Copyright 2008 Society of Photo-Optical Instrumentation Engineers.
This paper was published in Proceedings of SPIE, vol. 6915, Medical Imaging 2008: Computer Aided Diagnosis and is made available as an electronic reprint with permission of SPIE. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

The impact of pulmonary nodule size estimation accuracy on the measured performance of automated nodule detection systems

Sergei V. Fotin[a], Anthony P. Reeves[a],
David F. Yankelevitz[b] and Claudia I. Henschke[b]

[a]School of Electrical and Computer Engineering, Cornell University,
Ithaca, NY 14853, USA;
[b]Department of Radiology, New York-Presbyterian Hospital-Weill Cornell Medical Center,
New York, NY 10021, USA

ABSTRACT

The performance of automated pulmonary nodule detection systems is typically qualified with respect to some minimum size of nodule to be detected. Also, an evaluation dataset is typically constructed by expert radiologists with all nodules larger than the minimum size being designated as true positives while all other smaller detected "nodules" are considered to be false positives. In this paper, we consider the negative impact that size estimation error, either in the establishment of ground truth for the evaluation dataset or by the automated detection method for the size estimate of nodule candidates, has on the measured performance of the detection system. Furthermore, we propose a modified evaluation procedure that addresses the size estimation error issue.

The impact of the size measurement error was estimated for a documented research image database consisting of whole-lung CT scans for 509 cases in which 690 nodules have been documented. We compute FROC curves both with and without size error compensation and we found that for a minimum size limit of 4 mm the performance of the system is underestimated by a sensitivity reduction of 5% and a false positive rate increase of 0.25 per case. Therefore, error in nodule size estimation should be considered in the evaluation of automated detection systems.

Keywords: Automated nodule detection, algorithm evaluation and validation, performance measurement, computer-assisted diagnosis (CAD), CT

1. INTRODUCTION

In the domain of computerized detection, the trade-off between sensitivity and number of raised false positives is an important indicator of an algorithm's performance.[1,2] Both metrics are derived from a comparison of the expert decision, the ground truth, to the decision of the algorithm. The aim for the majority of automated detection systems is to approach the ground truth: to maximize sensitivity and to minimize false positive rate. However, in nodule detection, nodule size is an important factor that affects the ground truth and complicates performance evaluation and comparison of various automated systems.

For any documented set of scans of pulmonary nodules, there is a minimum size limit due to image noise and reconstruction resolution below which nodules cannot be reliably seen. Very small nodules may be barely distinguishable on a CT scan and are likely to be missed or confused with image artifacts during radiological inspection. Therefore, even though the natural distribution of nodules in a general population contains far more small nodules than large nodules, not all of the small nodules are documented.

To illustrate this issue we may take a look at the histogram of nodule sizes in the subset of Weill Cornell Medical Center database shown in Figure 1. Notice that the shape of the graph has a single distinct peak in a nodule size range between 2 and 3 mm. In theory, we might expect to have increasing number of nodules to the left of the peak (as illustrated by the curve, connecting tops of the histogram bins), however only few of

Send correspondence to Sergei V. Fotin, e-mail: svf3@cornell.edu, phone: 1 607 255 0963

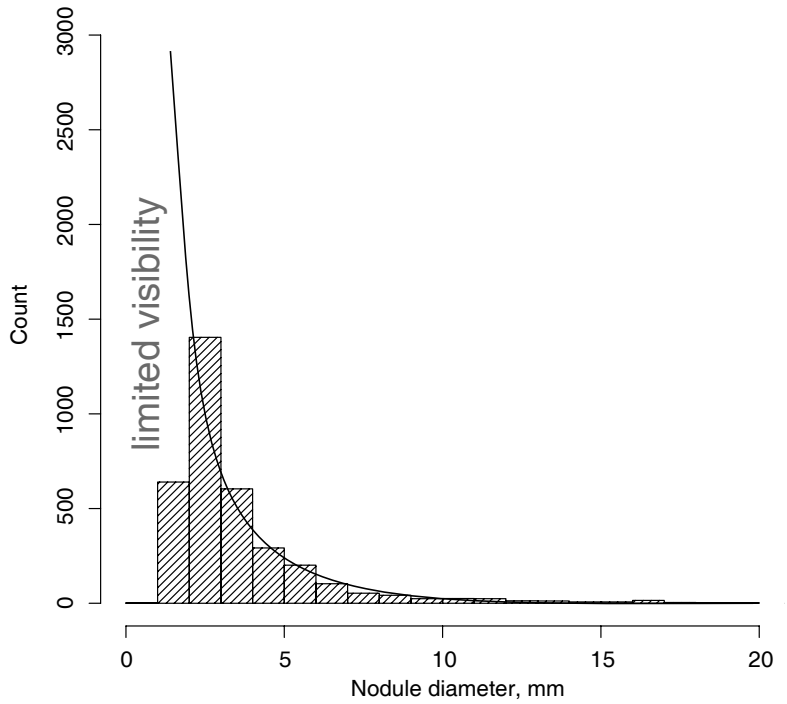


Figure 1. Histogram of diameters in Weill Cornell Medical Center database of 3503 nodules. The curve connecting the tops of histogram bars extrapolates to the small size range, showing the effect of limited visibility.

them are actually documented. It is important to note that the position of the histogram peak would greatly depend on image acquisition parameters, screening population and nodule documentation strategy and will vary for different datasets and institutions.

For practical applications of detection systems, a minimum size threshold is usually specified where the visibility is reliable and where a specific action will be taken for a detected nodule of that size; that is, this minimum size depends on the clinical protocol that is supported by the detection system. Consequently, detection system developers report the performance with respect to a predefined size range.[3–10] In a clinical environment, the size of a nodule is usually estimated by an expert using one of the standardized measurement procedures; however, even for a single size metric, the disagreement between various readers may be quite large.[11–13]

The task for the detection system is to detect only nodules with the size above this threshold and disregard smaller objects. Thus, the output of the algorithm, which is usually represented by a set of detected nodule candidates, should be separated into two size categories according to the estimated size. However, the disagreement between automatically measured and ground truth size values alters the performance evaluation subset and, subsequently, changes reported detection performance. Further, we will show how this disagreement affects both the detection sensitivity and the false positive rate of a detection system.

2. METHOD

Given a specific size cut-off threshold, the difference between the nodule size estimation by a radiologist providing the ground truth and the automated candidate generation system complicates the procedure of detection performance evaluation. To measure the performance of the system only on nodules above certain threshold D_{th}, the detection algorithm needs to be configured to disregard smaller nodule candidates.

In the hypothetical situation illustrated in Figure 2, when there is no disagreement between ground truth and automatic measurements, D_{th} clearly separates nodules and candidates into two size categories. Here the sensitivity $S_{x \geq D_{th}}$ and false positive rate $F_{x \geq D_{th}}$ of the system are easily calculated in terms of numbers of true

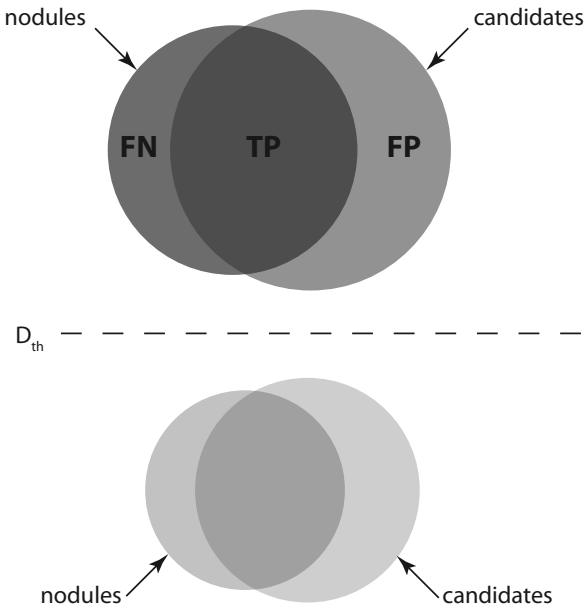


Figure 2. Ideal situation of zero size measurement error. Size threshold clearly separates nodules and candidates into two size categories.

and false candidates with the size x above the threshold D_{th}, while the smaller candidates are ignored:

$$S_{x \geq D_{th}} = \frac{N_{TP}}{N_{TP} + N_{FN}} = \frac{N_{TP}}{N_{x \geq D_{th}}}, \tag{1}$$

$$F_{x \geq D_{th}} = \frac{N_{FP}}{N_{cases}}, \tag{2}$$

where $N_{x \geq D_{th}}$ is the total number of "large" nodules and N_{cases} is the number of cases.

In a non-ideal world, the difference in size estimation leads to situations where some nodule sizes are either underestimated or overestimated. If a nodule of size $x_1 < D_{th}$ has corresponding candidate size $c(x_1) \geq D_{th}$, it must be considered as a false positive. Also, if a nodule of size $x_2 \geq D_{th}$ is measured as having size $c(x_2) < D_{th}$, it will be filtered out by the detection algorithm and recorded as false negative. This effect is illustrated in Figure 3 and definitions of true and false positives and negatives are shown in Table 1. In this case sensitivity of the detection system can be calculated as (strokes denote variables related to the non-ideal situation):

$$S'_{x \geq D_{th}} = \frac{N'_{TP}}{N_{x \geq D_{th}}} = \frac{N_{TP} - N'_{FN(UE)}}{N_{x \geq D_{th}}}. \tag{3}$$

If we denote the probability for nodule of fixed size x to be underestimated in size as $P(c(x) < D_{th})$, we can express the number of false negatives due to size underestimate using the conditional mean:

$$N'_{FN(UE)} = N_{TP} \cdot E\left[P(c(x) < D_{th}) \;\middle|\; x \geq D_{th}\right]. \tag{4}$$

After the substitutions, we get the following expression for the non-ideal sensitivity:

$$S'_{x \geq D_{th}} = S_{x \geq D_{th}} \cdot \left(1 - E\left[P(c(x) < D_{th}) \;\middle|\; x \geq D_{th}\right]\right). \tag{5}$$

Then, if we assume that for a nodule of size x there exists a random size measurement error $e(x)$ with the probability density function $f_{e(x)}$, so that $c(x) = x + e(x)$, then we can find the probability of the size underestimate:

$$P(c(x) < D_{th}) = P(e(x) < D_{th} - x) = \int_{-\infty}^{D_{th}-x} f_{e(x)}(z)\, dz. \tag{6}$$

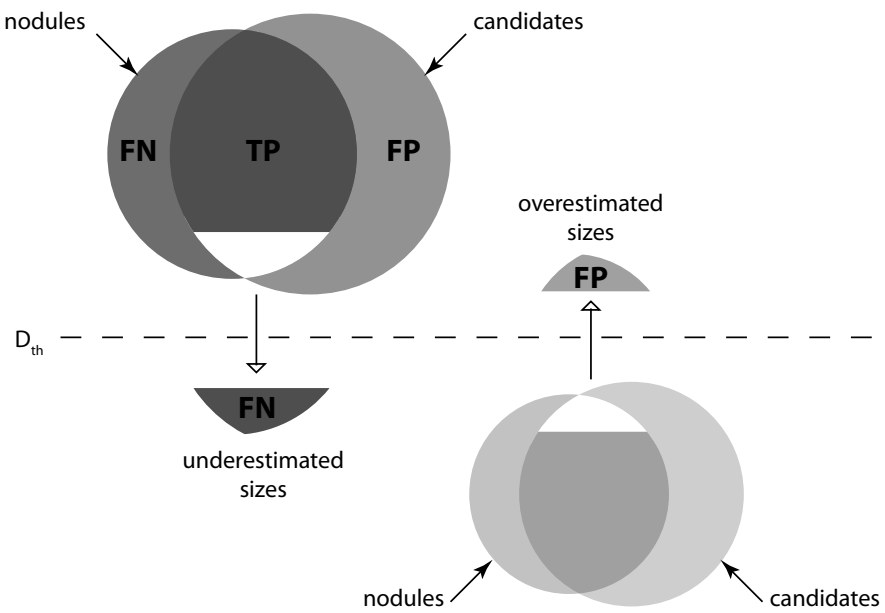


Figure 3. Diagram of nodule-candidate correspondence with the presence of non-zero size measurement error. Diameter threshold D_{th} separates nodules and candidates into two subsets, however because of the measurement error, some detected nodules have either underestimated or overestimated candidate size.

Table 1. Definition of true and false positives for a nodule detection system with a minimum nodule size limit of D_{th}. "OE" and "UE" stand for size overestimate and underestimate respectively.

		Nodule state		
		not exists	$x < D_{th}$	$x \geq D_{th}$
Candidate state	$c(x) \geq D_{th}$	FP	FP(OE)	TP
	$c(x) < D_{th}$	-	-	FN(UE)
	not exists	-	-	FN

Finally, considering that the size of the nodule x is drawn from the distribution f_x, the value of the sensitivity in the presence of measurement error can be expressed as:

$$S'_{x \geq D_{th}} = S_{x \geq D_{th}} \cdot \left(1 - \frac{\int_{D_{th}}^{\infty} \int_{-\infty}^{D_{th}-x} f_{e(x)}(z)\, dz \cdot f_x(x)\, dx}{\int_{D_{th}}^{\infty} f_x(x)\, dx} \right). \tag{7}$$

Similarly, we can calculate the fraction of nodules overestimated in size that should be counted as false positives. We see that, besides the natural distribution of nodules f_x, this fraction depends on the distribution of measurement error $f_{e(x)}$. An increase in this function variance decreases sensitivity and increases the false positive rate of the detection system.

There always will be discrepancy in size estimates due to real nodule size uncertainty and critical differences in human and machine measurements. We believe that this discrepancy, if small, should not lessen reported performance of the detection system. For this reason, we propose a Δ-size tolerance range method to compensate

Table 2. Compensation for size uncertainty using Δ radius for candidate sizes.

		Nodule state		
		not exists	$x < D_{th}$	$x \geq D_{th}$
Candidate state	$c(x) \geq D_{th} + \Delta$	FP	FP(OE)	TP
	$D_{th} \leq c(x) < D_{th} + \Delta$	FP	-	TP
	$D_{th} - \Delta \leq c(x) < D_{th}$	-	-	TP
	$c(x) < D_{th} - \Delta$	-	-	FN(UE)
	not exists	-	-	FN

for the size measurement error between the human expert and the computer system in which we do not count the candidates with sizes in the immediate proximity of a cut-off threshold as either false positives or false negatives. This concept is illustrated in Table 2. Given a small specific value for Δ, we treat a candidate as overestimated or underestimated only if its size lies outside the interval bounded by points: $D_{th} - \Delta$ and $D_{th} + \Delta$. By applying this technique, we directly reduce the chances for a nodule and its corresponding candidate to be on opposite sides of the cut-off threshold and increase reported sensitivity and reduce false positive rate.

In the case of Δ-compensation, the probability of the size underestimate with compensation for uncertainty becomes:

$$P(c(x) < D_{th} - \Delta) = \int_{-\infty}^{D_{th}-x-\Delta} f_{e(x)}(z)\, dz, \tag{8}$$

and as Δ approaches infinity, the fraction of false negatives due to size underestimate reduces to zero. As the result, sensitivity of such a detection system is equal to the one obtained with perfect ideal measurement:

$$S'_{x \geq D_{th},\ \Delta \to \infty} = S_{x \geq D_{th}}. \tag{9}$$

Generally any value of Δ greater than the maximum difference in measurement error would result in the same (maximum) value of sensitivity. It can be shown that larger values of Δ have a favorable influence on the false positive rate of a detection system as well.

However, it is not desirable to use large values for Δ, since the candidate size information would be completely disregarded. The candidates that have large size measurement disagreement and clearly lie outside the size range of interest should be not be counted as true positives.

To illustrate the concept of Δ-compensation, let us consider a hypothetical situation of a detection system that detects nodules of the size 4 mm and above. If the system generated a candidate of size 4.2 mm for a 3.8 mm nodule (as marked in the ground truth) this nodule is not counted as false positive, because human measurement could be done with error and, in fact, the nodule had the size of 4 mm. Similarly, when the system reports that a nodule has size 4.2 mm, while it was recorded in the ground truth as having size 3.8 mm, it is not counted as false negative. In the opposite situation, if a detection system reports a size of 4.2 mm on a 1 mm nodule, we must register a false positive, since the detection system has made a mistake and this nodule is clearly outside of the range of interest. Accordingly, if it reports 1 mm on a 4.2 mm nodule we must register a false negative. In order to implement this scheme, one may need to set a certain range of tolerance Δ for the difference between nodule and candidate sizes relative to the size cut-off threshold. In this case the value $\Delta = 1\ mm$ would be appropriate.

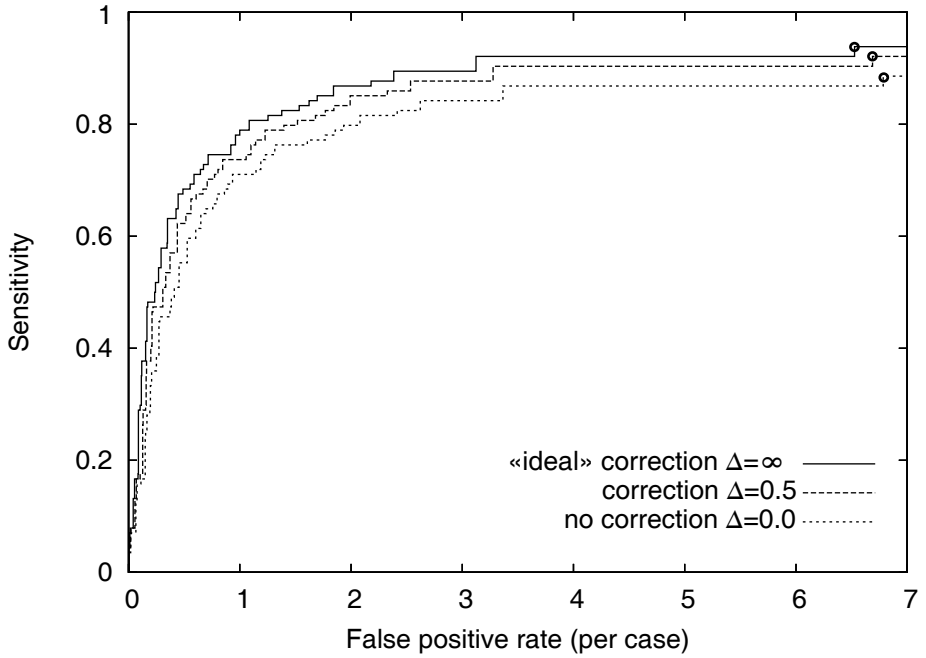


Figure 4. Effect of Δ on the FROC for detection of solid isolated nodules. The operating point, corresponding to the maximum sensitivity is denoted by circles.

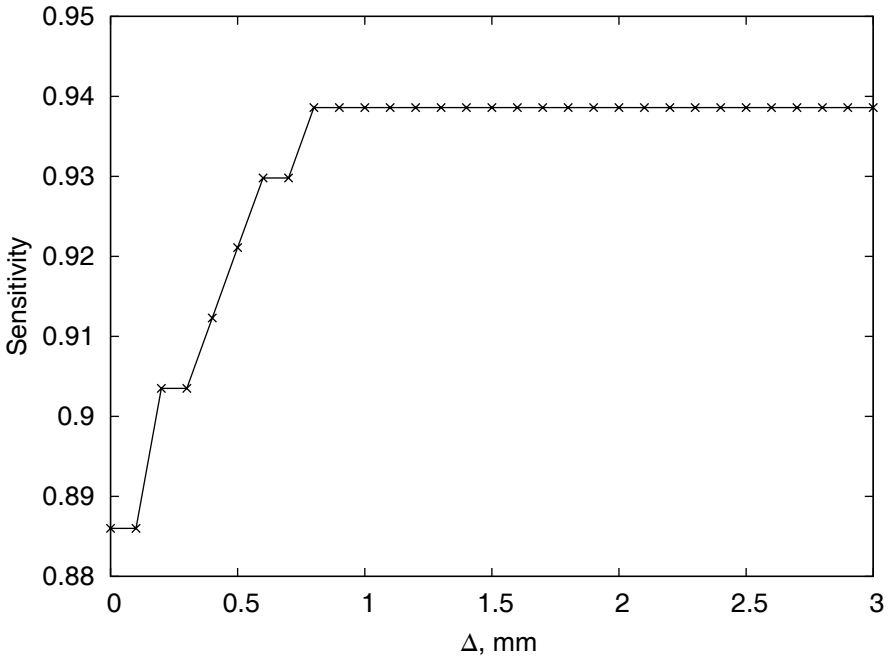
3. EXPERIMENT

In order to evaluate the impact of size measurement disagreement on detection performance, we used an experimental automated nodule detection system that targets solid isolated nodules.[4] The evaluation set consisted of 509 clinical cases from Weill Cornell Medical Center database in which 690 solid isolated nodules were identified and measured by at least two expert radiologists. In the ground truth nodule sizes were recorded as the average between the maximum axial diameter and its largest perpendicular. We selected the size cut-off threshold to be 4 mm and trained the system on the subset of 249 clinical cases containing 323 nodules. This system was then tested on the remaining 260 cases containing 367 nodules. We used a modified version of a performance evaluation procedure that allowed us to vary Δ and compute the resulting FROC over the test dataset. When $\Delta = 0$, the performance is the same as for the standard procedure that does not compensate for size measurement uncertainty. $\Delta = \infty$ results in the performance that would have been obtained by the detection system if the nodule size was measured perfectly without disagreement. The values of Δ in between corresponded to different degrees of compensation. Moreover, we picked the operating point on the FROC curve, that corresponds to the maximum sensitivity of detection and tracked the effect of Δ on the maximum sensitivity and false positive rate corresponding to this point.

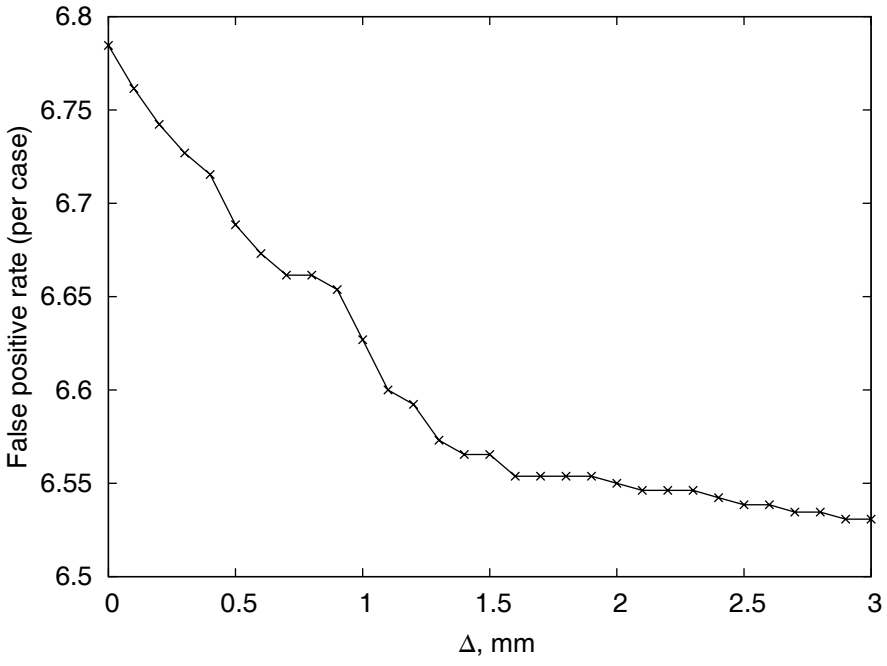
4. RESULTS

FROC curves were constructed for a documented dataset of 509 cases with the minimum size threshold set to 4 mm. Three curves, shown in Figure 4, correspond to different values of Δ compensation. The baseline performance corresponds to the case when there is no correction for size uncertainty, i.e. $\Delta = 0$. The best performance was recorded at $\Delta = \infty$, when we entirely ignore candidate size in performance evaluation. The middle curve corresponds to $\Delta = 0.5\,mm$.

The influence of Δ on sensitivity and specificity of the operating point, corresponding to the maximum sensitivity is shown in Figure 5. We discovered that difference in nodule size estimate accounts for over 5% loss in sensitivity from 0.938 to 0.886 and gain in false positive rate from 6.53 to 6.78. Any value of $\Delta > 0.8\,mm$ and $\Delta > 2.9\,mm$ did not help to improve sensitivity and false positive rate respectively. These values directly relate the maximum extent of nodule size disagreement due to size underestimate and overestimate.



(a)



(b)

Figure 5. Effect of Δ on reported sensitivity (a) and false positive rate (b) of the detection system for the maximum sensitivity operating point.

5. CONCLUSION

Nodule size disagreement between human and automated measurements can have a significant impact on the performance evaluation of automated pulmonary nodule detection systems due to the minimum size cut-off employed by these systems. In our study we found that this error had reduced reported sensitivity by about 5% and increased the false positive rate by about 0.25 per case. We have presented a modified evaluation method that compensates for this error and generates detection performance close to the ideal case that would have been obtained if the nodule size was measured without error.

REFERENCES

[1] Bowyer, K., "Validation of medical image analysis techniques," *Handbook of Medical Imaging* **2**, 567–607, 2000.

[2] Chakraborty, D. and Winter, L., "Free-response methodology: alternate analysis and a new observer-performance experiment," *Radiology* **174**, 873–81, 1990.

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

[4] Enquobahrie, A., Reeves, A. P., Yankelevitz, D. F., and Henschke, C. I., "Automated Detection of Small Solid Pulmonary Nodules in Whole Lung CT Scans from a Lung Cancer Screening Study," *Academic Radiology* **14**(5), 579–593, 2007.

[5] Zhang, X., McLennan, G., Hoffman, E., and Sonka, M., "Automated detection of small-size pulmonary nodules based on helical CT images," *Proceedings of Information Processing in Medical Imaging: 19th International Conference, IPMI* , 2005.

[6] Ge, Z., Sahiner, B., Chan, H., Hadjiiski, L., Cascade, P., Bogot, N., Kazerooni, E., Wei, J., and Zhou, C., "Computer-aided detection of lung nodules: False positive reduction using a 3D gradient field method and 3D ellipsoid fitting," *Medical Physics* **32**, 2443, 2005.

[7] Brown, M., Goldin, J., Suh, R., McNitt-Gray, M., Sayre, J., and Aberle, D., "Lung Micronodules: Automated Method for Detection at Thin-Section CT – Initial Experience," *Radiology* , 2002.

[8] Bae, K., Kim, J., Na, Y., Kim, K., and Kim, J., "Pulmonary nodules: automated detection on CT images with morphologic matching algorithm – preliminary results," *Radiology* **236**(1), 286–93, 2005.

[9] Armato, S., Roy, A., Macmahon, H., Li, F., Doi, K., Sone, S., and Altman, M., "Evaluation of automated lung nodule detection on low-dose computed tomography scans from a lung cancer screening program," *Academic Radiology* **12**(3), 337–46, 2005.

[10] Sahiner, B., Hadjiiski, L., Chan, H., Shi, J., Way, T., Cascade, P., Kazerooni, E., Zhou, C., and Wei, J., "The effect of nodule segmentation on the accuracy of computerized lung nodule detection on CT scans: comparison on a data set annotated by multiple radiologists," *Proceedings of SPIE* **6514**, 65140L, 2007.

[11] Reeves, A. P., Biancardi, A. M., Apanasovich, T. V., Meyer, C. R., MacMahon, H., van Beek, E. J., Kazerooni, E. A., Yankelevitz, D., McNitt-Gray, M. F., McLennan, G., Armato III, S. G., Henschke, C. I., Aberle, D. R., Croft, B. Y., and Clarke, L. P., "The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements," *Academic Radiology* **14**(12), 1475–1485, 2007.

[12] Armato III, 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., 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., and Croft, B. Y., "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, 2007.

[13] Reeves, A. P., Biancardi, A. M., Apanasovich, T. V., Meyer, C. R., MacMahon, H., van Beek, E. J., Kazerooni, E. A., Yankelevitz, D., McNitt-Gray, M. F., McLennan, G., III, S. G. A., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Croft, B. Y., and Clarke, L. P., "The lung image database consortium (LIDC): Pulmonary nodule measurements, the variation and the difference between different size metrics," *SPIE International Symposium on Medical Imaging* , 2007.