



UvA-DARE (Digital Academic Repository)

Advances in digital chest radiography: impact on reader performance

De Boo, D.W.

Publication date

2012

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

De Boo, D. W. (2012). *Advances in digital chest radiography: impact on reader performance*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

**Advances in digital
chest radiography:
impact on reader performance**



Diederick Willem De Boo

Advances in digital chest radiography:
impact on reader performance
Thesis, University of Amsterdam, the Netherlands

Copyright © 2012 Diederick De Boo, Amsterdam, the Netherlands
No part of this thesis may be reproduced, stored, or transmitted in
any form or by any means, without prior permission of the author.

Cover	Searching for Utopia, Jan Fabre
Layout	Creative waves
Printed by	Ridderprint
ISBN	978-90-5335-522-0

Publication of this thesis was supported by:
Department of Radiology, Academic Medical Centre, the Netherlands
the University of Amsterdam, the Netherlands

Advances in digital chest radiography: impact on reader performance

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op 29 maart 2012 te 10:00 uur

door

Diederick Willem De Boo
geboren te Naarden

PROMOTIECOMMISSIE:

Promotor: Prof. dr. J.S. Laméris
Co-promotor: Dr. C.M. Schaefer-Prokop

Overige leden: Prof. dr. E.H.D. Bel
Dr. O.M. van Delden
Prof. dr. G.J. den Heeten
Prof. dr. M.B. van Herk
Prof. dr. ir. N. Karssemeijer
Prof. dr. J.A. Verschakelen

Faculteit der Geneeskunde

'Life is what happens to you,
while you're busy making other plans'

John Lennon

TABLE OF CONTENTS:

Chapter 1	Introduction and outline <i>Adapted from European Journal of Radiology 2009</i>	1
Chapter 2	Computed radiography versus mobile direct radiography for bedside chest radiographs: impact of dose on image quality and reader performance <i>Clinical Radiology 2011</i>	7
Chapter 3	Gray-scale reversal for the detection of pulmonary nodules on a PACS workstation <i>AJR American Journal of Roentgenology 2011</i>	23
Chapter 4	Computer-aided detection (CAD) of lung nodules and small tumours on chest radiographs (a review) <i>European Journal of Radiology 2009</i>	37
Chapter 5	Computer-aided detection of lung cancer on chest radiographs: effect on observer performance <i>Radiology 2010</i>	57
Chapter 6	Computer-aided detection of small pulmonary nodules in chest radiographs: an observer study <i>Academic Radiology 2011</i>	75
Chapter 7	Observer training for computer-aided detection of intrapulmonary nodules in chest radiography <i>Accepted for publication in European Radiology</i>	93
Chapter 8	Summary and general discussion	107
Chapter 9	Nederlandse samenvatting	119
	List of publications	122
	Dankwoord	123
	Curriculum Vitae	126

Introduction



Introduction

1

Chest radiography is still the most commonly used imaging technique to rule out cardiopulmonary disease, to study the effect of treatment, and to follow-up patients. Up to thirty years ago the radiographic technique was based on conventional film-screen, however, in the late 80ties the first digital radiographic technique, so called computed radiography (CR), was introduced. CR systems use a storage phosphor plate that has a comparable function as the amplifier screen in conventional film-screen radiography. The storage phosphor plate retains the information of the incident photons as a latent image, which is later retrieved by stimulation by a read-out laser. The emitted light is amplified, digitized and transferred into grey values for the radiographic image. About 10 years later a second digital radiographic technique was introduced: direct radiography (DR). DR uses flat-panel detectors which are characterized by a direct read-out matrix of electronic elements that are made of thin layers of amorphous silicon thin-film transistors (aSi-TFT elements) that are deposited on a piece of glass. This TFT layer is coupled with an X-ray absorption medium. The absorbed X-ray energy is either directly transferred into electronic charge (direct systems) or via an intermediate conversion into visible light (indirect or opto-direct systems) before transferred to grey values to produce a radiographic image. Compared to conventional film-screen radiography, both CR and DR have a much wider dynamic range for displaying attenuation differences and thereby offer an improved image quality^(1,2). Additionally, DR allows for reducing acquisition dose as compared to both, film-screen radiography and CR, based on its higher dose efficiency⁽⁴⁻⁷⁾. The options for dose reduction on one side and improvement of image quality on the other side have to be outweighed and determined with respect to the clinical indication and the diagnostic requirements. Together with the rapid development of computer technology and digital storage capacity, digital radiography facilitated the upcoming of Picture Archiving and Communication Systems (PACS). Digital images are now centrally archived and are accessible throughout the hospital by all clinicians. Besides this organizational advantage associated with PACS, the digital format of the images also induced the development of several processing tools. Nowadays basic tools such as magnification, window / level adjustments, and grey-scale reversal are available on every PACS workstation. Recently more elaborate post-processing tools designed to support the readers' detection and diagnostic performance were introduced. Factors contributing to detection errors in chest radiographs include image

quality, type of pathology, superposition of anatomical structures (e.g., ribs), the presence of accompanying abnormalities, and last but not least the radiologists' experience and perception capacity. Several processing techniques were introduced to lower the effects of anatomical structures or 'anatomic noise'. Energy subtraction produces images of the chest without overlapping osseous structures by subtracting two datasets recorded at the same time with low and high photon energies. Similar effects but without the expense of two exposures, can be achieved by digitally suppressing ribs and clavicles (SoftView; Riverain, Miamisburg, Ohio). Temporal subtraction aims to selectively enhance interval changes by subtracting the previous radiograph from the current one. Though all techniques delivered promising results under study conditions, they are not (yet) in broad clinical use⁽⁸⁻¹²⁾. In order to reduce the amount of perception errors made by radiologists computer-aided detection software (CAD) was developed. The software marks candidate regions to alert radiologists towards a potential lesion. Several algorithms have been developed of which only a minority is approved by the United States Food and Drug Administration (FDA). They all have in common that they analyze the chest radiographs in the background and show candidate lesions on demand. Subsequently the reader can either accept the candidate as a true positive or dismiss it as a false positive. The potential of CAD to increase the radiologist's sensitivity for pulmonary nodules has been acknowledged. Two studies reported that 35% and 47% of bronchogenic tumors missed in the original reports were correctly marked by a FDA approved CAD^(13,14). Both studies, however, tested only the CAD software alone (stand-alone performance) without taking the reader – CAD interaction into account. The effects of CAD on actual reader performance, however, will decide over its clinical utility and last but not least its acceptance by the radiologists' community.

Outline of this thesis

3

The aim of this thesis is to evaluate how actual reader performance is influenced by some of the advances offered by digital chest radiography. **Chapter 2** focuses on bedside chest radiographs of patients admitted to the Intensive Care Unit. Mobile CR and DR units were compared with respect to image quality and the potential for dose reduction. Detectability of monitor material and interobserver agreement were used as criteria to assess the effects of image quality on reader performance. In **Chapter 3** the effect of gray-scale reversal on nodule detection in chest radiography was tested. Inexperienced and experienced observers participated in a reader study to determine whether there would be any benefit from this rather simple processing tool available with a single mouse click on the PACS workstation. **Chapter 4** provides an overview of the published data on the performance of CAD for the detection of pulmonary nodules and small bronchogenic tumors. It also critically reviews the limitations of those studies and the considerations that have to be taken into account when drawing conclusions from the results. The effects of two FDA approved CAD systems available for chest radiography were tested in two observer studies. In **Chapter 5** chest radiographs of the Dutch-Belgian lung cancer screening trial (NELSON) with proven primary lung cancer were included, whereas in **Chapter 6** the test lesions were small pulmonary nodules seen on radiographs of older patients with so called "dirty lungs". Finally in **Chapter 7** the effect of short-term feedback on readers' ability to discriminate true from false positive CAD candidates was tested.

References

- 1 Chotas HG, Dobbins JT 3rd, Ravin CE. Principles of digital radiography with large-area, electronically readable detectors: a review of the basics. *Radiology* 1999; 210:595-599
- 2 Kotter E, Langer M. Digital radiography with large-area flat-panel detectors. *Eur Radiol* 2002; 12:2562-2570
- 3 Gruber M, Uffmann M, Weber M, Prokop M, Balassy C, Schaefer-Prokop C. Direct detector radiography versus dual reading computed radiography: feasibility of dose reduction in chest radiography. *Eur Radiol* 2006; 16:1544-1550
- 4 Strotzer M, Volk M, Frund R, Hamer O, Zorger N, Feuerbach S. Routine chest radiography using a flat-panel detector: image quality at standard detector dose and 33% dose reduction. *AJR Am J Roentgenol* 2002; 178:169-171
- 5 Herrmann A, Bonel H, Stabler A, et al. Chest imaging with flat-panel detector at low and standard doses: comparison with storage phosphor technology in normal patients. *Eur Radiol* 2002; 12:385-390
- 6 Bacher K, Smeets P, Bonnarens K, De Hauwere A, Verstraete K, Thierens H. Dose reduction in patients undergoing chest imaging: digital amorphous silicon flat-panel detector radiography versus conventional film-screen radiography and phosphor-based computed radiography. *AJR Am J Roentgenol* 2003; 181:923-929
- 7 Bacher K, Smeets p, Vereecken L, et al. Image quality and radiation dose on digital chest imaging: comparison of amorphous silicon and amorphous selenium flat-panel systems. *AJR Am J Roentgenol* 2006; 187:630-637
- 8 Szucs-Farkas Z, Patak MA, Yuksel-Hatz S, Ruder T, Vock P. Single-exposure dual-energy subtraction chest radiography: detection of pulmonary nodules and masses in clinical practice. *Eur Radiol* 2008; 18:24-31
- 9 Gilkeson RC, Sachs PB. Dual energy subtraction digital radiography: technical considerations, clinical applications, and imaging pitfalls. *J Thorac Imaging* 2006; 21:303-313
- 10 Aoki T, Oda N, Yamashita Y, Yamamoto K, Korogi Y. Usefulness of computerized method for lung nodule detection in digital chest radiographs using temporal subtraction images. *Acad Radiol* 2011; 18:1000-1005
- 11 Kano A, Doi K, MacMahon H, et al. Digital image subtraction of temporally sequential chest images for detection of interval change. *Med Phys* 1994; 21:453-461
- 12 Freedman MT, Lo SC, Seibel JC, Bromley CM. Lung nodules: improved detection with software that suppresses the rib and clavicle on chest radiographs. *Radiology* 2011; 260:265-273
- 13 Li F, Engelmann R, Metz CE, Doi K, MacMahon H. Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. *Radiology* 2008; 246:273-280
- 14 White CS, Flukinger T, Jeudy J, Chen JJ. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology* 2009; 252:273-281

Computed Radiography versus Mobile Direct
Radiography for Bedside Chest Radiographs:
Impact of Dose on Image Quality and
Reader Agreement

Diederick W De Boo
Michael Weber
Eline E Deurloo
Geert J Streekstra
Nicole J Freling
Dave A Dongelmans
Cornelia M Schaefer-Prokop

Abstract

Objective

To assess the image quality and potential for dose reduction of mobile direct detector (DR) chest radiography as compared with computed radiography (CR) for intensive care unit (ICU) chest radiographs (CXR).

Material and methods

Three groups of age-, weight- and disease-matched ICU patients (n=114 patients; 50 CXR per acquisition technique) underwent clinically indicated bedside CXR obtained with either CR (single read-out powder plates) or mobile DR (GOS-TFT detectors) at identical or 50% reduced dose (DR_{50%}). Delineation of anatomic structures and devices used for patient monitoring, overall image quality and disease were scored by four readers. In 12 patients pairs of follow-up CR and DR images were available, and in 15 patients pairs of CR and DR_{50%} images were available. In these pairs the overall image quality was also compared also side-by-side.

Results

Delineation of anatomy in the mediastinum was scored better with DR or DR_{50%} than with CR. Devices used for patient monitoring were seen best with DR, with DR_{50%} being superior to CR. In the side-by-side comparison, the overall image quality of DR and DR_{50%} was rated better than CR in 96% (46/48) and 87% (52/60), respectively. Inter-observer agreement for the assessment of pathology was fair for CR and DR_{50%} ($\kappa = 0.33$ and $\kappa=0.39$, respectively) and moderate for DR ($\kappa = 0.48$).

Conclusion

Mobile DR units offer better image quality than CR for bedside chest radiography and allow for 50% dose reduction. Inter-observer agreement increases with image quality and is superior with DR while DR_{50%} and CR are comparable.

Introduction

Digital radiography using storage phosphor plates (computed radiography, CR) is widely available and generally accepted for both, upright and bedside chest radiography. More recently flat-panel X-ray detectors (direct radiography, DR) using amorphous Silicon thin-film-transistor with different types of converter arrays have become increasingly used for upright chest radiography. Direct radiography offers improved dose quantum efficiency compared to computed radiography. This higher dose quantum efficiency can either be used to reduce patients' dose while maintaining image quality or to increase signal-to-noise ratio and image quality. Both features have proven validity for upright chest radiography: DR achieved an increased image quality compared with CR when acquired with the same dose^(1,2); at the same, a possible dose reduction of up to 60% was found without loss of diagnostically relevant image quality^(1,3-6). Mobile DR units use a different scintillator material (gadolinium oxysulphide = GOS) that has a lower quantum efficiency than cesium iodide = CsI, which is used for most upright radiography units^(7,8). Therefore, the dose reduction rates reported for upright radiography may not necessarily be transferable to bedside radiography. Conversely, from a diagnostic point-of-view image quality requirements are different at the bedside as compared with upright radiography. So far publications on mobile DR chest radiography are limited to pediatric applications. They reported a dose reduction of 50% possible with DR compared to CR without decrease of relevant image quality^(9,10). The aim of the present study was, therefore to, compare mobile DR (Gd_2O_2S) with mobile CR (storage phosphor plates) for adult bedside chest radiography with respect to image quality and the potential for dose reduction and the impact on diagnostic performance.

Material and Methods

Study group

One hundred and fifty anteroposterior bedside chest radiographs (CXR) were obtained of patients admitted to an adult mixed medical-surgical intensive care unit (ICU) of a large university hospital. All CXR were obtained due to clinical indication. Patient informed consent was, therefore, not necessary, and approval for the study had been obtained by the local ethic committee (071711598). The images were randomly obtained with one of the following three acquisition techniques: CR, DR or DR with 50% dose reduction ($DR_{50\%}$). Fifty CXRs were obtained per acquisition technique. For organizational reasons images included in the study were obtained only on selected days per week. Which patient would be radiographed depended on the clinician's indication; the assignment to one of the three techniques was random. None of the patients underwent a CXR using different techniques on the same day. Twenty-seven patients, however, underwent several CXRs using two or three of the above mentioned techniques, leading to a study group of 150 bedside CXRs obtained in 114 patients.

Detector systems

Images were obtained using storage phosphor technique (Agfa, Mortsels, Belgium) or direct detector flat-panel technology (MOBILETT XP Digital, Siemens, Erlangen, Germany). The storage phosphor plates (Agfa) had a size of 35 x 43 cm, a pixel matrix of 2048 x 2500 with a pixel size of 0.17 mm. They were based on powder phosphor plates, and were read-out by a focal spot laser only from one side. Images were processed using multifrequency processing (MUSICA, Agfa), processing parameters were used as recommended by the manufacturer. The direct detector unit uses a scintillator (Gd_2O_2S) layered on top of an array with light-sensitive photodiodes with thin-film transistors. The 35 x 43 cm detector plate had a 2208 x 2688 pixel matrix with a pixel size of 0.16 mm. Standard post-processing available on the detector unit (Siemens) was used.

Superimposition of catheter material

The presence of monitoring devices was simulated by superimposing fragments of varying types of catheters (chest tube, central venous line, naso-gastric tube, Swan-Ganz catheter) over the upper abdomen of the patients. Thirteen templates were made containing different combinations of three to five catheter fragments fixed on the lower end of hardcopies of posteroanterior chest radiographs. The hardcopies

were used as carriers and were taped on the cassette before it was placed underneath the patient for acquisition of the anteroposterior CXR. Each fragment had a length ranging from 3 to 5 cm. The fragments were positioned in such a way as to be randomly imaged over the area of the upper abdomen. The area was chosen to avoid interference of the superimposed material with the clinical purpose of the radiograph, but to be able to test the detectability in a high attenuation area of the body in a standardized manner.

Image acquisition

Image acquisition parameters were manually set by three technicians. All images were obtained without an antiscatter grid at 90 kVp. The milliampere-second (mAs) depended on the patient's weight and the visual assessment of the patients' constitution by the technicians and varied from 1.0 to 1.4 for standard CR / DR and was divided by a factor of two for DR_{50%} (0.5-0.7 mAs). All images were obtained using the same tube system, which automatically recorded dose-area product (DAP). From the DAP, patient entrance skin dose and effective dose were calculated (Table 1).

The acquisition technique used in each patient was determined by chance. Care, however, was taken during image inclusion that the three different groups were matched with respect to patient weight, body mass index (BMI) and total number of superimposed catheter fragments. For each patient weight, height, mAs, DAP and the template of superimposed catheter material used were documented. In 27 patients, pairs of follow-up images using two different acquisition techniques were available: there were 12 pairs of CR and DR images and 15 pairs of CR and DR_{50%} images available for direct comparison.

Image evaluation

Images were evaluated in three reading sessions using a dedicated picture archiving and communication system (PACS) system (Agfa Impax, version 4.5) equipped with high resolution LCD monitors (Barco, MDCG 2121-CB, pixel matrix of 1.2K x 1.9K) that were regularly controlled for their DICOM conformity. Two radiology residents (1st and 4th year training) and two certified radiologist (both more than 15 years of experience) independently interpreted the images in different random order. Reading conditions with subdued ambient light were kept constant throughout all reading sessions. Readers were allowed to use processing tools, such as windowing or magnification, at their preference. The readers' assessment referred to image quality criteria as well as to the diagnostic definition of the presence of disease, and compiled

the following tasks:

1. Assessment of the visibility of nine anatomic landmarks / lines using a three-point scale (3 = visible, sharply demarcated contours; 2 = moderately visible with partially blurred contours; and 1 = poorly visible with blurred contours). All anatomic landmarks were located within the high attenuation area of the chest radiograph or represented interfaces between low and high attenuation areas and included the trachea, the carina, retro-cardiac vessels, the upper / lower thoracic spine, both hemi-diaphragms, the esophageal-diaphragmatic recess and the descending aorta-diaphragmatic recess.
2. Assessment of the visibility of the superimposed catheter fragments using a similar three-point scale (3 = sharp contours throughout whole length; 2 = blurred contours but visible throughout whole length; and 1 = not visible throughout whole length).
3. Determination of the presence or absence of four different types of frequently encountered diseases in an ICU setting applying a five-point confidence scale ranging from 5 = definitely present; 4 = probably present; 3 = equivocal; 2 = probably not present; and 1 = definitely not present. The four diseases included: a) pulmonary consolidations, b) vascular congestion, c) pleural effusion and d) atelectasis. Criteria for the presence of these four pathologies followed generally accepted morphological criteria. Consolidations could be based on pneumonia, edema, or acute respiratory distress syndrome (ARDS), no specific interpretation was required. Generally, consolidations had blurred borders, could be patchy or confluent, and showed no signs of volume loss. There was no further subdivision with respect to the size of the consolidations. Conversely atelectasis were sharply demarcated, showed signs of volume loss, and were mostly located in the dependant parts of the lower lungs.
4. Assessment of the overall image quality using the Radlex ® image quality scoring criteria ⁽¹¹⁾: 0 = non-diagnostic (unacceptable for diagnostic purposes); 1 = limited (acceptable, with some technical defects but still adequate for diagnostic purposes); 2 = diagnostic (acceptable, with no technical defects likely to impair using the images for diagnosis); 3 = exemplary (good, most adequate for diagnostic purposes).

All images were individually scored. If pairs of images from the same patient using two different techniques were available, images were first assessed individually and in a separate reading session in a side-by-side comparison using a forced-decision-model meaning that readers were requested to rank the images differently according to their image quality.

Statistical analysis

Power analysis of statistics

A power analysis was performed (nQuery Adviser, version 5.0, statistical solutions, Cork, Ireland) to compare the image quality scores of the different acquisition techniques. For an alpha error value of 0.05, a power of 0.82 was calculated.

Match of patient groups

Statistical significance of differences between the three patient groups with respect to age, BMI, acquisition dose and prevalence of disease was tested using Pearson's Chi² test. There was no standard of truth for the definition of disease in the study group. To be able to assess differences between the three patient groups with respect to prevalence of disease, a disease category was assumed to be present if at least two of the four readers had rating scores of 4 or 5 ("probably" or "definitely") for its presence.

Visibility of catheter fragments and of anatomical landmarks

The statistical significance of differences between the three acquisition techniques with respect to the visibility of catheter fragments and anatomical landmarks was assessed using an analysis of variance with repeated measures. The dependant variables in this test included the visibility scores with the three techniques: independent variables were acquisition techniques and readers. The analysis of variance tested the significance of influence of the acquisition technique and reader on the visibility scores and whether there were significant interactions between the independent variables. Hochberg's post hoc test was calculated to determine the differences between the three acquisition techniques. Statistical analysis was carried out separately for the visibility of catheters, the visibility of the five landmarks located in the high attenuation areas of the chest (i.e., trachea, the carina, retro-cardiac vessels and the upper / lower thoracic spine) and of the four landmarks located at density interfaces (both hemi-diaphragms, the esophageal-diaphragmatic recess and the descending aorta-diaphragmatic recess). SPSS version 11.5 was used for statistical analysis.

Subjective assessment of overall image quality

The statistical significance of differences between the three acquisition techniques with respect to overall image quality ratings was assessed using a logistic regression analysis with repeated measures. The number of rating scores 2 and 3 (diagnostic and exemplary image quality) per technique and reader were chosen as characteristic values.

The side-by-side comparison ratings of overall quality of image pairs obtained in the same patient were assessed by descriptive statistics.

Interobserver agreement

As no standard of truth for the presence of disease was available, the interobserver agreement was used as a surrogate to evaluate the impact of image quality on diagnostic performance⁽¹²⁾. Interobserver agreement for the four groups of pathology was quantified for pairs of readers using weighted Cohen Kappa statistics. A κ value of more than 0.4 was considered to represent moderate but clinically acceptable observer agreement, a kappa value of more than 0.6 was considered to represent good observer agreement, a value below 0.4 to represent only fair agreement⁽¹³⁾. The calculation was based on 5 reading categories (1 = definitely not present, 2 = probably not present, 3 = equivocal, 4 = probably present, 5 = definitely present).

Results

Matching of patient groups

A total of 150 bedside CXRs (n=50 CXR per acquisition technique) were obtained in 114 ICU patients. There was no significant difference between the three patient groups with respect to gender, age, BMI, patient entrance skin dose and effective dose (Table 1). Pulmonary consolidations were present in 13%, vascular congestion in 27%, pleural effusion in 37% and atelectasis in 54% of the patients. A significantly different prevalence was not found for any of the four diseases between the three different acquisition groups ($p = 0.32$, $p = 0.66$, $p = 0.12$ and $p = 0.65$, respectively; Table 2). None of the patients was suffering from a pneumothorax.

Visibility of anatomical landmarks

The landmarks located in the high attenuation area of the chest (trachea, carina, upper / lower thoracic spine and retro-cardiac vessels) were best seen with DR, followed by DR_{50%}; CR was the most inferior technique. Analysis of variance found significant differences between acquisition techniques ($p < 0.001$) and readers ($p < 0.001$), but no significant interactions ($P = 0.33$). The post hoc test found that all three techniques performed differently. For the remaining 4 landmarks located at density interfaces (bilateral diaphragmatic contours, azygo-esophageal recess and descending aorto-diaphragmatic recess) the post hoc test found a superiority of DR over DR_{50%}, but no significant difference between DR_{50%} and CR.

Table 1

Characteristics of the patients of the three study groups.

Acquisition technique	Age (years)	Gender (m/f)	BMI	mAs	Entrance skin dose (mGy)	Effective dose (mSv)
CR	61	32/18	26.3	1.31	0.18	0.031
DR	59	30/20	26.3	1.16	0.16	0.029
DR _{50%}	60	33/17	25.9	0.62	0.09	0.015

CR: computed radiography, DR: direct radiography, DR_{50%}: direct radiography at 50% dose reduction, BMI: body-mass index (weight / length²)

Table 2

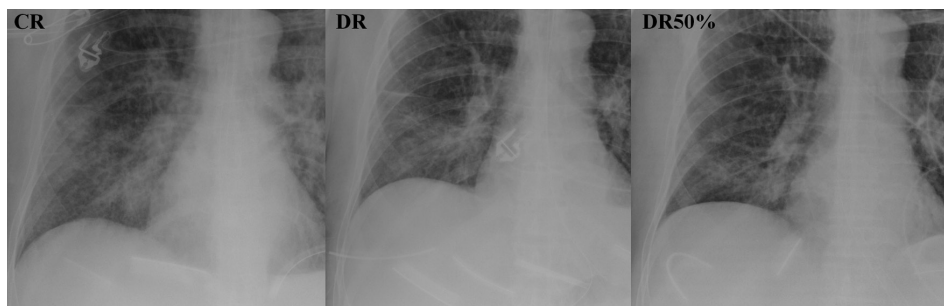
Presence of disease in the three study groups.

	Parenchymal consolidations *	Vascular congestion	Pleural effusion	Atelectasis
CR	10%	23%	48%	50%
DR _{50%}	16%	32%	38%	54%
DR	12%	26%	36%	58%
Mean	13%	27%	37%	54%

=* Could be caused by a pneumonia, edema, ARDS, etc. Atelectasis was specifically excluded and rated separately. CR: computed radiography, DR: direct radiography, DR_{50%}: direct radiography at 50% dose reduction

Visibility of superimposed catheter fragments

The visibility of the catheter fragments superimposed over the upper abdomen was rated best with DR, followed by DR_{50%}; CR was the most inferior technique. An example is given in Figure 1. Analysis of variance found significant differences between techniques ($p < 0.001$) and readers ($p = 0.001$), but no significant interactions ($p = 0.75$). The post hoc test found that all three techniques performed differently.

Figure 1

Region of interest of bedside chest radiographs of one patient obtained on three different days with all three techniques. There is improved transparency of the mediastinum with DR and DR_{50%} compared with CR with a superior delineation of the thoracic spine and the nasogastric tube. Catheter fragments superimposed over the upper abdomen are best delineated with DR followed by DR_{50%}; CR is the most inferior technique.

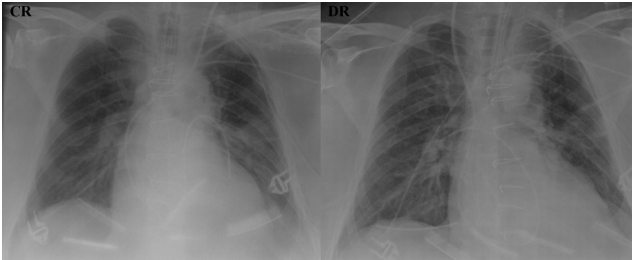
Subjective assessment of overall image quality

The overall image quality of DR was significantly superior to CR ($p = 0.006$), there was no difference between overall image quality of DR_{50%} and CR ($p = 0.85$). Differences and significance of differences remained the same whether only scores 3 (exemplary image quality) or scores 2 and 3 (diagnostic and exemplary image quality) were included into the logistic regression analysis. Side-by-side comparison of DR versus CR was possible in 12 image pairs (48 ratings by four readers) and of DR_{50%} versus CR in 15 image pairs (60 ratings by four readers). Using the forced-decision model, the image quality of DR was ranked higher than CR in 96% (46/48) of ratings (Figure 2) and of DR_{50%} over CR in 87% (52/60) of ratings (Figure 3).

Interobserver agreement

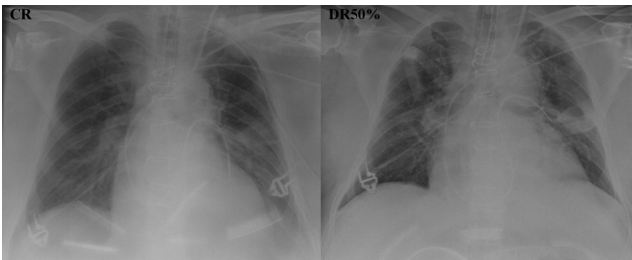
With a κ -value of 0.48 the interobserver agreement with DR was moderate but clinically acceptable. The interobserver agreement with DR_{50%} was just below the threshold of clinically acceptable with a κ -value of 0.39; while CR was considerably below the threshold with a κ -value of 0.33 when averaged over all 4 readers. In general, averaged κ -values were fair for the assessment of consolidations ($\kappa = 0.34$) and vascular congestion ($\kappa=0.39$) and moderate for the assessment of pleural effusions ($\kappa = 0.57$) and atelectasis ($\kappa = 0.56$).

Figure 2



CR versus DR images of the same patient in a side-by-side comparison. All four readers ranked the overall image quality of DR higher than that of the CR image. Note the superior delineation of the Swan-Ganz catheter and of the catheter fragments with DR.

Figure 3



CR and DR_{50%} images of the same patient as in Figure 2 in a side-by-side comparison. Although there is increased noise within the high-attenuation area of the mediastinum in the DR_{50%} image, the delineation of the Swan-Ganz catheter and of the catheter fragments is still superior with DR_{50%} as compared with CR. All four readers rated the image quality superior with DR_{50%}.

Discussion

Multiple studies have proven that dose reduction is feasible for upright chest radiography using flat-panel direct radiography (aSi TFT - DR) as compared with storage phosphor radiography (CR) without significant loss of image quality^(1,4,5,14-16). All clinical studies focussed on the delineation of anatomical landmarks and subjective scoring of image quality and uniformly report that DR, if obtained with the same dose as CR, provides superior image quality. When acquisition dose was reduced by a factor of two, DR achieved at least equivalent image quality compared with CR. Despite this dose reduction, DR still outperformed CR for some imaging aspects (Table 3). In order to being able to place these dose considerations into perspective, it is important to state not only relative but also absolute dose levels: reduction of a relatively high "standard dose" by a factor of 2 seems to be less of an achievement compared with operating on a generally lower dose level. Table 3 summarizes the clinical studies published recently for comparison of CR and DR. In only two of the six studies were absolute dose levels indicated: one referred to the entrance skin dose and the second to the detector dose. In the present study, the "standard dose" amounted to a mean of 0.18 mGy entrance skin dose for CR and DR and to 0.09 mGy entrance skin dose for the DR_{50%}. These dose levels are comparable to skin entrance dose levels published by Bacher et al. for upright chest radiography; 0.17 mGy for CR and 0.07 mGy for DR⁽³⁾. More than 10 years ago Leppik et al. had published two to three times higher entrance skin doses for conventional film-screen bedside chest radiography with a mean of 0.42 mGy (range 0.16-0.69 mGy)⁽¹⁶⁾. Although the single acquisition dose is rather low in bedside chest radiography and substantially reduced with the CR technique, as compared with conventional film-screen, dose reduction remains a relevant issue, since those patients often require multiple follow-up studies. Image quality requirements are different in bedside chest radiography compared with routine upright chest radiography. Contrast resolution has to be high, especially in high-attenuation areas of the mediastinum, to display the multiple types of catheter materials, whereas requirements for spatial resolution are generally low. The aim of the present study was to test whether dose reduction with DR is feasible for bedside chest radiographs of adults. To test the influence of dose reduction on image quality a number of reading measures, which included subjective scoring of visibility of anatomic landmarks and a relatively simple detection task, were applied. The latter referred to one of the major imaging tasks of bedside chest radiography, namely the assessment of monitor

Table 3 List of clinical studies comparing direct radiography (DR) and computed radiography (CR) to assess the influence of potential dose reduction with DR.

Author, journal and year of publication	n	Reference for comparison	Studied parameters	Results
Goo AJR 2000	46	CR vs. SE-DR	11 anat. landmarks	Residents: DR > CR (n=6 landmarks) CR > DR (n=2 landmarks) Radiologist: DR > CR (n=8 landmarks) Postero-anterior: DR > CR Lateral: DR = CR DR _{50%} = CR
Biemans Invest Rad 2002	63	CR vs. SE-DR	21 anat. landmarks	
Ganten AJR 2002	30	CR vs. CsI-DR	10 anat. landmarks	DR _{50%} = CR
Herrmann Eur Rad 2002	75	CR vs. CsI-DR	8 anat. landmarks	DR = DR _{50%} = CR
Bacher AJR 2003	100	CR vs. CsI-DR	Overall image quality	DR _{40%} > CR
Gruber Eur Rad 2006	50	CR vs. DR	Overall image quality 8 anat. landmarks Visual perception of image noise	Image quality: DR > DR _{50%} > CR Landmarks in high attenuation area: DR and DR _{50%} > CR Landmarks in low attenuation area: DR > CR; DR _{50%} = CR Image noise: DR < CR; DR _{50%} = CR

Indices (e.g., DR_{50%}) indicate the reduced level of acquisition dose in percentage of the “standard” dose. CR: computed radiography; SE-DR: flat-panel direct radiography using a selenium detector; CsI-DR: flat-panel direct radiograph using a Cesium-Iodine TFT detector.

devices, and was based on an absolute standard of truth. For all criteria that referred to image quality, DR was significantly superior to both DR_{50%} and CR. DR_{50%} was superior to CR for the delineation of catheter fragments and the visibility of anatomic landmarks in the high-attenuation area of the mediastinum, suggesting a strong superiority of DR over CR with respect to the signal-to-noise ratio in high-attenuation areas based on its higher dose efficiency. When overall image quality was evaluated separately per image, DR was found to be superior to CR whereas DR_{50%} and CR achieved equivalent image quality scores. During the side-by-side comparisons of image pairs using a forced-decision model, image qualities of both DR and DR_{50%} were rated superior to CR in 96% and 87%, respectively. These results support the fact that when judging image quality, a side-by-side comparison is far more sensitive in detecting subtle quality differences when compared with separate scoring of image quality per image. Regarding differences of image quality, in the absence of a pathological standard, we used interobserver agreement as a surrogate for diagnostic performance of the three different techniques. Four diseases frequently encountered at bedside chest radiography were tested. As no standard of truth was available, readers were not specifically asked to determine the underlying disease, but were only asked to assess the presence or absence of this type of imaging finding irrespective of severity and extent of findings. Only a fair to moderate interobserver agreement was found for the assessment of these four categories of intrapulmonary disease. It is common clinical experience that interobserver agreement is lower for bedside chest radiography than for upright radiography, which is most likely due to the sometimes vastly differing image quality and the non-specific nature of radiographic findings. It has to be noted that images were evaluated without the availability of previous radiographs and without any clinical context. As stated before, no specification of disease severity was required, which is likely to have caused some disagreement, especially for images with subtle findings. All of these aspects are likely to have contributed to the generally low interobserver agreement rates. In fact, the kappa values in the present study are probably lower than normally encountered under clinical conditions and are likely to underestimate the diagnostic potential of the bedside chest radiographs. However, under comparable (study) conditions DR was the only technique that on average achieved an agreement rate beyond the threshold of 0.4, which is considered as clinically acceptable. In that respect DR outperformed CR and also DR_{50%}. DR_{50%} also achieved at least the same agreement rate as CR. These results confirm previous experiences that DR either allows for improved image quality and potentially improved diagnostic

performance if obtained with the same dose as that used for CR or alternatively allows for dose reduction on a slightly lower but unaltered diagnostic level comparable to CR. An advantage of the present study was that the image quality obtained via different methods was assessed, including subjective scoring of quality, grading of the delineation of anatomic features, and using visual tasks that are typical of the radiological evaluation methods of bedside chest radiographs of ICU patients, such as the delineation of catheter fragments and the assessment of the presence of disease. However, the present study suffers from the following limitations: first, the catheter fragments were relatively short and superimposed over the area of the upper abdomen and not within the mediastinum. Thus, position and length of the catheters lacked clinical characteristics; however, this technique allowed the visualisation of a range of monitoring devices under controlled conditions. Second, image pairs of the same patient obtained with two acquisition techniques were available in only a subgroup of the study patients. It should be noted requests for ICU bedside chest radiographs were strictly based on clinical indications; thus, patients did not undergo chest radiography daily. In addition, only a group of three technicians were involved in the acquisition of the study images with the advantage of minimizing external effects on image quality. As result, image acquisition was spread over a longer period of time. To minimize the effect of different patient groups, the study groups were statistically matched with respect to age, BMI, acquisition dose, and the presence of disease. Differences that may have potentially remained e.g., with respect to the extent of disease or the effects of different inspiration levels, cannot be excluded but are likely to be small as the choice of acquisition technique was randomly assigned. Third, acquisition parameters were set by the technicians according to the patient's weight and the technician's visual assessment of the patient's constitution and condition as it is common practice in an ICU environment where an automatic exposure control is not available. Although it cannot be proved that every single DR image was acquired with exactly 50% less dose, on average DR images had been obtained either with the "standard" or with a dose lowered by 50%. From the present results it can be concluded that DR generally provides superior image quality to CR also at the bedside. A dose reduction of 50% can be applied without risking the loss of diagnostic information. At the same dose as CR, DR provides better delineation of landmarks and catheters and increases interobserver agreement on the assessment of disease categories.

References

1. Gruber M, Uffmann M, Weber M, Prokop M, Balassy C, Schaefer-Prokop C. Direct detector radiography versus dual reading computed radiography: feasibility of dose reduction in chest radiography. *Eur Radiol* 2006; 16:1544-1550
2. Redlich U, Hoeschen C, Effenberger O, et al. Comparison of four digital and one conventional radiographic image systems for the chest in a patient study with subsequent system optimization. *ROEFO* 2005; 177:272-278.
3. Strotzer M, Volk M, Frund R, Hamer O, Zorger N, Feuerbach S. Routine chest radiography using a flat-panel detector: image quality at standard detector dose and 33% dose reduction. *AJR Am J Roentgenol* 2002; 178:169-171
4. Herrmann A, Bonel H, Stabler A, et al. Chest imaging with flat-panel detector at low and standard doses: comparison with storage phosphor technology in normal patients. *Eur Radiol* 2002; 12:385-390
5. Bacher K, Smeets P, Bonnarens K, De Hauwere A, Verstraete K, Thierens H. Dose reduction in patients undergoing chest imaging: digital amorphous silicon flat-panel detector radiography versus conventional film-screen radiography and phosphor-based computed radiography. *AJR Am J Roentgenol* 2003 ;181:923-929
6. Bacher K, Smeets p, Vereecken L, et al. Image quality and radiation dose on digital chest imaging: comparison of amorphous silicon and amorphous selenium flat-panel systems. *AJR Am J Roentgenol* 2006; 187:630-637
7. Samei E. Performance of digital radiographic detectors: quantification and assessment methods. *Advances in digital radiography: RSNA 2003: Categorical course in diagnostic Radiology Physics*, pp. 37-47.
8. Metz S, Damoser P and Hollweck R, et al. Chest radiography with a digital flat-panel detector: experimental receiver operating characteristic analysis. *Radiology* 2005 ;234:776-784
9. Rapp-Bernhardt U, Bernhardt TM, Lenzen H, et al. Experimental evaluation of a portable indirect flat-panel detector for the pediatric chest comparison with storage phosphor radiography at different exposures by using a chest phantom. *Radiology* 2005 ;237:485-491
10. Rapp-Bernhardt U, Roehl FW, Esseling R, et al. Portable flat-panel detector for low-dose imaging in a pediatric intensive care unit comparison with an asymmetric film-screen system. *Invest Radiol* 2005; 40:736-741
11. RSNA Radlex: image quality. Available at: <http://www.radlex.org/RID/RID10> [accessed 07.03.2011]
12. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? *Radiology* 2010; 257:14-17
13. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992 ;304:1491-1494
14. Goo JM, Im JG, Kim JH, Seo JB, Kim TS, Shine SJ, Lee W. Digital chest radiography with a selenium-based flat-panel detector versus a storage phosphor system: comparison of soft-copy images. *AJR Am J Roentgenol* 2000; 175:1013-1018
15. Biemans JM, Van Heesewijk JP, Van Der Graaf Y. Digital chest imaging: selenium radiography versus storage phosphor imaging. Comparison of visualization of specific anatomic regions of the chest. *Invest Radiol* 2002; 37:47-51
16. Ganten M, Radeleff B, Kampschulte A, Daniels MD, Kauffmann GW, Hansmann J. Comparing image quality of flat-panel chest radiography with storage phosphor radiography and film-screen radiography. *AJR Am J Roentgenol* 2003; 181:171-176
17. Leppik R, Bertrams SS, Höltermann W, Klose KJ. Radiation exposure during thoracic radiography at the intensive care unit. Dose accumulation and risk of radiation-induced cancer in long-term therapy. *Radiologe* 1998; 38:730-736

Gray-scale Reversal for the Detection of Pulmonary
Nodules on a PACS workstation

A large, stylized number '3' graphic is positioned on the right side of the page. It is rendered in a dark gray color with a slight gradient, giving it a three-dimensional appearance. The number is centered vertically relative to the author list.

Diederick W De Boo
Martin Uffmann
Shandra Bipat
Eelco FA Boorsma
Maeke J Scheerder
Michael Weber
Cornelia M Schaefer-Prokop

Abstract

Objective

The purpose of this article is to evaluate the impact of gray-scale reversal on the detection of small solid pulmonary nodules in two-view chest radiography (CXR).

Material and methods

One hundred and twenty-eight patients (mean age, 62y) who underwent CT and chest radiography within 6 weeks were retrospectively selected for this study. Seventy-three percent of patients showed variable degrees of radiographic findings of a 'dirty lung'. A total of 129 solid pulmonary nodules were present in 74 patients (nodule diameter range, 5-30mm; mean 13mm). The remaining 54 patients served as negative control subjects. Six readers of varying experience levels evaluated the images without and with the availability of gray-scale reversal in two separate reading sessions. Figure of merit (FOM), sensitivity per lesion, mean number of false positive marks per image and accuracy were calculated.

Results

Five of the six readers showed a slight increase of sensitivity with the use of gray-scale reversal, but on average, the difference was not significant (48% vs. 50%; $p>0.05$). The mean number of false positive marks per image also nonsignificantly increased from 0.20 to 0.23. The increases in both sensitivity and mean number of false positive marks translated into nonsignificant decrease in averaged FOM (0.79 vs. 0.77) and accuracy (72% vs. 71%). Data analysis of subgroups of nodules or different reader groups, dependant on level of experience, did not reveal significant differences.

Conclusion

Using PACS display of digital chest radiographs, gray-scale reversal does not help the radiologists in detecting pulmonary nodules.

Introduction

Modern digital radiography offers a number of processing tools to improve the detection of focal lung lesions. Signal normalization, gradation adjustment and multi-frequency edge enhancement are performed automatically in the background without further interaction from the radiologist. They represent steps within the process which are designed to produce images of constant quality and optimal display; their parameter sets have been separately optimized. Other more elaborate processing tools, such as temporal or energy subtraction and computer-aided detection, are increasingly available but demand specific software and represent tools specifically required by the radiologists. In addition to online window and level adjustments, gray-scale reversal represents a rather simple tool that is a built-in feature on all PACS workstations. In the literature, little evidence is available for gray-scale reversal in chest radiography⁽¹⁻⁵⁾. All studies applied techniques that are obsolete today, such as hard-copy radiographs, low-resolution soft-copy display, or secondary digitized images and low resolution digital radiographs. It is known from optical physiology that optical contrast perception is increased when a dark object is presented on a white background⁽⁶⁾. Consistent with that idea, we found in clinical practice that evaluation of chest radiographs in the positive mode ("bones black") in addition to the usual negative mode ("bones white") is helpful for the detection of focal lung densities. We therefore decided to reevaluate with modern digital image and monitor equipment whether inversion of grey-scale values would facilitate detection of nodular lung lesions. To assess the effects of reader experience and increased interstitial markings in the images, we included six readers with varying experience and a study group of elderly patients, the majority of whom were smokers.

Material and Methods

Study group

We retrospectively selected 128 patients from our institution's data archive who had undergone both a posteroanterior (PA) and lateral chest radiograph and a chest CT within 6 weeks for clinical purposes. Approval for this study was obtained by our institution's ethic committee (registration number 08170465). The study group included 74 patients with nodules and 54 without intrapulmonary nodules. Patients' mean age was 62 years (range, 15-89 years), 74 were males and 54 were females. Patient records showed a smoking history in 74 patients (58%) and no prior or current tobacco use in 24 patients (19%). For the remaining 30 patient (23%) no data were available. Both smoking and increase of age led to a variable increase of parenchymal markings on the chest radiograph, also known as "dirty lung"⁽⁷⁾. Its presence was subjectively scored on the chest radiographs by an experienced chest radiologist and the researcher, ranging from none to mild, moderate and severe. Hereafter, we will refer to this score as "anatomic noise" (AN).

Intrapulmonary lesions

CT findings were used as the reference standard for the definition of nodule size, type and location. A total of 129 CT-proven nodular opacities with diameters between 5 and 30mm (mean, 13 mm; median, 11 mm) were present in 74 patients. Eighty-six nodules were round, had smooth margins on CT, and had well-defined contours on the chest radiographs. None of these nodules were calcified. Forty-three nodules were patchy with spiculated margins on CT and ill-defined contours on the CXR.

Lesion conspicuity

Lesion conspicuity was subjectively graded on the conventional negative mode chest radiography ("bones white") by a board-certified chest radiologist (with > 15 years of experience) and by the researcher after the reading had been completed and ranged from high to moderate, low and very low.

Chest radiography

Chest radiographs were obtained in digital technique using a dedicated chest stand (Thoravision, Philips Healthcare). Images were processed using nonlinear multifrequency processing (Unique, Philips Healthcare). Processing parameters were chosen according to the recommendations of the manufacturer and represent the standard processing used in routine clinical studies.

Gray-Scale Reversal

Gray-scale reversal was accomplished by inverting the slope of the lookup table and transcribing the original gray-scale values (bones white) to their inverted counterparts. No additional processing was applied. To produce gray-scale reversed images, we used the facilities of the workstation (a single mouse click).

Image evaluation

Images were evaluated using a dedicated PACS system (Impax version 4.5) equipped with high-resolution (1.2K x 1.9K) liquid crystal display monitors (MDCG 2121-CB, Barco). Three radiology residents (first to third year training) and three board-certified radiologists (all with > 10 years of experience) independently interpreted the images. The 128 PA and lateral radiographs were evaluated twice in two separate reading sessions, once without and once with the availability of gray-scale reversal. Approximately half of the cases were seen first without gray-scale reversal, and the other half was interpreted first with the availability of gray-scale reversal. During both sessions, readers were allowed to use processing tools, such as windowing or magnification, according to their preferences. There was an interval of at least 6 weeks between the two reading sessions, and images were evaluated in different random orders. The presence or absence of an intrapulmonary opacity was scored using a 5-point scale of confidence, as follows: 5, definite pulmonary lesion; 4, probable pulmonary lesion; 3, unequivocal; 2, probably no lesion; and 1, definitely no lesion. For the confidence ratings 3, 4 and 5, readers were asked to indicate the anatomic location of the suspected lesion on a separate data sheet. This was done separately for each patient; per patient more than one lesion could be marked. The readers were informed that images could contain none, a solitary, or multiple lesions but did not know the percentage of each subgroup. They were instructed to ignore lesions smaller than 5 mm. Reading time was documented per reader and reading session.

Statistical analysis

The patients with and patients without intrapulmonary focal lesions were compared with regard to gender and smoking history by chi-square test, with regard to age by Student t-tests because the data were normally distributed, and with regard to anatomic noise by chi-square test for trend. For smoking history, the missing data (19%) were excluded. All reader markings, without and with gray-scale reversal, were determined to be true or false positive by comparing the markings on the separate data sheet with the original chest radiograph and the corresponding CT. This was done in consensus by the researcher and

an experienced chest radiologist after all readings had been completed. The data were analyzed using the jackknife free-response receiver operating characteristic (JAFROC) method⁽⁸⁻¹⁰⁾. JAFROC software was used to calculate a figure of merit (FOM). The FOM is defined as the probability on a scale from 0 to 1.0 that true positive marks for lesions are rated with higher confidence than false positive marks (nonlesions) on control chest radiographs⁽⁹⁾. An FOM of 1.0 describes the ideal situation in which all nodules are correctly marked with high confidence and there are no false positive marks on the control images; an FOM of 0.5 means that the confidence for true positive marks for lesions on abnormal chest radiographs is equal to the confidence for false positive marks on negative control chest radiographs. Additionally, we calculated descriptive statistical measures, such as sensitivity, mean number of false positive marks per image, and accuracy. Sensitivity was calculated on a per-lesion basis. Reader ratings 4 and 5 were considered true positive if made for correctly localized lesions. The mean false positive marks per image was calculated by dividing the total number of false positives per reader by the number of study cases ($n = 128$). Reader ratings 4 and 5 were considered false positive if made for nonexistent lesions. Accuracy was calculated per image: an image was considered true positive if at least one nodule was correctly identified and was considered true negative if there was no false positive mark on a negative control image. Sensitivity, the mean number of false positive marks per image, and accuracy were calculated for all observers, for experienced, and for inexperienced observers. For all observers, the data of all six observers were counted as separate observations. Similarly, for the subgroups of experienced or inexperienced readers, data of the respective group were also counted as separate observations. Differences in sensitivity, mean number of false positive marks per image and accuracy were compared with the McNemar test. All analyses were performed in SPSS software (version 15.0.1, SPSS) Significance was assumed at p less than 0.05.

Results

Study groups and lesion characteristics

An increase in anatomical noise was scored present in 93 patients (73%). The severity of anatomic noise varied from mild in 43% of patients, to moderate in 20% of patients, and severe in 10% of patients. Forty patients had a solitary nodular lesion, 20 patients had two lesions and 14 patients had three or more lesions. The lesions were located in the upper lobes in 57%, in the middle lobe in 9% of patients, and in the lower lobes in 35% of patients.

Table 1

Reader outcome by figure of merit, sensitivity, mean number of false positive marks per image, and accuracy.

Readers	Reading	Figure of merit, mean (95%, CI)	Sensitivity, % (95%, CI)	Number of false positives per image, mean (total no. of false positive marks by readers / no. of study cases)	Accuracy, % (95%, CI)
All	Baseline	0,79 (0,66-0,92)	48% (45%-52%)	0,20 (156/768)	72% (68%-75%)
	Grey-scale reversal	0,77 (0,61-0,92)	50% (46%-54%)	0,23 (179/768)	71% (68%-74%)
Inexperienced	Baseline	0,72 (0,36-1,0)	42% (37%-47%)	0,25 (97/384)	66% (61%-71%)
	Grey-scale reversal	0,66 (0,34-0,99)	45% (40%-50%)	0,30 (115/384)	64% (59%-69%)
Experienced	Baseline	0,85 (0,78-0,92)	55% (49%-59%)	0,15 (59/384)	77% (73%-82%)
	Grey-scale reversal	0,87 (0,73-1,0)	55% (50%-60%)	0,17 (64/384)	79% (75%-83%)

Baseline reading was performed without gray-scale reversal.

Lesion conspicuity was rated very low in 24% of patients, low in 19% of patients, moderate in 37% of patients, and high in 20% of patients. More patients with nodules had a positive smoking history compared to the patients without nodules ($p < 0.001$). There were no differences with regard to age ($p = 0.83$), gender ($p = 0.78$) and AN ($p = 0.14$).

Reader performance without gray-scale reversal

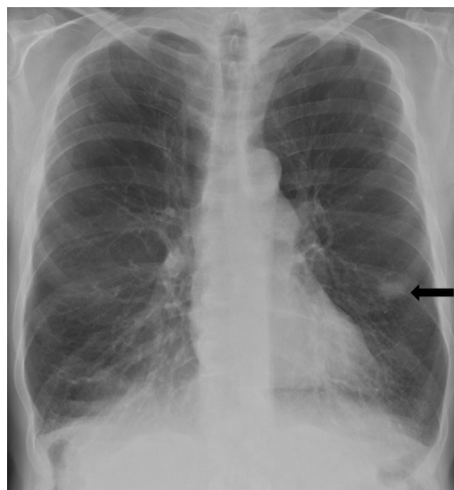
The average FOM was 0.79, with a mean FOM of 0.72 for the inexperienced readers and a mean FOM of 0.85 for the experienced readers. The mean sensitivity for the inexperienced readers was 42%, with a mean number of false positive marks per image of 0.25, whereas the mean sensitivity for the experienced readers was 55% with a mean number of false positive marks per image of 0.15. For both reader groups, there was no significant association between the mean number of false positive marks per image and the degree of anatomic noise. The accuracy varied widely among the readers, with a mean of 72% (Table 1). The group of missed lesions comprised 74% of the lesions of low to very low conspicuity and 75% of the lesions smaller than 10 mm (Table 2).

Reader performance with the availability of gray-scale reversal

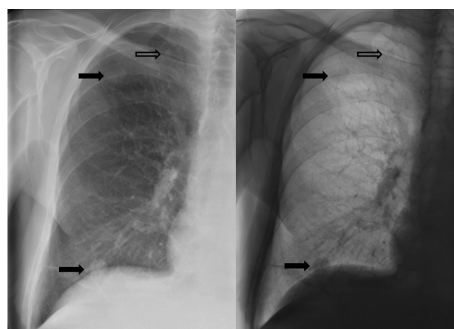
For the inexperienced readers, the average FOM decreased from 0.72 to 0.66 with the availability of gray-scale reversal; for the experienced readers, the mean FOM slightly increased from 0.85 to 0.87, though both differences did not reach significance. Five out of six readers showed an increase in sensitivity with the use of gray-scale reversal, but differences were not significant (43% vs. 43%; 40% vs. 44%; 45% vs. 47%; 51% vs. 47%; 55% vs. 57% and 57% vs. 61%). The mean sensitivity of the inexperienced readers nonsignificantly increased from 42% to 45%; the mean sensitivity of experienced readers remained unchanged at 55%. Both inexperienced and experienced readers showed a nonsignificant increase in mean number of false positive marks per image (0.25 vs. 0.30 and 0.15 vs. 0.17, respectively). Mean accuracy decreased from 66% to 64% for the inexperienced readers and slightly increased from 77% to 79% for experienced readers. Again, differences did not reach statistical significance (Table 1). Examples are given in Figure 1 - 3.

Reading time

Reading time was documented per reading session. On average, reading time per image increased by 5 seconds (8%) with the availability of gray-scale reversal (59 vs. 64 seconds per image).

Figure 1

Example of a patient with moderate increased parenchymal markings. The nodule in the left lower lobe (arrow) was detected by all readers without and with gray-scale reversal.

Figure 2

Three nodules, two in the upper lobe and one in the lower lobe. None of the readers detected the two larger nodules (black arrows) when using gray-scale reversed images complimentary. One reader saw the small nodule (open arrows) only with the use of the gray-scale reversed image.

Table 2

Sensitivity of readers without and with gray-scale reversal for subgroups of lesions, by conspicuity and size.

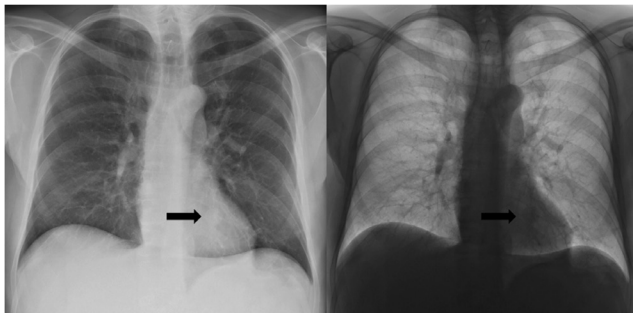
Reader group	reading	Conspicuity		Size	
		High or moderate	Low or very low	< 10 mm	≥ 10 mm
All	Baseline	79 (77 - 81)	26 (24 - 28)	25 (22-28)	60 (57-62)
	Gray-scale reversal	82 (80 - 84)	26 (24 - 28)	25 (23-28)	62 (60-64)
Inexperienced	Baseline	72 (68 - 75)	21 (18 - 23)	19 (16-23)	54 (51-57)
	Gray-scale reversal	78 (74 - 81)	21 (18 - 23)	20 (16-23)	57 (54-60)
Experienced	Baseline	86 (83 - 89)	31 (28 - 34)	31 (27-35)	67 (63-68)
	Gray-scale reversal	86 (83 - 89)	32 (28 - 35)	32 (27-35)	67 (64-70)

Data are percentage (95% CI). Baseline reading was performed without gray-scale reversal.

Discussion

To our knowledge, only few studies evaluating gray-scale reversal have been published so far⁽¹⁻⁵⁾. They all are more than 15 years old and used image and display techniques that are outdated with respect to contrast and spatial resolution and are not applied any more. As known from optical physiology, the contrast sensitivity of the human eye is greater for the detection of dark structures on bright background than vice versa⁽⁶⁾. We therefore wanted to reevaluate the usefulness of gray-scale reversal for the detection of nodular opacities using modern state-of-the-art equipment. Gray-scale reversal can be performed in different ways, dependant on whether the slope of the lookup table that determines the transfer of pixel values into gray levels is preserved or changed to shift contrast resolution from the high to the low absorption area or vice versa. In our study, the gray levels were transcribed to their inverted counterparts, preserving a higher contrast in the area of the lung (originally black) compared with the area of the mediastinum (originally white). No further processing was obtained. We chose so to maintain contrast characteristics more closely to the original display. The gray-scale reversed images were used as an adjunct to the normal negative displays. In this way, the gray-scale reversed images served as a complimentary display, similar to when radiologists change the window or level settings to focus on the detection of subtle density differences. Images were evaluated side-by-side on two monitors or by toggling from one to the other on a single monitor, dependant on the reader's preference. Though there was a tendency for an improved sensitivity for the detection of nodular opacities (e.g. five out of six readers showed an increase of sensitivity with gray-scale reversal), the differences failed to reach statistical significance. The increase of sensitivity

Figure 3



One nodule in the left lower lobe (arrow). Three readers missed the nodule and detected it with the complimentary use of the gray-scale inverted image. One reader detected the nodule without gray-scale and missed it with the availability of gray-scale reversal. For the remaining two readers there was no difference without or with the availability of gray-scale reversal.

was counteracted by loss of specificity and therefore did not lead to a change of FOM or accuracy. Also, when comparing subgroups of nodules as a function of nodule size or subgroups of readers as function of experience, differences did not reach significance. These results somewhat contradict our clinical experience that suggested a more positive impact of gray-scale reversal on the detection of focal lesions. A possible explanation for our results refers to the overall small diameter of our test lesions with a median of 11 mm and mean of 13 mm. In addition, most chest radiographs showed a varying degree of increased parenchymal markings, which further impaired detection. Both aspects certainly contributed to the loss of specificity due to misinterpretation of focal densities or vascular markings that became more pronounced in the gray-scale reversed images. Lack of familiarity with gray-scale reversed images might also have played a role. Sheline et al. found an overall improved detection of pulmonary nodules using secondary digitized radiographs as compared to conventional screen-film radiographs but failed to prove a significant difference comparing "bones white" and "bones black" soft-copy display⁽²⁾. Three other studies evaluating gray-scale reversal in chest radiography all reported a decrease in performance when gray-scale reversal was applied⁽³⁻⁵⁾. It should be mentioned, however, that all studies evaluated the gray-scale reversal as stand-alone display and not as an adjunct to the normal negative ("bones white") display, as we did in our study. Kheddache et al. were the only ones to report improved detection of simulated nodules projected over the area of the mediastinum of an anthropomorphic chest phantom applying the same type of gray-scale reversal we used⁽¹⁾. However, the tested study images had been obtained with an image intensifier with a much lower spatial and contrast resolution as compared with the chest radiographs we evaluated. In our clinical study group, the number of nodules projecting over the mediastinum and the retrocardiac area on PA radiographs was too low to detect significant differences. Besides, both PA and lateral images were available for evaluation. Our study has some limitations. We evaluated a selected group of patients with a higher prevalence of lesions than normally seen in clinical routine. In addition readers were specifically asked to look for small nodular densities, which probably lowered their overall specificity. Images were evaluated without and with gray-scale reversal in two reading sessions, introducing the additional factor of intrareader variability. We chose to do this to keep reading conditions possibly realistic and equal for both conditions and not to interrupt the readers' visual analysis by requiring lesion documentation without and also with gray-scale reversal. Intra-reader variability may have contributed to the nonsignificant performance differences, however,

this again appears to reflect more closely clinical reality. We conclude that, regardless of experience of radiologist or nodule characteristics, complimentary use of gray-scale reversal does not improve the detection of radiologists for small nodules in chest radiographs. Further studies are needed to assess whether gray-scale reversal is advantageous for the detection of larger ill-defined lesions.

References

1. Kheddache S, Månsson LG, Angelhed JE, Denbratt L, Gottfridson B, Schlossman D. Digital chest radiography: should images be presented in negative or positive mode? *Eur J Radiol* 1991; 13:151-155
2. Sheline ME, Brikman I, Epstein DM, Mezrich DM, Kundel HL, Arenson RL. The diagnosis of pulmonary nodules: Comparison between standard and inverse digitized images and conventional chest radiographs. *AJR* 1989; 152:261-263
3. MacMahon H, Metz CE, Doi K, Kim T, Giger ML, Chan HP. Digital chest radiography: effect on diagnostic accuracy of hard copy, conventional video and reversed gray scale video display formats. *Radiology* 1988; 168:669-673
4. Oestmann JW, Kuser DC, Bourgouin M, Llewellyn HJ, Mockbee BW, Greene R. Subtle lung cancers: Impact of edge enhancement and gray scale reversal on detection with digitized chest radiographs. *Radiology* 1988; 167:657-658
5. Schaefer CM, Greene R, Hall DA, et al. Mediastinal abnormalities: detection with storage phosphor radiography. *Radiology* 1991; 178:169-173
6. Blackwell H. Contrast thresholds of the human eye. *J Opt Soc Am* 1946; 36:624-643
7. Gückel C, Hansell DM. Imaging the 'dirty lung' -has high resolution computed tomography cleared the smoke? *Clin Radiology* 1998; 53:717-722
8. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004; 31:2313-2330
9. Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. *Acad Radiol* 2006; 13:1187-1193
10. Vikgren J, Zachrisson S, Svalkvist A, et al. Comparison of chest radiography for the detection of pulmonary nodules: human observer study of clinical cases. *Radiology* 2008; 249:1034-1041

Computer-Aided Detection (CAD) of Lung Nodules and Small Tumours on Chest Radiographs



Diederick W De Boo
Mathias M Prokop
Martin Uffmann
Bram van Ginneken
Cornelia M Schaefer-Prokop

European Journal of Radiology 2009; 72:218-225

Abstract

Detection of focal pulmonary lesions is limited by quantum and anatomic noise and highly influenced by variable perception capacity of the reader. Multiple studies have proven that lesions - missed at time of primary interpretation - were visible on the chest radiographs in retrospect. Computer-aided detection (CAD) schemes do not alter the anatomic noise but aim to decrease the intrinsic limitations and variations of human perception by alerting the reader to suspicious areas in a chest radiograph when used as a "second reader". Multiple studies have shown that the detection performance can be improved using CAD, especially for less experienced readers, at a variable amount of loss of specificity. There seems to be a substantial learning process for both experienced and inexperienced readers, to be able to optimally differentiate between false positive and true positive candidates. Readers have to build up sufficient "trust" in the capabilities of these systems to be able to use them at their full advantage. Studies so far focussed on stand-alone performance of the CAD schemes to reveal the magnitude of potential impact, or on retrospective evaluation of CAD as second reader for selected study groups. Further research is needed to assess the performance of these systems in clinical routine and to determine the trade-off between performance increase in terms of increased sensitivity, decreased interobserver variability, loss of specificity, and secondary indicated follow-up examinations for further diagnostic work up.

Background

Lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer-related deaths in the United States. Despite increasing awareness of the deleterious effects of smoking and continuous research in lung cancer diagnosis and treatment, lung cancer mortality is very high and has not changed substantially over the last decade. Yet, it is proven that bronchogenic carcinoma has a much better chance of cure if diagnosed at an early (localized) stage⁽¹⁾. Lung cancer screening using CT is presently subject to intense research with a number of randomized and non-randomized trials still ongoing. While CT is very sensitive for detection of small pulmonary lesions, the vast majority of detected nodules are in fact benign⁽²⁾. The verdict is not yet out whether CT screening is reducing lung cancer-related mortality or in fact leads to a large number of unnecessary interventions and ultimately no survival benefit. Lung cancer screening efforts in the 1980s had focused on the use of chest radiographs. While more cancers were detected and more resections were performed, there was no survival benefit for screened participants. Since then the technique of chest radiography has seen vast improvements including the transition to digital radiography with better contrast resolution and better depiction of previously difficult areas, such as the lung recesses and the retrocardiac space. Detection performance for small lung tumours should therefore be improved. Whether this translates into better patient survival, however, remains to be seen. Lung cancers missed on chest radiographs, however, remain one of the major reasons for lawsuits against radiologists⁽³⁾. Investigators of the Mayo Lung Project reported 75% of the perihilar (12/16) and 90% of peripheral nodules (45/50) had been missed at time of primary interpretation, but were visible on the chest radiographs in retrospect⁽⁴⁾. Similar results were reported by other publications^(5,6). For this reason, detection of suspicious nodules remains a main task for any radiologist reading chest radiographs, even if the primary diagnostic question is not cancer-related. Though its inferior sensitivity to CT, chest radiography represents a fast and relatively cheap imaging modality. It remains a popular modality for the surveillance of pulmonary metastatic disease in patients with known malignancies. Also, in most cases it is the primary method of choice to screen for intrapulmonary abnormalities.

Rationale for CAD

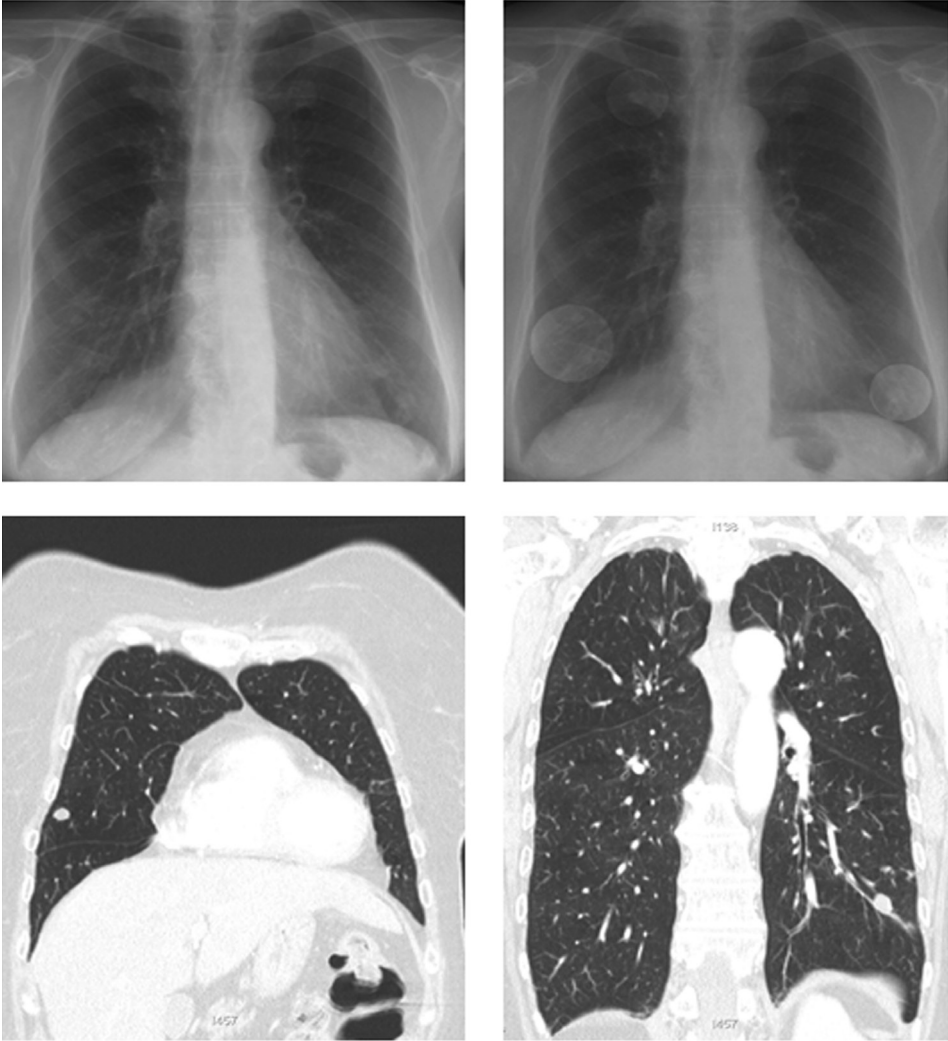
Detection of subtle focal pulmonary opacities on radiographs remains a challenge. It is limited by two categories of noise: 1) the radiographic noise (mottle) related to the quantum nature of radiation^(7,8) and 2) the anatomic noise⁽⁹⁾ which refers to surrounding and overprojecting anatomic structures, such as ribs, abnormalities in the lung parenchyma and vascular structures. The term “conspicuity” of a lesion describes the relation of feature contrast to surrounding complexity, thus including the contributions of both noise categories⁽¹⁰⁾. In chest radiography, anatomic noise appears to have far greater influence on the detection of pulmonary nodules^(11,12). The complexity of the surrounding anatomic noise greatly influences the perception of the radiologists. In an experiment, in which the authors slightly varied the location of simulated nodules, Samei et al. found strong correlations between nodule size, nodule location and observer detection performance⁽⁹⁾. Anatomic structures such as ribs and pulmonary vessels superimposing on a subtle lung nodule on a chest radiograph influenced nodule detection, measured as the area under the receiver operating characteristics (ROC) curve (A_z), by as much as 28%. The effects of distracting anatomic noise are aggravated by the intrinsic limitations of human perception. Perception is influenced by predictable measures such as training and experience, but also by less predictable effects such as concentration, distraction and fatigue. Radiologists are not consistent in what they detect and what they diagnose. The less conspicuous the findings, the greater the variability will be. When confronted with a set of radiographs a second time, the reader may detect different lesions than during the first time. Studies have repeatedly shown that missed lesions were evident in retrospect⁽⁴⁻⁶⁾. To improve interpretation in chest radiography, various technical innovations have been introduced, among which image processing techniques represent the most important tools. Image processing includes techniques such as dual energy subtraction, temporal subtraction and tomosynthesis all aim to increase the conspicuity of a lesion by reducing the impact of anatomic noise. Computer-aided detection (CAD) algorithms have a different approach; they do not alter the anatomic noise. By alerting the reader to suspicious areas in the chest radiograph CAD aims to decrease the intrinsic limitations and variations of human perception. In addition to prototypes only applied under research conditions, there are currently two US Food and Drug Administration (FDA)-approved systems available on the market (IQQA-chest, EDDA technology, Inc Princeton Junction, NJ, USA; Figure 1) and ONGUARD; Riverain Medical, Miamisburg, Ohio, USA; Figure 2), more systems from other manufacturers are expected to follow.

For mammography, the first FDA approved clinical system was introduced already in 1998. Since then systems were continuously improved and received a widespread adoption in the USA, where there is additional reimbursement for the use of CAD⁽¹³⁾. Though being in use already for a longer time, the clinical value of CAD is still being debated⁽¹⁴⁾.

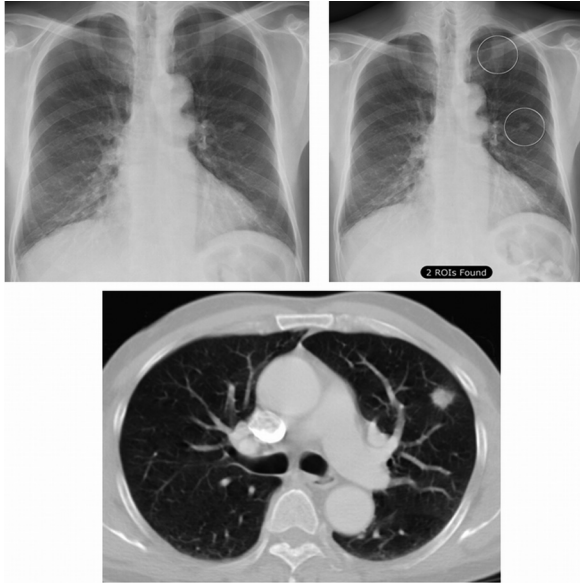
Prerequisites for routine implementation of CAD

Computerized methods for automatic detection of lung nodules were published already in the 1970s. None of these methods, however, were applied in clinical studies at this point, probably because of large numbers of false positive findings generated with these early methods⁽¹⁵⁾. A study from 1993 postulated the potential usefulness of CAD schemes for nodule detection if the false positive rate could be reduced to a level of approximately one (!) false positive detection per radiograph at a sensitivity of 75% ⁽¹⁶⁾. More than 15 years later, even the most advanced CAD algorithms do not live up to this postulated requirement. In the meantime CAD has become clinically available for mammography, and it now appears to be ready also for broader clinical application in the chest. The reason for this difference in acceptance between CAD for mammography and CAD for chest radiography may be based on the different tasks. In mammography the detection task is two-fold and refers to the localisation of small, but high-contrast microcalcifications on one side and the detection of ill-defined soft-tissue masses on the other side. While for the first, the sensitivity of modern CAD systems is reported to be very high (98%), the sensitivity for mammographic masses is significant lower⁽¹⁷⁾. Pulmonary nodules and masses, however, seem to face similar problems with respect to conspicuity and their vulnerability towards the effect of overlying and distracting anatomic structures. All CAD systems are designed to be used as a second reader: they are meant as complementary tool that draws the radiologists' attention to certain image areas that need further evaluation. They are not designed to detect all potential lesions, which would allow the radiologist only to focus on the areas identified by the CAD system. The present systems do not provide a 100% sensitivity, which makes it necessary for the radiologist still to evaluate the whole image. However, the systems can detect additional lesions that might escape the radiologist's attention. In clinical practice, the CAD program runs in the background while the radiologist completes visual interrogation of the posteroanterior (PA) and lateral radiograph. Subsequently suspicious lesions detected by the CAD are revealed and the radiologist has the opportunity to accept or discard the CAD findings.

Figure 1



Two intrapulmonary metastases (see CT), both correctly assigned by the CAD (IQQA-chest, EDDA).

Figure 2

Correctly assigned CT proven T1 tumor (ONGUARD, Riverain) in the left upper lobe and false positive candidate in the left long apex due to crossing of clavicle and first rib.

Usually the findings are indicated by a region of interest (ROI) around the suspicious lesion or by highlighting the suspicious areas. Comparing the original image with CAD candidates can be done side-by-side or by toggling the highlighted areas / circles on and off. Whatever implementation, the application of CAD results in additional reading effort and increased reading time at least for a substantial number of cases, certain readers, and within an initial time period of adaptation. Whether reading time will be generally increased or only before adaptation to the system remains to be evaluated. In order to make these additional investments worth their while, the number of false positive candidates has to be low and the true positive candidate lesions should complement the lesions found by the radiologists: there is no advantage if the lesions found by CAD completely overlap the lesions found by human observers. Another prerequisite for clinical practice is the seamless integration of the CAD into a picture archiving and communication system (PACS). CAD results have to be available on-demand and need to be presented in a fashion that disturbs the normal workflow as little as possible. In addition, CAD should work independently of which digital imaging system has been used for the acquisition of the digital chest radiographs. A few logistic and potentially medico-legal issues have not yet been resolved; whether and for how long candidate lesions have to be stored, and whether CAD candidates should be visible only for radiologists or also for the referring physicians⁽¹⁸⁾. CAD will most inevitably detect lesions that turn out to be true tumours but have been refuted by the

radiologist at time of the evaluation. This may lead to more defensive medicine with unnecessary follow-up radiographs or CT examinations if lawyers are allowed to interpret such a situation as an instance of malpractice. In addition, it will be difficult for less experienced clinical colleagues to differentiate between true and false positives candidates. Again this holds potential for improper follow-ups and conflict between radiologists and referring physicians. For all these reasons, it is probably most appropriate not to store the CAD results in a PACS environment.

CAD results as published so far

Study designs

There are various study setups and statistical models that have been applied to assess the effects of CAD (Table 1). Selection and prevalence of lesions, number and experience of readers, standard of truth, and statistical analysis, represent factors that influence the results and should be considered when drawing conclusions and comparing performance of various CAD systems. The majority of studies are based on a set of radiographs that was selected to include a certain number of exams containing nodules or tumours and a certain number of controls (Table 1). Ideally, the presence or absence of lesions is proven by CT as the superior gold standard. However, such a process introduces a selection bias that depends on the selected type of abnormality and does not reflect the normal clinical situation in which there is no superior standard available for most patients. CAD performance has two components, the stand-alone performance of CAD without human interaction and the effect of CAD on reader performance. Stand-alone performance is a good indicator of the magnitude of the potential effects of CAD; it determines how many and what kind of lesions can be detected and how many false positives per image are produced. Sensitivity and false positives are interrelated: higher sensitivity for a specific CAD algorithm always comes with a higher number of false positive CAD candidates per image. For this reason a compromise has to be made between these two factors. In some studies the operating point was variable and allowed for analysis of this interrelation, in other studies the operating point is fixed and provides only one set of sensitivity and number of false positive CAD candidates per image. Because CAD algorithms cannot detect all potential lesions in a radiograph at a reasonable number false positive candidates, they need to be used as second reader. How much a CAD algorithm is able to improve the performance of a reader depends on the following factors:

- The reader experience; less experience will lead to a potentially greater increase in performance.
- The stand-alone performance of CAD; the higher the detection rate of nodules that are typically missed by human observers, the more effect a CAD will have.
- The ability of observers to distinguish between a true positive and a false positive CAD candidates. Especially the latter has not yet been extensively studied, but is an important factor that will ultimately determine whether CAD will mainly increase the numbers of true lesions found by an observer or whether it will also increase the number of false positives by the observer. It is therefore important that readers become familiar with the behaviour of a "their" CAD system so that they learn the optimum cut-off for differentiating true and false positive candidates. As a result of the unfamiliarity with CAD, observers may dismiss true positive candidates or lose confidence in CAD, because of too many false positive candidates. Statistical analysis of data is important for interpreting the outcome of CAD studies; sensitivity and specificity are influenced by the prevalence of disease within the study group, especially if a high prevalence is known or suspected by the readers. ROC incorporates the interrelation of sensitivity and specificity of human observers depending on their confidence level. Therefore, ROC analysis are to be preferred. Besides that, ROC statistics work on a per-region basis; any positive reading of a region containing a true lesion is counted as true positive, independent of whether the reader had actually identified the lesion or had read a false positive contained in the same region. Localized ROC (L-ROC) analysis forces the reader to indicate a lesion and thus also incorporates the correct localization of a lesion. Any type of ROC statistics, however, requires the availability of a superior gold standard, usually provided by CT examinations.

CAD for nodule detection

First publications⁽¹⁹⁾ are more than 10 years old and were based on the comparison of conventional radiographs and digitised versions of these radiographs that included the superimposed CAD output. Although it can be expected that CAD software has further improved in the meantime, already this first study reported excellent results; diagnostic accuracy improved significantly with CAD (ROC area 0.906 vs. 0.948) while reading time did not increase significantly. The authors also reported a larger benefit for inexperienced radiologists using CAD and demonstrated a subsequent decrease in variability of accuracy across readers of different experience levels. The first study that brought the potential of CAD for nodule detection

to the attention of a large group of radiologists was a large-scale observer test conducted during the 1996 RSNA scientific assembly in Chicago⁽²⁰⁾. Radiologists at the RSNA were invited to evaluate 22 abnormal and 20 normal digitised chest radiographs in random order without and then with CAD. The CAD algorithm had a stand-alone performance of 70% with a mean of only one false positive CAD candidate per radiograph. The 146 observers included chest radiologists, general radiologists and residents. For all categories of readers an improved detection rate was seen with the application of CAD. Az varied from 0.697 to 0.825 without CAD and from 0.80 to 0.88 with CAD. The correct diagnosis was shown after each case, which is unrealistic under clinical conditions, but is likely to have introduced a learning effect. Learning how to use CAD, is crucial. Understanding which lesions CAD is likely to detect and which it is likely to miss is important for a reader to make the decision whether to accept or dismiss a CAD candidate. The reduction in variability between radiologists of different experience was further confirmed by other studies; the less experienced the observer, the greater the improvement in lung nodule detection with CAD will be⁽²¹⁻²³⁾. In 2008 Bley et al. published the stand-alone performance of one of the first commercial CAD systems (IQQA-chest, EDDA technology) for the detection of CT-proven nodules with a mean diameter of 7.5 mm \pm 2.2 mm⁽²⁴⁾. CAD yielded a sensitivity of 39% (stand-alone performance) compared to sensitivities between 18% and 30% for the radiologists. Most interestingly, the agreement among radiologists was larger ($\kappa = 0.64-0.73$) than between radiologists and the CAD algorithm ($\kappa = 0.45-0.52$). This indicates that CAD indeed detected different lesions than radiologists. An important step towards an increased comparability of the performance of various CAD algorithms would be their evaluation using the same study cases; Schilham et al. assessed their own CAD algorithm developed in academia using the Japanese Society of Radiological Technology (JSRT) database, a large and publicly available database. They reported a stand-alone sensitivity of 51% with 2 false positive CAD candidates per image, and a stand-alone sensitivity of 67% with 4 false positive CAD candidates per image⁽²⁵⁾. Other authors detected 78% of the nodules, but at 4 false positive CAD candidates per image⁽²⁶⁾. These results nicely demonstrate the inevitable trade-off between sensitivity and specificity. Kasai et al. evaluated the effect of CAD on the detection of subtle nodules in PA and lateral chest radiographs⁽²⁷⁾. About 50% of the nodules (15/31) were visible only on the PA views. The CAD software indicated candidate lesions on both views. The authors found an increased sensitivity (65% vs. 68%) with CAD, but no significant increase of the Az (0.804 vs. 0.816, $p = 0.297$) indicating that the increased sensitivity was

counteracted by an increased number of false positive marks with CAD. The only study so far that tried to assess the usefulness of CAD in a clinical environment was published in 2008 by van Beek et al⁽²⁸⁾. They investigated the benefit of CAD for nodule detection on chest radiographs obtained in daily practice. The study group exclusively consisted of patients with known extra-pulmonary primary malignancies, who were under surveillance for metastatic disease. In 214 of the 324 chest radiographs a standard of reference could be established by CT or follow-up of at least 6 months. Nodules were present on 55 of the 214 chest radiographs. The sensitivity significantly increased with CAD from 64% to 93%; with only a nonsignificant decrease in specificity (93% vs. 96%).

Table 1

List of publications using FDA approved and prototypes of CAD schemes for the detection of nodules and T1 bronchogenic carcinomas.

Author	System	n	Prevalence	CAD stand-alone	FP per image by CAD	Reader vs. reader with CAD
CAD for pulmonary nodules						
v Beek (2008)	EDDA	214	26%	-	-	sensitivity 64% vs. 93%*
Bley (2008)	EDDA	117	36%	39%	2.7	-
Hardie (2008)	not FDA	154	100%	78%	4	-
He (2008)	EDDA	116	50%	67%	2.4	-
Kasai (2008)	not FDA	60	52%	52%	4.2	sensitivity 65% vs. 68%* Az 0.804 vs. 0.816
Shiraishi (2007)	not FDA	106	100%	71%	4.9	-
Schillham (2006)	not FDA	247	62%	51%	2	-
				67%	4	-
Shiraishi (2006)	not FDA	48	75%	62%	-	Az 0.724 vs. 0.778*
Shiraishi (2006)	Riverain	459	100%	70%	5	
Song (2005)	EDDA	232	36%	71%	2.8	sensitivity 49% vs. 80%*
Coppini (2003)	Riverain	128	-	60% **	4,3	
				75% **	10.2	
Shiraishi (2003)	not FDA	90	60%	100% **	3.1	Az 0.682 vs. 0.808*
Freedman (2002)	not FDA	240	67%	66%	5	Az 0.835 vs. 0.865*
MacMahon (1999)	not FDA	40	50%	80% **	1 **	Az 0,825 vs. 0.889*
Xu (1997)	not FDA	200	50%	70%	1.7	-
Kobayashi (1996)	not FDA	120	50%	-	-	Independent Az 0.894 vs 0.940*
Matsumoto (1992)	not FDA	198	48%	60%	15	-
CAD for T1 lungcarcinoma						
White (2009)	Riverain	114	100%	47%	3.9	-
Li (2008)	Riverain	34	100%	34%	5.9	-
Sakai (2006)	not FDA	100	50%	74%	2.2	Az 0.896 vs. 0.923*
Kakeda (2004)	not FDA	90	50%	-	3.2	Az 0.924 vs. 0.986*

FP: false positives; Az: area under ROC curve.

* significant difference

** fixed

In summary, CAD for nodule detection has shown to increase reader performance with better sensitivity, especially of inexperienced readers, and to decrease inter- and intra-reader variability.

CAD for detection of early stage bronchogenic cancer

The first studies evaluating the effects of CAD on the detection of T1 tumours were published by Japanese colleagues. This task may differ from detecting metastases because nodule characteristics of T1 tumours are often different from those of metastases. A significantly increased detection rate for early lung tumours ($n = 45$, mean diameter 18 mm; range between 8 and 25 mm) was described by Kakeda et al. who evaluated CAD using 8 radiologists of varying experience⁽²¹⁾. The Az increased from 0.924 to 0.986, and thus started already at a very high baseline. Board certified radiologists performed better and profited less from CAD than residents, but even residents had a baseline performance without CAD that exceeded 0.90. These results were confirmed by Sakai et al. who also assessed the detection of T1 tumours by 8 readers of varying experience⁽²⁹⁾. Four chest radiologists and 4 residents evaluated PA radiographs of 50 patients with T1 tumours (< 3 cm in diameter) and 50 controls. Stand-alone performance of CAD included a sensitivity of 74% and a mean number of false positive CAD candidates of 2.3. Az significantly increased from 0.896 to 0.924 with CAD. CAD might help to reduce the number of T1 bronchogenic tumours originally missed in the reports at the time of clinical evaluation. White et al. reassessed the detectability of missed lung cancers in 89 patients⁽³⁰⁾. They included all prior images of missed cancers in their study and were able to include 114 positive images. The CAD system (ONGUARD; Riverain Medical) correctly identified 53 (47%) of the originally missed lung cancers, which indicates the potential of CAD to reduce the number of missed cancers. However, this study only evaluated the stand-alone performance of CAD, which does not include reader response to CAD candidate. It is not clear how many of the tumours detected by CAD would be accepted by readers and at what cost in terms of accepting false positive CAD candidates. Li et al. published the stand-alone performance of the same CAD scheme for a selected group of 34 CT proven T1 cancers that had been missed in the original chest radiography reports⁽³¹⁾. CAD detected 35% of these originally missed lesions. Sensitivity was 45% for the more obvious and 30% for the more subtle cases. The mean number of false positive CAD candidates per image 5.9.

Differentiation of benign from malignant lesion

Differentiating benign from malignant pulmonary nodules on chest radiographs is a difficult and in many cases impossible task. Usually patients with newly identified non-calcified nodules undergo further diagnostic workup that may include CT, PET or biopsy. The superiority of PET-CT for detecting metabolic activity and the ability to perform accurate growth measurements with CT have made early attempts to use morphological features on radiography for nodule differentiation less attractive⁽³²⁾. However, in 2003 Shiraishi et al. presented a CAD algorithm trained to discriminate benign nodules from malignancies with a performance better than even experienced radiologists ($p < 0.002$)⁽³³⁾. In addition to clinical features such as age, sex, and history of the patients, a computerized analysis of morphological features, such as irregularity, density, and contrast had formed the base for this CAD scheme. With CAD the increase in performance was higher for experienced than for inexperienced radiologists, which suggests that CAD and personal skills are complementing effects. Most interestingly, the stand-alone performance of CAD was higher than the performance of each radiologist independent of experience. This demonstrates that radiologists may be reluctant to accept the suggestions of a CAD⁽³³⁾. In a second study the same group of authors tested the combined -detection and classification - task using an advanced CAD scheme and reported similarly significant improvement for both, detection and classification with application of CAD (Az 0.724 vs. 0.778, $p = 0.008$)⁽³⁴⁾. In summary, both detection and classification of T1 tumors has been shown to increase with the application of CAD which is especially appealing for lesions that had been missed during the primary reports.

Limitations of present studies

The vast majority of studies assessed the performance of CAD for a selected study group with a varying prevalence and a varying conspicuity of lesions. The selection of the lesions, with respect to size, location, and overall conspicuity, is tremendously important when interpreting the results. Additional information, such as the fact that the lesions had been originally missed or a relatively low sensitivity below 40%, serve as indicators for a generally low conspicuity of the majority of lesions. Inclusion of many small nodules (5 – 10 mm, based on CT) of which many might be hardly or not at all visible on a chest radiograph limits the assessment of CAD, as well as the inclusion of so obvious lesions that the performance without CAD is already exceeding an Az of 0.9. As recently pointed out in an editorial by D. Gur, the higher the baseline performance is to start with, the more difficult it becomes

to prove a clinically relevant improvement by a certain technique e.g., CAD) ⁽³⁵⁾. The consequence is that the various studies show widely varying levels of improvement, due to the varying level of baseline performances. He also stressed the important point how difficult the appreciation of a 'clinically relevant performance difference' is as one approaches a perfect performance level. While statistics may show significance already for small differences, the perspective in terms of clinical relevance and importance remain frequently unanswered. In studies with a limited and selected group of readers (mostly four to six) the chance that outliers have a great influence on the outcome results is high. Freedman graphically demonstrated the diversity across readers and cases without and with the use of CAD by so called "heat maps" ⁽³⁶⁾. These "heat maps" display three-dimensional data in two dimensions with the third dimension represented by colour. They were found well suited to document the complexity of reader variability as function of lesion conspicuity, experience and level of training, and the interaction between these facts with CAD. This diversity, however, is easily missed in summary statistics, but represents the determinant factor in the individual (clinical) situation. An equally important point is the integration of readers' behaviour. For organizational reasons it is obviously easier to assess the stand-alone performance of a CAD system without including readers in the study. But, it should not be forgotten that these systems are thought to be used as second reader. The highest performance can be expected when readers use CAD as complimentary tool. To be able to use a CAD to its full advantage, readers show a learning curve; they have to become familiar with the potential and limitations of the used CAD algorithm. Readers have to build up a "trust" in order to be able to accept the true positive candidates without a too big loss of specificity by correctly dismissing false positive candidates. More experience has to be gained with respect to the influence of the processing on the performance of CAD. Though there is a tendency towards a consensus to what represents a "good quality" chest radiograph with respect to processing, there remains still room for individually customized version that may drastically differ from the image display in other institutions. Most manufacturers use some type of multi-scale multi-frequency processing which has similar, but not identical effects on the image display. He et al. found a significant impact on the performance of CAD for the detection of nodules when comparing a default processing (parameter 0.0 for structure preference) with a high-pass filtered image (parameter 0.4 for structure preference) and a low-pass filtered image (parameter -0.4 for structure preference) ⁽³⁷⁾. There is so far only one study that aimed to investigate the effect of

CAD in daily practice⁽²⁸⁾. It is thus far the only study which provides some information about the order of magnitude of potential improvement through CAD which can be expected in clinical routine and how many radiographs have to be evaluated for that. The results are valid for a specific group of oncology patients with known extra-pulmonary malignancies. In a clinical setup, establishment of truth remains a problem because not all of these patients can undergo CT and vice versa inclusion of only patients with CT would represent a selection bias. It is therefore not surprising that only for 66% of the original study images, a truth could be established based on follow-up and CT. Prevalence of disease was unusually high with 26% (55/214) of patients with lung nodules. According to the authors, pathology (e.g., nodules) was established in 19 of the 55 patients only with the help of CAD, resulting in an increase of sensitivity from 64% to over 90%. Further clinical evaluation, however, confirmed nodules in only 16 patients, and in only 5 of the 16 patients the nodules were proven to be malignant.

Discussion

Sensitivity versus specificity

Increase in sensitivity and decrease of specificity are somewhat correlated to each other. This is easily understandable if we consider the concept of each reader's "individual threshold" to accept or dismiss a lesion. This concept is true for lesions the reader detected himself as well as for CAD candidates. Even though many, if not most of the false positive CAD candidates, may be easily dismissed, there is a substantial amount of lesions which are difficult to differentiate between projection effects, scarring and a potential malignancy. These lesions increase in number in patients with pre-existing lung disease, such as chronic obstructive lung disease, chronic airways disease or interstitial lung disease. Especially for these patient groups it seems inevitable that a substantial increase in sensitivity associated with CAD goes along with a decrease in specificity. The opposite effect is seen when readers keep the threshold high: a substantial number of correctly indicated lesions by CAD will not be accepted as such. In summary learning effects to get adjusted to the characteristics, e.g., capacities as well as limitations of the CAD algorithm seem to be very important. The detrimental effect of decreasing specificity, or at least substantially decreasing reader confidence, is best studied in lesion-free images ("controls"). Images, previously called not suspicious will now be called suspicious or cases, previously only being characterized as being of low suspicion will now be called "highly suspicious", when

also identified by CAD. The ability of differentiating false from true positive CAD candidates is significantly correlated with the experience of the reader and the amount of anatomic background noise in the source image. Yet, a decrease in specificity with the application of CAD not only applies for inexperienced readers. A quantification of the secondarily induced diagnostic work-up seems warranted with respect to radiation dose, financial aspects, number and invasiveness.

Decrease of intra- and interobserver variability

Freedman et al. demonstrated that most of the lesions newly identified by one radiologist using CAD were actually cases that the radiologist himself or other radiologists would have identified had they interpreted the images without CAD at another point of time⁽³⁶⁾. Even when the application of CAD did not succeed in significantly increasing the sensitivity, in most studies so far it could be shown that it decreases intra- and interobserver variability⁽³⁶⁾.

Value of a high negative predictive value of CAD

The majority of radiographs are indeed done to exclude pathology. It is conceivable that the increase in confidence by CAD to exclude pathology ("I did not see a nodule and CAD also did not find a nodule") will play a substantial role in clinical practice. This depends, of course, on the prevalence of disease in the patient group. However, it has to be noted that the less than 100% sensitivity of CAD requires its use as "second reader".

Options for the future

Computer-aided detection (CAD) has become one of the major research topics in medical imaging and diagnostic radiology. It has been applied to various imaging modalities including CT, MRI, ultrasound, and radiography. Current CAD schemes appear to have the potential to increase the sensitivity for detection of focal lung lesions, such as small tumours and metastases. CAD appears to work best for less experienced readers. At the same time CAD will slightly increase the number of false positive readings. Up to now there is little evidence about how CAD performs in a clinical environment with regard to diagnostic accuracy on one side and negative effects, such as unnecessary follow-up, on the other side. Publications so far tended to focus on the effects of CAD on diagnostic accuracy, its effects on productivity may eventually prove to be equally important. Development and evaluation of these techniques require large databases with validated clinical images. Ideally, common databases should become

available, which can be used to train and test CAD systems, and improve comparison of future commercial systems. This issue is being addressed successfully by several initiatives with the goal to provide a database of web accessible cases for use in CAD research⁽³⁸⁾.

References

- 1 Henschke CI, McMauley DI, Yankelevitz DF, et al. Early lung cancer action project: overall design and findings from baseline screening. *Lancet* 1999; 354:99-105
- 2 Swensen SJ, Jett JR, Hartmann TE, et al. CT screening for lung cancer: five year prospective experience. *Radiology* 2005; 235:259-265
- 3 White CS, Sali AI, Meyer CA. Missed lung cancer on chest radiography and computedtomography:imagingandmedico-legalissues.*JThoracImag*1999;14:63-68
- 4 Muhm JR, Miller WE, Fontana RS, Sanderson DR, Uhlenhoop MA. Lung cancer detected during a screening program using 4-month chest radiographs. *Radiology* 1983; 148:609-615
- 5 Quekel LG, Kessels AG, Goei R, Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest* 1999; 115:720-724
- 6 Shah PK, Austin JH, White CS, et al. Missed non-small cell lung cancers: Radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology* 2003; 226:235-241
- 7 Gavelli G, Giampalma E. Sensitivity and specificity of chest x-ray screening for lung cancer. *Proceedings of the International Conference on Prevention and Early Diagnosis of Lung Cancer*. 1998: 103-108
- 8 Burgess AE, Wagner RF, Jennings RJ. Human signal detection performance for noisy medical images. *IEEE Computer Soc Int Workshop Med Imaging* 1982: 88-105
- 9 Samei E, Flynn MJ, Eyler WR. Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs. *Radiology* 1999; 213:727-734
- 10 Kundel HL, Revesz G. Lesion conspicuity, structured noise, and film reader error. *AJR Am J Roentgenol* 1976; 126:1233-1238
- 11 Boynton RM, Bush WR. Recognition of forms against a complex background. *J Opt Soc Am* 1956; 46:758-764
- 12 Revesz G, Kundel HL, Graber MA. The influence of structured noise on the detection of radiologic abnormalities. *Invest Radiol* 1974; 9:479-486
- 13 Bick U, Diekmann F. Digital mammography : what do we and what don't we know. *Eur Radiol* 2007 ; 17 :1931-1942
- 14 Astley SM, Gilbert FJ. Computer-aided detection for screening mammography (review). *Clin Radiol* 2004 ; 59:390-399
- 15 Van Ginneken B, te Haar Romeny BM, Viergever MA. Computer-aided diagnosis in chest radiography: a survey. *IEEE Trans Med Imaging* 2001 ; 20:1228-1241
- 16 Matsumoto T, Doi K, Kano A, Nakamura H, Nakanishi T. Evaluation of the potential benefit of computer-aided (CAD) for lung cancer screenings using photofluorography: analysis of an observer study. *Nippon Acta Radiol* 1993 ; 53:1195-1207
- 17 Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000 ; 215:554-562
- 18 Chen JJ and White CS. Use of CAD to evaluate lung cancer on chest radiography. *J Thorac Imaging* 2008; 23:93-96
- 19 Kobayashi T, Xu XW, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology* 1996; 199:843-848
- 20 MacMahon H, Engelmann R, Behlen FM, et al. Computer-aided diagnosis of pulmonary nodules: results of a large scale observer test. *Radiology* 1999; 213:723-726
- 21 Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. *AJR Am J Roentgenol* 2004; 182:505-510

- 22 Shiraishi J, Katsuragawa S, Ikezoe J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristics analysis of radiologists' detection of pulmonary nodules. *AJR Am J Roentgenol* 2000; 174:7-74
- 23 Song W, Fan L, Xie Y, Qian JZ, Jin Z. A study of inter-observer variations of pulmonary nodule marking and characterizing on DR images. *Proc SPIE* 2005; 5749:272-280
- 24 Bley TA, Baumann T, Saueressig U, et al. Comparison of radiologist and CAD performance in the detection of CT-confirmed subtle pulmonary nodules on digital chest radiographs. *Invest Radiol* 2008; 43:343-348
- 25 Schilham A, van Ginneken B, Loog M. A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database. *Med Image Anal* 2006; 10:247-258
- 26 Hardie RC, Rogers SK, Wilson T, Rogers A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Med Image Anal* 2008; 12:240-258
- 27 Kasai S, Li F, Shiraishi J, Doi K. Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs. *AJR Am J Roentgenol* 2008; 191:260-265
- 28 Van Beek EJ, Mullan B, Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images: a prospective study. *Acad Radiol* 2008; 15:571-575
- 29 Sakai S, Soeda H, Takahashi N, et al. Computer-aided detection on digital chest radiography: validation test on consecutive T1 cases of resectable lung cancer. *J Digit Imaging* 2006; 19:376-382
- 30 White CS, Flukinger T, Jeudy J, Chen JJ. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology* 2009; 252:273-281
- 31 Li F, Engelmann R, Metz CE, Doi K, MacMahon H. Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. *Radiology* 2008; 246:273-280
- 32 Gietema HA, Schaefer-Prokop CM, Mali W, Groenewegen G, Prokop M. Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT- influences of inspirations level, nodule size, and segmentation performance. *Radiology* 2007; 245:388-394
- 33 Shiraishi J, Abe H, Engelmann R, et al. Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance – initial experience. *Radiology* 2003; 227:469-474
- 34 Shiraishi J, Hiroyuki A, Li F, et al. Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs ROC analysis of radiologists' performance. *Acad Radiol* 2006; 13: 995-1003
- 35 Gur D. Imaging technology and practice assessment studies: importance of the baseline or reference performance level. *Radiology* 2008; 247:8-11
- 36 Freedman M, Osicka T. Heat maps: an aid for data analysis and understanding of ROC CAD experiments. *Acad Radiol* 2008; 15:249-259
- 37 He Q, He W, Wang K, Ma D. Effect of multislice processing in digital chest radiography on automated detection of lung nodule with a computer assistance system. *J Digit Imaging* 2008; 21:164170
- 38 MacMahon H. Advanced image processing and computer-aided diagnosis: are we there yet? *J Thorac Imaging* 2008; 23:75-76

Computer-aided Detection of Lung Cancer on Chest Radiographs: Effect on Observer Performance

Bartjan de Hoop
Diederick W De Boo
Hester A Gietema
François van Hoorn
Banafsche Mearadji
Laura Schijf
Bram van Ginneken
Mathias Prokop
Cornelia M Schaefer-Prokop

Abstract

Objective

To assess how computer-aided detection (CAD) affects reader performance in detecting early lung cancer on chest radiographs.

Material and Methods

In this ethics committee-approved study, 46 individuals with 49 computed tomographically (CT)-detected and histologically proven lung cancers and 65 patients without nodules at CT were retrospectively included. All subjects participated in a lung cancer screening trial. Chest radiographs were obtained within 2 months after screening CT. Four radiology residents and two experienced radiologists were asked to identify and localize potential cancers on the chest radiographs, first without and subsequently with the use of CAD software. A figure of merit was calculated by using free-response receiver operating characteristic analysis.

Results

Tumor diameter ranged from 5.1 to 50.7mm (median, 11.8 mm). Fifty-one percent (22 of 49) of lesions were subtle and detected by two or fewer readers. Stand-alone CAD sensitivity was 61%, with an average of 2.4 false positive annotations per chest radiograph. Average sensitivity was 63% for radiologists at 0.23 false positive annotations per chest radiograph and 49% for residents at 0.45 false positive annotations per chest radiograph. Figure of merit did not change significantly for any of the observers after using CAD. CAD marked between five and 16 cancers that were initially missed by the readers. These correctly CAD-depicted lesions were rejected by radiologists in 92% of cases and by residents in 77% of cases.

Conclusion

The sensitivity of CAD in identifying lung cancer depicted with CT screening was similar to that of experienced radiologists. However, CAD did not improve cancer detection because, especially for subtle lesions, observers were unable to sufficiently differentiate true positive from false positive CAD annotations.

Introduction

Chest radiography is still the most commonly used technique in clinical routine to rule out chest disease, to study the effect of treatment, and to follow up patients. Missing a lung cancer on a chest radiograph is one of the most frequent causes for malpractice lawsuits in radiology⁽¹⁾. However, the task of detection focal lung lesions is challenging: sensitivity for detecting bronchopulmonary malignancies with chest radiography ranges only 36%-84%, depending on study population and tumor size⁽²⁻⁶⁾. Several authors have reported that many missed lesions can be detected in retrospect⁽⁷⁻¹⁰⁾. Pulmonary lesions can be missed for two reasons: either they are overlooked or they are misinterpreted as normal structures. To increase the sensitivity of chest radiographs in depicting pulmonary nodules, computer-aided detection (CAD) systems are currently being developed. The goal of CAD is to identify lesions that might be overlooked and missed by the reader. The current standard paradigm for the use of CAD systems is to use CAD as a second reader. After the radiologist has evaluated the image, CAD offers a number of candidate lesions, which have to subsequently be accepted by the radiologist as true positive (TP) or rejected as false positive (FP) annotations. The final detection rate by the reader will be influenced by the interaction between the reader's perception, the performance of the CAD system, and the reader's capability to differentiate TP from FP candidate lesions. CAD is most useful when it is able to depict lung cancer in patients who have an increased risk but who do not have disease-related symptoms, meaning that the tumor is detected at such an early stage that it is characterized by a better prognosis⁽¹¹⁾. We therefore conducted a case-control study that included only cases of lung cancer that were detected during a CT-screening study. CAD systems are constantly being improved with the aim of increasing sensitivity, while simultaneously decreasing the number of FP lesions. The purpose of this study was to assess how CAD affects reader performance in the task of detecting early lung cancer on chest radiographs.

Material and Methods

Study population

All chest radiographs used in this study were retrospectively collected from participants from two centers (Utrecht and Groningen, the Netherlands) of the Dutch-Belgian Randomized Lung Cancer Screening, or NELSON, trial⁽¹²⁾. This trial was approved by the Ministry of Health and by the ethics committee of each participating hospital. Participants were aged between 50 and 75 years and were current or former heavy smokers, reflecting a population with high risk of developing cancer. In this population of 4938 participants at the two centers, chest radiographs may be ordered for preoperative routine work-up or for follow up of screening-detected lesion and also for clinical causes unrelated to screening. We included all chest radiographs obtained between April 2004 and January 2008 in this group of subjects under the following conditions: In patients whom a pulmonary malignancy was detected at screening CT and was histologically confirmed (cases), chest radiographs had to be performed within 6 weeks after screening CT; in the other subjects (control subjects), chest radiography had to be performed within 2 months of screening CT and no nodules larger than 5 mm in diameter had to be present at the screening CT. In total, the subjects had 43 nodules that were smaller than 5 mm at CT. These nodules had an average diameter of 3.9 mm (range, 3.1-5.0 mm) and did not show malignant growth during follow-up in the CT lung cancer screening study. None of the control subjects developed lung cancer during this follow-up period. Chest radiographs for which the radiology report mentioned pulmonary abnormalities other than chronic obstructive pulmonary disease were excluded.

Acquisition of images

All chest radiographs were obtained by using a cesium iodine amorphous silicon flat-panel detector unit (DigitalDiagnost, Philips, Best, the Netherlands). Images were processed by using nonlinear multifrequency-band processing⁽¹³⁾; parameters recommended by the manufacturer were used. For all patients, posteroanterior and lateral projections were available. The screening CT examinations were performed with 16 x 0.75-mm collimation at 30 mAs and 120-140 kV, depending on weight. Sections of 1 mm thickness were reconstructed every 0.7 mm.

Standard of reference

In the cancer-positive cases, the exact location of each nodule on a chest radiograph was determined by two observers who did not participate as reader (B.d.H., radiology researcher with 3 years experience in reading CT lung cancer screening studies). In case of doubt, he consulted an independent chest radiologist (M.P.). This chest radiologist also judged whether lesions were retrospectively visible on a chest radiograph. Both had access to the chest radiograph, as well as screening CT scans. Lesions that were, even with knowledge of the CT findings, not visible on the chest radiograph were excluded from analysis. Findings of all screening CT examinations were evaluated for nodules according to the criteria set by the lung cancer screening program⁽¹⁴⁾. Volumetric software (Lung Care 5 VB10A-W; Siemens Medical Solutions, Erlangen, Germany) was used to assess nodule volume. This volume was used to calculate the diameter on the basis of the assumption of a perfect sphere.

CAD system

We used a commercially available CAD system (Onguard 5.0; Riverain, Miamisburg, Ohio). The software highlights regions suspicious for containing a focal lung lesion by placing a circle of 5 cm in diameter around the suspicious area (Figure 1). Images are automatically processed in the background so that results are immediately available on demand when the chest radiograph is being read by a radiologist. The program only analyzes the posteroanterior or anteroposterior projection. According to the manufacturer, the algorithm was optimized to detect nodules of 9-30 mm in diameter, although in clinical practice, it also marks larger and smaller nodules.

Stand-alone CAD performance

To assess stand-alone performance of the CAD system, annotations were labeled TP if the suspicious lesion was located at least partially within the central 50% of the circular CAD annotation.

Observer study

Images were evaluated on Digital Imaging and Communications System in Medicine-calibrated liquid crystal display monitors (MFGD 3220D; Barco, Kortrijk, Belgium) with a matrix size of 2048 x 1536). Options for magnification and adaption of window settings were available. All chest radiographs were anonymized. Posteroanterior and lateral images were available for evaluation. Chest radiographs were shown in alphabetical order on the basis of patient name to six

61

independent observers. The observer varied in their level of experience: one general radiologist with six years of experience (observer A), one chest radiologist with more than 20 years of experience (observer B) and four radiology residents with experience that varied from 1 to 4 years (observer C-F). Observers knew that the study group was chosen from a lung cancer screening trial and they were also told that some patients might have more than one malignant lesion. Two of the observers (reader B and E) had used the CAD system before during other reading studies, but none of the readers had routine experience. To familiarize the observers with the CAD system, five cancer cases that were not included in the observer study were shown to the observers without and with CAD annotations before the start of the study. Each chest radiograph was first evaluated without and subsequently with CAD results, and observer readings were recorded separately. On a per-patient basis, the observers were asked to document all potentially malignant focal abnormalities seen on the chest radiograph on a separate paper printout with respect to the anatomic lesion locations and the readers' confidence scores by using a four-point scale (score 1: potential lesion, very low degree of suspicion; score of 2: dubious lesion; score of 3: probable lesion; and score of 4, definite lesion). Observers were allowed to mark multiple suspicious lesions on each chest radiograph. They were instructed, however, to ignore nodules smaller than 5 mm in diameter. The researcher (B.d.H.) and the experienced radiologist (M.P.) who had not been involved in the readings analyzed all paper printouts, with the chest radiograph and CT scans being available. The readers' markings were considered TP if the centers of the markings were within the boundaries of the nodules on the chest radiograph. Locations that did not match with a lesion were classified as FP.

Data analysis

Free-response receiver operating characteristics (FROC) analysis of the observer study was performed as described by Swensson⁽¹⁵⁾ on a per-marker basis. Jackknife FROC, especially developed to analyze free-response tasks⁽¹⁶⁻¹⁸⁾, was used to analyze the FROC data. Jackknife FROC software (JAFROC, version 2.3a; <http://www.devchakraborty.com>)⁽¹⁶⁻¹⁹⁾ was used to compute a figure of merit (FOM). The FOM is defined as the probability that lesions (including unmarked lesion) are rated higher than nonlesion marks on control chest radiographs⁽¹⁷⁾, or, in other words, that lesions are give a higher confidence rating for the presence of malignancy than normal findings. Normal images with no marks and unmarked lesions are assigned a zero rating. The level of significance was corrected for multiple comparisons by using Bonferroni correction. Sensitivity

was calculated as the number of TP markings divided by the total number of malignancies. All observer markings, even those that were scored with low confidence, were included to calculate the sensitivity and FP rate. Since it is controversial whether application of CAD as second reader also allows for discharge of candidates seen without CAD⁽²⁰⁾, we also evaluated a situation in which the observers could only increase their suspicion with CAD while preserving all lesion locations seen without CAD. In an effort to understand the effect of lesion conspicuity on our results, we performed a separate jackknife FROC analysis on conspicuous nodules, defined as lesions that were detected by three or more readers. To test for demographic differences between case and the control subjects, we compared both groups with respect to sex by using a chi-square test and age by using a student t-test. P values less than 0.05 were considered to indicate a significant difference.

Results

Sample characteristics

A total of 46 participants with 49 histologically proven pulmonary malignancies met the criteria for cancer-positive case. Sixty-five subjects met the criteria for control cases. Indications for acquisition of the chest radiograph in the control group were exclusion of acute cardiovascular disease (n = 18), chronic obstructive pulmonary disease (n = 18), screening of lung abnormalities because of rheumatoid arthritis (n = 10), preoperative screening (n = 10), unexplained fever (n = 4), chronic cough (n = 3), malaise (n = 1), and trauma (n = 1). Cases did not differ significantly from the control subjects with respect to age and sex (Table 1). Tumor diameter ranged from 5.1 to 50.7 mm (median, 12.0 mm), with two lesions being larger than 30 mm. Conspicuity of malignancies was very variable: Ten of 49 (20%) malignancies were detected by all six observers without the use of CAD. Furthermore, 11 malignancies were detected by five observers, two were detected by four observers, six were detected by three observers, five were detected by two observers, and seven were detected by only one observer without the use of CAD. Eight (16%) malignancies were not detected by any of the observers without or with the use of CAD. None of the 43 small benign nodules in the control group was marked by either the CAD system or any of the observers.

Table 1

Demographic characteristics of study participants.

	Cases (n = 46)	Controls (n = 65)	p value
Mean age, yrs	64.0 (6.0)*	62.5 (5.3)*	0.18
No of men/women	41/5	54/11	0.37

*Data in parenthesis are standard deviations.

CAD stand-alone performance

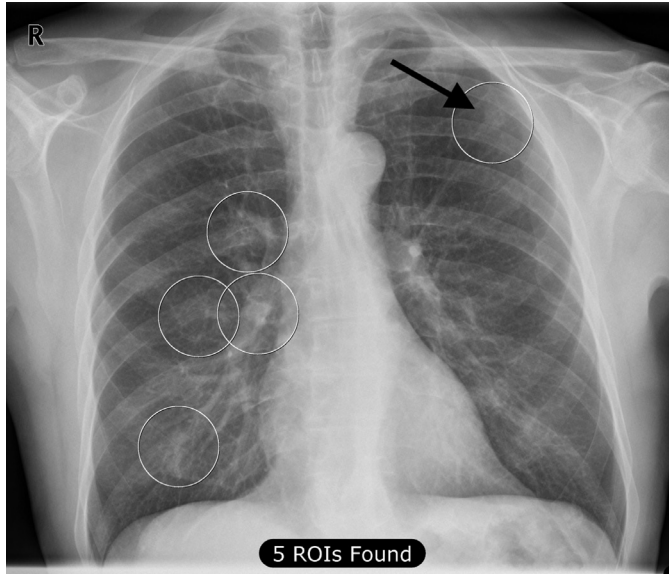
The CAD stand-alone sensitivity was 61% (30 of 49), with an average of 2.4 FP annotations (range, zero to five) per chest radiograph. CAD depicted three malignancies that were initially not detected by any of the observers. The diameter of the CAD-depicted malignancies ranged from 7.0 to 50.7 mm.

Observer performance without CAD

Without CAD, the FOM was 0.72 for radiologists and 0.58 for residents (Table 2, Figure 2). The radiologists had an average sensitivity of 63%, with 0.23 FP annotations per chest radiograph. The residents had an average sensitivity of 49%, with 0.45 FP annotations per chest radiograph. Twenty-seven lesions were detected by at least three observers. In this subselection of more conspicuous lesions, the average FOM was 0.93 for radiologists and 0.76 for residents, with an average sensitivity of 96% for radiologists and 75% for residents.

Observer performance with CAD when lowering of confidence scores was allowed

When the readers were allowed to change their ratings depending on CAD suggestions, average FOM for the radiologists did not change (0.72, $p = 0.98$). Average FOM for the residents increased from 0.58 to 0.61, but the improvement was not significant ($p = 0.08$) (Table 2). With CAD, the average sensitivity of radiologists and residents remained virtually unchanged, 61% and 51%, respectively. Specificity improved, from 0.23 to 0.19 FP annotations per chest radiograph for radiologists and from 0.45 to 0.36 FP annotations for residents. In the subselection of conspicuous lesions, average FOM remained 0.93 for radiologists, but significantly improved for residents (from 0.76 to 0.82, $p < 0.001$). Sensitivity remained 96% for radiologists, but improved from 75% to 84% for residents.

Figure 1

Chest radiographs shows TP (arrow) and FP CAD annotations in a patient with a malignancy in the left upper lobe. ROIs = regions of interest.

Observer performance with CAD when lowering of confidence scores was not allowed

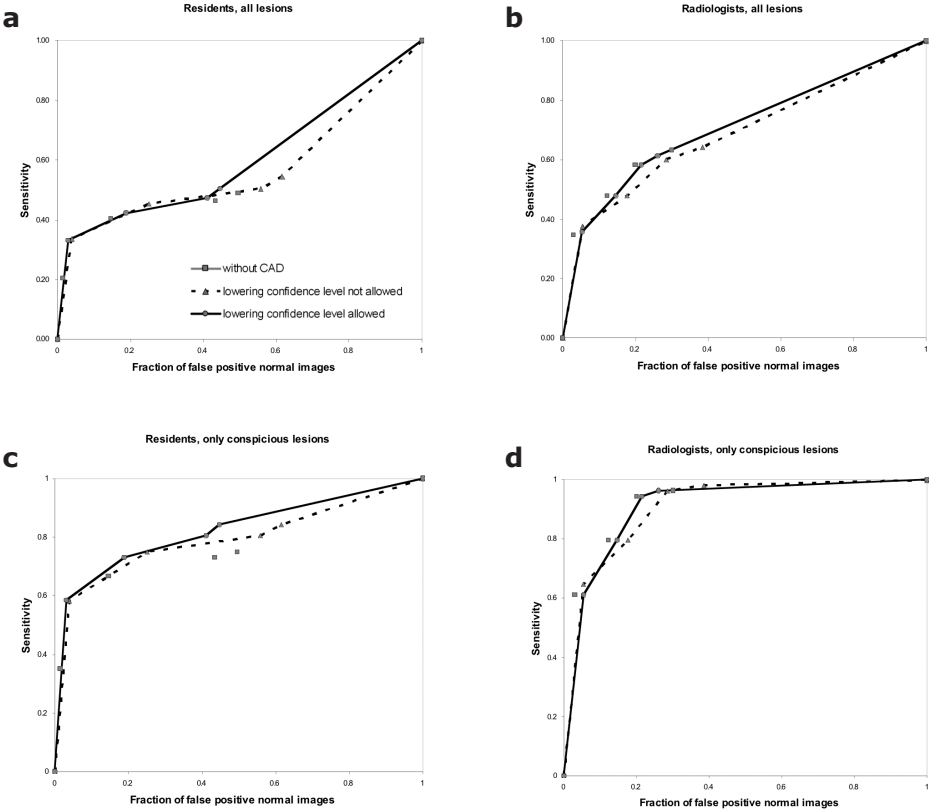
When readers were only allowed to increase their confidence scores after CAD results, average FOM decreased from 0.72 to 0.70 for radiologists ($p < 0.001$) and from 0.58 to 0.57 for residents ($p = 0.60$). Average sensitivity increased from 63% to 64% (range, 63%-65%) for radiologists and from 49% to 55% (range, 41%-69%) for residents, but the average number of FP annotations per chest radiograph also increased from 0.23 to 0.31 and 0.45 to 0.54, respectively.

Interaction between CAD and readers

Together, the six observers placed a total of 66 new markings after having CAD results: 12 for TP CAD annotations and 54 for FP CAD annotations. The number of additionally detected malignancies following TP CAD annotations ranged from zero to six for the various observers (Table 3). The residents benefited more from CAD than did the radiologists, but they also accepted more FP CAD annotations, on average one per 11 chest radiographs versus one per 19 chest radiographs for the radiologists. Observers A, B, C, D, E and F, respectively, dismissed 23, 4, 35, 35, 3, and 17 of their own initial markings because CAD had not annotated these regions (Table 4). The number of malignancies initially not seen by the observers but correctly annotated by CAD varied between five and 16 per observer. Eighty percent (47 of 59) of these TP CAD annotations were rejected by the observers (Table 3). An example is shown in Figure 3.

The average confidence levels were generally low for new TP markings, new FP markings, and markings that were initially called but later dismissed after seeing CAD annotations, with confidence levels of 1.9, 1.8, and 1.6, respectively.

Figure 2



Alternative FROC curves for the detection of pulmonary malignancies by residents (a, c) and radiologists (b, d). Separate analysis for all lesions (a, b) and more conspicuous lesions that were seen by more than 2 observers (c, d). The FOM, which is the area under the AFROC curve, improved significantly for detection of more conspicuous lesions by residents if they were allowed to freely adjust their level of confidence after being provided with the CAD output. The remaining AFROC curves did not significantly change with the use of CAD.

Table 2

Individual outcome of the observer study without and with CAD when lowering of the confidence score was allowed.

Reader	FOM		Sensitivity (%)		FP markings per CXR	
	without CAD	with CAD	without CAD	with CAD	without CAD	with CAD
Radiologist						
A	0.73	0.75	63	57	0.25	0.11
B	0.71	0.70	63	65	0.22	0.28
<i>Average</i>	<i>0.72</i>	<i>0.72</i>	<i>63</i>	<i>61</i>	<i>0.23</i>	<i>0.19</i>
Resident						
C	0.47	0.53	39	41	0.59	0.41
D	0.60	0.63	69	65	0.75	0.58
E	0.62	0.65	37	41	0.13	0.12
F	0.62	0.62	51	55	0.32	0.34
<i>Average</i>	<i>0.58</i>	<i>0.61</i>	<i>49</i>	<i>51</i>	<i>0.45</i>	<i>0.36</i>

Table 3

Potential of CAD to improve observer performance.

	Radiologist			Resident				
	A	B	Average	C	D	E	F	Average
No. of TP CAD annotations initially not detected by observers	5	7	6	13	5	16	13	11.8
No. of rejected TP CAD annotations	5	6	5.5	10	5	14	7	9

CAD correctly annotated 30 malignancies. Most TP CAD annotations were rejected by the observers.

Table 4

Effect of CAD on the number of TP, FP, true negative and false negative markings of the observers.

Effect of CAD*	Radiologist			Resident				
	A	B	Average	C	D	E	F	Average
Positive								
FN to TP markings	0	1	0.5	3	0	2	6	2.8
FP to TN markings	20	4	12	33	33	3	13	20.5
Negative								
TN to FP markings	3	9	6	13	13	1	15	10.5
TP to FN markings	3	0	1.5	2	2	0	4	2

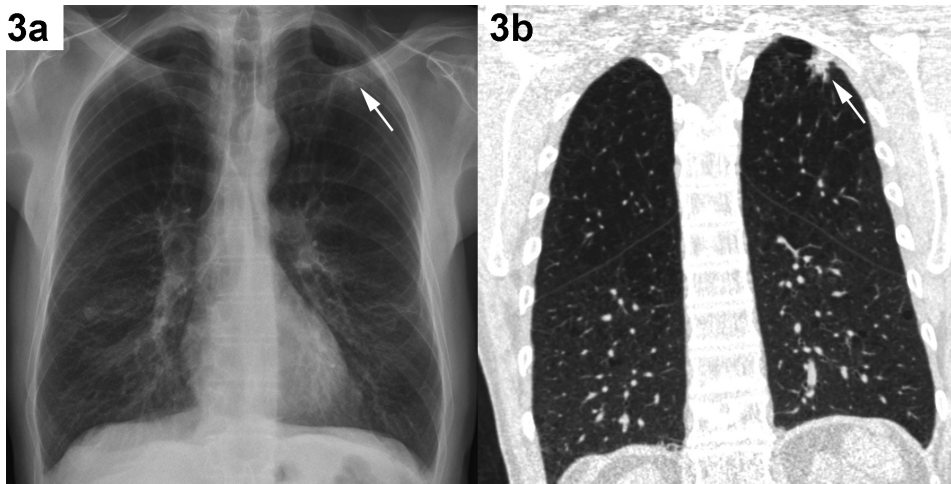
*FN = false negative, TN = true negative

Discussion

In this study we assessed how recently released, commercially available CAD software affected reader performance in detecting early lung cancer on chest radiographs. Stand-alone sensitivity of CAD was virtually identical to that of experienced radiologists: 61% in a dataset where 16% of the nodules were not detected by any of the observers. However, the number of FP annotations per chest radiograph was, on average, 10 times higher with CAD than with the two radiologists. The number of CAD-annotated malignancies that were initially not detected by observers varied between five and 16 per observer, out of a total of 49 malignancies, which indicates a vast potential for CAD to improve reader performance. Still, no significant improvement in observer performance could be demonstrated with the use of CAD as second reader in the detection of nodules on chest radiographs. An interesting observation is that in the current study, CAD did not improve observer performance. The reason is not that the observers disregarded the CAD annotations; on the contrary, in total, 66 CAD annotations were accepted and 117 initial observer markings were removed because CAD did not annotate the corresponding region. The 66 accepted annotations, pooled over all observers, were 12 TP CAD annotations of lesions initially missed and 54 FP CAD annotations. Among the 117 removed markings were 11 TP lesions. This shows that the observers had difficulties differentiating TP from FP CAD annotations. This principle has previously been described in a chest radiograph nodule detection study in which eye-tracking was used. In that study, only a minority of the lesions were missed due to inefficient search. The dominant cause of unreported nodules proved to be incorrect decision making⁽²¹⁾. This has also been described in a study that used CAD for detection, as well as classification of suspicious regions⁽²²⁾. The detection function of that CAD system annotated suspicious regions, but only slightly increased the number of lung cancers detected by the observers. Similar to our study, cancers initially missed by the observers but correctly annotated by CAD were frequently rejected by the observers. The authors report that the missed cases were mainly subtle lesions. The reported improvement in radiologists' performance was mainly due to the classification function that computed the likelihood of malignancy for regions indicated by the observer. Using this information, the observer could then change his or her initial decision. All malignancies included in our study were depicted with CT during lung cancer screening. Malignancies detected during CT screening are usually in an early stage and consequently more difficult to recognize on chest radiographs⁽²³⁾, a fact that reflected in this study by the relatively low sensitivity of the observers. In a previous study⁽²⁴⁾

analyzing CAD, pulmonary malignancies that were inadequately visible on chest radiographs were excluded from analysis. A very high area under the ROC curve of 0.92 was reported without the use of CAD. Observer performance was reported to improve significantly with CAD, and detection became almost flawless. When we excluded subtle lesions from the analysis, we also found excellent performance for radiologists (FOM, 0.93) and significant improvement in FOM to 0.82 for residents. These results show that classification is less problematic in nonsubtle lesions and the benefit of CAD is larger in more conspicuous cases, although such obvious lesions are less likely to be missed in the first place by experienced radiologists. We showed that to improve observer performance for subtle lesions, observers need to learn to better differentiate between TP and FP CAD annotations. Observer training to recognize FP CAD annotations or a change in how CAD presents results might lead to this goal. In that respect, the lack of training of our observers might have contributed to the low positive effect of CAD. On the other hand, it is, to date, unknown how much training would be necessary and how strong such learning effects would be.

Figure 3



(a) Chest radiograph and (b) CT scan of correctly CAD-annotated adenocarcinoma (arrow). Both radiologists detected the tumor without CAD, but none of the four residents marked the region, even after seeing CAD results

CAD systems of the future may not only provide annotations, but also assign likelihood that an annotation is a true lesion. Alternatively, CAD may just display the likelihood of a suspicious area of demand. This approach has been shown to improve detection of cancer on mammograms⁽²⁵⁾. However, it requires a low threshold for radiologists to query potentially suspicious regions marked by CAD, and it also will not prevent missing detection errors by the radiologist. No consensus exists whether it is allowed for observers to reduce their suspicion when using CAD. Some state that radiologists should never reduce their initial level of suspicion for markings, irrespective of the CAD results⁽²⁰⁾. However, the interaction between the radiologists' confidence and the CAD markings is unavoidable in clinical practice, and the final diagnosis will be the results of the interaction between the individual reader and CAD. Our separate analysis under the condition that observer ratings could not be reduced after seeing CAD results demonstrated that this approach actually resulted in a 2%-6% higher sensitivity, however, at the cost of such an increase of FP markings that the FOM decreased significantly with this approach. The relative high number of FP markings in our study can be explained by the low threshold for calling a marking positive. Even the lowest rating was already counted as a positive. This threshold is also used in other studies that evaluated CAD^(22,26) and ensures that all changes made owing to CAD are accounted for in the evaluation, because observers had a low confidence in most markings that were placed after seeing the CAD results. Our study was limited by the fact that the observers were explicitly asked to search for lung nodules. In clinical practice, chest radiographs are often requested for other reasons than lung cancer screening. In such cases, the search for lung nodules by radiologists is potentially less thorough, and nodules may be overlooked more easily. CAD may be more beneficial in such a situation. In addition, there was a bias toward calling suspicious abnormalities a lesion, because observers knew that the presence of cases was higher than in a normal screening situation. In practice, lower detection rates for the observer may therefore be likely. How far that will affect their attitude toward positive CAD markings is unknown. Finally, although we did not find significant improvement in FOM with the use of CAD, we did find a strong trend for residents. All residents showed an equal or higher FOM with the use of CAD, with an improvement that reached a p level of 0.08. It is likely that this improvement would have yielded statistical significance when more cases or observers had been included. More research is needed to confirm this trend for the use of this CAD system by residents. We conclude that the detection rate of pulmonary malignancies

on chest radiographs is comparable for current CAD software and experienced radiologists. However, the positive predictive value of CAD was limited by the high FP rate. Because observers were unable to sufficiently differentiate TP from FP annotations, CAD did not significantly improve detection of more conspicuous lesions by less experienced observers. For subtle lesions, however, additional measures are needed to be able to take advantage of lesions that were missed by observers but were annotated by CAD. Special training of readers might help them differentiate TP from FP CAD annotations. As an alternative, CAD findings might be presented so that they also provide an estimation of the probability of malignancy.

References

- 1 Spring DB, Tennenhouse DJ. Radiology malpractice lawsuits: California jury verdicts. *Radiology* 1986; 159:811-814
- 2 Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer: review article. *Cancer* 2000; 89:2453-2456
- 3 Li F, Arimura H, Suzuki K, et al. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. *Radiology* 2005; 237:684-690
- 4 Potchen EJ, Cooper TG, Sierra AE, et al. Measuring performance in chest radiography. *Radiology* 2000; 217:456-459
- 5 Quekel LG, Kessels, Goei R, van Engelshoven JM. Detection of lung cancer on chest radiograph: a study on observer performance. *Eur J Radiol* 2001; 39:111-116
- 6 Toyoda Y, Nakayama T, Kusunoki Y, Iso H, Suzuki T. Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography. *Br J Cancer* 2008; 98:1602-1607
- 7 Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with potentially resectable lesion evident in retrospect. *Radiology* 1992; 182:115-122
- 8 Quekel LG, Kessels AG, Goei R, Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest* 1999; 115:720-724
- 9 Monnier-Cholley L, Arrivé L, Porcel A, et al. Characteristics of missed lung cancer on chest radiographs: a French experience. *Eur Radiol* 2001; 11:597-605
- 10 Shah PK, Austin JH, White CS, et al. Missed non-small cell lung cancers: Radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology* 2003; 226:235-241
- 11 International Early Lung Cancer Action Program Investigators, Henscke CI, Yankelevitz DF, et al. Survival of patients with stage I lung cancer detected on CT screening. *N Eng J Med* 2006; 335:1763-1771
- 12 van Iersel CA, de Koning HJ, Draisma G, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomized lung cancer multi-slice CT screening (NELSON). *Int J Cancer* 2007; 120:868-874
- 13 Stahl M, Aach T, Dippel S. Digital radiography enhancement by nonlinear multiscale processing. *Med Phys* 2000; 27:56-65
- 14 Xu DM, Gietema H, de Koning H, et al. Nodule management protocol of the NELSON randomized lung cancer screening trial. *Lung Cancer* 2006; 54:177-184
- 15 Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996; 23:1709-1725
- 16 Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004; 31:2313-2330
- 17 Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. *Acad Radiol* 2006; 13:1187-1193
- 18 Vikgren J, Zachrisson S, Svalkvist A, et al. Comparison of chest thomosynthesis and chest radiography for the detection of pulmonary nodules: human observer study of clinical cases. *Radiology* 2008; 249:1034-1041
- 19 Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur D. A comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. *Med Phys* 2005; 1031-1034

- 20 Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative analysis—the image role of medical physics and AAPM. *Med Phys* 2008; 35:5799-5820
- 21 Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from posteroanterior chest radiograph. *Br J Radiol* 2004; 77:231-235
- 22 Shiraishi J, Abe H, Li F, Engelsmann R, MacMahon H, Doi K. Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs ROC analysis of radiologists' performance. *Acad Radiol* 2006; 13:995-1003
- 23 Henscke CI; for the International Early Lung Cancer Action Program Investigators. Survival of patients with clinical stage I lung cancer diagnosed by computed tomography screening for lung cancer. *Clin Cancer Res* 2007; 13:4949-4950
- 24 Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. *AJR Am J Roentgenol* 2004; 182:505-510
- 25 Karssemeijer N, Otten JD, Verbeek AL, et al. Computer-aided detection versus independent double reading of masses on mammograms. *Radiology* 2003; 227:192-200
- 26 Sakai S, Soeda H, Takahashi N, et al. Computer-aided detection on digital chest radiography: validation test on consecutive T1 cases of resectable lung cancer. *J Digit Imaging* 2006; 19:376-382

Computer-aided Detection of Small Pulmonary Nodules
in Chest Radiographs: an Observer Study



Diederick W De Boo
Martin Uffmann
Michael Weber
Shandra Bipat
Eelco F Boorsma
Maeke J Scheerder
Nicole J Freling
Cornelia M Schaefer-Prokop

Abstract

Objective

To evaluate the impact of computer-aided detection (CAD, IQQA-Chest; EDDA Technology, Princeton Junction, NJ) used as second reader on the detection of small pulmonary nodules in chest radiography (CXR).

Material and Methods

A total of 113 patients (mean age 62 years) with CT and CXR within 6 weeks were selected. Fifty-nine patients showed 101 pulmonary nodules (diameter 5-15mm); the remaining 54 patients served as negative controls. Six readers of varying experience individually evaluated the CXR without and with CAD as second reader in two separate reading sessions. The sensitivity per lesion, figure of merit (FOM), and mean false positive per image (mFP) were calculated. Institutional review board approval was waived.

Results

With CAD, the sensitivity increased for inexperienced readers (39% vs. 45%, $p < 0.05$) and remained unchanged for experienced readers (50% vs. 51%). The mFP nonsignificantly increased for both inexperienced and experienced readers (0.27 vs. 0.34 and 0.16 vs. 0.21). The mean FOM did not significantly differ for readings without and with CAD irrespective of reader experience (0.71 vs. 0.71 and 0.84 vs. 0.87). All readers together dismissed 33% of true positive CAD candidates. False positive candidates by CAD provoked 40% of all false positive marks made by the readers.

Conclusion

CAD improves the sensitivity of inexperienced readers for the detection of small nodules at the expense of loss of specificity. Overall performance by means of FOM was therefore not affected. To use CAD more beneficial, readers need to improve their ability to differentiate true from false positive CAD candidates.

Introduction

In clinical practice, small primary lung carcinomas and pulmonary metastases may be missed on two-view chest radiographs (CXR) though they were frequently visible in retrospect. Depending on the study design, miss rates between 20% and 90% have been reported for primary lung carcinomas⁽¹⁻⁷⁾. Factors that contribute to detection errors include image quality, size and type of nodules, superposition of anatomical structures, the presence of accompanying abnormalities, and the radiologists' variable experience and perception capacity. The goal of computer-aided detection (CAD) software is to reduce the effects of the latter, namely to lower the number of perception errors. Variable sensitivities of CAD in chest radiography (34%-78%) have been reported⁽⁸⁻¹³⁾. Several publications only refer to the CAD stand-alone performance^(8,10,11,13). In clinical practice, however, CAD is designed to be used as second reader, meaning that the ultimate impact of CAD on the diagnostic outcome will be determined by both, the CAD performance and the readers' diagnostic judgment. In this study we investigated a US Food and Drug Administration (FDA)-approved CAD system. We evaluated how CAD, when used as second reader, affects the detection performance of readers with vastly varying experience. Study group and nodules were selected to challenge perception capabilities: most nodules were of low conspicuity and the majority of the elderly study patients showed increased parenchymal markings on the CXR from smoking history and aging and described as "dirty lung" in the literature⁽¹⁴⁾.

Material and Methods

Study group

From our institution's data archive, we retrospectively selected 113 patients who had undergone both, a two-view (CXR) and a chest CT within six weeks. All images had been acquired for clinical purposes only. Institutional review board approval was therefore waived by our institution's ethic committee (registration number 08170465). Patients' mean age was 62 years (16-89 years), 65 were male and 48 were female. According to patients' records, 61 patients (54%) had a positive history of smoking, 27 patients (24%) were nonsmokers and for the remaining 25 patients (22%) no data was available. Both smoking and increase of age lead to a variable increase of parenchymal markings on the CXRs also described as "dirty lung" ⁽¹⁴⁾. Its presence was subjectively scored on the CXRs in consensus by a board certified chest radiologist (>15 years of experience) and the researcher (third-year resident in radiology) using a score from 0 (none) to 1 (minimal), 2 (moderate) and 3 (severe). In the following, we will refer to this score as "anatomic noise" (AN).

Intrapulmonary nodules

We followed the glossary of terms for thoracic imaging by the Fleischner society that defines a nodule as a well or poorly defined nodular opacity with a diameter between 3 and 30 mm. Based on this definition, 113 nodules were present in 59 patients. The tested CAD software is optimized for detecting nodules between 5 and 15 mm, therefore the 12 nodules exceeding 15 mm in diameter were excluded from further data analysis. Thus the study group comprised 101 nodules present in 59 patients with a mean and median diameter of 10 mm and a range of 5-15 mm. The 54 patients that had no nodules, as proven by CT, served as normal controls. Seventy-four of the 101 nodules had smooth margins on CT and well-defined contours on the CXR. Twenty-seven nodules had spiculated margins on CT and poorly defined contours on the CXR. None of these nodules were calcified. Histological proof was available in 24% (18/74) of the well-defined nodules (M) and referred to primary non-small-cell lung cancer (8), metastases of renal cell carcinoma (2) or osteosarcoma (1); seven biopsies revealed no malignancy. The remaining 56 well-defined nodules were interpreted as metastases because of a known history of extrathoracic malignancy and the fact that follow-up studies had revealed nodule growth. Histological proof was available in 78% (21/27) of the poorly defined nodules (T) and referred to primary non-small-cell lung cancer (11),

bronchoalveolar cell carcinoma (1), neuro-endocrine tumors (2) and a nonmalignant histology (7). The remaining six poorly defined nodular opacities were considered nonmalignant because they decreased in size over time under antibiotic and / or anti-inflammatory treatment (Table 1).

Nodule conspicuity

Nodule conspicuity on the CXR was subjectively graded by a board-certified chest radiologist (>15 years of experience) and the researcher (third-year resident in radiology) after the reading had been completed and ranged from 1) high to 2) moderate, 3) low and 4) very low.

Chest radiography

CXRs were obtained in digital technique using a dedicated chest stand (Thoravision Philips Medical Systems, Hamburg, Germany). Images were processed using non-linear multifrequency processing (Unique, Philips Medical Systems). Processing parameters were chosen according to the recommendations of the manufacturer and represented the standard processing used in clinical routine at our institution.

CAD

The CAD software (IQQA-Chest; EDDA Technology, Princeton Junction, NJ) is designed to detect nodules in a size range from 5 to 15 mm in diameter. Only the posteroanterior (PA) radiographs are analyzed. The software is running in the background and only on demand between zero and five candidates are marked by semitransparent circles that are centered on a focal density. By visual side-by-side comparison between the marked and unmarked radiograph or by toggling the semitransparent overlay the reader can decide to accept or dismiss the CAD candidate.

Image evaluation

Three radiology residents (first-, second- and third-year training) and three board certified radiologists (two general radiologist, one chest radiologist, all with more than 10 years of experience) independently interpreted all 113 CXR using high resolution LCD monitors (Barco, MDCG 2121-CB, 1.2K x 1.9K matrix) that are subject to regular quality control. PA and lateral radiographs of the 113 patients were evaluated twice in two separate reading sessions, once without CAD and once with the availability of CAD. Half of the cases were seen first without CAD; the other half was first interpreted with the use of CAD. When using CAD, readers were specifically instructed to first analyze the images unassisted before taking the result of the CAD analysis into account and having the candidate circles available.

Readers were asked to document localization of nodules and diagnostic confidence after consideration of CAD. For both conditions, readers were allowed to use processing tools such as windowing or magnification according to their preferences. There was an at least 6 weeks time interval between the two reading sessions and images were evaluated in different random orders. The presence or absence of a nodule was scored using a five point scale of confidence ranging from 5 = definite pulmonary nodule, 4 = probable pulmonary nodule, 3 = equivocal, 2 probably no nodule, and 1 = definitely no nodule. For the confidence ratings 3, 4 and 5, readers were asked to indicate the anatomic location of the suspected nodule on a separate data sheet. This was done separately for each patient; per patient more than one nodule could be marked. The readers were informed that images could contain none, a solitary or multiple nodules but did not know the percentage of each subgroup. They were instructed to ignore calcified nodules and nodules smaller than 5 mm. All readers evaluated 10 training cases with CAD before conducting the study; none of the readers had experience with CAD in daily routine.

Statistical analysis

The patients with and patients without intrapulmonary nodules were compared with regard to gender and smoking history by chi-square test, with regard to age by student t-tests as the data were normally distributed and with regard to AN by chi-square test for trend. For smoking history, the missing data (22%) were excluded. All reader markings, without and with CAD, were determined to be true or false positive by comparing the markings on the separate data sheets with the original CXR and the corresponding CT. This was done in consensus by the researcher and an experienced chest radiologist after all readings had been completed. The data were analyzed using the jackknife free-response receiver operating characteristic (JAFROC) method especially developed to analyze observer free-response tasks⁽¹⁵⁻¹⁷⁾. JAFROC software 2.3a was used to calculate a figure of merit (FOM). The FOM is defined as the probability on a scale from 0 to 1.0 that true positive marks for nodules are rated with higher confidence than false positive (non-lesions) marks on control CXRs⁽¹⁶⁾. A FOM of 0.5 means that the confidence of marks for nodules on abnormal CXR are equal to marks for nonlesions (false positive marks) on negative control CXR. A FOM of 1.0 describes the ideal situation where all nodules are correctly marked with no false positives in the control images. Sensitivity was calculated on a per-nodule basis. For these calculations, reader ratings 4 and 5 were considered true positive if made for correctly

localized nodules. CAD candidates were considered true positive if the candidate was centralized above a CT proven nodule. Specificity was described as mean number of false positive marks per image (mFP). For these calculations, reader ratings 4 and 5 were considered false positive, if made in locations, where CT did not show a nodule. CAD candidates only partially covering a nodule or indicating a nodule not proven by CT were considered false positive. Sensitivities and mFPs were calculated for CAD as stand-alone and for each reader separately without and with use of CAD. Significance of differences was tested using McNemar's test. We also assessed the number of rating differences made in the two reading sessions without and with the availability of CAD. A rating difference was considered beneficial if a nodule was missed during reading without CAD but was correctly detected during the reading with CAD. The scenario, that the reader made a false positive mark without CAD but made no false positive mark with the availability of CAD, was also considered a beneficial rating difference. It was considered a detrimental rating difference if the reader detected the nodule during the reading without CAD but missed it in the reading session with the availability of CAD or if the reader made a new false positive mark during the reading with CAD. Agreement between the readers and the stand-alone performance of the CAD-system was calculated pairwise using non-weighted Kappa statistics. A Fleiss Kappa statistics was used to evaluate the interobserver agreement during readings without and readings with CAD as second reader. All analyses were performed in SPSS 15.0.1. Significance was assumed at $p < 0.05$.

Table 1
Characteristics of the 101 nodules.

Characteristics		
Shape	M	27
	T	74
Conspicuity	high (1)	21
	moderate (2)	39
	low (3)	20
	very low (4)	21
Size	<10mm	42
	≥10mm	59

M: round nodules, T: nodules with irregular margins

Results

Study groups and nodule characteristics

An increased AN was scored present in 69% of all patients and varied from mild in 42% to moderate in 18% and severe in 9% of patients. Of the 59 patients with nodules 38 had a solitary nodule, 10 had two nodules and 11 had more than two nodules. The nodules were located in the upper lobes in 58%, in the middle lobe in 10% and in the lower lobes in 32%. Nodule conspicuity was rated very low in 21%, low in 39%, moderate in 20% and high in 21%. Nodule conspicuity was negatively correlated with AN ($p = 0.04$) and nodule size ($p = 0.024$). There were no differences between patients with and without nodules with regard to age, gender, AN or history of smoking ($p = 0.85, 0.52, 0.49, 0.20$, respectively).

Stand-alone performance of CAD

The CAD system demonstrated a stand-alone sensitivity of 47%. The sensitivity was 70% for nodules with moderate and high conspicuity and 30% for nodules with low and very low conspicuity. The CAD stand-alone sensitivity was not associated with nodule shape, nodule size, anatomic location or AN. A total of 194 false positive candidates were produced with a mFP of 1.7. There was no association between the number of produced false positives and the AN.

Table 2

Individual reader outcome without and with use of CAD.

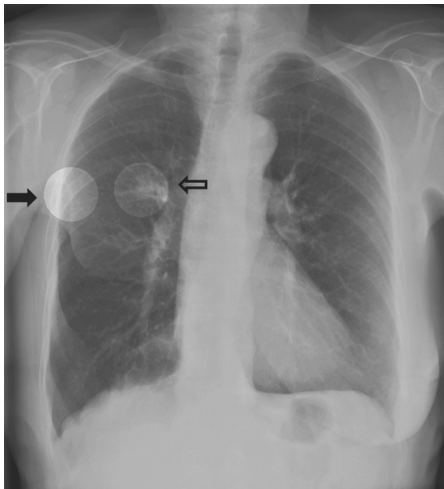
	FOM		Sensitivity (in%)		mFP	
	Baseline	CAD	Baseline	CAD	Baseline	CAD
Inexperienced readers (n = 3)	0.71	0.71	39	45	0.27	0.34
1	0.52	0.53	39	46	0.54	0.67
2	0.77	0.79	35	39	0.08	0.13
3	0.84	0.80	42	50	0.18	0.21
Experienced readers (n = 3)	0.84	0.87	50	51	0.16	0.21
4	0.85	0.91	45	44	0.15	0.18
5	0.80	0.83	53	54	0.20	0.24
6	0.87	0.88	53	57	0.12	0.20

Baseline: reading without CAD, CAD: reading with computer-aided detection as second reader, FOM: figure of merit, mFP: mean false positives per image

Reader performance without CAD

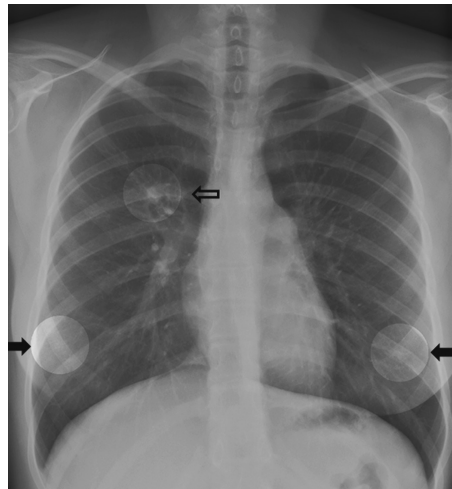
The mean sensitivity for the inexperienced readers was 39% with a mFP of 0.27. The mean sensitivity for the experienced readers was 50% with a mFP of 0.16. The average FOM for the inexperienced readers was 0.71 compared with 0.84 for the experienced readers (Table 2). For both reader groups, there was no significant association between the number of false positives and AN. Looking at subgroups of nodules as function of conspicuity, size and anatomic background noise, the readers' sensitivity was generally lower for nodules of low conspicuity (22% vs. 76%) and smaller size (25% vs. 58%) and for nodules located in lungs with increased anatomic noise (31% vs. 52%; Table 3).

Figure 1



The three inexperienced readers missed the right perihilar nodule without computer-aided detection (CAD) but accepted the true positive CAD candidate (open arrow). All readers dismissed the false positive CAD candidates located more laterally (black arrow).

Figure 2



Four readers dismissed the true positive computer-aided detection (CAD) candidate for the right perihilar nodule (open arrow). The CAD candidates in the right and left lower zone represented false positives.

Reader performance with CAD as second reader

For the group of the inexperienced readers, the mean sensitivity increased from 39% to 45% ($p = 0.008$; example in Figure 1) with a nonsignificant increase of the mFP from 0.27 to 0.34. For the experienced readers the mean sensitivity remained unchanged (50% vs. 51%) with a mFP that also nonsignificantly increased from 0.16 to 0.21. When separately analyzed, sensitivity increased for five of six readers but differences were not significant. The mean FOM was higher for the experienced readers but for both reader groups, the mean FOM did not significantly change with the availability of CAD (0.71 vs. 0.71 and 0.84 vs. 0.87, respectively).

Reader-CAD interaction

Altogether, 33% (94/282) of the true positive CAD candidates were dismissed by the readers, meaning readers did not realize that CAD correctly indicated existing nodules which had been primarily missed by the readers. The readers subsequently disregarded those CAD candidates though they were true positive (Figure 2). Sixty-eight percent of these dismissed true positive CAD candidates were for nodules of low to very low conspicuity. Inexperienced readers declined 51, whereas experienced readers declined 43 true positive CAD candidates (Table 4). When comparing the readings without and with CAD we found a total of 286 rating differences that affected the correct diagnosis. 54% (154/286) of them were detrimental and 46% were beneficial. 40% (61/154) of the detrimental rating differences represented false positive ratings likely to have been provoked by a false positive CAD candidate in identical location (Figure 3). Maximum benefit from CAD would be established if the readers accepted all true positive CAD candidates and declined only false positive CAD candidates. In this theoretical scenario all readers would show a substantial increase in sensitivity to a mean overall sensitivity of 70% with a slight inferiority of the inexperienced readers (66%) compared to the experienced readers (74%).

Interobserver agreement and agreement with CAD

The agreement between the stand-alone performance of CAD and the six readers was poor to fair with Kappa values ranging from 0.15 to 0.32. The interobserver agreement for the readers during reading without CAD was 0.36 (fair) and increased slightly to 0.39 (fair) when CAD was used as second reader.

Table 3
Sensitivity (in %) of CAD stand-alone and of readers without and with CAD for subgroups of nodules, dependant on shape, conspicuity, size and presence of anatomic noise.

	Shape		Conspicuity			Size			AN	
	T	M	1+2	3+4	<10mm	≥10mm	0+1	2+3		
CAD stand-alone	48	46	70	30	40	51	51	39		
All readers (n=6)										
baseline	38	46	76	22	25	58	52	31		
CAD	42	50	84	23	29	61	56	31		
Inexperienced readers (n=3)										
baseline	30	41	69	17	19	52	45	27		
CAD	37	47	82	19	25	58	52	30		
Experienced readers (n=3)										
baseline	46	51	83	27	31	63	58	34		
CAD	47	52	85	28	33	64	61	33		

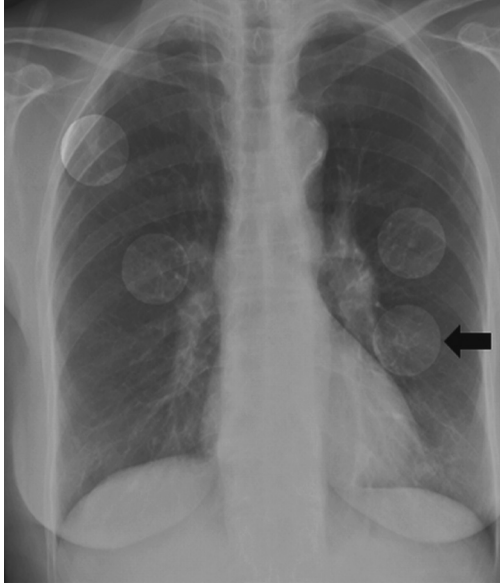
AN: anatomic noise, Baseline: reading without CAD, CAD: reading with computer-aided detection as second reader,

M: round nodules, T: nodules with irregular margins

Table 4
Dismissed true positive CAD candidates (in absolute numbers and in percentage of all CAD true positives between brackets) pooled over all readers and over subgroups of experienced and inexperienced readers.

	Shape		Conspicuity			Size			AN	
	T	M	1+2	3+4	<10mm	≥10mm	0+1	1+2		
All readers (n=6)	34 (44%)	60 (29%)	21 (12%)	73 (68%)	52 (51%)	42 (23%)	50 (25%)	44 (52%)		
Inexperienced readers (n=3)	19 (49%)	32 (31%)	11 (13%)	40 (74%)	28 (55%)	23 (26%)	29 (29%)	22 (52%)		
Experienced readers (n=3)	15 (38%)	28 (27%)	10 (11%)	33 (61%)	24 (47%)	19 (21%)	21 (21%)	22 (52%)		

AN: anatomic noise, M: round nodules, T: nodules with irregular margins

Figure 3

Four false positives were generated by the computer-aided detection (CAD) software. Three readers accepted the left lower false positive (arrow). The other false positives were dismissed by all readers.

Discussion

Various CAD systems with and without FDA approval have been developed for the detection of nodules in CXRs. Reported stand-alone sensitivities of these systems vary from 34% to 78% dependant on lesion selection and study group⁽⁸⁻¹³⁾. In our study, the stand-alone sensitivity of the tested CAD software was 47% and this was slightly better than the mean sensitivity of 44% achieved by the readers. Thus both readers and CAD showed a relatively low baseline performance, which is likely to be the result of the low to very low conspicuity of the majority of the studied nodules. For a successful application of CAD, it seems warranted that CAD detects nodules, which the radiologist tends to miss. As indicated by the Kappa analysis, we found a relatively lower agreement between CAD and the readers as between the readers, supporting the fact that CAD has the potential to find nodules, the readers have not seen. Also Bley et al., who evaluated the same CAD software, reported an only moderate agreement between CAD and the readers underlining the capacity of this CAD software to detect different nodules as the readers⁽¹⁰⁾. Application of CAD ideally leads to an increase of sensitivity without loss of specificity. We found a significant increase in sensitivity for the inexperienced readers with the use of CAD. Experienced readers, however, did not take any advantage of CAD. Yet, the increase of sensitivity of the inexperienced readers went along with a loss of specificity that, though not statistically significant, impeded

an increase of the FOM. With regard to the localization of false positive marks by the readers, an interaction with CAD candidates took place: 40% of the detrimental rating differences referred to false positive marks that corresponded to false positive CAD candidates in these locations and thus were likely to have been provoked by CAD. On the other hand our study could not prove a significant increase of the mFP. These findings suggest that false positives seen by the readers without CAD did not attract the same grade of suspicion when CAD indicated different candidate areas. In fact, in a substantial amount of readings (61/154) CAD candidates entrapped the readers to call an area suspicious. The number of false positive marks possibly provoked by CAD were equally distributed between experienced and inexperienced readers and showed no significant association with the amount of anatomic noise. Though, our results could not prove that CAD improves detection performance for small intrapulmonary nodules, there are some findings indicating that CAD may have potential to increase detection rates if applied differently. In that context, it is noteworthy that all six readers together declined 33% of the true positive CAD candidates. The majority of the dismissed true positive candidates referred to the low conspicuous nodules and to nodules smaller than 10 mm in diameter. These nodules represent indeed the type of pathology that are most difficult to detect by the readers and might take the highest advantage of a CAD system. If the readers would have accepted all true positive CAD candidates and declined only false positive CAD candidates, all readers would have shown a substantial increase in performance from a baseline sensitivity of 47% to a mean overall sensitivity of 70%, with a slight inferiority of the inexperienced readers (66%) compared to the experienced readers (74%). These findings indicate the potential of CAD if used optimally but also indicate the readers' severe difficulty to distinguish true positive from false positive CAD candidates. For further improvement of reader performance with CAD, two aspects seem important. Firstly, the number of false positive CAD candidates should be decreased. Recent publications described a significantly reduced mFP when CAD was applied on energy subtracted soft-tissue images^(18,19). Whether this indeed will increase reader performance yet remains to be proven by an observer study. Second, additional tools appear to be needed to help the observer to accept true positive candidates even when they indicate low conspicuous nodules. Whether this can be achieved by means of additional displays such as grey-scale reversal or rib subtraction or whether it requires additional software tools, such as indication of the grade of suspicion based on the CAD analysis, is subject of ongoing research. The latter was found quite promising for reading mammography: reader detection performance could be further

increased when the CAD software indicated a likelihood of suspicion for an area of interest selected by the observer instead of showing the reader a certain amount of equally weighted ROIs⁽²⁰⁾. Previous literature reports controversial results about the effect of CAD on the detection of nodules. While Song et al. only state the impact on sensitivity without considering the specificity⁽²¹⁾, other studies more precisely report changes of Az that way statistically taking into consideration both sensitivity and specificity. Most interestingly, all studies that report a significantly increased performance (increase of Az) also stated a higher CAD stand-alone performance with sensitivities between 52% and 100% as compared to the stand-alone sensitivity of 47% found in our study^(12,22,23,24,25). The majority of the CAD systems evaluated were not FDA approved at time of study conductance. There are four studies with a stand-alone sensitivity below 50%^(8,10,22,26): two of them assessed reader performance and found no increase of reader performance measured as Az or FOM, conform our results. We conclude from this that in addition to reader experience the conspicuity of nodules represents a determining factor with respect to the impact of CAD on reader performance. As seen in our results (Table 3), more conspicuous nodules (here classified as 1 and 2), if missed by the reader because of perception error, were more easily accepted by the observer when indicated by CAD. Low conspicuous nodules (classified as 3 and 4), however, that require high perception and interpretation skills, were less easily accepted by the observer even when correctly indicated by CAD. Our study suffers from a number of limitations:

- a) We evaluated a selected group of patients with a higher prevalence of nodules than normally seen in clinical routine. Readers were specifically asked to look for small nodular densities which probably lowered their overall specificity.
- b) Images were evaluated without and with CAD as second reader in two reading sessions introducing the additional factor of intrareader variability. We chose to do so to keep reading conditions possibly realistic and not to interrupt the readers' visual analysis by requiring nodule documentation separately without and with CAD. However, we cannot exclude that intrareader variability may have partially ameliorated the effect of CAD on reader performance.
- c) Though the readers were appropriately introduced to the use of CAD by a number of training cases, which were not included in the study, none of them had experience with CAD for a longer period of time or within routine application. For mammography a substantial positive impact of learning effects on the performance with CAD is assumed to be two years⁽²⁷⁾; whether this also holds for chest radiography remains to be evaluated.

d) We evaluated a specific type of CAD algorithm. Our results only pertain to that particular CAD application and to that particular study group and are not directly transferrable to another CAD software and to other study groups of lesions.

In summary we conclude that the sensitivity of the inexperienced readers significantly increased with the use of CAD as second reader. Overall performance (FOM) for both inexperienced and experienced readers, however, was not affected, meaning that the increase in sensitivity came with a decrease in specificity. The impact of CAD on the detection of small and low conspicuous nodules seems to be largely impeded by the readers' inability to differentiate true from false positive CAD candidates. There is potential for further improvement of reader performance with CAD, given the high rate of dismissed true positive CAD candidates and the number of accepted CAD false positives. Our findings underline the importance to further decrease the number of false positive CAD candidates in the future and the need for further research to find out whether a longer learning process or additional visual or analytic tools that come along with the CAD output can further improve the readers' ability to distinguish true from false positive CAD candidates.

References

- 1 Muhm JR, Miller WE, Fontana RS, Sanderson DR, Uhlenhoop MA. Lung cancer detected during a screening program using 4-month chest radiographs. *Radiology* 1983; 148:609-615
- 2 Heelan RT, Flehinger BJ, Melamed MR, et al. Non-small-cell lung cancer: results of the New York screening program. *Radiology* 1984; 151:289-293
- 3 Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with potentially resectable lesion evident in retrospect. *Radiology* 1992; 182:115-122
- 4 Quekel LG, Kessels AG, Goei R, Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest* 1999; 115:720-724
- 5 Monnier-Cholley L, Arrivé L, Porcel A, et al. Characteristics of missed lung cancer on chest radiographs: a French experience. *Eur Radiol* 2001; 11:597-605
- 6 Shah PK, Austin JH, White CS, et al. Missed non-small cell lung cancers: Radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology* 2003; 226:235-241
- 7 Wu HM, Gotway MB, Lee TJ, et al. Features of non-small cell lung carcinomas overlooked at digital chest radiography. *Clinical Radiology* 2008; 63:518-528
- 8 Li F, Engelmann R, Metz CE, Doi K, MacMahon H. Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. *Radiology* 2008; 246:273-280
- 9 van Beek EJ, Mullan B, Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images: a prospective study. *Acad Radiol* 2008; 15:571-575
- 10 Bley TA, Baumann T, Saueressig U, et al. Comparison of radiologist and CAD performance in the detection of CT-confirmed subtle pulmonary nodules on digital chest radiographs. *Invest Radiol* 2008; 43:343-348
- 11 White CS, Flukinger T, Jeudy J, Chen JJ. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology* 2009; 252:273-281
- 12 Sakai S, Soeda H, Takahashi N, et al. Computer-aided detection on digital chest radiography: validation test on consecutive T1 cases of resectable lung cancer. *J Digit Imaging* 2006; 19:376-382
- 13 Hardie RC, Rogers SK, Wilson T, Rogers A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Med Image Anal* 2008; 12:240-258
- 14 Gückel C, Hansell DM. Imaging the 'dirty lung' - has high resolution computed tomography cleared the smoke? *Clinical Radiology* 1998; 53:717-722
- 15 Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004; 31:2313-2330
- 16 Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. *Acad Radiol* 2006; 13:1187-1193
- 17 Vikgren J, Zachrisson S, Svalkvist A, et al. Comparison of chest thomosynthesis and chest radiography for the detection of pulmonary nodules: human observer study of clinical cases. *Radiology* 2008; 249:1034-1041
- 18 Szucs-Farkas Z, Patak MA, Yuksel-Hatz S, Ruder T, Vock P. Improved detection of

- pulmonary nodules on energy-subtracted chest radiographs with a commercial computer-aided diagnosis software: comparison with human observers. *Eur Radiol* 2010; 20:1289-96
- 19 Balkman JD, Mehandru S, DuPont E, Novak RD, Gilkeson RC. Dual energy subtraction digital radiography improves performance of a next generation computer-aided detection program. *J Thorac Imaging* 2010; 25:41-47
 - 20 Samulski M, Hupse R, Boetes C, Mus RD, den Heeten GJ, Karssemeijer N. Using computer-aided detection in mammography as a decision support. *Eur Radiol* 2010; 20:2323-2330
 - 21 Song W, Fan L, Xie Y, Qian JZ, Jin Z. A study of inter-observer variations of pulmonary nodule marking and characterizing on DR images. *Proc SPIE* 2005; 5749:272-280
 - 22 Kasai S, Li F, Shiraishi J, Doi K. Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs. *AJR Am J Roentgenol* 2008; 191:260-265
 - 23 MacMahon H, Engelmann R, Behlen FM, et al. Computer-aided diagnosis of pulmonary nodules: results of a large scale observer test. *Radiology* 1999; 213:723-726
 - 24 Freedman MT, Lo SCB, Osicka T, et al. Computer-aided detection of lung cancer on chest radiographs: effect of machine CAD false-positive locations on radiologists' behavior. *Proc SPIE* 2002; 4684:1311-1319
 - 25 Shiraishi J, Abe H, Li F, Engelmann R, MacMahon H, Doi K. Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs: ROC analysis of radiologists' performance. *Acad Radiol* 2006; 13:995-1003
 - 26 de Hoop B, De Boer DW, Gietema HA, et al. Computer-aided detection of lung cancer on chest radiographs: effect on observer performance. *Radiology* 2010; 257:532-540
 - 27 Nishikawa R. Increased CAD use prompts look at advantages, drawbacks. In *RSNA news* February 2007. http://www.rsna.org/Publications/rsnanews/feb07/upload/RSNANews_Feb07_CAD_Usage.pdf

Observer Training for Computer-aided Detection of Pulmonary Nodules in Chest Radiography

Diederick W De Boo

François van Hoorn

Joost van Schuppen

Laura Schijf

Maeke J Scheerder

Nicole J Freling

Onno Mets

Michael Weber

Cornelia M Schaefer-Prokop

Abstract

Objective

To assess whether short-term feedback helps readers to increase their performance using computer-aided detection (CAD) for nodule detection in chest radiography.

Material and Methods

The study group of 140 CXRs (56 with a solitary CT-proven nodule and 84 negative controls) was divided into four subsets of 35 examinations each that were read in a different order by six readers. Lesion presence, location and diagnostic confidence were scored without and with CAD (IQQA-Chest, EDDA Technology) as second reader. Readers received individual feedback after each subset. Sensitivity, specificity and area under the receiver operating characteristics curve (AUC) were calculated for readings with and without CAD with respect to change over time and impact of CAD.

Results

CAD had a stand-alone performance of 59% with 1.9 false-positives per image. Mean AUC slightly increased over time with and without CAD (0.78 vs. 0.84 with and 0.76 vs. 0.82 without CAD) but differences did not reach significance. The sensitivity increased (65% vs. 70% and 66% vs. 70%) and specificity decreased over time (79% vs. 74% and 80% vs. 77%) but no significant impact of CAD was found.

Conclusion

Short-term feedback does not increase the ability of readers to differentiate true- from false-positive candidate lesions and to use CAD more effectively.

Introduction

Various computer-aided detection (CAD) systems for chest radiography with and without FDA approval have been developed. The latest reported stand-alone sensitivities of these systems for the detection of small focal opacities vary from 34% to 78% dependent on lesion selection and study group⁽¹⁻⁶⁾. Results of studies evaluating the effects of CAD on actual observers' performances are not homogeneous and range from significant improvement^(2,7,8) to lack of any impact⁽⁹⁾. Results seem to be influenced by the type of CAD algorithm used, reader experience and the conspicuity of the study lesions. The potential of CAD to increase the radiologist's sensitivity for pulmonary nodules was described earlier: two studies reported that 35% and 47% of bronchogenic tumours missed in the original reports were correctly marked by CAD^(1,4). A third study found a lower agreement between CAD and observers' detection compared with the agreement between observers indicating the ability of CAD to mark lesions the radiologists tend to miss⁽³⁾. All three studies, however, compared the CAD performance with previously made observer readings and did not assess the actual influence of CAD on the readers' decisions. The importance of this interaction between CAD and the observers for a successful implementation of CAD was shown by de Hoop et al⁽⁹⁾. They had not been able to show a positive effect of the CAD algorithm they tested because readers had difficulties in differentiating true-positive from false-positive CAD candidates. More than two-thirds of the true-positive CAD candidates in whom CAD was given for lesions that were originally missed were not accepted by the readers. The authors suspected that this inability of the readers to use the CAD more beneficially was at least partly due to lack of experience with the performance of CAD and subsequently a lack of confidence in the CAD analysis. For CT colonography, it was reported that a 1-day training period already resulted in increased readers' sensitivity, but at the expense of decreased specificity and increased reading time⁽¹⁰⁾. In mammography sensitivity, specificity and area under the receiver operating characteristics curve (AUC) significantly increased after a 4-week training period⁽¹¹⁾. The ultimate learning curve for CAD in mammography has been estimated to be around 2 years but this has never been evaluated in a structured manner⁽¹²⁾. Up to now there have been no studies evaluating the effect of observer training on the application of CAD in chest radiography. The purpose of the current study was to test whether short-term feedback to readers on their own performance when using CAD would increase the readers' confidence in the CAD analysis and thus their ability to differentiate true- from false-positive CAD candidates.

Material and Methods

Study population

For this retrospective study we selected 140 patients from our institution's data archive. Patients were included if a two-view chest radiograph (CXR) and thoracic CT were obtained within 6 weeks and revealed no or a single nodular opacity. The diameter ranged from 5 mm to 15 mm (measured on axial CT images) and none of the nodules showed calcifications. Of the 140 patients 56 had a solitary CT-proven nodule and 84 served as negative controls. Patients with more than one nodular opacity or a pathological feature other than COPD on the CXR were excluded. Ethics committee approval was obtained and because of the retrospective nature of the study patient informed consent was waived (registration number 10171150).

Pulmonary nodules

Thoracic CT served as a reference standard and revealed a solitary nodular opacity in 56 patients (40%). Conspicuity of the lesion on the CXR was subjectively graded in consensus by a board-certified chest radiologist (> 15 years of experience) and the researcher (fifth year resident) who were not involved in the readings and ranged from high (1) to moderate (2), low (3) and very low (4).

Image acquisition

All CXRs were obtained using a digital technique with a dedicated chest stand (Thoravision Philips Medical Systems, Hamburg, Germany). Images were processed using non-linear multifrequency processing (Unique, Philips Medical Systems, Hamburg, Germany). Processing parameters were implemented following the recommendations of the manufacturer and represented the same as those used as standard processing in our institution. Both the posteroanterior and the lateral views were available for evaluation.

CAD

We used a commercially available CAD system (IQQA-Chest; EDDA Technology, Princeton Junction, NJ, USA). This system is designed to detect nodules within the range from 5 mm to 15 mm in diameter on the PA radiograph. Images are automatically analysed in the background after acquisition of the images, thus results are immediately available when the radiographs are read but are only shown on demand. The CAD algorithm marks between zero and five suspicious areas with semitransparent circles (candidates).

Image evaluation

Six observers of vastly varying experience participated in this study: five radiology residents with 0 to 5 years of training (R1 zero years; R2 and R3 two years, R4 three years and R5 five years) and one board-certified radiologist (R6) with more than 15 years' experience in reading chest films. Two observers (R1 and R3) had no previous experience at all with using CAD in chest radiography, the other four observers had served as observers in previous studies evaluating CAD, two using the same CAD system as in this study and two using a different CAD system. None of the observers had experience with CAD in clinical routine work. The 140 pairs of PA and lateral radiographs were divided into four subsets of 35 each. Each subset consisted of 14 cases with a solitary nodular opacity and 21 negative control cases. Care was taken that the distribution of nodule conspicuity and the CAD stand-alone sensitivity were equal for the four subsets. Each observer individually interpreted the 35 PA and lateral chest radiographs of one subset first without and subsequently with the availability of the CAD markings within a single reading session. Observers were asked to determine the presence or absence of an intrapulmonary opacity using a five-point scale of confidence ranging from 5 – pulmonary nodule definitely present, to 3 – equivocal with respect to the presence of a nodule and 1 – definitely no nodule present. For a confidence rating > 1, observers were asked to indicate the anatomical location of the suspected lesion on a separate data sheet. Readings with and without the availability of the CAD results were separately documented. Readers were allowed to modify their confidence levels after CAD became available, also for lesions seen during unassisted reading. The observers were informed that images contained only a single nodular opacity and were instructed to ignore calcified lesions and lesions smaller than 5 mm in diameter. Magnification, window/level adjustment and grey-scale reversal were allowed during both readings with and without CAD results. All observers read the four subsets of CXRs in a different order. After completion of each of the four subsets of cases the observer received individual feedback on his/her performance by the researcher. During this feedback, the observer and researcher discussed on a case by case basis the location of the lesions if present, the CAD marks with respect to whether they were true- or false-positive and the observer's individual response.

Data analysis

The stand-alone performance of CAD was determined by calculating the sensitivity and mean false-positive per image (mFP). A one-way ANOVA with Tukey post hoc test was used to test differences among the four subsets. Sensitivity, specificity and AUC were calculated per observer and per subset for the readings with and without the availability of CAD results. For calculating the sensitivity, ratings 1–3 were considered negative and ratings 4 and 5 were considered positive if the lesions were correctly localised. The literature is controversial with regard to whether application of CAD as a second reader should allow for the discharge of lesions located during primary unassisted reading (as done in our study) or should only be used to add potential lesions (add-on mode) ⁽¹³⁾. We therefore also analysed the data for the add-on scenario with preservation of all originally indicated lesions. Comparisons of all three methods were made using Cochran Q tests as well as logistic regressions for repeated measurements (GEE). Pairwise comparisons were carried out using McNemar's test for each reader separately. To assess the impact of the feedback on reader performance we compared the results of all readers for the first two subsets with the results for the last two subsets. All analyses were performed in SPSS 17. Statistical significance was assumed at $P < 0.05$.

Results

Study group

Patient mean age was 61 years with no significant difference among the four subsets (59.7, 58.8, 61.9 and 60.3). All subsets consisted of 14 patients with a solitary nodule and 21 negative control patients. Of all nodules 30% (17) had a high conspicuity, 23% (13) a moderate, 30% (17) a low and 16% (9) a very low conspicuity (Table 1). None of the patients of the diseased or of the control group showed any relevant pathological feature other than the effects of smoking and the focal study lesion.

Table 1

Distribution of false positive CAD candidates and nodule conspicuity per subset.

Subset	Total CAD FP	Nodule conspicuity			
		1	2	3	4
A	56	5	4	4	4
B	70	3	4	3	3
C	52	4	4	5	4
D	82	2	2	2	3

Nodule conspicuity: 1 = high, 2 = moderate, 3 = low and 4 = very low

CAD stand-alone

Stand-alone CAD detected 32 out of 56 nodules leading to a mean sensitivity and a sensitivity per subset of 57%. Sensitivity was 100%, 54%, 44% and 22% for nodules of high, moderate, low and very low conspicuity, respectively. CAD generated a total of 260 FP candidates with a mFP of 1.9. The mFP for the different subsets amounted to 1.6, 2.0, 1.5 and 2.3 with a significant difference between subset three and four ($P = 0.024$). The positive predictive value was 0.31 (32/103) whereas the negative predictive value was 0.35 (13/37).

Reader performance for subsets 1 and 2

Sensitivity, specificity and AUC were not significantly affected by the use of CAD (Table 2). Mean AUC increased from 0.76 without CAD to 0.78 with CAD, but the difference did not reach significance. Use of CAD in the add-on scenario led to an increase of sensitivity (65% versus 69%) but at the expense of loss of specificity (79% versus 76%). If discharge of lesion candidates was allowed with the use of CAD, sensitivity changed from 65% to 66% and specificity from 79% to 80%. None of the differences listed reached statistical significance.

Reader performance for subsets 3 and 4

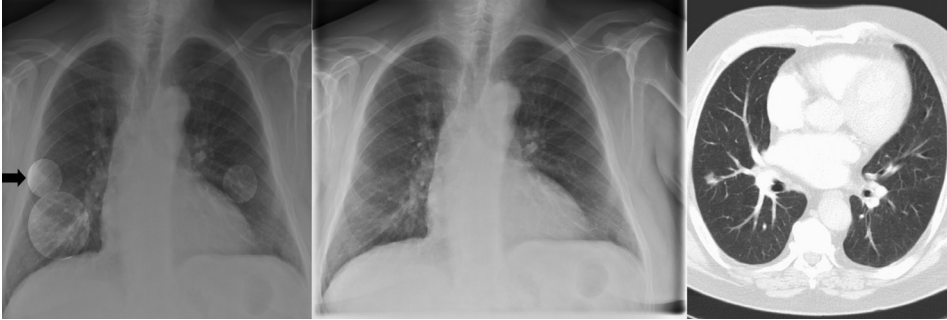
Reading sessions 3 and 4, sensitivity, specificity and AUC were not significantly affected by the use of CAD (Table 2). Readers increased their baseline sensitivity to a mean of 70% at the expense of a slightly lower specificity of 74% compared with the first two reading sessions. With CAD the AUC slightly increased from 0.82 to 0.84, but the differences did not reach significance. While the sensitivity remained unchanged with CAD (70 and 71%, respectively), the specificity slightly increased with CAD when discharge of lesions was allowed (77 and 74%, respectively). None of the differences listed reached statistical significance.

Reader-CAD interaction

Analysing pooled data, all six readers dismissed 17% (32 out of 192) of the TP CAD candidates. There was no difference between the first two and last two subsets (16 vs. 16). Of the dismissed TP candidates 50% (16 out of 32) referred to nodules of low conspicuity and 28% (9 out of 32) to lesions of very low conspicuity with the remaining 22% (7 out of 32) referring to moderately conspicuous nodules. The number of dismissed true positive CAD amounted to 1/1, 0/0, 6/4, 1/0, 1/0 and 1/1 for subsets 1 and 2 versus subsets 3 and 4, respectively for readers 1 to 6. Based on pooled data analysis the number of accepted FP CAD

candidates non-significantly decreased from a total of 10 made in the first two readings to 6 made in the last two readings. An example of reader-CAD interaction is given in Figure 1.

Figure 1



The nodule of low conspicuity in the right lower lobe was missed by three readers without the CAD results. The true-positive CAD candidate was accepted by two readers and dismissed by one reader (arrow). None of the readers accepted the false-positive CAD candidate

Table 2

Individual reader outcome without and with use of CAD.

	Sensitivity		Specificity		Az	
	1+2	3+4	1+2	3+4	1+2	3+4
No CAD	65% (58% - 73%)	70% (63% - 77%)	65% (74% - 84%)	70% (69% - 80%)	0.76 (0.71 - 0.81)	0.82 (0.78 - 0.86)
CAD with possible discharge	66% (59% - 73%)	70% (63% - 77%)	80% (75% - 85%)	77% (72% - 82%)	0.78 (0.73 - 0.83)	0.84 (0.79 - 0.88)
CAD add-on	69% (62% - 76%)	71% (65% - 78%)	76% (71% - 81%)	74% (69% - 80%)	0.78 (0.73 - 0.83)	0.85 (0.82 - 0.89)

CAD: computer-aided detection, Az: area under receiver operating characteristics curve

Discussion

It is common clinical practice that small primary lung carcinomas are missed on two-view chest radiographs although they were frequently visible in retrospect. Miss rates of between 20% and 90% have been reported for primary lung carcinomas⁽¹⁴⁻²⁰⁾. Two recent papers reported a CAD stand-alone sensitivity of 35% and 47% for pulmonary tumours initially overlooked by the radiologist⁽¹⁻⁴⁾ indicating the potential of CAD to improve the readers' detection performance. Both papers, however, did not include observer performances, thus to which

extent radiologists would have taken advantage of CAD and accepted these true-positive candidates remains unanswered. Equally the risk of accepting false-positive candidates with unnecessary follow-up diagnostics cannot be quantified. The importance of the interaction between CAD and observer has been previously illustrated by a study that tested the impact of CAD on observer performance for the detection of T1 tumours on chest radiographs of patients who had been part of a CT screening trial: the number of malignancies initially not seen by the observers but correctly annotated by CAD varied between 5 and 16 per observer but 80% (46 out of 59) of these correctly annotated lesions were not accepted by the readers and subsequently dismissed⁽⁹⁾. One underlying reason for the readers' inability to differentiate true- from false-positive lesions might have been the lack of experience with the CAD algorithm and subsequently insufficient trust in its performance. In the current study we therefore tested the effect of short-term feedback on the detection of pulmonary nodules on digital chest radiography. Our hypothesis was that individual feedback after the interpretation of each of the four subsets would help readers to build up more confidence in the performance of the CAD algorithm, eventually resulting in an increased ability to distinguish true- from false-positive candidates and in a higher acceptance of true-positive CAD candidates. We found a slight but not significant increase in baseline performance between sessions 1 and 2 compared with sessions 3 and 4, meaning that there was a small overall training effect and that readers detected more nodules in the last two readings than in the first two readings. For none of the paired sessions, however, could we prove a significant influence of reader performance by CAD. Neither the acceptance of true-positive CAD candidates nor the ability to dismiss false-positive candidates was significantly influenced by the feedback information. There was an overall tendency towards increased performance with CAD but the differences were too small to reach significance. Seventy-eight percent of the dismissed true-positive CAD candidates referred to lesions of low and very low conspicuity indicating that readers had difficulties in assigning sufficient credibility to CAD candidates, indicating nodules with low conspicuity. Feedback did not help in this respect because the number of dismissed true-positive candidates was the same in sessions 1 and 2 compared with sessions 3 and 4. There are very few studies in the literature evaluating the impact of training on the use of CAD and those referred for CT colonography and mammography. Results were quite different in the respect that a short, 1-day period of training already affected observer performance in CT colonography as opposed to mammography, which had a lower learning curve requiring at least 4 weeks⁽¹⁰⁻¹¹⁾. It seems that the effect of CAD follows different perception and learning rules in CT

versus radiography and that a beneficial use of CAD to reduce perception errors of well-defined and more conspicuous lesions (e.g. colon polyps) can be learned faster than the use of CAD to detect lesions of low conspicuity that require differentiation from obscuring background noise (e.g. lesions on mammography). This is also supported by our result that most of the dismissed true-positive CAD candidates referred to lesions of low and very low conspicuity. It is very likely that lesions of high or moderate conspicuity that have been missed by the readers due to “inattentional blindness” meaning that the lesions missed during routine reporting take more effectively and more easily advantage of the availability of CAD. Unfortunately it is much more difficult to prove this effect under study conditions because readers tend to analyse the radiographs with an especially high degree of alertness than under normal conditions. Even though the readers in our study generally had a low level of experience in reading chest radiographs – 5 out of 6 were residents – the baseline sensitivity was relatively high with a mean of 65% for the first two subsets and 70% for the last two subsets. It is possible that this high baseline sensitivity impeded further increase in sensitivity with the availability of CAD results. Lesions of low or very low conspicuity, however, have different diagnostic requirements: correct diagnosis requires not only visual localisation but also correct differentiation from surrounding “anatomical” noise. For this type of lesions, our results suggest that CAD has no significant impact on reader behavior and short-term feedback has no effect. Whether a longer learning period would be more efficient as postulated for mammography⁽¹²⁾ remains to be proven for chest radiography. Our results also underline the need to further decrease the number of false-positive CAD candidates. A lower number of false-positive candidates will not only decrease reading time but also increase readers’ confidence in the reliability of CAD and will help them to focus on the presence of underlying lesions in the circled areas of interest. The number of false-positive calls provoked by CAD was quite low in this study: 10 and 6, respectively, pooled over all readers for the first two and last two readings. It might also be the case that the pure presentation of CAD candidates alone is not sufficient and more information is needed to help the reader to correctly differentiate true- from false-positive lesions. In that context the availability of likelihood calculations together with an active localisation procedure by the reader him/herself had been found to be very effective for mammography⁽²¹⁾. Our study suffers from the following limitations. Nodule incidence in the study population was higher than usually seen under clinical conditions. Although the readers did not know the exact distribution of positive and negative cases they certainly were more alert to detecting focal lesions than in a usual clinical setting.

The number of CAD false-positive candidates was not equally distributed over the four subsets. Although we consider it unlikely that this had an effect on our results as all readers interpreted the subsets in a different order and there was a generally low number of accepted CAD false-positive candidates. This study used a specific type of CAD algorithm. It has to be noted that these results are not necessarily transferable to other CAD algorithms: a different or updated CAD algorithm with a different performance may yield different results in a context in which perception, reader experience and confidence as well as lesion conspicuity form a complicated framework. Our study represents the experience of one institution; whether a different reader group or involving readers from various institutions would have yielded different results remains speculative. We conclude that short-term feedback does not significantly increase the ability of readers to differentiate true- from false-positive candidate lesions in chest radiography in order to use CAD more effectively for the detection of nodular lesions. Further research is needed to determine whether a longer training period or additional processing and display tools such as temporal subtraction, rib suppression or likelihood calculations can increase the benefits of CAD for reader performance in chest radiography.

References

- 1 Li F, Engelmann R, Metz CE, Doi K, MacMahon H. Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. *Radiology* 2008; 246:273-280
- 2 Van Beek EJ, Mullan B, Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images: a prospective study. *Acad Radiol* 2008; 15:571-575
- 3 Bley TA, Baumann T, Saueressig U et al. Comparison of radiologist and CAD performance in the detection of CT-confirmed subtle pulmonary nodules on digital chest radiographs. *Invest Radiol* 2008; 43:343-348
- 4 White CS, Flukinger T, Jeudy J, Chen JJ. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology* 2009; 252:273-281
- 5 Sakai S, Soeda H, Takahashi N et al. Computer-aided detection on digital chest radiography: validation test on consecutive T1 cases of resectable lung cancer. *J Digit Imaging* 2006; 19:376-382
- 6 Hardie RC, Rogers SK, Wilson T, Rogers A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Med Image Anal* 2008; 12:240-258
- 7 Xu Y, Ma D, He W. Assessing the use of digital radiography and a real-time interactive pulmonary nodule analysis system for large population lung cancer screening. *Eur J Radiol* 2011 May 27 [Epub ahead of print]
- 8 Song W, Fan L, Xie Y, Qian JZ, Jin Z. A study of inter-observer variations of pulmonary nodule marking and characterizing on DR images. *Proc SPIE* 2005; 5749:272-280
- 9 de Hoop B, De Boo DW, Gietema HA et al. Computer-aided detection of lung cancer on chest radiographs: effect on observer performance. *Radiology* 2010; 257:532-540
- 10 Taylor SA, Burling D, Roddie M, Heneyfield L, McQuillan J, Bassett P, Halligan S. Computer-aided detection for CT colonography: incremental benefit of observer training. *Br J Radiol* 2008; 81:180-186
- 11 Luo P, Qian W, Romilly P. CAD-Aided mammogram training. *Acad Rad* 2005; 12:1039-1048
- 12 Nishikawa R. Increased CAD use prompts look at advantages, drawbacks. *Radiological Society of North America Web site*. http://www.rsna.org/Publications/rsnanews/feb07/upload/RSNANews_Feb07_CAD_Usage.pdf. Published February 2007. Last accessed September 18, 2011
- 13 Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative analysis—the image role of medical physics and AAPM. *Med Phys* 2008; 35:5799-5820
- 14 Muhm JR, Miller WE, Fontana RS, Sanderson DR, Uhlenhoop MA. Lung cancer detected during a screening program using 4-month chest radiographs. *Radiology* 1983; 148:609-615
- 15 Heelan RT, Flehinger BJ, Melamed MR, et al. Non-small-cell lung cancer: results of the New York screening program. *Radiology* 1984; 151:289-293
- 16 Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with potentially resectable lesion evident in retrospect. *Radiology* 1992; 182:115-122
- 17 Quekel LG, Kessels AG, Goei R, Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest* 1999; 115:720-724
- 18 Monnier-Cholley L, Arrivé L, Porcel A, et al. Characteristics of missed lung cancer

- on chest radiographs: a French experience. *Eur Radiol* 2001; 11:597–605
- 19 Shah PK, Austin JH, White CS, et al. Missed non-small cell lung cancers: Radiographic findings of potentially resectable lesions evident only in retrospect. *Radiology* 2003; 226:235–241
 - 20 Wu HM, Gotway MB, Lee TJ, et al. Features of non-small cell lung carcinomas overlooked at digital chest radiography. *Clin Radiol* 2008; 63:518-528
 - 21 Samulski M, Hupse R, Boetes C, Mus RD, den Heeten GJ, Karssemeijer N. Using computer-aided detection in mammography as a decision support. *Eur Radiol* 2010; 20:2323-2330

Summary
and
General Discussion



Summary

Since the introduction of digital chest radiography continuous advances have been made with respect to detector dose efficiency and processing tools. In this thesis the effect of various aspects of advances in digital chest radiography on actual reader performance was studied. In **Chapter 2** we compared mobile direct radiography (DR) and mobile computed radiography (CR) units for bedside chest radiography of patients admitted to an intensive care unit (ICU). Overall image quality, delineation of anatomical landmarks, and of devices for monitoring were scored better with DR as compared with CR. Even with 50% dose reduction DR outperformed CR with respect to delineation of mediastinal landmarks and devices for monitoring. Image quality was scored equally when assessed individually, in a side-by-side comparison, however, 87% of radiographs obtained with DR at 50% dose reduction were rated superior to CR. Interobserver agreement for the assessment of pathology was used as surrogate to test whether improved image quality would have an effect on diagnostic performance. Only DR achieved an agreement rate of 0.48, which is considered as clinically acceptable. DR_{50%} and CR performed equally (0.39 and 0.33, respectively).

Chapter 3 focuses on the effects of grey-scale reversal for the detection of small pulmonary nodules in chest radiography. Grey-scale reversal is a very simple processing tool, available on PACS workstations with a single mouse click. From optical physiology it is known that optical contrast perception is increased when a dark object is presented on a white background. We suspected that a positive image ("bones black") would help the reader to detect small nodules. Three residents and three radiologists did not benefit from grey-scale reversal for the detection of the nodules (mean diameter 13 mm; median 11 mm) when offered as additional image. We concluded that grey-scale reversal is not a helpful processing tool for the detection of small pulmonary nodules.

Chapter 4 provides an overview of the publications dealing with CAD for the detection of intrapulmonary nodules and T1 lung carcinoma's up to the time when our observer studies were carried out. Very variable results were reported for the various prototypes and FDA approved CAD systems. In all studies, the prevalence of nodules was higher as one would normally see under clinical conditions. Most studies only assessed the stand-alone performance, which is strongly influenced by prevalence, lesion conspicuity and lesion size. Similarly as for the assessment of the stand-alone performance, reader and lesion selection play an important role in observer studies: the level of reader experience and the distribution of lesion conspicuity will strongly

influence the potential effect of CAD. Besides, one CAD algorithm functions differently from the other. All these aspects make it very difficult to compare the results of the various studies with each other and one has to be very careful to draw conclusions on the impact of CAD in general. Experts therefore request to unify evaluation forces and to use "common databases" of clinically validated images.

In **Chapter 5** and **Chapter 6** the two, currently commercially available and FDA approved, CAD systems were tested. Besides their underlying algorithm analysis, they differ in the display of the CAD candidates and the integration into the workflow.

Chapter 5 presents the results of an observer study testing the impact of CAD (Onguard 5.0; Riverain, Miamisburg, Ohio) on the detection of CT proven T1 tumors in participants of the Dutch-Belgian lung cancer screening trial (NELSON). The stand-alone sensitivity of CAD and the sensitivity of the experienced radiologists were equal; however, the mean number of false positives per image was ten times higher for CAD as compared with the radiologists (2.4 vs. 0.24). Use of CAD did not improve the detection performance of the readers though CAD correctly identified tumors, the readers originally had missed. Out of 55 true positive CAD candidates between 5 to 16 were dismissed by the radiologists. Especially true positive CAD candidates for subtle, low conspicuous tumors were dismissed, indicating the potential for CAD, but also the difficulties readers had differentiating true from false positive CAD candidates.

In **Chapter 6** a different patient group and lesion type were selected to test the second, FDA approved, CAD software (IQQA-Chest; EDDA Technology, Princeton Junction, NJ). The study-group compiled of elderly patients, the majority of which had a positive smoking history. Both smoking and aging led to increased interstitial markings described as "dirty lungs", hampering the detection of focal lesions. Small and low conspicuous pulmonary nodules were included to challenge perception capabilities. Sensitivity significantly increased for inexperienced readers from 39% without CAD to 45% with CAD ($p < 0.05$). A nonsignificant increase of the mean false positives per image (0.27 vs. 0.34) impeded a significant increase of the figure of merit (0.71 vs. 0.71). The experienced readers performed better than the inexperienced, but showed no significant difference in sensitivity (50%) mean false positive per image (0.16 vs. 0.21) or figure of merit (0.84 vs. 0.87) when using CAD as second reader. With 33% of true positive CAD candidates dismissed

and 40% of false positive marks by the readers provoked by a false positive CAD candidate, readers again showed difficulties to differentiate true from false positive CAD candidates.

Based on these results we hypothesized that a lack of “trust” in the performance of CAD obviates a more beneficial use. In **Chapter 7** we therefore studied whether observer training would increase reader performance with CAD for pulmonary nodules. The CAD stand-alone sensitivity was 59%, which was slightly lower than the 65% of the radiologists. Each reader received individual feedback on his / her performance after having read a subset of images. The hypothesis was that this would cause a learning curve, and induce a more beneficial use of CAD. Short-term feedback resulted in an overall increase of reader performance without an improvement by CAD. This was true for both ways CAD can be applied: either with the possibility to discharge marks for lesions located during primary unassisted reading, or for the add-on scenario with preservation of all originally indicated lesions. A total of 32 true positive CAD candidates were dismissed by all readers without a difference over time (16 vs. 16). The dismissed true positive CAD candidates were for low and very low conspicuous nodules in 78% (25/32). These results indicate that short term feedback did not increase the ability of readers to differentiate between true and false positive candidate lesions in order to use CAD more beneficially.

General discussion

Similar as with other imaging techniques (e.g., CT or MRI) the continuous technical development of digital radiography requires constant adaptation of protocols. Scientific evaluation of these advances is necessary in order to estimate the impact on image quality and the value for diagnostic performance. This thesis covers three aspects of recent advances made in the area of digital chest radiography:

- increased dose efficiency for mobile chest radiography
- grey-scale reversal for image display
- computer-aided detection aimed to support the detection of focal pulmonary opacities.

Increased dose efficiency for mobile chest radiography

DR, as opposed to CR, offers higher dose efficiency which had been extensively evaluated in upright chest radiography. Dose reductions up to 60% have been reported⁽¹⁻⁵⁾. Mobile DR units appropriate for bedside chest radiography became technically available only years after those for upright radiography. Until the study, presented in **Chapter 2**, mobile DR units had only been tested for pediatric applications^(6,7). We compared mobile DR and CR units for bedside chest radiography in an adult ICU. One of the major indications for ICU chest radiography, delineation of monitor materials, was scored significantly better with DR as compared with CR, even when DR images were obtained with 50% of the "standard" acquisition dose. The underlying reason is that especially in high absorption areas, such as the mediastinum and upper abdomen, the increased dose efficiency of the DR provides a better signal-to-noise, thus contrast-to-noise ratio. Therefore also the anatomical landmarks within the mediastinum were seen superiorly with DR. Equally for the assessment of overall image quality, DR outperformed CR in the side-to-side comparison. The clinically relevant question is whether this increased image quality can be transferred into higher diagnostic performance. In the absence of a reference standard for pathology, the interobserver agreement was calculated as surrogate for diagnostic performance. Thereby we could demonstrate that the increased image quality of DR indeed transferred into a superior interobserver agreement. DR obtained with 50% dose reduction was still equal to CR. Though the dose of a single radiograph is very low (effective dose: 0.029 mSv), ICU patients frequently undergo multiple follow-up radiographic studies. Whether the increased detector efficiency should be invested into dose reduction for the patient or should be used to increase diagnostic

performance cannot be generally answered and should be determined also as function of the clinical situation and the diagnostic purpose of the examination which is different for upright and bedside radiography. In our institution we chose for patients' dose reduction for bedside radiography.

Grey-scale reversal for image display

Both, clinical experience and the known superior optical contrast perception of the human eye for dark objects on a white background⁽⁸⁾ motivated us to reevaluate grey-scale reversal for the detection of nodular opacities in chest radiography. Gray-scale reversal is a rather simple processing tool and has already been evaluated in the era of hardcopies, more than 20 years ago. At that point, not surprisingly, results had been disappointing, since extra hardcopies with a fixed reversed gradation curve had been produced and tested separately from radiographs with usual gradation characteristics. Nowadays, radiographs are exclusively read as softcopies and gray-scale reversal is available by a simple mouse click on the PACS workstation. The study presented in **Chapter 3** revealed no benefit from grey-scale reversal for the detection of small pulmonary nodules which contradicted our hypothesis. The nodules included in the study had all been rather small with a mean diameter of 13 mm and a median diameter of only 11 mm. We suspected that grey-scale reversal does not help for these types of nodules, due to the interference with vascular and interstitial structures that equally become more prominent with grey-scale reversal. Larger and more ill defined geographic lesions might take more advantage of grey-scale reversal, but this has not yet been tested.

Computer-aided detection aimed to support the detection of focal pulmonary opacities.

Computer-aided detection (CAD) has evolved from prototypes with low sensitivity and high mean false positives per image to FDA approved algorithms with reported stand-alone sensitivities of 35% and 47% for initially missed bronchogenic carcinomas^(9,10). A lower agreement rate between CAD and radiologists compared to the agreement between radiologists furthermore underscores the potential of CAD to detect nodules that radiologists tend to miss⁽¹¹⁾. The literature lists some publications that report a promising increase of reader performance with the use of CAD as second reader⁽¹²⁻¹⁴⁾. These results are in contrast to the data presented in **Chapter 5** and **Chapter 6**. Both studies failed to show a significantly improved detection performance of observers for small focal lung lesions. Nevertheless we observed an interaction between observers and CAD.

In the first study 82% (54/66) of new markings made by observers were due to false positive CAD candidates. In the second study 40% (61/154) of detrimental rating differences were possibly provoked by a false positive CAD candidate in the same location. True positive CAD candidates were dismissed in 80% and 33% respectively, meaning the observer interpreted the region pointed out by CAD as normal. The majority of these dismissed true positive candidates were for low to very low conspicuous lesions. We concluded from these results that CAD has potential to have a positive impact on reader performance since it detects different lesions than the observers. A more beneficial use of CAD was undone by the inability of the observers to discriminate true positive from false positive CAD candidates. It is likely that high and moderate conspicuous lesions that have been missed by the readers due to "inattentive blindness" would take more easily advantage of the availability of CAD. This effect, however, is very difficult to prove under study conditions, since observers will always analyze images with special diligence under study conditions. Lesions of low or very low conspicuity, however, have different diagnostic requirements; correct diagnosis requires not only visual localization but also correct differentiation from surrounding "anatomic noise". None of the observers in our study had a broader clinical experience with the CAD algorithm. Though we had introduced them to the software by a number of training cases, we suspected that the inability to correctly discriminate the candidates could have been also caused by a lack of familiarity with the system. We therefore undertook the study evaluating the effects of short term feedback to the readers described in **Chapter 7**. All readers received an individual feedback on their performance without and with CAD immediately after having read a subset of the study images. Results, however, found no increase of readers' ability to differentiate true from false positive CAD candidates. There was an overall training effect for the detection of nodules but not an increased benefit of CAD. The ability to discriminate true from false positive candidates is apparently more complex and requires more information than increasing the familiarity with the CAD algorithm itself. Limited evidence is available about observer training for CAD. A one-day-training for CAD in CT colonography resulted in an increase of sensitivity, but also in a decrease of specificity⁽¹⁵⁾. For CAD applied in mammography an increase of sensitivity, specificity and Az was reported after a four week training period⁽¹⁶⁾. However, the ultimate learning curve for CAD in mammography was estimated to be around 2 years⁽¹⁷⁾. Whether this applies for CAD in chest radiography is not yet known. Further on it is very likely that an increase of confidence into CAD can be achieved by decreasing the number of false positive CAD candidates. Li et al. reported that 68% of

the false positives CAD candidates were solely or in part triggered by a bony structure⁽⁹⁾. First studies combining CAD with bone suppression techniques, like energy subtraction, provided promising results^(18,19). Studies under way using an upgraded software version of the same CAD algorithm used in chapter 6 (Onguard; Riverain), showed a marked performance improvement with a sensitivity of around 70-80% with a mean number of false positives per image of 0.5. Whether this indeed transfers to an improved observer performance is currently under evaluation.

Another approach is to alter the presentation of the CAD candidates: currently a maximum of 5 ROIs are presented on demand without any additional information. An option could be to add the likelihood per CAD candidate. Another option could be to leave the visual interrogation of the chest radiograph to the observer and only provide CAD results for locations the observer is pointing out. The latter approach was found successful for the detection of tumors in mammography⁽²⁰⁾ and produced promising results for nodule detection in a preliminary study using a public data base⁽²¹⁾ and a number of observers not trained in radiology⁽²²⁾.

All of these above mentioned approaches are currently under evaluation and upcoming results will determine the future role of CAD for nodule detection in chest radiography.

In summary this thesis demonstrates that

- The introduction of mobile direct radiography at the bedside allows for 50% dose reduction, as compared to computed radiography, without loss of clinically relevant image quality. Alternatively, the improved image quality obtained at unaltered dose can be used to uniform diagnostic performance.
- Using PACS display of digital chest radiographs, gray-scale reversal does not help the radiologists in detecting small pulmonary nodules.
- The potential of CAD to reduce detection errors by radiologists is not fully established.
- Despite short-term observer training, radiologists still have difficulties differentiating true positive from false positive CAD candidates.

References

- 1 Gruber M, Uffmann M, Weber M, Prokop M, Balassy C, Schaefer-Prokop C. Direct detector radiography versus dual reading computed radiography: feasibility of dose reduction in chest radiography. *Eur Radiol* 2006; 16:1544-1550
- 2 Strotzer M, Volk M, Frund R, Hamer O, Zorger N, Feuerbach S. Routine chest radiography using a flat-panel detector: image quality at standard detector dose and 33% dose reduction. *AJR Am J Roentgenol* 2002; 178:169-171
- 3 Herrmann A, Bonel H, Stabler A, et al. Chest imaging with flat-panel detector at low and standard doses: comparison with storage phosphor technology in normal patients. *Eur Radiol* 2002; 12:385-390
- 4 Bacher K, Smeets P, Bonnarens K, De Hauwere A, Verstraete K, Thierens H. Dose reduction in patients undergoing chest imaging: digital amorphous silicon flat-panel detector radiography versus conventional film-screen radiography and phosphor-based computed radiography. *AJR Am J Roentgenol* 2003; 181:923-929
- 5 Bacher K, Smeets p, Vereecken L, et al. Image quality and radiation dose on digital chest imaging: comparison of amorphous silicon and amorphous selenium flat-panel systems. *AJR Am J Roentgenol* 2006; 187:630-637
- 6 Rapp-Bernhardt U, Bernhardt TM, Lenzen H, et al. Experimental evaluation of a portable indirect flat-panel detector for the pediatric chest comparison with storage phosphor radiography at different exposures by using a chest phantom. *Radiology* 2005; 237:485-491
- 7 Rapp-Bernhardt U, Roehl FW, Esseling R, et al. Portable flat-panel detector for low-dose imaging in a pediatric intensive care unit comparison with an asymmetric film-screen system. *Invest Radiol* 2005; 40:736-741
- 8 Blackwell H. Contrast thresholds of the human eye. *J Opt Soc Am* 1946; 36:624-643
- 9 Li F, Engelmann R, Metz CE, Doi K, MacMahon H. Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. *Radiology* 2008; 246:273-280
- 10 White CS, Flukinger T, Jeudy J, Chen JJ. Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology* 2009; 252:273-281
- 11 Bley TA, Baumann T, Saueressig U et al. Comparison of radiologist and CAD performance in the detection of CT-confirmed subtle pulmonary nodules on digital chest radiographs. *Invest Radiol* 2008; 43:343-348
- 12 Van Beek EJ, Mullan B, Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images: a prospective study. *Acad Radiol* 2008; 15:571-575
- 13 Xu Y, Ma D, He W. Assessing the use of digital radiography and a real-time interactive pulmonary nodule analysis system for large population lung cancer screening. *Eur J Radiol* 2011 May 27 [Epub ahead of print]
- 14 Song W, Fan L, Xie Y, Qian JZ, Jin Z. A study of inter-observer variations of pulmonary nodule marking and characterizing on DR images. *Proc SPIE* 2008; 5749:272-280
- 15 Taylor SA, Burling D, Roddie M, Heneyfield L, McQuillan J, Bassett P, Halligan S. Computer-aided detection for CT colonography: incremental benefit of observer training. *Br J Radiol* 2008; 81:180-186
- 16 Luo P, Qian W, Romilly P. CAD-Aided mammogram training. *Acad Rad* 2005; 12:1039-1048.
- 17 Nishikawa R. Increased CAD use prompts look at advantages, drawbacks. *Radiological Society of North America Web site*. <http://www.rsna.org/Publications/>

- rsnanews/feb07/upload/RSNANews_Feb07_CAD_Usage.pdf. Published February 2007. Last accessed September 18, 2011
- 18 Szucs-Farkas Z, Patak MA, Yuksel-Hatz S, Ruder T, Vock P. Improved detection of pulmonary nodules on energy-subtracted chest radiographs with a commercial computer-aided diagnosis software: comparison with human observers. *Eur Radiol* 2010; 20:1289-1296
 - 19 Balkman JD, Mehandru S, DuPont E, Novak RD, Gilkeson RC. Dual energy subtraction digital radiography improves performance of a next generation computer-aided detection program. *J Thoracic imaging* 2010; 25:41-47
 - 20 Samulski M, Hupse R, Boetes C, Mus RD, den Heeten GJ, Karssemeijer N (2010) Using computer-aided detection in mammography as a decision support. *Eur Radiol* 20:2323-2330
 - 21 Shiraishi J, Katsuragawa S, Ikezoe J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR Am J Roentgenol* 2000; 174:71-74
 - 22 Samulski MRM, Snoeren PR, Platel B, et al. Computer-aided detection as a decision assistant in chest radiography. *Proc. SPIE* 2011; 796614



Samenvatting



Samenvatting

Sinds de introductie van digitale thorax radiografie zijn er voortdurend verbeteringen geïntroduceerd op het gebied van dosis efficiëntie van de detectoren. Tevens maakte de digitale thorax radiografie de weg vrij voor de ontwikkeling van diverse processing tools. Dit proefschrift behandelt de effecten van de verschillende voordelen die digitale thorax radiografie biedt en dan vooral de invloed hiervan op de prestaties van radiologen.

In **hoofdstuk 2** wordt mobiele direct radiografie (DR) en mobiele computed radiografie (CR) vergeleken voor thorax foto's gemaakt aan bed bij patiënten opgenomen op een volwassen intensive care afdeling. De algehele beeldkwaliteit, de afgrensbaarheid van anatomische structuren en van lijnen / tubes werden beter beoordeeld met DR in vergelijking met CR. Zelfs DR met 50% dosis reductie ($DR_{50\%}$) was beter dan CR zowel met betrekking tot de afgrensbaarheid van mediastinale structuren als die van de lijnen / tubes. De beeldkwaliteit werd tijdens de separate beoordeling als gelijk beoordeeld. Echter tijdens de directe vergelijking werd de beeldkwaliteit van $DR_{50\%}$ in 87% van de gevallen als beter beschouwd dan die van CR. Interobserver variabiliteit voor de beoordeling van pathologie werd als surrogaat gebruikt om te onderzoeken of verbeterde beeldkwaliteit een effect zou hebben op de diagnostische prestaties. Alleen DR had een variabiliteit van 0.48, welke gezien wordt als klinisch acceptabel. $DR_{50\%}$ en CR hadden een lagere, vergelijkbare variabiliteit (respectievelijk, 0.39 en 0.33).

Hoofdstuk 3 richt zich op de effecten van "grey-scale reversal" op de detectie van kleine long noduli in de thorax radiografie. "Grey-scale reversal" is een simpel hulpmiddel dat met één muisklik beschikbaar is op elk PACS werkstation. Men weet vanuit optische fysiologie dat contrast perceptie beter is wanneer een donker object op een witte achtergrond wordt gepresenteerd. Wij vermoedden dat een geïnverteerde thoraxfoto ("bones black") de beoordelaar zou helpen in het detecteren van kleine noduli. Drie radiologen in opleiding en drie ervaren radiologen hadden echter geen baat bij "grey-scale reversal" voor de detectie van de noduli (gemiddelde diameter 13mm; mediaan 11mm). Wij hebben hieruit geconcludeerd dat "grey-scale reversal" geen bruikbaar hulpmiddel is voor de detectie van kleine long noduli.

Hoofdstuk 4 biedt een overzicht van de publicaties over computer-aided detection (CAD) voor de detectie van long noduli en T1 longcarcinomen ten tijde van de uitvoering van onze twee

observer studies. Zeer uitlopende resultaten worden gerapporteerd voor de verschillende prototypes en FDA goedgekeurde CAD systemen. De prevalentie van noduli was in alle studies hoger dan normaal. In vele studies werd slechts de stand-alone performance beoordeeld. Deze wordt echter sterk beïnvloed door prevalentie, detecteerbaarheid en grootte van de studie laesies. Net als bij de beoordeling van de CAD stand-alone performance, bepalen waarnemers en laesie selectie grotendeels de uitkomst bij observer studies. De ervaring van de waarnemers en distributie van detecteerbaarheid van de laesies zal het potentiële effect van CAD sterk beïnvloeden. Daarnaast presteert het ene CAD systeem anders dan het andere. Al deze aspecten maken het vergelijken van de tot nu gepubliceerde studies lastig en men dient voorzichtig te zijn met het trekken van algemene conclusies over CAD. Experts pleiten derhalve ook voor een uniforme vergelijking door middel van publiekelijk beschikbare databases met gevalideerde thorax foto's.

In **hoofdstuk 5** en **hoofdstuk 6** worden de twee huidig beschikbare FDA goedgekeurde CAD systemen onderzocht. Naast de daadwerkelijke analyse verschillen de systemen onderling zowel met betrekking tot de weergave van de CAD markeringen als de integratie in de dagelijkse praktijk.

Hoofdstuk 5 beschrijft de resultaten van de eerste observer studie waarin de invloed van CAD (Onguard 5.0; Riverain, Miamisburg, Ohio) op de detectie van T1 tumoren van deelnemers aan de Nederlands-Belgische longkanker screening studie (NELSON) wordt onderzocht. De sensitiviteit van CAD en die van ervaren radiologen was gelijk, echter het gemiddeld aantal fout positieve markeringen per foto was tien maal hoger voor CAD in vergelijking met die van de radiologen (2.4 vs. 0.24). Ondanks dat CAD tumoren markeerde die de radiologen initieel hadden gemist, leidde het gebruik van CAD niet tot een betere detectie van T1 tumoren. Van de 55 terecht positieve CAD markeringen werden er 5 tot 16 door de radiologen verworpen. Vooral terecht positieve CAD markeringen voor subtiele tumoren werden verworpen. Dit benadrukt het potentieel van CAD, maar tegelijkertijd ook dat gebruikers ervan moeite hebben om terecht positieve van fout positieve CAD markeringen te onderscheiden.

In **hoofdstuk 6** werd een studiegroep met andere studie laesies geselecteerd om het tweede FDA goedgekeurde CAD systeem (IQQA-Chest; EDDA Technology, Princeton Junction, NJ) te onderzoeken. De studiegroep bestond uit oudere patiënten waarvan de meerderheid rookte of gerookt had. Zowel ouderdom als roken leidt tot een toename van interstitiële longafwijkingen die in de literatuur

beschreven wordt als "dirty lungs". Kleine noduli met een lage detecteerbaarheid werden geselecteerd om de detectie capaciteit te testen. De sensitiviteit van onervaren waarnemers steeg significant van 39% naar 45% met gebruik van CAD als second reader ($p < 0.05$). Een niet significante stijging van het gemiddeld aantal fout positieve markeringen per foto (0.27 vs. 0.34) verhinderde echter een significante stijging van de figure of merit (0.71 vs. 0.71). De ervaren waarnemers presteerden beter dan de onervaren waarnemers, maar de verschillen zonder en met gebruik van CAD als second reader waren niet significant voor de sensitiviteit (50% vs. 51%), het gemiddeld aantal fout positieve per foto (0.16 vs. 0.21) en de figure of merit (0.84 vs. 0.87). Er werd 33% van de terecht positieve CAD markeringen verworpen en 40% van de fout positieve markeringen door de radiologen werd uitgelokt door een fout positieve CAD markering op dezelfde plaats. Dit liet wederom de moeite zien die gebruikers van CAD hebben om de terecht positieve van fout positieve CAD markeringen te onderscheiden.

Op basis van deze resultaten veronderstelden wij dat een gebrek aan vertrouwen in CAD een beter gebruik van de CAD resultaten beperkt. In **hoofdstuk 7** wordt onderzocht of training in het gebruik van CAD tot een betere detectie van long noduli zou leiden. De sensitiviteit van CAD was 59% en dit was iets lager in vergelijking met de sensitiviteit van 65% van de radiologen. Elke waarnemer kreeg individuele feedback na het beoordelen van een subset van foto's zonder en met CAD als second reader. De hypothese was dat deze feedback zou leiden tot een leercurve waardoor uiteindelijk de prestaties van de radiologen met CAD zouden verbeteren. De feedback resulteerde in een betere detectie van noduli zonder dat de CAD resultaten beter geïnterpreteerd werden. Dit gold voor beide manieren waarop CAD gebruikt kan worden: 1) met de mogelijkheid om markeringen te verwijderen die tijdens het beoordelen zonder CAD zijn gemaakt en 2) met alleen de mogelijkheid om extra markeringen te plaatsen tijdens het gebruik van CAD (add-on). In totaal werden door alle radiologen 32 terecht positieve CAD markeringen verworpen zonder verschil in resultaat tussen de eerste twee en laatste twee beoordelingssessies (16 vs. 16). Van de verworpen terecht positieve CAD markeringen was 78% voor slecht en zeer slecht detecteerbare noduli. Deze resultaten laten zien dat deze manier van training de moeite die gebruikers van CAD hebben om de terecht positieve van fout positieve CAD markeringen te onderscheiden niet verbetert.

List of publications:

Observer Training for Computer-aided Detection of Pulmonary Nodules in Chest Radiography.

De Boo DW, van Hoorn F, van Schuppen J, Schijf L, Scheerder MJ, Freling NJ, Mets O, Weber M, Schaefer-Prokop CM

Accepted for publication in European Radiology

Gray-scale Reversal for the Detection of Pulmonary Nodules on a PACS workstation

De Boo DW, Uffmann M, Bipat S, Boorsma EFA, Scheerder MJ, Weber M, Schaefer-Prokop CM.

AJR American Journal of Roentgenology 2011; 197:1096-1100

Computer-Aided Detection of Small Pulmonary Nodules in Chest Radiographs: An Observer Study

De Boo DW, Uffmann M, Weber M, Bipat S, Boorsma EFA, Scheerder MJ, Freling N, Schaefer-Prokop CM

Academic Radiology 2011; 197:1507-14

Computed Radiography versus Mobile Direct Radiography for Bedside Chest Radiographs: Impact of Dose on Image Quality and Reader Performance

De Boo DW Weber M, Deurloo EE, Streekstra GJ, Freling N, Dongelmans DA, Schaefer-Prokop CM.

Clinical Radiology 2011; 66:826-32

Computer-aided Detection of Lung Cancer on Chest Radiographs: Effect on Observer Performance.

de Hoop B, **De Boo DW** Gietema HA, van Hoorn F, Mearadji B, Schijf L, van Ginneken B, Prokop M, Schaefer-Prokop C.

Radiology. 2010; 257:532-40

Computer-aided detection (CAD) of lung nodules and small tumours on chest radiographs (a review).

De Boo DW, Prokop M, Uffmann M, van Ginneken B, Schaefer-Prokop CM.

European Journal of Radiology 2009; 72:218-25

DR and CR: recent advances in technology (a review).

Schaefer-Prokop C, **De Boo DW**, Uffmann M, Prokop M.

European Journal of Radiology 2009; 72:194-201

Dankwoord

“If you put your mind to it, you can accomplish anything”. Echter het tot stand komen van dit proefschrift zou volstrekt onmogelijk zijn geweest zonder de hulp van velen. Iedereen die op enigerlei wijze heeft bijgedragen aan het voltooien van dit proefschrift wil ik onwijs bedanken. Een aantal echter in het bijzonder.

Allereerst mijn promotor **Prof. dr. J.S. Laméris**. Beste Han, je eerste woorden tijdens mijn sollicitatie waren: “Nederhorst den Berg, daar heb ik veel foto’s van.” Jij bleek al jaren ‘s ochtends Nederhorst vast te leggen vanaf de andere kant van de Vecht; om vervolgens, net als ik, op de fiets naar het AMC te rijden. Je rust en vertrouwen hebben me gesterkt in het starten, maar zeker ook in het voltooien van dit proefschrift, waarvoor dank. Ook dank ik je voor de vele tips en trics in de interventieradiologie die ik momenteel van je mee krijg.

Uiteraard wil ik mijn copromotor **Dr. C.M. Schaefer-Prokop** bedanken. Lieve Cornelia, je tomeloze enthousiasme voor de thoraxradiologie maakte dat ik een klein project kon uitbouwen tot dit proefschrift. Een weg met meerdere ups and downs. Die laatste werden door jou vaak gemakkelijk weg gerelativeerd: that does not change our lives! Ik heb genoten van onze discussies over wetenschap, maar ook het leven buiten de radiologie. Altijd ging dit in die heerlijke mengelmoes van Duits-Nederlands-Engels. Ons gezamenlijke doel, ik promoveren en jij inaugureren, is voor de helft bereikt. Ik hoop dat er voor jou nog een leerstoel staat, want die verdien je!

Mijn opleiders **Dr. O.M. van Delden** en **Dr. A.M. Spijkerboer**. Beste Otto en Anje, jullie zijn net als ik gepromoveerd tijdens je opleiding. Hierdoor begrepen jullie dat ik soms meer met het ‘één’ dan met het ‘ander’ bezig was. Mede dankzij jullie vertrouwen en begrip is het gelukt. Otto, ik vind het mooi dat je tevens plaats wilt nemen in mijn commissie.

De overige **commissieleden**: Prof. dr. E.H.D. Bel, Prof. dr. G.J. den Heeten, Prof. dr. M.B. van Herk, Prof. dr. ir. N. Karssemeijer en Prof. dr. J.A. Verschakelen wil ik hartelijk bedanken voor het beoordelen van mijn proefschrift en het plaatsnemen in mijn commissie.

Zonder mijn medeauteurs / lezers geen data en dus geen artikelen. Dank voor de tijd en moeite van het zoeken naar al die nodi. **Michael Weber** en **Shandra Bipat** dank ik in het bijzonder. Data genereren is één, maar daar relevante conclusies uit trekken is twee. Zonder jullie hulp zou dit nooit gelukt zijn.

Alle **assistenten radiologie** dank ik niet zozeer voor hun bijdrage aan dit proefschrift, maar zeker wel voor mijn mooie assistententijd. Ik heb genoten van de heerlijke tijd die we als AIOS hebben gehad, zowel in het ziekenhuis, maar meer nog tijdens de vele borrels, (verkleed)feesten en weekendjes weg.

Mijn paranimfen **Sjoerd Frantzen**, **Rende Oudhof** en **Bas Polle**.

Sjoerd, sinds ons eerste jaar geneeskunde hebben we een mooie vriendschap opgebouwd. Het blindelings vertrouwen in elkaar heeft tot mooie avonturen geleid. Een voorstel van de één, kon een vast antwoord van de ander verwachten: is goed! Zo kom je nog eens ergens en maak je veel mee (of juist niet...). Tijdens mijn verdediging zit je "down under", maar gelukkig ben je er in gedachten wel bij.

Rende: dat is een barman die ook wel aardig kan tennissen.

Zo begonnen we bij Chip & Charge in team 3. Dat we stijf laatste werden maakte ons niet uit, want de basis voor een fijne vriendschap werd hier gelegd. Mooie dubbels hebben we gespeeld en deze werden vaak afgesloten in de Gieter. Naast lief, hebben we door het overlijden van Marlous ook het nodige leed gedeeld. Dat jij aan mijn zijde staat tijdens mijn verdediging is dan ook niet meer dan logisch.

Bas, of beter gezegd Polle, wij kennen elkaar van het onderwijs geven op de snijzalen en de daarbij horende donderdagmiddagen in de Epstein-Bar. Als twee enorme betweters hielden wij elkaar scherp, zowel op als buiten de snijzaal. Zo kon ik jou nog eens mooi de vrouwelijke cyclus uitleggen. Jij ging mij voor naar het weledelzeergeleerde en ik was als paranimf daar getuige van. Ik vind het mooi dat jij erbij bent als ik ook "over ga".

In willekeurige volgorde dank ik iedereen die mee was naar de **Alpe de Lous** (2009) en **Alpe d'Huez** (2010): Helmy, Peter, Steven, Cilia, Daan, Hans, Willemijn, Luuk, Bart, Menno, Rene, Marije, Marloes, Maarten, Rende, Anke, Bob, Nina, Ralf, Leonie, Koen, Eline, Irene, Catelijne, Jeroen, Jasper, Lotte, Rolf, Rik, Jeremy en Marieke. Samen op de Alpe d'Huez. Jullie weten wat dit voor mij betekend heeft!

De Bergers aka Melle, Thijs, Vincent, Mark en Vincent. Van ongein uithalen op de basisschool, gigantische feesten tijdens onze puberteit tot volwassen kerels, onze vriendschap is alleen maar beter geworden. Gabbers voor het leven dus, ook al zien we er niet meer zo uit.

Tot slot mijn **ouders**. Lieve pap en mam, de belangrijkste daad voor het tot stand komen van dit boekje leverden jullie alweer ruim 33 jaar geleden. Jullie hebben me in de daarop volgende jaren alle vrijheid, liefde en vertrouwen gegeven die nodig waren om te worden wat ik nu ben. Ik hou van jullie!

Beter een oorlog dan gewapende vrede, Beter op zoek gaan dan altijd gewacht, Beter gevallen dan nooit gesprongen, En beter de liefde verloren dan nooit lieggedad



**Diederick
Willem
De Boo**

10 februari 1979, drie pinken, Nederhorst den Berg

Opleiding tot radioloog AMC (2006-2011) professor J.S. Laméris en dr. O.M. van Delden. Fellowship interventieradiologie: AMC (2011-2012) professor J.A. Reekers, Straatsburg (2012) professor A. Gangi en Melbourne (2012-2013) professor K.R. Thomson.

Geneeskunde aan de Universiteit van Amsterdam, onderwijzen van anatomie, Chip & Charge, co-schap cardiovasculair chirurgie en TExaShear Institute Houston

