more features of higher complexity in order to classify the more ambiguous remaining candidates as accurately as possible. A total of 135 features were initially considered as being potentially useful. For both classification steps, sequential forward floating selection was employed in the training stage to identify the most useful features. A total of eight features were selected for the initial classification, with 19 features being chosen for the final classifier.

The training data consisted of data from 722 scans from the NELSON screening program, which is the same data source as the ANODE09 data set, giving this method a unique advantage over the other methods considered in this paper. The complete set of NELSON annotations were used as reference for training the CAD system. The ANODE09 scans are from different subjects that those

in the training data. In previous work (Murphy et al., 2009), the method was tested on 813 NELSON scans and detected 80% of annotated nodules at an average of 4.2 false positive detections per scan.

The method is implemented in C++ and the required computation time per scan is about 2 min for lung segmentation and 2 min for nodule analysis using a single core of a 3 GHz processor in a PC with at least 2GB RAM.

4.6. Method F: Philips Lung Nodule CAD

The final method in this paper is a commercially available algorithm. We evaluated the Philips Lung Nodule CAD option that is part of the Lung Nodule Assessment (LNA-K023785) application that runs in the CT workstation called Extended Brilliance Workspace or EBW (Philips Medical Systems, Cleveland, OH). This is a general purpose viewing and processing workstation for medical images with several packages for dedicated CT image analysis on-board of which nodule analysis is one. The software produces a number of markers per CT scan. This number cannot be adjusted (as is the case in most commercial CAD software). The software also does not return voxel coordinates. The markers were presented in a list box as CAD1, CAD2, CAD3, and so on, depending on how many marks were available for a particular case. It was our impression that items higher in this list corresponded to more suspicious findings. In order to convert the software output to a format suitable for ANODE09, we therefore proceeded as follows. Two medical students processed the 50 ANODE09 cases and located the markers in a separate software program to obtain approximate voxel locations. A finding that was listed as CAD1 received a likelihood of 1, a finding that was listed as CAD2 received a likelihood of 1/2, and so on. In this way, the first point of the FROC curve that will be generated consists of only the CAD1 marks of all cases, the second point on the curve consists of CAD1 and CAD2 marks, and so on. Note that it cannot be taken for granted that the first finding of one scan has the same absolute relevance of the first finding of another scan. Thus the real FROC of the system will necessary be unpredictably different, and will probably be slightly better than what is reported here. The only point that we know for sure to be correct is the point with the lowest false positive rate at the highest sensitivity, where all marks are used.

The Philips Lung Nodule CAD comes with extensive documentation on its use and describes the results of clinical studies to investigate its effectiveness. The documentation is brief, however, about the working of the algorithm. It states that the scheme consists of four principal parts. First, the lungs are segmented. Next, seed points are determined from 2D analysis of slices. From these seeds, 3D features and metrics are computed. Finally, the list of candidates is reduced by filtering each candidate on its features and metrics and the application of simple accept/reject rules. From this description, we believe the method is comparable to the algorithm outlined in two publications by Wiemker and co-workers (Wiemker et al., 2002; Wiemker et al., 2005). The characteristics and total number of training scans that were used to develop the accept/reject rules is unknown.

In a clinical study where four sites contributed 110 subject cases, the system was found to yield between 5 and 8 false positive markers per case and have a sensitivity of around 60% for nodules that were determined by a consensus panel and around 36% for all nodules indicated by the radiologists in the study. These results are reported in the documentation of the EBW but a literature reference is not given.

The algorithm takes 40–60 s per scan, running on a central server where the web portal version of EBW resides. From the status messages provided by the softweare, it seems about half of that time is spent on the lung segmentation.

4.7. System combination

Like many tasks in medical image analysis, nodule detection is a complicated problem that can be approached in many different ways. The detection algorithms outlined above indeed appear substantially different. If multiple methods focus on different aspects of a problem, it is not unlikely that a proper combination of their output would yield a higher performance than any of the methods stand-alone.

To investigate this, we propose a way to combine the results of multiple nodule CAD systems without access to their internals, like the feature values of candidates that are input to an internal classifier. The proposed blending method employs only the findings (coordinates and degree of suspicion p for each finding) and information about the performance of individual systems. It uses this performance information in such a way that systems with better performance are implicitly weighed more heavily in the combination. Without such knowledge, making a proper combination of systems with widely different performance levels is difficult.

More precisely, we assume that, before combining, the results of a CAD system on an evaluation set with known truth are available. Let p_i, $i = 1,\ldots,n$ denote the likelihood of each CAD finding. Every unique value of p in the set of n findings corresponds to a point on the FROC curve of the system, as explained in Section 3.3. For every unique p value we can compute the number of true positives TP when we consider all findings with $p_i \geqslant p$ as positive. We can also compute the number of false positives FP we obtain at this threshold (disregarding irrelevant findings). Now we associate with each p a value

$$f(p) \rightarrow \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{TP}+1}, \tag{1}$$

where the factor +1 has been added in the denominator to avoid division by zero in the exceptional situation that all findings are irrelevant, in which case both TP and FP equal zero. The values $f(p)$ are approximately equal to the probability that a finding in the evaluation set with likelihood p or higher represents a true nodule. Such probabilities are natural measures to combine.

To combine systems, we compute $f(p)$ for every finding from every system. All findings are sorted so that we have f_i, $i = 1,\ldots,n$ and $f_i \geqslant f_j$ if $i < j$. Starting at f_i with $i = 1$, it is checked for all findings f_j, $j = i + 1,\ldots,n$ if they correspond with f_i. In this study we used the simple rule that findings within five voxels of each other (and obviously located in the same scan) are corresponding. A more elaborate criterion, such as the one used to compute the FROC curves in this study, could be used instead, but this is not possible as no segmentations or effective diameters of the input findings are available. If two findings f_i and f_j correspond, we set

$$f_i \rightarrow f_i + f_j,$$

remove f_j from the list of findings and continue the procedure. It is easy to see that this is conceptually similar to averaging the probabilities for each finding across all systems, where undetected findings correspond to a zero probability: we add up the findings we are able to match across systems and if a system does not detect a particular finding, nothing will be added. Note that systems with low performance have f values that are close to zero for (nearly) all their findings, and these systems are therefore automatically weighed less heavily in the combination.

5. Results

The results for all systems are tabulated in Table 1. There is considerable variation in the overall scores. System E clearly outperforms the other schemes. The results for the different classes of

Table 1
Results for systems A–F. For each of the nodule categories and for all nodules, sensitivity is provided at seven levels of average numbers of false positives per scan, $2^{-3,\ldots,+3}$. In the final column, the average of the sensitivities at the seven false positive levels is given. The number in the bottom-right of each table can be considered an overall score for the system.

FPs/scan	1/8	1/4	1/2	1	2	4	8	Average
System A								
Small nodules	0.154	0.171	0.231	0.282	0.299	0.316	0.316	0.253
Large nodules	0.111	0.122	0.144	0.178	0.178	0.189	0.189	0.159
Isolated nodules	0.238	0.262	0.310	0.381	0.381	0.381	0.381	0.333
Vascular nodules	0.116	0.140	0.186	0.209	0.221	0.244	0.244	0.194
Pleural nodules	0.051	0.051	0.068	0.119	0.136	0.153	0.153	0.104
Peri-fissural nodules	0.171	0.171	0.286	0.314	0.314	0.314	0.314	0.269
All nodules	0.135	0.150	0.193	0.237	0.246	0.261	0.261	0.212
System B								
Small nodules	0.111	0.171	0.222	0.299	0.453	0.538	0.581	0.339
Large nodules	0.111	0.122	0.144	0.222	0.278	0.344	0.367	0.227
Isolated nodules	0.214	0.262	0.310	0.476	0.595	0.667	0.667	0.456
Vascular nodules	0.105	0.116	0.163	0.198	0.267	0.337	0.360	0.221
Pleural nodules	0.017	0.017	0.017	0.085	0.220	0.339	0.390	0.155
Peri-fissural nodules	0.171	0.314	0.371	0.457	0.600	0.686	0.743	0.478
All nodules	0.111	0.150	0.188	0.266	0.377	0.454	0.488	0.291
System C								
Small nodules	0.009	0.017	0.077	0.205	0.342	0.530	0.624	0.258
Large nodules	0.089	0.111	0.222	0.267	0.322	0.356	0.378	0.249
Isolated nodules	0.024	0.048	0.119	0.333	0.476	0.595	0.667	0.323
Vascular nodules	0.070	0.093	0.151	0.198	0.302	0.442	0.488	0.249
Pleural nodules	0.034	0.034	0.153	0.203	0.220	0.356	0.441	0.206
Peri-fissural nodules	0.057	0.057	0.171	0.286	0.457	0.514	0.571	0.302
All nodules	0.043	0.058	0.140	0.232	0.333	0.454	0.517	0.254
System D								
Small nodules	0.107	0.205	0.299	0.393	0.462	0.564	0.624	0.379
Large nodules	0.017	0.022	0.089	0.144	0.222	0.333	0.444	0.182
Isolated nodules	0.149	0.214	0.405	0.571	0.571	0.667	0.690	0.467
Vascular nodules	0.055	0.116	0.198	0.256	0.372	0.453	0.547	0.285
Pleural nodules	0.013	0.034	0.068	0.153	0.220	0.356	0.475	0.188
Peri-fissural nodules	0.089	0.171	0.229	0.257	0.286	0.429	0.514	0.282
All nodules	0.068	0.126	0.208	0.285	0.357	0.464	0.546	0.293
System E								
Small nodules	0.470	0.491	0.573	0.658	0.711	0.761	0.778	0.634
Large nodules	0.423	0.483	0.567	0.611	0.714	0.778	0.822	0.628
Isolated nodules	0.548	0.595	0.595	0.619	0.619	0.643	0.643	0.609
Vascular nodules	0.570	0.573	0.616	0.686	0.757	0.802	0.849	0.693
Pleural nodules	0.052	0.140	0.322	0.475	0.630	0.695	0.729	0.435
Peri-fissural nodules	0.629	0.643	0.743	0.771	0.804	0.886	0.886	0.766
All nodules	0.450	0.488	0.570	0.638	0.712	0.768	0.797	0.632
System F								
Small nodules	0.019	0.038	0.075	0.133	0.186	0.278	0.359	0.155
Large nodules	0.053	0.106	0.195	0.306	0.395	0.539	0.711	0.329
Isolated nodules	0.044	0.088	0.152	0.222	0.260	0.381	0.429	0.225
Vascular nodules	0.038	0.077	0.145	0.246	0.334	0.437	0.558	0.262
Pleural nodules	0.012	0.025	0.057	0.112	0.136	0.229	0.424	0.142
Peri-fissural nodules	0.032	0.063	0.155	0.295	0.418	0.543	0.571	0.297
All nodules	0.034	0.067	0.127	0.208	0.276	0.392	0.512	0.231

nodules reveal more subtle differences between the systems. For example, system F scores much better for larger nodules compared to smaller ones, but for other systems the opposite holds. In general, isolated nodules seem easier to detect than peri-fissural and vascular nodules, and pleural nodules are the hardest. But for some systems this general trend does not hold.

Table 3 shows the results for all 57 possible combinations that can be made from six systems. It is evident that blending the output of CAD systems can lead to spectacular improvements in performance. The combination of systems B and C, with individual scores of 0.291 and 0.254, leads to a system with a score of 0.437, an increase of 0.146 compared to B alone. An even larger improvement is obtained when systems C and D are combined. This leads to a system with a score of 0.471 and the results of this system are also given in Table 2 where it can be seen that for some categories of nodules performance almost doubles. Combining the best performing system (E) with one other system mostly leads to smaller improvements, and even some slight deteriorations. But, in the case of combining E with C, performance improves to from 0.632 to 0.659, the best combination of two systems. Combining E with D scores lower, although D scores higher than C. When all systems are combined an overall score of 0.685 is obtained, compared to 0.632 for system E alone. The best combination without system E is the combination of all remaining systems. This system is also tabulated in Table 2. It leads to the highest improvement compared to any single system in the combination, scoring 0.592, which is 0.299 higher than system D, the best single system in this combination, alone.

Fig. 2 shows the FROC curves for all nodules for all systems, including three combinations. Fig. 3 shows the same, but for all

Table 2
Results for three combined systems. System C + D is the best performing combinations of two systems excluding system E. System A + B + C + D + F has the largest overall performance improvement compared to any of its composing systems. The best result is obtained for the combination of all systems (A + B + C + D + E + F). For each of the nodule categories and for all nodules, sensitivity is provided at seven levels of average numbers of false positives per scan, $2^{-3,\ldots,+3}$. In the final column, the average of the sensitivities at the seven false positive levels is given. The number in the bottom-right of each table can be considered an overall score for the system.

FPs/scan	1/8	1/4	1/2	1	2	4	8	Average
System C + D								
Small nodules	0.393	0.436	0.479	0.547	0.615	0.650	0.726	0.549
Large nodules	0.200	0.289	0.322	0.356	0.389	0.489	0.533	0.368
Isolated nodules	0.548	0.619	0.667	0.667	0.690	0.714	0.738	0.663
Vascular nodules	0.302	0.349	0.372	0.419	0.465	0.535	0.616	0.437
Pleural nodules	0.153	0.203	0.288	0.407	0.508	0.576	0.661	0.400
Peri-fissural nodules	0.314	0.457	0.457	0.514	0.543	0.629	0.629	0.506
All nodules	0.309	0.372	0.411	0.464	0.517	0.580	0.643	0.471
System A + B + C + D + F								
Small nodules	0.453	0.513	0.598	0.650	0.702	0.795	0.821	0.647
Large nodules	0.344	0.389	0.456	0.511	0.556	0.656	0.722	0.519
Isolated nodules	0.619	0.619	0.738	0.738	0.742	0.762	0.786	0.715
Vascular nodules	0.360	0.419	0.477	0.512	0.593	0.709	0.779	0.550
Pleural nodules	0.254	0.271	0.407	0.525	0.542	0.695	0.763	0.494
Peri-fissural nodules	0.514	0.657	0.657	0.686	0.771	0.829	0.829	0.706
All nodules	0.406	0.459	0.536	0.589	0.638	0.734	0.778	0.592
System A + B + C + D + E + F								
Small nodules	0.496	0.573	0.684	0.761	0.803	0.821	0.872	0.716
Large nodules	0.389	0.411	0.578	0.678	0.778	0.811	0.867	0.644
Isolated nodules	0.595	0.619	0.643	0.738	0.786	0.810	0.810	0.714
Vascular nodules	0.430	0.465	0.616	0.721	0.802	0.826	0.907	0.681
Pleural nodules	0.254	0.356	0.542	0.627	0.695	0.746	0.831	0.579
Peri-fissural nodules	0.629	0.657	0.771	0.829	0.886	0.914	0.914	0.800
All nodules	0.449	0.502	0.638	0.725	0.792	0.816	0.870	0.685

nodule categories separately. Note that the false positive rate plotted on the horizontal axis in this Figure comprises all false positives, not only false positives in the respective categories.

6. Discussion

The six systems considered in this comparison show remarkably different results. This supports the notion that comparisons on the same database are important. There are three possible main reasons for performance difference between systems: the underlying algorithm or architecture of the CAD system; the training data that is used to train the classifiers or to set the internal model parameters of the CAD system; and the characteristics of the test data and the protocol that was used to set the reference on the test data. The ANODE09 data set does not supply a separate set for training, as was done for example in Heimann et al. (2009). As a result, some of the systems included in this comparison may have been trained with data with different characteristics and a different protocol for determining what constitutes a relevant nodule. It would be interesting to compare systems that use identical training data, however, this limits the possibilities for including certain systems that are used in clinical practice or that have been used in previously published studies in the comparison.

Clearly the training data that has been used by the systems varies considerably. In particular, system E has a distinct advantage over the other studies in that it has used a large training set, originating from the same lung cancer trial, using the same scanners and scan protocol. This system was trained with the NELSON annotations, which are comparable, but slightly different from the annotation protocol adopted for ANODE09. It is unclear how much of its better performance can be attributed to the difference in training data. The performance of system E reported here is roughly comparable to that reported in Murphy et al. (2007), Murphy et al. (2009). Method A and B both used the five example cases in the ANODE09 date set for training. Clearly this is a small training set, although it is representative of the ANODE09 test data. Methods C and D used the same training data, originating from an Italian lung cancer screening trial. This training set was also small compared to the set used by system E. The results obtained by systems C and D on their training data, tested by cross-validation are substantially better than those obtained on the ANODE09 data set. This indicates that there may be important differences between the Italian data and the ANODE09 data, which can be related to the scans or to the type of annotations. The Japanese team (method A) has investigated the effect of changing training databases when using the five ANODE09 example cases for testing and found substantial differences depending on which training database was used. It is likely that all methods A–D would improve if they would have more training data available. It is therefore impressive that the combination of all systems except E approaches the score of E so closely. LIDC has announced that a database with over 1000 CT scans will become publicly available, and this will greatly facilitate investigations into the effect of type and size of training databases on nodule detection performance.

The categorization of relevant and irrelevant nodule findings is also specific to the ANODE09 study. This categorization is not universal and it is perhaps unfair to compare systems trained with data in which other definitions of what constitutes an actionable nodule were adopted. It is important in studies like these to carefully consider the definition of 'truth'. The study of Armato et al. (2009) shows that even experienced thoracic radiologists may not perform well when measured against the 'truth' established by other experienced thoracic radiologists.

The commercial system, F, does not achieve a very high score. It is at a disadvantage compared to all other systems because the actual degree of suspicion used internally in the algorithm was not accessible to the researchers who applied the system to the ANODE09 data. The strategy used to construct intermediate points (see Section 4.6) is not optimal, and the shape of the FROC curve suggests that as well. On the other hand, it is unlikely that knowing the proper p values for the findings of this system would have resulted in much increased detection rates at lower false positive levels. System F, and system E and A as well, might have achieved slightly higher scores if more findings had been included in their

Table 3
Results of all combinations that can be obtained from six systems. The filled and open squares indicate which systems have and have not been included in the combination, so for example □ ■■□□■ is the combination of systems B, C, and F. The score is the average sensitivity at the seven false positive levels $2^{-3,\ldots,+3}$. The best score of any single system included in the combination is also given, and the difference between the combination score and the best score of a single system in the combination is listed under Δ.

Combination	Score	Best single	Δ
■□□□□□		0.212	
□■□□□□		0.291	
□□■□□□		0.254	
□□□■□□		0.293	
□□□□■□		0.632	
□□□□□■		0.231	
■■□□□□	0.371	0.291	0.080
■□■□□□	0.336	0.254	0.082
■□□■□□	0.372	0.293	0.079
■□□□■□	0.606	0.632	−0.026
■□□□□■	0.330	0.231	0.099
□■■□□□	0.437	0.291	0.146
□■□■□□	0.468	0.293	0.175
□■□□■□	0.604	0.632	−0.028
□■□□□■	0.413	0.291	0.122
□□■■□□	0.471	0.293	0.178
□□■□■□	0.659	0.632	0.027
□□■□□■	0.361	0.254	0.107
□□□■■□	0.636	0.632	0.004
□□□■□■	0.445	0.293	0.152
□□□□■■	0.634	0.632	0.002
■■■□□□	0.471	0.291	0.180
■■□■□□	0.498	0.293	0.205
■■□□■□	0.607	0.632	−0.025
■■□□□■	0.451	0.291	0.160
■□■■□□	0.477	0.293	0.184
■□■□■□	0.649	0.632	0.017
■□■□□■	0.418	0.254	0.164
■□□■■□	0.625	0.632	−0.007
■□□■□■	0.452	0.293	0.159
■□□□■■	0.640	0.632	0.008
□■■■□□	0.522	0.293	0.229
□■■□■□	0.625	0.632	−0.007
□■■□□■	0.494	0.291	0.203
□■□■■□	0.637	0.632	0.005
□■□■□■	0.560	0.293	0.267
□■□□■■	0.641	0.632	0.009
□□■■■□	0.668	0.632	0.036
□□■■□■	0.526	0.293	0.233
□□■□■■	0.678	0.632	0.046
□□□■■■	0.668	0.632	0.036
■■■■□□	0.546	0.293	0.253
■■■□■□	0.640	0.632	0.008
■■■□□■	0.518	0.291	0.227
■■□■■□	0.636	0.632	0.004
■■□■□■	0.568	0.293	0.275
■■□□■■	0.651	0.632	0.019
■□■■■□	0.664	0.632	0.032
■□■■□■	0.528	0.293	0.235
■□■□■■	0.687	0.632	0.055
■□□■■■	0.663	0.632	0.031
□■■■■□	0.659	0.632	0.027
□■■■□■	0.585	0.293	0.292
□■■□■■	0.664	0.632	0.032
□■□■■■	0.666	0.632	0.034
□□■■■■	0.689	0.632	0.057
■■■■■□	0.668	0.632	0.036
■■■■□■	0.592	0.293	0.299
■■■□■■	0.672	0.632	0.040
■■□■■■	0.677	0.632	0.045
■□■■■■	0.702	0.632	0.070
□■■■■■	0.690	0.632	0.058
■■■■■■	0.685	0.632	0.053

result set so that the point of eight false positive detections per scan had been reached.

System F is peculiar in that it detects larger nodules much better than smaller ones, whereas for the other systems this is the other way around, with C and E showing comparable performance for large and small nodules. Intuitively, one would expect better performance for large nodules, but one reason for better performance for smaller nodules could be that these are more often isolated. Also smaller nodules are more common so probably occur more in CAD systems' training data, and smaller nodules are more likely to have the classical appearance of a simple sphere, whereas larger nodules are more likely to be lobulated or spiculated. It could also be a pre-determined setting of system F to give smaller potential nodules a lower degree of suspicion. From a clinical point of view, this makes sense as larger nodules are much more likely to represent malignancies.

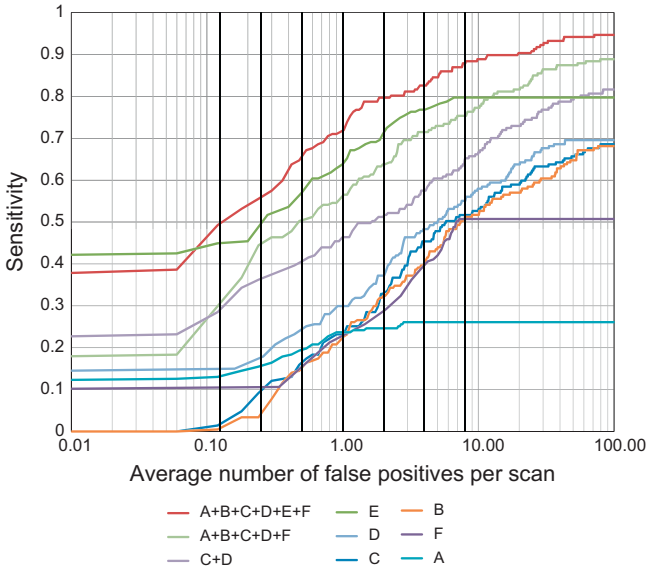

Fig. 2. FROC curves of all six systems and three combinations. The horizontal axis is logarithmic and covers four orders of magnitude.

The results for the different categories of nodules reveal interesting information. Some systems are particularly suited to detecting isolated nodules (systems D and B, for example) which could be the result of a high prevalence of these type of nodules in their training data. Some systems are not very adept at detecting perifissural or pleural nodules. Such weaknesses of systems to handle particular types of nodules can be compensated by other systems when combining them.

We believe that much more than identifying 'good' and 'bad' systems, the real value of this study lies in the demonstration that the combination of systems yields such spectacular improvements. As we noted, the methods have different strengths and weaknesses. The effect of combining systems reveals how complementary they are. System F is not a very good system in terms of overall performance, and adding it to system E, the best performing system, leads only to minor improvements (0.632–0.634), but when putting all systems together, leaving out system F decreases performance from 0.685 to 0.668. Apparently system F is in some ways different from the other systems. Note that this complementarity is not observed for systems A and B. Leaving them out of the total combination even slightly improves results. However, leaving both systems out decreases performance slightly.

One general explanation for the improvements gained by many combinations listed in Table 3 is that CAD systems contain many elements, and therefore the designer of a CAD system faces many choices and a combinatorial explosion of possibilities. There is a wide array of possible features to compute for lesion candidates. Moreover, the widely different maximum sensitivity levels reached by the various systems suggest that the cadidates detectors of the

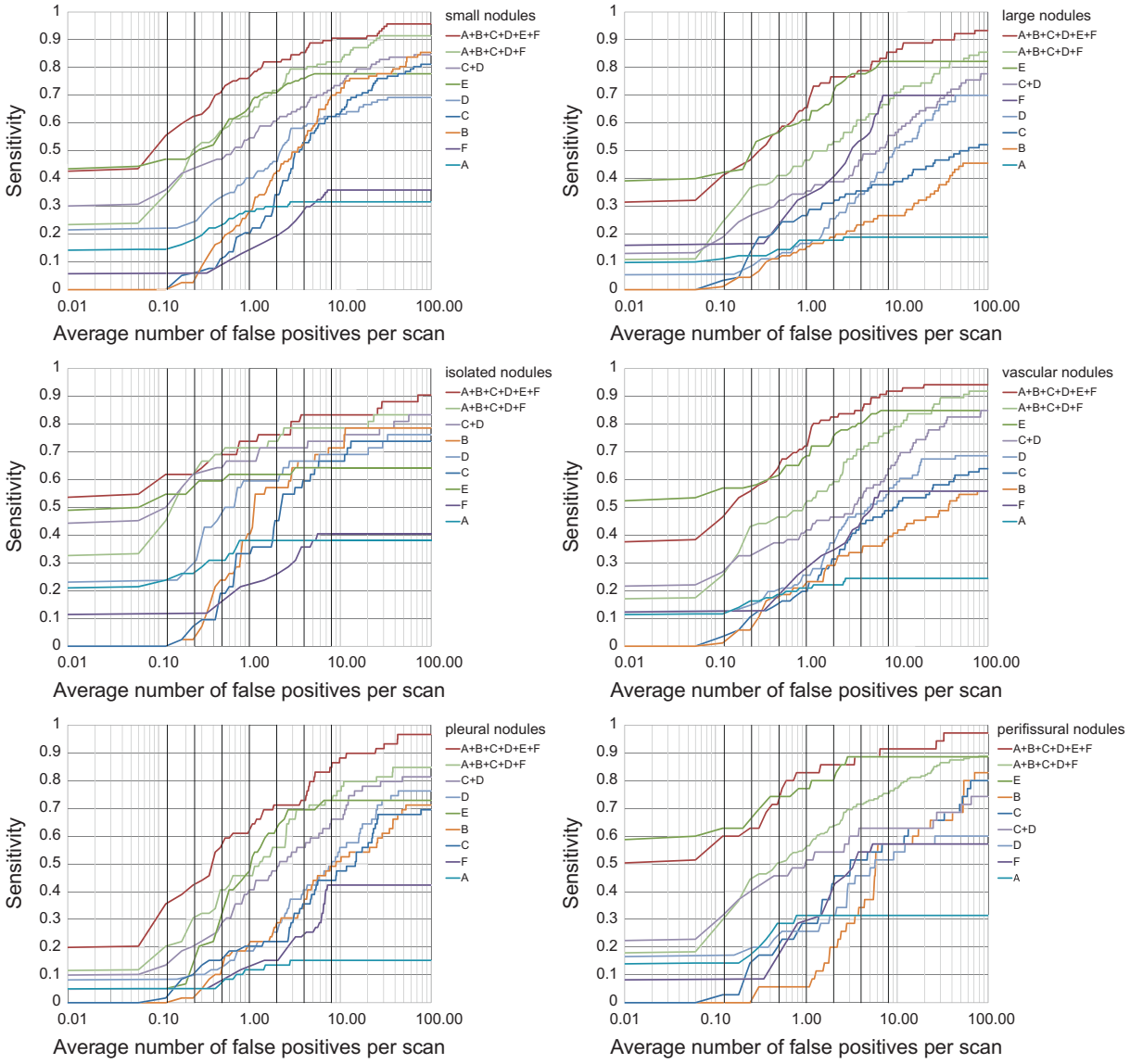

Fig. 3. FROC curves of all six systems and three combinations for each of the nodule categories (small, large, isolated, vascular, pleural and peri-fissural).

systems are quite different. Although a single system may fail to give individual nodules a high probability, or may not even detect them, if several methods analyze a scan very few nodules escape detection.

More sophisticated blending strategies could be devised than the one employed here. The technique we applied is similar to the averaging rule in classifier combination strategies (Kittler et al., 1998). It requires knowledge about the performance of a system on a reference database, in the form of an FROC curve. That curve is used in a look-up table to convert the degree of suspicion as reported by a system, which has an arbitrary scale, to an indication of the probability that a finding with that degree of suspicion or higher is truly a nodule in the reference database. This transformation is given by Eq. (1). The requirement that an FROC curve of each system is needed may seem a limiting factor to use this combination strategy in a clinical setting. However, we believe this is not the case. If an institution would have access to multiple CAD systems, the only thing that would be required is to establish a reference for a test set representative for that clinic. After running the CAD systems on that test set, which is a useful exercise to get a feeling for the performance of the CAD systems anyway, they can be readily combined using Eq. (1) and the algorithm in Section 4.7. Developing more effective ways to combine multiple CAD systems is a promising direction for future research. It should also be noted that the test database itself is used to measure the FROC curve of each individual system, needed to compute the combined results. This may have introduced statistical bias. Experiments with more complicated cross-validation procedures to estimate the transformation in Eq. (1) showed that this effect is small.

Although the combined system performs quite well, it is important to analyze what could be done to further improve results. Two approaches are possible: focus on further reduction of false positives at the left end of the FROC curve or improve sensitivity. For the latter it can be insightful to inspect the missed nodules at the right end of the FROC curve. We visually inspected nodules that were missed or only detected at very high false positive levels (Fig. 4, last row), and compared them with nodules that are detected at very low false positive levels (Fig. 4, middle row). The very suspicious nodules are indeed clear, prototypical examples of nodules. The difficult nodules were somewhat less conspicuous,

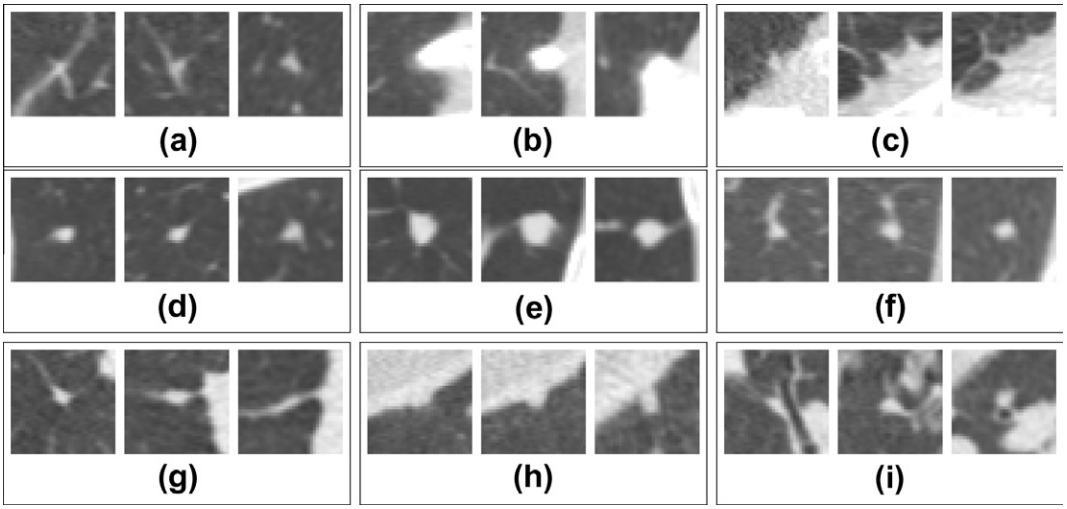

Fig. 4. Examples of false positives and easy and hard to detect nodules. In every box a finding is displayed in a sagittal, coronal and axial view, 35 voxels (approximately 25 mm) around the center point using a lung window (center −600 HU, width 1600 HU). The top row shows false positives with a high degree of suspicion in the combined system A + B + C + D + E + F: (a) is a point where multiple vessels meet as is especially apparent from the sagittal view; (b) is an apparent protrusion caused by bony structures close to the lung pleura; (c) is an apical scar. Many of these scars were listed as irrelevant findings, but this one is not very nodular in appearance and was not marked. The last two rows show actual nodules that were either very suspicious, and thus easy to detect (d–f) or very hard to detect (g–i).

but it was hard to detect any clearly identifiable characteristics among these cases that could be used as an inspiration to improve the performance of CAD systems.

To analyze the characteristics of the false positives, we visually inspected the 100 most suspicious FPs of the combined system. Although the variation among those findings was large, some broad categories could be discerned. It appeared that vessel branchings were the most common cause for false positives. This is in agreement with observations in Gori et al. (2007), Lee et al. (2005), Das et al. (2006) regarding the false positives produced by other commercial systems not included in this study. Interestingly, there were quite a few findings where two or more vessels seemed to be in contact with each other, forming a point that looked nodular to anyone not tracing the vessel tree through several sections. Fig. 4a shows an example. Especially on the sagittal view it is clear that multiple vessels meet at the location of the perceived nodule. An accurate dedicated vessel segmentation algorithm might be employed to reduce the occurrence of such false positives. Methods B and C include vessel segmentation and this may be one reason why they work well in combination with other methods.

Another common source of false positives are apparent protrusions mimicking pleural nodules at locations where high density bony structures, such as ribs, vertebrae and the sternum are close to the pleural surface of the lung or even pressing against it. An example is given in Fig. 4b. Using the output of a separate segmentation of the bony anatomy might prevent such false positives. Another common source of false positives are other lesions such as small scars.

Many false positives are small and this indicates that systems could eliminate them by adding an algorithm that segments the findings and discards findings that are below 4 mm, or gives them a lower degree of suspicion. It is likely that some systems were not designed to discard such small findings. This is suggested by the curves in Fig. 5 where the results of all systems are given if the relevant and irrelevant findings are swapped in the reference standard. Most irrelevant findings are nodules smaller than 4 mm in diameter, and Fig. 5 shows that some methods still detect quite a few of these very small nodules while others do not. At 1 FP per scan, all systems are more sensitive for the detection of relevant findings than for the detection of irrelevant ones.

The ANODE09 study is the first to compare and combine a large group of CAD systems for nodule detection on a single database, but the study also has some limitations. Most importantly, all data originates from a single hospital where all scans have been acquired with scanners from one manufacturer, with a single acquisition protocol. Moreover all scans are from subjects from a particular screening population. In clinical practice, CAD systems should be capable of operating with diverse input data. Also the reading protocol and characteristics of findings are particular to this study and this influences the reported results. For example, a hypothetical system that has been particularly designed to not display any markers on nodules under 5 mm diameter is clearly at a disadvantage, although it should obtain good results for the large nodule category. Fotin et al. (2008) proposed a different evaluation strategy where the implicit inaccuracy for measuring the size of smaller lesions is taken into account in the evaluation strategy. This has not been done in the current work.

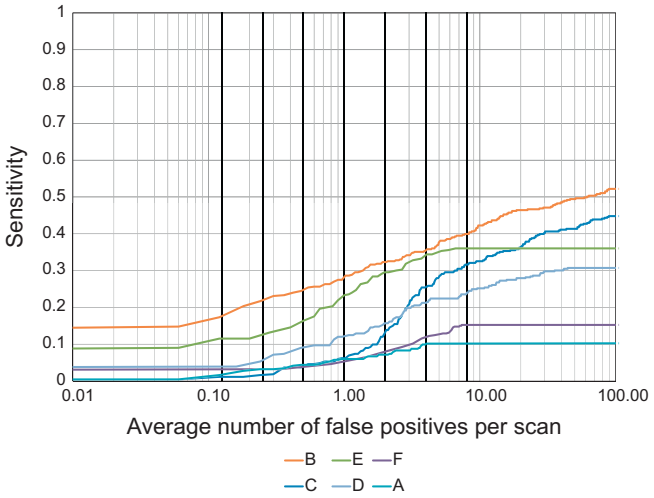

Fig. 5. FROC curves of all six systems for the detection of irrelevant findings. In this analysis the relevant findings are ignored (relevant and irrelevant findings have been switched) and therefore false positive levels are directly comparable to those in Fig. 2.

There are few large lesions in this data set and one could argue that those are actually the most important for a CAD system to detect as they are most likely to represent cancer. This would be especially true if CAD were used as a first reader or as a pre-screening system where it would select cases or locations to be inspected by radiologists. On the other hand, this is currently not the usual mode of operation for a CAD system and some radiologists actually prefer that CAD finds especially small and subtle nodules. They do not mind if some large and obvious nodules are missed, because they are able to find these themselves. How small and subtle those cases that CAD should detect should be will likely vary across users. Different users have different preferences. In this study, the discussion of which nodules a CAD system should detect is somewhat circumvented by the introduction of the category of irrelevant findings. In the future it would be interesting to repeat the study on a larger data set, containing more variety in nodules, and with data originating from multiple hospitals, different populations, multiple scanner types from different vendors and a reasonable variety of scanning protocols.

In this study we have not addressed the question of whether the output of the CAD system is actually beneficial for human experts. This has been researched extensively in clinical studies, and may be investigated in the future for the ANODE09 set.

Finally, only a small number of published and commercially available systems have been applied to the ANODE09 data set as of yet. We hope that in the future other groups will upload the results of their algorithms to help the research community in the identification of open challenges regarding this important CAD application.

7. Conclusions

A publicly available database and web-based framework for the evaluation of CAD algorithms for nodule detection in thoracic CT scans has been presented. The results of six algorithms are compared and combined. The introduction of irrelevant findings ensures that false positives of the algorithms are true errors and not nodules that do not meet the particular requirements of the study. A simple but effective method for the combination of various systems has been proposed. This combination method requires knowledge about the performance of the systems to be combined, in the form of an FROC curve on a data set with a known number of positive findings. Combining the findings of different systems appears to be a very powerful method to improve the performance of CAD systems. The combination of six CAD algorithms is able to detect 80% of all nodules at the expense of only two false positive detections per scan and 65% of all nodules with only 0.5 false positives. This suggests that blending detection algorithms is a promising direction for future research in CAD.

Acknowledgments

We would like to thank the NELSON study for making the ANODE09 data set available. We would like to thank the organizers of SPIE Medical Imaging for allowing us to organize a special session devoted to ANODE09 at the SPIE Medical Imaging 2009 conference. The Pisa team acknowledges Dr. F. Falaschi and Dr. C. Spinelli (U.O. Radiodiagnostica dell'Azienda Ospedaliera Universitaria Pisana), Prof. D. Caramella and Dr. M. Barattini (Divisione di Radiologia Diagnostica e Interventistica del Dipartimento di Oncologia, Trapianti e Nuove Tecnologie in Medicina dell'Università di Pisa), and Dr. M. Mattiuzzi (Bracco Imaging S.p.A.). The three Italian teams are grateful to all members of the MAGIC-5 Italian Collaboration funded by Istituto Nazionale di Fisica Nucleare (INFN) and Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR). The Gifu team would like to thank Dr. Shoji Okura and Dr. Takuya Tomida.

References

Arimura, H., Katsuragawa, S., Suzuki, K., Li, F., Shiraishi, J., Sone, S., Doi, K., 2004. Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening. Academic Radiology 11 (6), 617–629.

Armato, S.G., Giger, M.L., MacMahon, H., 2001. Automated detection of lung nodules in CT scans: preliminary results. Medical Physics 28 (8), 1552–1561.

Armato, S.G., Li, F., Giger, M.L., MacMahon, H., Sone, S., Doi, K., 2002. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. Radiology 225 (3), 685–692.

Armato, S.G., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Yankelevitz, D., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Reeves, A.P., Croft, B.Y., Clarke, L.P., 2004. Lung image database consortium: developing a resource for the medical imaging research community. Radiology 232 (3), 739–748.

Armato, S.G., McNitt-Gray, M.F., Reeves, A.P., Meyer, C.R., McLennan, G., Aberle, D.R., Kazerooni, E.A., MacMahon, H., van Beek, E.J.R., Yankelevitz, D., Hoffman, E.A., Henschke, C.I., Roberts, R.Y., Brown, M.S., Engelmann, R.M., Pais, R.C., Piker, C.W., Qing, D., Kocherginsky, M., Croft, B.Y., Clarke, L.P., 2007. The lung image database consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Academic Radiology 14 (11), 1409–1421.

Armato, S.G., Roberts, R.Y., Kocherginsky, M., Aberle, D.R., Kazerooni, E.A., Macmahon, H., van Beek, E.J.R., Yankelevitz, D., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Caligiuri, P., Quint, L.E., Sundaram, B., Croft, B.Y., Clarke, L.P., 2009. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". Academic Radiology 16 (1), 28–38.

Austin, J.H., Müller, N.L., Friedman, P.J., Hansell, D.M., Naidich, D.P., Remy-Jardin, M., Webb, W.R., Zerhouni, E.A., 1996. Glossary of terms for CT of the lungs: recommendations of the nomenclature committee of the Fleischner society. Radiology 200 (2), 327–331.

Bae, K.T., Kim, J.S., Na, Y.H., Kim, K.G., Kim, J.H., 2005. Pulmonary nodules: automated detection on CT images with morphologic matching algorithm – preliminary results. Radiology 236, 286–294.

Bellotti, R., De Carlo, F., Gargano, G., Tangaro, S., Cascio, D., Catanzariti, E., Cerello, P., Cheran, S.C., Delogu, P., De Mitri, I., Fulcheri, C., Grosso, D., Retico, A., Squarcia, S., Tommasi, E., Golosio, Bruno, 2007. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. Medical Physics 34 (12), 4901–4910.

Bellotti, R., Cerello, P., Tangaro, S., Bevilacqua, V., Castellano, M., Mastronardi, G., De Carlo, F., Bagnasco, S., Bottigli, U., Cataldo, R., Catanzariti, E., Cheran, S.C., Delogu, P., De Mitri, I., De Nunzio, G., Fantacci, M.E., Fauci, F., Gargano, G., Golosio, B., Indovina, P.L., Lauria, A., Lopez Torres, E., Magro, R., Masala, G.L., Massafra, R., Oliva, P., Preite Martinez, A., Quarta, M., Raso, G., Retico, A., Sitta, M., Stumbo, S., Tata, A., Squarcia, S., Schenone, A., Molinari, E., Canesi, B., 2007. Distributed medical images analysis on a grid infrastructure. Future Generation Computer Systems 23 (3), 475–484.

Brown, M.S., Goldin, J.G., Suh, R.D., McNitt-Gray, M.F., Sayre, J.W., Aberle, D.R., 2003. Lung micronodules: automated method for detection at thin-section CT—initial experience. Radiology 226, 256–262.

Buscema, P.M., 2004. A method for encoding image pixels, a method for processing images and a method for processing images aimed at qualitative recognition of the object reproduced by one or more image pixels. Patent EP1483721.

Cerello, P., Cheran, S.C., Bagagli, F., Bagnasco, S., Bellotti, R., Bolanos, L., Catanzariti, E., De Nunzio, G., Fiorina, E., Gargano, G., Gemme, G., Lopez Torres, E., Masala, G., Peroni, C., Santoro, M., 2008. The channeler ant model: object segmentation with virtual ant colonies. In: IEEE Nuclear Science Symposium, pp. 3147–3152.

Das, M., Mühlenbruch, G., Mahnken, A.H., Flohr, T.G., Gündel, L., Stanzel, S., Kraus, T., Günther, R.W., Wildberger, J.E., 2006. Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance. Radiology 241 (2), 564–571.

Dehmeshki, J., Ye, X., Lin, X., Valdivieso, M., Amin, H., 2007. Automated detection of lung nodules in CT images using shape-based genetic algorithm. Computerized Medical Imaging and Graphics 31 (6), 408–417.

de Hoop, B., Gietema, H., van Ginneken, B., Zanen, P., Groenewegen, G., Prokop, M., 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. European Radiology 19 (4), 800–808.

Dolejsí, M., Kybic, J., 2009. The lung TIME annotated lung nodule dataset and nodule detection framework. In: Proceedings of the SPIE, vol. 7260, pp. 72601U1–72601U8.

Enquobahrie, A.A., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I., 2007. Automated detection of small pulmonary nodules in whole lung CT scans. Academic Radiology 14, 579–593.

Farag, A., El-Baz, A., Gimel'farb, G.G., Falk, R., Hushek, S.G., 2004. Automatic detection and recognition of lung abnormalities in helical CT images using deformable templates. In: Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, vol. 3217, pp. 856–864.

Fotin, S.V., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I., 2008. The impact of pulmonary nodule size estimation accuracy on the measured performance of automated nodule detection systems. In: Proceedings of the SPIE, vol. 6915, pp. 6915G1–6915G8.

Fotin, S.V., Reeves, A.P., Biancardi, A.M., Yankelevitz, D.F., Henschke, C.I., 2009. A multiscale Laplacian of Gaussian filtering approach to automated pulmonary nodule detection from whole-lung low-dose CT scans. In: Proceedings of the SPIE, vol. 7260, pp. 72601Q1–72601Q8.

Ge, Z., Sahiner, B., Chan, H., Hadjiiski, L.M., Cascade, P.N., Bogot, N., Kazerooni, E.A., Wei, J., Zhou, C., 2005. Computer-aided detection of lung nodules: false positive reduction using a 3D gradient field method and 3D ellipsoid fitting. Medical Physics 32 (8), 2443–2454.

Godoy, M.C.B., Cooperberg, P.L., Maizlin, Z.V., Yuan, R., McWilliams, A., Lam, S., Mayo, J.R., 2008. Detection sensitivity of a commercial lung nodule CAD system in a series of pathologically proven lung cancers. Journal of Thoracic Imaging 23 (1), 1–6.

Gohagan, J., Marcus, P., Fagerstrom, R., Pinsky, P., Kramer, B., Prorok, P., 2004. Baseline findings of a randomized feasibility trial of lung cancer screening with spiral CT scan vs. chest radiograph: the lung screening study of the national cancer institute. Chest 126 (1), 114–121.

Gori, I., Mattiuzzi, M., 2008. Method for coding pixels or voxels of a digital image and a method for processing digital images. Patent EP1880364.

Gori, I., Bagagli, F., Fantacci, M.E., Preite Martinez, A., Retico, A., De Mitri, I., Donadio, S., Fulcheri, C., Gargano, G., Magro, R., Santoro, M., Stumbo, S., 2007. Multi-scale analysis of lung computed tomography images. Journal of Instrumentation 2 (09), P09007.

Gori, I., Fantacci, M.E., Preite Martinez, A., Retico, A., 2007b. An automated system for lung nodule detection in low-dose computed tomography. In: Proceedings of the SPIE, vol. 6514, pp. 65143R1–65143R9.

Gori, I., Bagagli, F., Camarlinghi, N., Fantacci, M.E., Retico, A., Barattini, M., Bolanos, L., Falaschi, F., Gargano, G., Massafra, A., Spinelli, C., 2009. Methodology for automated detection of parenchymal and juxtapleural lung nodules in computed tomography images. In: Proceedings of CARS.

Gruden, J.F., Ouanounou, S., Tigges, S., Norris, S.D., Klausner, T.S., 2002. Incremental benefit of maximum-intensity-projection images on observer detection of small pulmonary nodules revealed by multidetector CT. American Journal of Roentgenology 179 (1), 149–157.

Heimann, T., van Ginneken, B., Styner, M., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P., Chi, Y., Cordova, A., Dawant, B., Fidrich, M., Furst, J., Furukawa, D., Grenacher, L., Hornegger, J., Kainmuller, D., Kitney, R., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S., Meinzer, H-P., Nemeth, G., Raicu, D., Rau, A-M., van Rikxoort, E., Rousson, M., Rusko, L., Saddi, K., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen, P., Sorantin, E., Soza, G., Susomboon, R., Waite, J., Wimmer, A., Wolf, I., 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Transactions on Medical Imaging 28 (8), 1251–1265.

Henschke, C.I., the International Early Lung Cancer Action Program Investigators, 2007. Survival of patients with clinical stage I lung cancer diagnosed by computed tomography screening for lung cancer. Clinical Cancer Research 13 (17), 4949–4950.

Henschke, C.I., Yankelevitz, D.F., Naidich, D.P., McCauley, D.I., McGuinness, G., Libby, D.M., Smith, J.P., Pasmantier, M.W., Miettinen, O.S., 2004. CT screening for lung cancer: suspiciousness of nodules according to size on baseline scans. Radiology 231 (1), 164–168.

Infante, M., Cavuto, S., Romano Lutman, F., Brambilla, G., Chiesa, G., Ceresoli, G., Passera, E., Angeli, E., Chiarenza, M., Aranzulla, G., Cariboni, U., Errico, V., Inzirillo, F., Bottoni, E., Voulaz, E., Alloisio, M., Destro, A., Roncalli, M., Santoro, A., Ravasi, 2009. A randomized study of lung cancer screening with spiral CT (the Dante trial): three-year results. American Journal of Respiratory and Critical Care Medicine (June 11. Epub ahead of print).

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3), 226–239.

Ko, J.P., Betke, M., 2001. Chest CT: automated nodule detection and assessment of change over time-preliminary experience. Radiology 218 (1), 267–273.

Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I., 2003. Three-dimensional segmentation and growth rate estimation of small pulmonary nodules in helical CT images. IEEE Transactions on Medical Imaging 22 (10), 1259–1274.

Kung, J.W., Matsumoto, S., Hasegawa, I., Nguyen, B., Toto, L.C., Kundel, H., Hatabu, H., 2004. Mixture distribution analysis of a computer assisted diagnostic method for the evaluation of pulmonary nodules on computed tomography scan. Academic Radiology 11 (3), 281–285.

Lee, Y., Hara, T., Fujita, H., Itoh, S., Ishigaki, T., 2001. Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique. IEEE Transactions on Medical Imaging 20 (7), 595–604.

Lee, I.J., Gamsu, G., Czum, J., Wu, N., Johnson, R., Chakrapani, S., 2005. Lung nodule detection on chest CT: evaluation of a computer-aided detection (CAD) system. Korean Journal of Radiology 6 (2), 89–93.

Li, Q., 2007. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Computerized Medical Imaging and Graphics 31 (4-5), 248–257.

Li, Q., Sone, S., Doi, K., 2003. Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. Medical Physics 30 (8), 2040–2051.

Li, Q., Li, F., Doi, K., 2008. Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier. Academic Radiology 15 (2), 165–175.

MacMahon, H., Austin, J.H.M., Gamsu, G., Herold, C.J., Jett, J.R., Naidich, D.P., Patz, E.F., Swensen, S.J.the Fleischner Society, 2005. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner society. Radiology 237 (2), 395–400.

Matsumoto, S., Kundel, H.L., Gee, J.C., Gefter, W.B., Hatabu, H., 2006. Pulmonary nodule detection in CT images with quantized convergence index filter. Medical Image Analysis 10 (3), 343–352.

McCulloch, C.C., Kaucic, R.A., Mendonça, P.R.S., Walter, D.J., Avila, R.S., 2004. Model-based detection of lung nodules in computed tomography exams. Academic Radiology 11 (3), 258–266.

Mendonça, P.R.S., Bhotika, R., Zhao, F., Miller, J.V., 2007. Lung nodule detection via Bayesian voxel labeling. In: Information Processing in Medical Imaging. Lecture Notes in Computer Science, vol. 4584, pp. 134–146.

Murphy, K., Schilham, A.M.R., Gietema, H., Prokop, M., van Ginneken, B., 2007. Automated detection of pulmonary nodules from low-dose computed tomography scans using a two-stage classification system based on local image features. In: Proceedings of the SPIE, vol. 6514, pp. 651410-1–651410-12.

Murphy, K., van Ginneken, B., Schilham, A.M.R., de Hoop, B.J., Gietema, H.A., Prokop, M., 2009. A large scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Medical Image Analysis 13, 757–770.

Novak, C.L., Shen, H., Odry, B.L., Ko, J.P., Naidich, D.P., 2004. A system for automatic detection of lung nodules exhibiting growth. In: Proceedings of the SPIE, vol. 5370, pp. 10–21.

Ochs, R., Kimb, H.J., Angel, E., Panknin, C., McNitt-Gray, M., Brown, M., 2007. Forming a reference standard from LIDC data: impact of reader agreement on reported CAD performance. In: Proceedings of the SPIE, vol. 6514, pp. 65142A-1–65142A-6.

Receiver Operating Characteristic Analysis in Medical Imaging. Journal of the ICRU 8.

Osman, O., Ozekes, S., Ucan, O.N., 2007. Lung nodule diagnosis using 3D template matching. Computers in Biology and Medicine 37 (8), 1167–1172.

Paik, D.S., Beaulieu, C.F., Rubin, G.D., Acar, B., Jeffrey Jr., R.B., Yee, J., Dey, J., Napel, S., 2004. Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. IEEE Transactions on Medical Imaging 23 (6), 661–675.

Pegna, A.L., Picozzi, G., Mascalchi, M., Carozzi, F.M., Carrozzi, L., Comin, C., Spinelli, C., Falaschi, F., Grazzini, M., Innocenti, F., Ronchi, C., Paci, E., 2009. Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung Cancer 64 (1), 34–40.

Peloschek, P., Sailer, J., Weber, M., Herold, C.J., Prokop, M., Schaefer-Prokop, C.M., 2007. Pulmonary nodules: sensitivity of maximum intensity projection versus that of volume rendering of 3D multidetector CT data. Radiology 243 (2), 561–569.

Retico, A., Delogu, P., Fantacci, M.E., Gori, I., Preite Martinez, A., 2008. Lung nodule detection in low-dose and thin-slice computed tomography. Computers in Biology and Medicine 38 (4), 525–534.

Retico, A., Bagagli, F., Camarlinghi, N., Carpentieri, C., Fantacci, M.E., Gori, I., 2009. A voxel-based neural approach (VBNA) to identify lung nodules in the ANODE09 study. In: Medical Imaging 2009: Computer-Aided Diagnosis, vol. 7260. SPIE, Lake Buena Vista, FL, USA, pp. 72601S–8.

Schneider, C., Amjadi, A., Richter, A., Fiebich, M., 2009. Automated lung nodule detection and segmentation. In: Proceedings of the SPIE, vol. 7260, pp. 72601T1–72601T8.

Sluimer, I.C., Schilham, A.M.R., Prokop, M., van Ginneken, B., 2006. Computer analysis of computed tomography scans of the lung: a survey. IEEE Transactions on Medical Imaging 25 (4), 385–405.

Suzuki, K., Armato, S.G., Li, F., Sone, S., Doi, K., 2003. Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. Medical Physics 30 (7), 1602–1617.

van den Bergh, K.A.M., Essink-Bot, M.-L., Bunge, E.M., Scholten, E. Th., Prokop, M., van Iersel, C.A., van Klaveren, R.J., de Koning, H.J., 2008. Impact of computed tomography screening for lung cancer on participants in a randomized controlled trial (NELSON trial). Cancer 113 (2), 396–404.

van Iersel, C.A., de Koning, H.J., Draaisma, G., Mali, W.P.T.M., Scholten, E. Th., Nackaerts, K., Prokop, M., Habbema, J.D.F., Oudkerk, M., van Klaveren, R.J., 2006. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). International Journal of Cancer 120 (4), 868–874.

van Rikxoort, E.M., de Hoop, B., Viergever, M.A., Prokop, M., van Ginneken, B., 2009. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Medical Physics 36 (7), 2934–2947.

Wei, G.-Q., Fan, L., Qian, J., 2002. Automatic detection of nodules attached to vessels in lung CT by volume projection analysis. In: Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, vol. 2488, pp. 746–752.

Wiemker, R., Rogalla, P., Zwartkruis, A., Blaffert, T., 2002. Computer aided lung nodule detection on high resolution CT data. In: Proceedings of the SPIE, vol. 4684, pp. 677–688.

Wiemker, R., Rogalla, P., Blaffert, T., Sifri, D., Hay, O., Shah, E., Truyen, R., Fleiter, T., 2005. Aspects of computer-aided detection (CAD) and volumetry of pulmonary nodules using multislice CT. British Journal of Radiology 78 (1), S46–S56.

Xu, D.M., Gietema, H., de Koning, H., Vernhout, R., Nackaerts, K., Prokop, M., Weenink, C., Lammers, J., Groen, H., Oudkerk, M., van Klaveren, R., 2006. Nodule

management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 54 (2), 177–184.

Ye, X., Lin, X., Beddoe, G., Dehmeshki, J., 2007. Efficient computer-aided detection of ground-glass opacity nodules in thoracic CT images. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, vol. 1, pp. 4449–4452.

Zhang, X., Stockel, J., Wolf, M., Cathier, P., McLennan, G., Hoffman, E.A., Sonka, M., 2007. A new method for spherical object detection and its application to computer aided detection of pulmonary nodules in CT images. In: Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science, vol. 4791, pp. 842–849.

Zhao, B., Gamsu, G., Ginsberg, M.S., Jiang, L., Schwartz, L.H., 2003. Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm. Journal of Applied Clinical Medical Physics 4 (3), 248–260.