COMPUTED TOMOGRAPHY

READING STRATEGIES IN

LUNG CANCER SCREENING

Arjun Nair

Department of Radiology

Royal Brompton Hospital

London

A thesis submitted for the Degree of

Doctor of Medicine

The University of Edinburgh

2014

DECLARATION

This is to certify that the work contained within has been composed by me and is entirely my own work. No part of this thesis has been submitted for any other degree or professional qualification.

Arjun Nair

ABSTRACT

Numerous studies investigating low-dose computed tomography (LDCT) as a screening tool for lung cancer have either recently been completed or are ongoing. However, the optimum strategy for detecting nodules in a CT screening programme is still unknown. To date screening trials have varied substantially in their reading strategies. Each of these strategies may lead to different rates of true and false positive detection. The ideal strategy would maximise detection of lung cancers while minimising unnecessary costly and potentially harmful investigations. The type of strategy chosen also has significant implications for the number of radiologists required for a screening programme, and their workload.

The investigations contained in this thesis are aimed at identifying an optimal and pragmatic reading strategy for LDCT screening.

First, the potential role of radiographers as readers for LDCT screening was investigated. Following training and an assessment of continuous feedback learning, the performance of radiographers reading LDCT screening examinations was prospectively compared against radiologists; the performance of these radiographers was comparable to that of radiologists in the published literature, but inferior to that of radiologists reading the same scans. However, using radiographers as concurrent readers helped to improve radiologists' sensitivities in nodule detection, with an increase in false positive detections that is still below that reported for computer-aided detection (CAD) systems.

When evaluated as first readers against a clinically-approved CAD system, radiographers showed that they could achieve sensitivities comparable to or exceeding that of CAD, with a lower number of average false positive detections for the majority.

The impact of double- and triple-reading strategies using different methods of arbitration for discordant findings was compared. Using more than one reader did not invariably improve pulmonary nodule detection accuracy for experienced thoracic radiologists, and resulted in increased false positive detections when double-reading with independent arbitration or triple-reading were used.

This effort is dedicated with all my love and affection to my amazing wife Gopika.

Thank you so much for spurring me on.

Whenever the going got tough, I only had to think of you...

...and the rest was easy.

ACKNOWLEDGEMENTS

Several individuals have been instrumental in helping me realise this thesis. First and foremost, I am forever indebted to my supervisors, Professor David Hansell and Dr Anand Devaraj, for their support, encouragement, inspiration, and mostly, for their leadership by example. In particular, I would like to thank Professor Hansell for the privilege of serving as his thoracic imaging fellow at the Royal Brompton Hospital. I would also like to thank Dr John Murchison for helping me complete this thesis, in his capacity as my educational advisor and liaison with the University of Edinburgh.

Furthermore, I thank Professor John Field, Dr David Baldwin, and Professor Athol Wells for all their advice and approachability. I must also thank Dr Nick Screaton and Dr John Holemans for their help during the course of the UKLS pilot trial, and all the staff of the NELSON trial who assisted me, especially Dr Mathias Prokop, Dr Bram van Ginneken, and Dr Rozemarijn Vleigenthart. I am also grateful to Dr Michael Rubens, Dr Simon Padley, Dr Dennis Carr, and Dr Ed Nicol for their help. In addition, I am deeply indebted to all the radiographers who took part in the investigations contained in this thesis. Of course, my sincerest thanks go to all the participants in the UKLS pilot who made this thesis possible.

I reserve special thanks for Dr and Mrs Sujal Desai, who have tirelessly encouraged me in my medical and radiology career- none of this would have been possible without them. Last but not least, I thank my family, especially my parents, parents-in-law, sister-in-law, and my cousin Sanjay, for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 4

Chapter 5

Chapter 6

Chapter 7

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

LIST OF ABBREVIATIONS

CT- and dose-related parameters

AEC	Automatic Exposure Control
ALARA	As Low As Reasonably Achievable
CT	Computed Tomography
CTDIvol	CT dose index over the volume scanned
DICOM	Digital Imaging and Communications
DICOM SR	Digital Imaging and Communications Structured Report
DLP	Dose-Length Product
HU	Hounsfield Units
kVp	Kilovolt potential
LDCT	Low-dose Computed Tomography
mA	milliamperes
mAs	Milliamperes-second (tube current-time product)
MDCT	Multidetector Computed Tomography
mGy-cm	milligray-centimetres
mSv	millisieverts
PACS	Picture Archiving and Communications

Image reconstruction and nodule reading terms

CAD	Computer-aided detection/diagnosis
MinIP	Minimum Intensity Projection
MIP	Maximum Intensity Projection
MPR	Multiplanar Reconstruction
ROI	Region of interest
VDT	Volume Doubling Time
IPLN	Intrapulmonary lymph node
NCN	Non-calcified nodule

Other imaging modalities

FDG	2-[^{18}F]-Fluoro2-deoxy-D-glucose
FDG-PET	2-[^{18}F]-Fluoro2-deoxy-D-glucose Positron Emission Tomography
MR	Magnetic Resonance (Imaging)
PET-CT	Positron Emission Tomography-Computed Tomography

Screening trials

COSMOS	Continuous Observation of Smoking Subjects
DANTE	Detection And Screening of early lung cancer by Novel imaging Technology and molecular Essays
DLCST	Danish Lung Cancer Screening Trial
ELCAP	Early Lung Cancer Action Project
HIP	Health Insurance Plan
I-ELCAP	International Early Lung Cancer Action Project
ITALUNG	Italian Lung Trial
JHLP	Johns Hopkins Lung Project
MILD	Multicentric Italian Lung Detection trial
MLP	Mayo Lung Project
MSKLP	Memorial Sloan Kettering Lung Project
NELSON	Nederlands-Leuvens Longkanker Screenings Onderzoek
NHS BSP	National Health Service Breast Screening Programme
NLST	National Lung Screening Trial
PLCO	Prostate, Lung, Ovarian and Colon trial
UKLS	UK Lung Cancer Screening Trial

Organisations

ACCP	American College of Chest Physicians
ATS	American Thoracic Society
ERS	European Respiratory Society
FDA	US Food and Drug Administration
IASLC	International Association for the Study of Lung Cancer
LIDC	Lung Image Database Consortium
UKCCR	UK Cancer Coordinating Committee for Research
WHO	World Health Organisation

Pathological terms

AAH	Atypical Adenomatous Hyperplasia
AIS	Adenocarcinoma-in-situ
BAC	Bronchioloalveolar Carcinoma
LPA	Lepidic Predominant Adenocarcinoma
MIA	Minimally Invasive Adenocarcinoma
NSCLC	Non-Small Cell Lung Cancer
TNM	Tumour, Node, Metastases staging

Diagnostic techniques

EBUS	Endobronchial ultrasound
FNA	Fine Needle Aspiration
TBNA	Transbronchial Needle Aspiration
VATS	Video-assisted Thorascopic Surgery

Lung function parameters

DLco	Diffusing capacity of carbon monoxide
FEV1	Forced expiratory volume in one second

Therapeutic procedures

MWA	Microwave ablation
PCT	Percutaneous cryoablation therapy
RFA	Radiofrequency ablation
SBRT	Stereotactic body radiation therapy

Statistical methods/ reading methods

CR	Concurrent Reading
FN	False Negative
FP	False Positive
ICER	Incremental Cost-Effectiveness Ratio
IR	Independent Reading
κ	Kappa coefficient
PPV	Positive Predictive Value
QALY	Quality-Adjusted Life Year
ROC	Receiver-Operator Characteristic
TN	True Negative
TP	True Positive

Miscellaneous

RCT	Randomised control trial

CHAPTER 1: INTRODUCTION

"On one point, however, there is nearly complete consensus of opinion, and that is that primary malignant neoplasms of the lungs are among the rarest forms of disease".

-Isaac Adler, Professor Emeritus at the New York Polyclinic,
"Primary malignant growths of the lungs and bronchi: a pathological and clinical study", Longmans, Green and Co., New York, 1912.

1.1 Lung cancer

1.1.1 Prevalence and incidence of lung cancer

Tumours of the lung were rare at the beginning of the 20th century. The renowned physician Sir William Osler devoted only two pages to "New Growths in the Lung" in the second edition of his "Principles and Practice of Medicine" in 1895, stating that "primary growths are rare" [1]. In the first comprehensive monograph on the topic, published in 1912, Isaac Adler, Professor Emeritus at the New York Polyclinic, found only 374 verifiable cases of lung cancer in the published literature worldwide [2]. However, the incidence of lung cancer in the Western world accelerated through the mid-20th century, almost exclusively due to an increase in smoking prevalence during the first half of that century, itself driven by the expansion of commercial cigarette production [3]. Lung cancer is now the most common cancer worldwide, with an estimated 1.61 million new cases, accounting for approximately 12.7% of all new cancer cases globally in 2008, and remains the most common cause of death from cancer for both men and women internationally [4].

In the UK, lung cancer is the second most common cancer, but remains the commonest cause of cancer death. In 2010, there were 34,859 deaths from lung cancer in the UK, accounting for 24% of all male cancer deaths and 21% of all female cancer deaths [5]. Survival in the UK is poorer than many countries in Europe [6], with only 8.2% of men and 9.3% of women alive at five years in England [7]. The late presentation of lung cancer in the UK [8, 9] is an important reason for this poor survival: according to the National Lung Cancer Audit Report 2005, 43% of patients with non-small cell lung cancer in England and Wales presented with stage IIIB or greater, and would have been unsuitable for surgical resection [10].

The peak incidence and mortality trends closely reflect the trends in maximum exposure to cigarette smoke over time. In the UK, smoking was at its most prevalent in the generation of men born around 1910-1911, and in women born around 1925-1930 [11]. Consequently, the overall incidence of lung cancer in males aged 60-69 rose to a peak in the late 1970s, with a peak of 440 cases per 100,000 men in 1977, while a delayed peak of 880 cases per 100,000 men older than 80 years of age was seen in 1985. Since then, however, the incidence of lung cancer in men in the UK has been declining, mirroring global trends: between 2009 and 2011, the European age-standardised rates for males aged 60-69 and older than 80 were 193 and 574 per 100,000 men, respectively [12].

Although the incidence of lung cancer in women is still lower than in men, there has been a worrying rise in the incidence amongst women across all age groups since the 1970s. In England, for example, the age-standardised incidence rate of lung cancer in women has increased by 9.3% in 10 years, from 34.3 per 100,000 population in 1999 to 37.5 per 100,000 population in 2009 [13]. Overall, lung cancer

rates increased by 62% between 1975 and 2011 among women aged 60-69 [12]. Much of the rise is again attributable to smoking prevalence in women amongst different birth cohorts. There is also evidence suggesting that lung cancer incidence in women who are lifelong never-smokers is higher than in men [14, 15]. However, a recent analysis concluded that while never-smoking women aged 50-54 and 55-59 years had a statistically significantly higher incidence of lung cancer than never-smoking men, the overall age-standardised incidence rates of lung cancer in never-smoking women of European descent are similar to, and not higher, than those in never-smoking men [16].

1.1.2 Risk factors for lung cancer

Risk factors for lung cancer include smoking, occupational or industrial exposure to carcinogens (including asbestos exposure), family history, and previous history of malignancy, especially treatment for previous Hodgkin's lymphoma.

Smoking is by far the greatest risk factor for the development of lung cancer. Cigarette smoke is known to contain over 60 carcinogens, such as polycyclic aromatic hydrocarbons, *N*-nitrosamines, aromatic amines, aldehydes, volatile organic hydrocarbons, and metals [17]. The now irrefutable causative link between smoking and lung cancer was suspected as early as the 1920s [3], but it was in 1950 that five case-control studies firmly established an association between lung cancer and smoking [18-22]. In a recent analysis, Parkin concluded that in 2010, 85% of lung cancers in men and 80% in women in the UK were attributable to smoking [23]. When environmental exposure to cigarette smoke is included, these proportions

increased to 87% and 84% for men and women, respectively. However, the contribution of tobacco exposure to global lung cancer incidence, especially in women, is thought to be decreasing; for example, up to 53% of lung cancer in women worldwide is now not attributable to tobacco use [24]. Unfortunately, such analyses of temporal changes in the proportional contribution of tobacco smoke exposure to lung cancer incidence have been hindered by the limited reliability of smoking information from population-based registries.

Intensity and duration of smoking are positively correlated with lung cancer risk and mortality. In one study, the cumulative risk of death from lung cancer by age 75 among current smokers was about 16%, increasing to 24% for those smoking at least 25 cigarettes per day [25]. Current smokers are 14.7 times more likely to die of lung cancer than lifelong non-smokers, and 3.6 times compared to former smokers [26].

Intensity and duration can be combined into a single measure of smoking activity by calculating pack-years (the number of cigarette packs smoked per day, multiplied by the number of years of smoking). However, smoking at a lower intensity for longer is more harmful than smoking at a higher intensity for a shorter period [27], and using pack-years as a measure of smoking activity can therefore mask the more important effect of duration.

1.1.3 Pathological classification of lung cancer

Histologically, lung cancer can broadly be divided into non-small cell and small cell cancer. Non-small cell lung cancer (NSCLC) is by far the most common

type worldwide, accounting for 85.3% of histologically confirmed lung cancer cases in the USA between 2004-2008 [28]. The most common subtypes of NSCLC are squamous cell carcinoma and adenocarcinoma, accounting for approximately 32% and 26% of all cancers in England and Wales in the period between 2006-2008, respectively [29]. However, there has been a change in the relative prevalence of these histological subtypes over the past three decades in particular. The prevalence of adenocarcinoma has been rising, such that it is now the most common subtype of lung cancer in the USA [28], and its incidence is rising in the UK and Europe [30]. This change has been predominantly attributed to smoking low-tar (i.e. filter) and low-nicotine cigarettes, resulting in smokers smoking more per day and inhaling more deeply, with a consequent increase in the proportions of carcinogenic *N*-nitrosamines within the inhaled smoke [31]. The tendency to inhale more deeply on such cigarettes also exposes the peripheral lung to higher doses of the carcinogens within cigarette smoke, with the effect that adenocarcinomas occur primarily in the periphery of the lung [31].

The first standardised classification of lung cancer by histological subtype was published by the World Health Organization (WHO) in 1967 [32]. There have been three further editions since then, with the most recent histological classification published in 2004 [33]. The classification has considerably evolved through the four iterations. The third edition, released in 1999, recognised pre-invasive lesions such as atypical adenomatous hyperplasia (AAH) and tumours of mixed histological subtype (Figure 1.1) [34]. A revised classification of adenocarcinoma was recently recommended by a joint committee consisting of members from the International Association for the Study of Lung Cancer (IASLC), American Thoracic Society

(ATS) and European Respiratory Society (ERS) in 2011 [35]. Chief among its recommendations was the introduction of new terms to replace the previously used terms "bronchioloalveolar carcinoma" (BAC) and "mixed type adenocarcinoma", as these latter categories had encompassed groups of tumours within the adenocarcinoma spectrum that were quite diverse in radiological and histological appearance, and in their natural history (see also section 1.5.3.3).

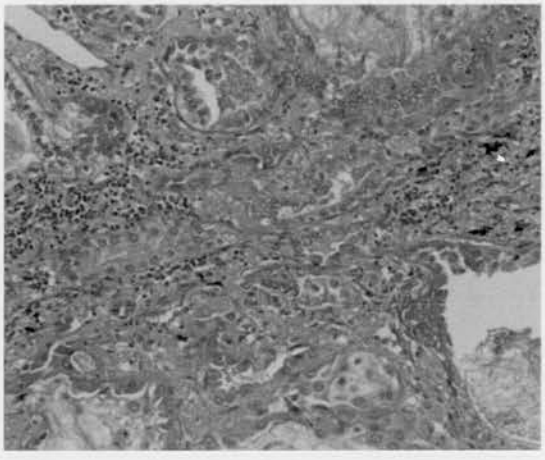

Figure 1.1. Adenocarcinoma formerly termed mixed subtype. Hematoxylin and eosin stain at X20 magnification demonstrates an invasive component comprising solid nests and acinar glands infiltrating fibrous stroma. The tumour also has non-invasive lepidic components. The term "mixed subtype" adenocarcinoma has now been discarded. (From reference [34].)

1.1.4 Clinical presentation of lung cancer

Cough is the most common presenting symptom of lung cancer, but occurs variably in 8-75% of patients [36]. Other symptoms include dyspnoea, chest pain, and haemoptysis. However, patients may commonly present with signs and symptoms that are not related to the primary tumour. Such symptoms may be constitutional (such as anorexia and weight loss), or secondary to paraneoplastic

manifestations or metastatic disease. Up to 43% of patients present late, with unresectable disease [10, 37]. On the other hand, 6-13% of patients may be asymptomatic at diagnosis [38-40].

The clinical presentation can also be related to the histological subtype [41]. For instance, small cell carcinoma is an aggressive tumour with early metastasis, while peripheral adenocarcinomas may grow silently and not present until they are either discovered incidentally on chest radiograph or computed tomography (CT), or if they present with local invasion, nodal and/or metastatic disease.

1.1.5 Investigations, staging and prognosis

Investigations in lung cancer are aimed at securing the diagnosis, obtaining staging information, and determining therapeutic options. Staging and diagnosis are performed using radiological imaging techniques or bronchoscopic techniques.

The chest radiograph remains the initial investigation of choice [29]. The chest radiograph may depict: (a) the primary tumour itself, as a solitary pulmonary lesion; (b) lobar or segmental collapse or consolidation, as a consequence of bronchial obstruction and inflammation; or (c) abnormalities related to the presence of local invasion, nodal disease or metastatic disease - for example rib erosion, hilar lymphadenopathy, or multiple pulmonary nodules. However, chest radiography is also unreliable in its detection of lung cancer, because it may not be able to detect small lesions (under 2cm), peripheral lesions may be obscured by bone, and central lesions arising in the trachea, main and lobar bronchi can be hard to detect due to the superimposition of mediastinal structures (Figure 1.2a) [42].

The advent of CT has helped to overcome many of the difficulties encountered in diagnosing lung cancer with the chest radiograph. With its markedly improved contrast and spatial resolution and its ability to resolve superimposed structures, thoracic CT avoids most of the obscuring effects that hamper chest radiography (Figure 1.2b) [43]. CT provides improved detection of lesions which are just a few millimetres in diameter [44], as well as staging information [36]. However, "central" lesions, especially lesions adjacent to the hilum or mediastinum, may still be easily missed on CT as compared to peripheral lesions [45], without the benefit of additional reconstruction techniques such as maximum intensity projections (see section 1.6.2.1).

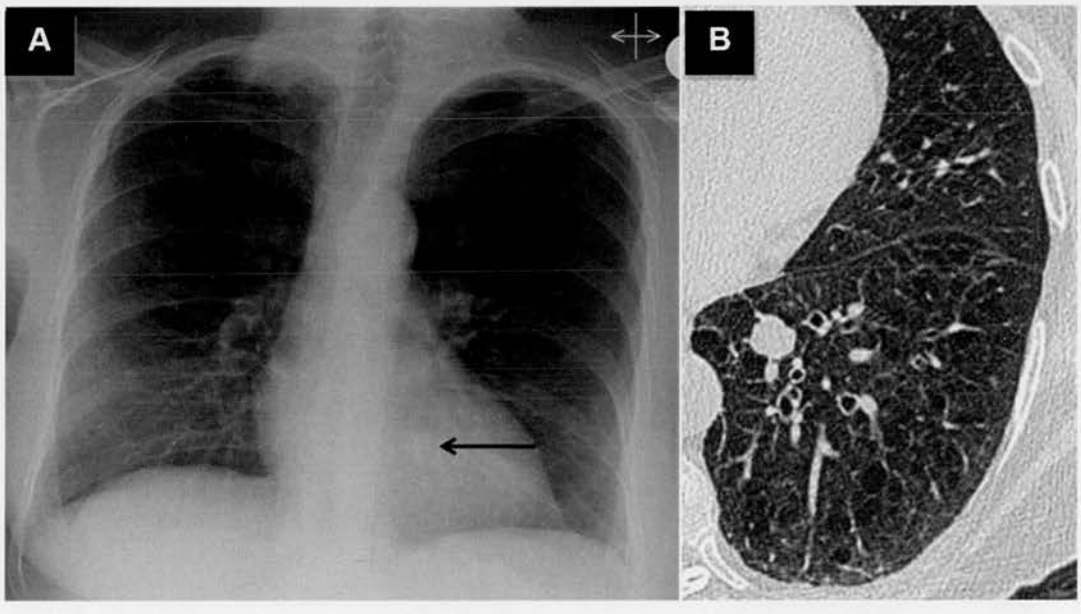

Figure 1.2. Left lower lobe cancer missed on chest radiograph. (A) Posteroanterior chest radiograph demonstrates a nodule in the left retrocardiac region that is only faintly perceptible (arrow). (B) CT section through the left lower lobe demonstrates the nodule clearly.

Fibreoptic bronchoscopy is an invaluable diagnostic tool for visualising and sampling central bronchial tumours. Rigid bronchoscopes have been in use since 1897 [46], but were gradually replaced as a primary means of airway visualisation

with the introduction of the flexible bronchoscope in 1967 [47]. Bronchoscopy enables direct sampling of an endobronchial lesion or indirect sampling by obtaining bronchial washings or brushings. Furthermore, bronchoscopy allows the simultaneous sampling of suspicious lymph nodes, using transbronchial needle aspiration (TBNA) if they are accessible. TBNA may be performed either unguided ("blind") [48] or more recently with ultrasound guidance, using endobronchial ultrasound (EBUS) [49].

2-[^{18}F]-Fluoro2-deoxy-D-glucose (FDG) positron emission tomography (PET) is an imaging tool which identifies malignant neoplasms by taking advantage of their increased glucose uptake and utilisation, which is proportional to their metabolic activity [50, 51]. The relative uptake of FDG can be calculated from a static FDG-PET image by the standardised uptake value (SUV). The SUV of a given tissue is given by:

$$\frac{\text{tracer activity in tissue}}{\text{(injected radiotracer dose/ patient weight)}}$$

where tissue tracer activity is in microcuries per gram, injected radiotracer dose is in millicuries, and patient weight is in kilograms [51]. FDG-PET integrated with CT (PET-CT) is now recommended for the non-invasive evaluation of lymph node involvement and identification of extranodal metastatic disease, if the patient is suitable for potentially curative treatment [29, 52]. Due to its ability to depict intra- and extrathoracic metastases, FDG-PET may help prevent unnecessary thoracotomy in up to 20% of patients previously considered operable [53].

CT-guided percutaneous biopsy can be undertaken to obtain histological tissue from peripheral lesions, provided the lesion is anatomically amenable to

percutaneous biopsy, and the patient has satisfactory lung function. Obtaining sufficient tissue for molecular and genetic typing is becoming increasingly important, with the growing number of molecular targeted therapies available for adenocarcinoma; thus, core biopsy is preferred over fine needle aspiration. Alternatively, diagnosis by video-assisted thoracoscopic surgery (VATS) biopsy, or open lung biopsy, may be performed [54].

As with any malignancy, the staging system used for lung cancer allows the grouping of tumours by their biological behaviour so that the prognostication and tumours of similar tumours can be standardised and compared. The Tumour, Node, Metastasis (TNM) staging system, first proposed by Denoix [55], incorporates different types of staging with different combinations of T, N and M descriptors forming stage groupings with similar prognoses. It was adapted for use in lung cancer staging by Dr Clifton Mountain in 1973 [56], with T descriptors broadly classifiable into size-based and non-size based descriptors. Since then, the TNM staging of lung cancer has undergone six iterations, with the latest (the 7th edition) providing the most changes to the system. In this latest edition, new size sub-divisions have been created, the status of intrapulmonary metastatic nodules have been modified, and the concept of intra- and extrathoracic non-nodal metastatic disease has been introduced, with pleural and pericardial dissemination now representing M, rather than T descriptors (Table 1.1) [57]. These changes have resulted in new stage groupings (Table 1.2) that have addressed the heterogeneity in prognosis seen in previous staging editions [58]. Also, unlike previous editions, the 7th edition TNM descriptors apply to both NSCLC and small cell carcinoma [59]; a

different system had previously been used for small cell carcinoma (limited versus extensive stage).

Most importantly, the latest edition of the TNM staging reinforces the importance of the size of tumour as an indicator of prognosis. Patients with tumours staged pathologically as T1a ($\leq$ 2cm) have a 77% 5-year survival rate, compared to those that are T3 on the basis of size (> 7cm), who have a 35% 5-year survival [60]. Patients with pathological stage IA disease (T1a or T1b, N0 M0) have a 73% 5-year survival, in comparison to a dismal 2% in those with clinical stage IV disease (i.e. M1a or M1b) [58].

Descriptors	Definition
Tumour	
Tx	Primary tumour cannot be assessed, OR primary tumour detected by sputum/ bronchial washings but not visualized on any imaging
T0	No evidence of primary tumour
Tis	Carcinoma in situ
T1	Tumour ≤ 3cm in greatest dimension, surrounded by lung or visceral pleura, without invasion more proximal than the lobar bronchus[1]
T1a	Tumour ≤ 2cm in greatest dimension
T1b	Tumour > 2cm but ≤ 3cm in greatest dimension
T2	Tumour > 3cm but ≤ 7cm OR tumour with any of the following features[2]: involves main bronchus, ≥ 2cm distal to the carina; invades visceral pleura; associated with atelectasis/obstructive pneumonitis that extends to the hilar region without involving entire lung
T2a	Tumour > 3cm but ≤ 5cm in greatest dimension
T2b	Tumour > 5cm but ≤ 7cm in greatest dimension
T3	Tumour > 7cm OR tumour that directly invades any of the following: chest wall (including superior sulcus tumours), diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium; OR tumour in the main bronchus < 2cm distal to the carina but without involvement of the carina; OR associated atelectasis or obstructive pneumonitis of the entire lung OR separate tumour nodule(s) in the same lobe
T4	Tumour of any size that invades any of the following: mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina; OR separate tumour nodule(s) in a different ipsilateral lobe
Node	
Nx	Regional lymph nodes cannot be assessed
N1	No regional lymph node metastasis Metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension
N2	Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)
N3	Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s)
Metastasis	
Mx	Distant metastasis cannot be assessed
M0	No distant metastasis
M1	Distant metastasis
M1a	Separate tumour nodule(s) in a contralateral lobe; OR tumour with pleural nodules or malignant pleural (or pericardial) effusion
M1b	Distant metastasis

Table 1.1. The 7th edition of the TNM staging of lung cancer [57].
[1]Includes any superficial spreading tumour (of any size) with its invasive component limited to the bronchial wall, but which may extend proximally to the main bronchus.
[2]T2 tumours with these features are classified T2a if ≤ 5cm, or if size cannot be determined, and T2b if > 5cm but ≤ 7cm.

		7th Edition N descriptor			
		N0	N1	N2	N3
7th edition T descriptor	T1a	IA	IIA	IIIA	IIIB
	T1b	IA	IIA	IIIA	IIIB
	T2a	IB	IIA	IIIA	IIIB
	T2b	IIA	IIB	IIIA	IIIB
	T3	IIB	IIIA	IIIA	IIIB
	T4	IIIA	IIIA	IIIB	IIIB
7th edition M descriptor	M1a	IV	IV	IV	IV
	M1b	IV	IV	IV	IV

Table 1.2. Stage groupings based on the tumour (T), node (N) and metastasis (M) descriptors of the 7th edition of the TNM staging of lung cancer [57].

1.1.6 Therapeutic strategies

Therapeutic advances relevant to small tumours and early lung cancer that may incidentally be discovered on screening asymptomatic high-risk individuals are discussed, as a detailed discussion of all therapeutic strategies for lung cancer is beyond the scope of this thesis.

Surgery with curative intent remains the standard of care for all patients with resectable lung cancer who are surgical candidates [29, 61]. In general, these are patients who have Stage I or II NSCLC. Surgery may also be contemplated for early stage (stage I or II) small cell lung carcinoma [62], but unfortunately such early presentation is rare.

Size of tumour in NSCLC is an independent prognostic factor for survival in patients undergoing surgical resection [63-68]. For example, in a retrospective

analysis of 6644 patients in the Japanese Lung Cancer Registry, Asamura et al. found that patients with pathologically staged node-negative, metastasis-negative tumours < 2cm had a 5-year survival of 83.7%, as opposed to tumours 2.1-3cm in size (76.0%) [63]. Indeed, the improved prognosis of tumours smaller than 2cm was a significant impetus for the new size subdivisions in the latest TNM staging edition [60].

The surgical method of choice remains lobectomy for patients who are able to tolerate it. However, lung-sparing operations, broadly termed sublobar resections, are now an option for patients with suboptimal fitness. Sublobar resections include wedge resections and segmentectomy, and can be performed via open thoracotomy or VATS. Wedge resections can be used with smaller peripheral tumours and N0 disease, but do not allow the intra-operative sampling of N1 nodes [69]. Interest in sublobar resection has been prompted by the more favourable prognosis of tumours smaller than 2cm. For instance, a recent analysis by Carr et al. has suggested that patients with (7^{th} edition TNM) stage IA who underwent segmentectomy had similar recurrence rates, mortality, and 5-year cancer-specific survival to those at a similar stage who underwent lobectomy, despite the fact that the segmentectomy cohort was older and had a significantly lower mean forced expiratory volume in 1 second [65].

There is an increasing array of non-surgical interventions available for patients who are either unfit for or have declined surgery. The emergence of stereotactic body radiation therapy (SBRT) has made radiotherapy with curative intent a viable option for such patients. SBRT is a technique in which high doses of radiation using numerous small, highly focussed, and accurate radiation beams are delivered in only 1 to 5 treatments over 1 to 2 weeks [70]. An evaluation of 59 patients by the Radiation Therapy Oncology Group, 44 of whom had T1 tumours

(i.e. up to 3cm) based on the 6th TNM edition, had revealed that SBRT could deliver 3-year disease-free and overall survival rates of 48.3% and 55.8% respectively [71]. A recent comparison of SBRT and wedge resection for stage I NSCLC showed that although wedge resection resulted in improved survival at 30 months, both modalities showed comparable cause-specific survival, and SBRT reduced the risk of local recurrence [72].

Platinum-based chemotherapy, as an adjuvant to surgical resection for patients with stage II NSCLC where there is N1 node involvement, is of proven benefit [73]. The role of adjuvant chemotherapy in patients with TNM 6th edition stage IB disease has remained controversial. An increase in disease-free survival may be seen in patients with tumours 4cm or greater, but no statistically significant survival difference at 74 months has been observed in patients receiving chemotherapy, as compared to those who have not [74]. Furthermore, patients who were classified as stage IB in the 6th edition of TNM staging have now been upstaged to later stages (stages IIA or IIB) in the 7th edition; the application of chemotherapy regimens based on the prior editions to patients who are stage IB according to the new edition is thus unproven, and, outside of clinical trials, such chemotherapy is no longer recommended [73].

Percutaneous ablative therapy with radiofrequency ablation (RFA), microwave ablation (MWA) and percutaneous cryoablation therapy (PCT) for small lung tumours have also been considered as options for patients unsuitable for surgery. RFA (Figure 1.3) and MWA are heat-based methods that achieve cell death through thermocoagulation, while PCT causes cell death by ice crystal formation in the target tissue. In general, technical success of ablation is more likely with tumours

that are 3cm or smaller in diameter. An analysis of 64 patients with clinical stage I NSCLC showed that while 3-year cancer-specific survival for a cohort of surgically unfit patients was best with sublobar resection (90.6%), a comparable survival benefit was still seen with RFA (87.5%) and PCT (90.2%) [75]. However, long-term data from larger populations regarding these modalities is not yet available. Given the larger and favourable evidence base for SBRT at present, percutaneous ablative therapies are not currently recommended as first choice non-surgical interventions, but may still be considered in patients who are not candidates for SBRT or sublobar resection [76].

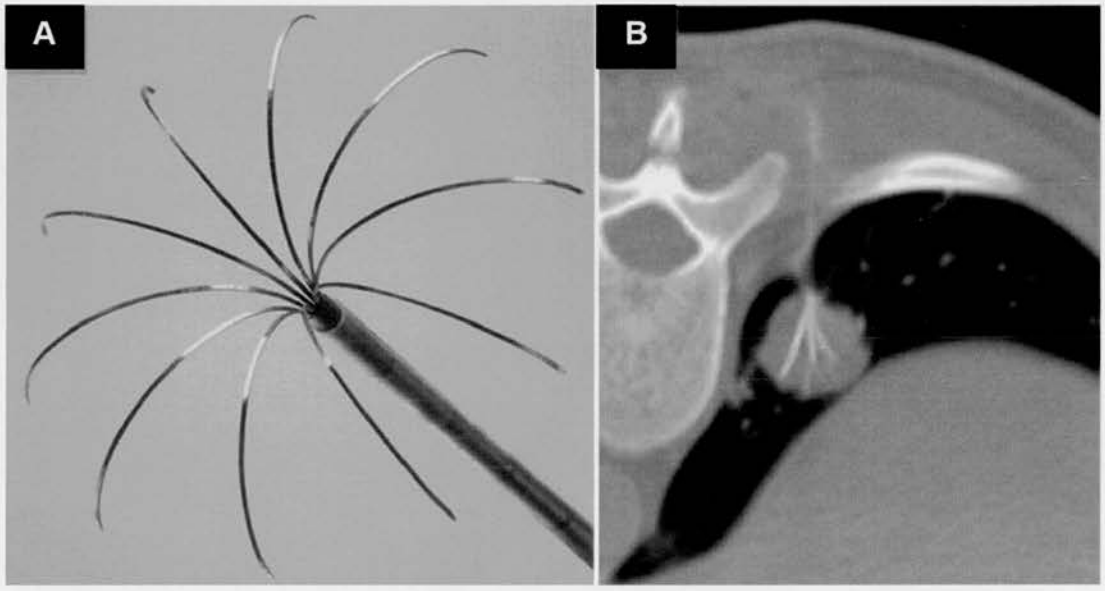

Figure 1.3. Radiofrequency ablation (RFA). (A) Radiofrequency ablation probe, through which radio waves are emitted. (B) CT section demonstrating the probe within a right lower lobe tumour during an RFA procedure. Images courtesy of Dr Sisa Grubnic, St George's Hospital, London.

1.1.7 Physiological considerations

Physiological evaluation is crucial to determining safety to undergo both diagnostic procedures and therapy. The forced expiratory volume in one second

(FEV1) and the diffusing capacity of the lung with carbon monoxide (DLco) are both important predictors of postoperative morbidity and death [77, 78]. However, FEV1 and DLco correlate poorly with each other [79], and a reduced DLco can predict morbidity in patients undergoing lobectomy and pneumonectomy, even in the absence of obstructive lung disease [80]. For these reasons, current guidelines recommend that both the predicted postoperative FEV1 and DLco should be calculated in patients being considered for lung cancer resection [81, 82].

1.2 Principles and practice of population-based screening for cancer

1.2.1 General principles

In 1951, the United States Commission on Chronic Illness defined screening as "the presumptive identification of unrecognised disease or defects by means of tests, examinations, or other procedures that can be applied rapidly" [83]. This definition recognises two components to screening:

1) A **modality** that can be used to identify an abnormality (or **biomarker**) that may signify pathology; and
2) The **speed** (and by implication, ease) with which that modality can be applied.

Screening is thus aimed at either the asymptomatic individual to detect a preclinical phase of the target condition, or the symptomatic individual in whom the condition has not yet been recognised. Screening can instigate diagnosis, but is

distinct from diagnosis [84], because the abnormality detected is not in itself diagnostic of the target condition.

In a seminal publication in 1968, Wilson and Jungner set forth 10 key principles that could be considered "guides to planning case-finding" (*sic*), that were applicable to screening at any level (Table 1.3) [85]. In addition, Wilson and Jungner elaborated some practical aspects of screening programmes, including data collection, handling and storage. In essence, these principles define criteria for the disease, test, treatment, case-finding and cost-effectiveness in an ethical framework, such that the Hippocratic principle of *primum non nocere* (first, do no harm), and the *prima facie* principle of justice - in this case, distributive justice as applied to limited resources that should be used fairly [86] - are adhered to.

Modern modifications to Wilson and Junger's criteria have been proposed [87, 88]. Andermann et al. summarised some of the modifications that have been suggested (Table 1.4). In the main, the updated criteria suggest that a target population be explicitly defined, and the aims, evaluation, and quality assurance methods of a screening programme be defined at the outset. Wilson and Jungner discussed mass screening (unselected screening of the whole population) as opposed to selective screening of high-risk groups, but did not make defining a target population a guiding principle.

1. The condition to be screened should be an important health problem.
2. An acceptable treatment should exist.
3. Facilities for diagnosis and treatment should be available.
4. There should be a recognisable latent or early symptomatic stage.
5. A suitable test or examination should exist.
6. The test should be acceptable to the population.
7. The natural history of the condition, including the evolution from latent to declared disease, should be adequately understood.
8. There should be agreement on whom to treat.
9. The costs of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole.
10. Case-finding should be a continuing process and not a once-off endeavour.

Table 1.3. Key principles of screening for disease, as described by Wilson and Junger [85].

1. The screening programme should respond to a recognized need.
2. The objectives of screening should be defined at the outset.
3. There should be a defined target population.
4. There should be scientific evidence of screening programme effectiveness.
5. The programme should integrate education, testing, clinical services and programme management.
6. There should be quality assurance, with mechanisms to minimize potential risks of screening.
7. The programme should ensure informed choice, confidentiality and respect for autonomy.
8. The programme should promote equity and access to screening for the entire target population.
9. Programme evaluation should be planned from the outset.
10. The overall benefits of screening should outweigh the harm.

Table 1.4. Modified principles of screening for disease, as summarised by Andermann et al [87].

In the UK, the principles described by Wilson and Jungner, and the subsequent modern modifications, have formed the basis for 22 contemporary criteria for the appraisal of a screening programme itself, as described by the UK National Screening Committee (Table 1.5) [89].

The Condition
1. The condition should be an important health problem
2. The epidemiology and natural history of the condition should be adequately understood and there should be a detectable risk factor, disease marker, latent period or early symptomatic stage.
3. All the cost-effective primary prevention interventions should have been implemented as far as practicable.
4. If the carriers of a mutation are identified as a result of screening the natural history of people with this status should be understood, including the psychological implications.
The Test
5. There should be a simple, safe, precise and validated screening test.
6. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.
7. The test should be acceptable to the population.
8. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.
9. If the test is for mutations the criteria used to select the subset of mutations to be covered by screening, if all possible mutations are not being tested, should be clearly set out.
The Treatment
10. There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment.
11. There should be agreed evidence based policies covering which individuals should be offered treatment and the appropriate treatment to be offered.
12. Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.

The Screening Programme
13. There should be evidence from high quality Randomised Controlled Trials that the screening programme is effective in reducing mortality or morbidity. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.
14. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.
15. The benefit from the screening programme should outweigh the physical and psychological harm.
16. The opportunity cost of the screening programme should be economically balanced in relation to expenditure on medical care as a whole (ie. value for money). Assessment against these criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.
17. All other options for managing the condition should have been considered (eg. improving treatment, providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available.
18. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.
19. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.
20. Evidence-based information, explaining the consequences of testing, investigation and treatment, should be made available to potential participants to assist them in making an informed choice.
21. Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.
22. If screening is for a mutation the programme should be acceptable to people identified as carriers and to other family members.

Table 1.5. Contemporary criteria used by the UK National Screening Committee for the appraisal of a screening programme [89].

1.2.2 Assessment of screening effectiveness

1.2.2.1 Outcome-related measures

Outcome-related measures of screening include survival, disease-specific mortality, all-cause mortality, absolute risk reduction and relative risk reduction.

Survival rates, usually expressed as the percentage of subjects alive at 1 or 5 years, are a convenient outcome measure, especially for studies assessing therapeutic benefits or prognosis. However, survival is subject to the multiple biases of lead-time, length and overdiagnosis (see section 1.2.3.1) and is hence potentially inaccurate when assessing screening [90].

Disease-specific mortality evaluates death that is directly attributable to the condition being screened, while all-cause mortality is a measure of deaths in a particular cohort as a whole. Intuitively, it would seem disease-specific mortality is the best measure of the effectiveness of screening. However, disease-specific mortality crucially relies on the accurate classification of the cause of death, without bias towards the particular condition under study. This has led to the idea that all-cause mortality could be a better measure of outcome, as it circumvents the problem of inaccurate recording of cause of death. For a given condition to be able to demonstrate a significant reduction in all-cause mortality, however, the prevalence of that condition in the population would have to be high, and for this reason screening rarely demonstrates such an effect [91], and all-cause mortality could be considered too stringent an outcome measure [92]. The debate regarding which measure is more accurate still continues, as reflected in two opposing articles in a recent issue of the British Medical Journal [92, 93].

Fatality rate (or case-fatality rate) is defined as the number of cancer deaths divided by the total number of cancers detected, and expressed as a percentage. Fatality rates thus provide a comparison of cancer deaths against those with the disease, as opposed to mortality, where the comparison is against the entire screened population.

Mortality rates and fatality rates can sometimes yield apparently conflicting data about the efficacy of screening, if the cumulative incidence of cancer differs among populations in a randomised trial [90]. Additionally, it is important to focus the measurement of efficacy on the time period during which a benefit from screening is likely to be evident. For example, the Malmö breast screening trial found no significant difference in cumulative mortality rates between the screened and unscreened populations [94]. However, using data from this trial, Henschke et al. argued that cumulative mortality is an insensitive measure that masks true benefit, and that annual fatality rate, which reflects the timing of death relative to the start of screening, is more sensitive [95]. Annual fatality rate in screened and non-screened cohorts would be equal during the early years of screening introduction, as the majority of cancers detected in the early rounds of screening would be cancers that in any case were about to manifest clinically. However, as screening continued, less aggressive cancers would be detected early, and so the annual fatality rate in the screened group would fall. Once screening ceased, the annual fatality rate in the screened group would again rise, until it equalled that of the non-screened cohort.

Absolute risk reduction is the difference in risk of developing lung cancer between the screened and non-screened cohorts [96]. Relative risk reduction is the

percentage reduction in risk between the screened and non-screened cohorts, given by:

$$\frac{(\text{risk in screened cohort}) - (\text{risk in unscreened cohort})}{(\text{risk in unscreened cohort})} \times 100$$

1.2.2.2 Detection-related measures

Detection-related performance measures include cancer detection rate (including early stage detection rate), sensitivity, specificity, positive predictive value, false positive rate, and false negative rate.

Cancer detection rate is the number of cancers detected as a proportion of the total number screened. Another meaningful metric of cancer detection is the proportion of early stage cancers (for example, Stage I lung cancers).

Sensitivity is the ability of the screening test to detect a positive case. The higher the sensitivity, the lower the proportion of false negative cases, and hence the lower the false negative rate (i.e. the probability of a false negative). Put another way, the false negative rate can be expressed by (1 - sensitivity).

Conversely, specificity is the ability of the screening test to detect a negative case. The higher the specificity, the lower the proportion of false positive cases, and hence the lower the false positive rate (i.e. the probability of a false positive). Put another way, the false positive rate can be expressed by (1 - specificity) [97].

Sensitivity and specificity are characteristics of the screening test itself, and are unaffected by the prevalence of the particular condition in the population [97]. However, they do rely on there being a gold standard for diagnosis. The ideal

screening test would have a high sensitivity (for a low false negative rate) and specificity (for a low false positive rate).

Finally, the positive predictive value of a test reflects the number of cases with a positive test that have the disease, as a proportion of the total number of positive tests. Unlike sensitivity and specificity, the positive predictive value is proportional to disease prevalence.

The measurement of the detection-related performance measures above is further detailed in section 2.8.1.

1.2.2.3 Cost-effectiveness

Cost-effectiveness describes the situation where the most benefit is gained from the least cost. In order to prove its cost-effectiveness, a medical intervention such as a screening programme must be able to justify its cost by providing an increased benefit in comparison to alternative strategies [98]. The cost of the intervention can thus be thought of in terms of an opportunity cost - that is, the benefit that could be obtained by the next best use of the resources that have instead been allocated to that intervention [99].

In its simplest form, the average cost-effectiveness ratio is calculated by dividing the cost of a given intervention by a measure of its effectiveness. However, such a ratio does not take into account the cost of alternative strategies. In comparison, the incremental cost effectiveness ratio (ICER) does take alternative strategies into account, and is given by:

$$\frac{\text{(cost of intervention) - (cost of alternative)}}{\text{(effectiveness of intervention) - (effectiveness of alternative)}}$$

[100].

The justification for an intervention also depends on the "utility" of the intervention - that is, the preference that individuals have for a given outcome [101]. In turn, these "preferences" are the levels of satisfaction, desirability or distress associated with a particular health outcome [102]. If increased life expectancy is the sole outcome, cost-effectiveness can be determined on the basis of whether or not increased life expectancy has been delivered by the intervention. However, many medical interventions (including screening) may affect measures other than life expectancy alone, such as pain or disability; as such, cost-effectiveness analyses increasingly incorporate measures of both life expectancy and preferences. A popular measure in this regard is quality-adjusted life years (QALYs). A QALY is an overall measure of health outcome that weighs life expectancy against an estimate of a person's health-related quality of life. QALYs attempt to address the trade-offs between mortality, morbidity, and the preferences of patients and society by combining these factors into a single measure [102]. In the UK, the National Institute for Health and Care Excellence generally views an intervention with an ICER of less than £20,000 per QALY gained as cost-effective [103].

Favourable cost-effectiveness has been found in breast screening [104, 105], while mixed results have been reported for lung cancer screening [106-109]. For example, Wisnivesky et al. reported that screening could be expected to increase survival by 0.1 year at an incremental cost of approximately 230 US dollars, using data from the Early Lung Cancer Action Project (ELCAP) [106]. In contrast, Mahadevia et al. found that the ICERs for current, quitting and former smokers were 116,300, 558,600, and 2,322,700 US dollars per QALY gained (approximately £73 261, £351,890 and £1,463,186) respectively. They concluded that barriers such as

cost of CT and participant anxiety over indeterminate nodules did not make screening with low-dose CT cost-effective at present [108].

1.2.3 Potential problems in screening

Like all medical interventions, screening has the potential for both benefit and harm. Unlike other medical interventions, however, it is performed on an individual in whom disease is not yet recognised. Thus, the difficulties that may be encountered in screening must be borne in mind when interpreting screening trial results and assessing its effectiveness.

1.2.3.1 Types of screening bias

Lead-time bias - Lead-time bias exists when the timing of detection/ diagnosis has not been accounted for when comparing survival rates. Lead-time bias is inherent in measurement of survival because screen-detected cases are by definition ones that are in a preclinical phase, in contrast to symptomatic cases. As such, if survival is measured from the time of detection, the survival of a screen-detected case will invariably be longer than that of a symptomatic case, simply by virtue of it being detected earlier, even if both screened and non-screened cases die from the disease at the same point in time (Figure 1.4) [91].

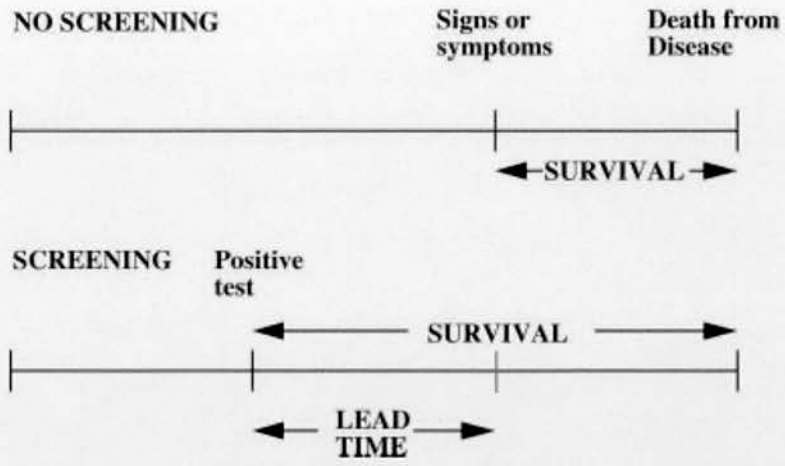

Figure 1.4. Depiction of lead-time bias. (From reference [91].)

Length bias - Whereas lead-time bias is a function of timing of detection, length bias could occur if screen-detected tumours are biologically less aggressive. These tumours would have intrinsically better survival rates than symptomatic cases.

Overdiagnosis bias - Overdiagnosis refers to the diagnosis of a cancer that would otherwise not cause death [110], either because the individual is more likely to die from other causes, or the tumour behaves less aggressively. Overdiagnosis can lead to higher incidence and higher survival in a screened cohort, but still have no effect on mortality, since the screen-detected cancers are not contributing to mortality [90].

1.2.3.2 Adherence and contamination

In any screening programme, the rate of adherence needs to be quantified - that is, the proportion of respondents who complied with the screening regimen and were followed up. A low rate of adherence can affect the power to detect a significant mortality effect. Similarly, in the context of a randomised control trial comparing screening with no screening, contamination refers to the proportion of

subjects in the non-screening arm who undergo the screening intervention (which they may do independent of the trial) [111]. High contamination also affects the power of a screening study because reduced mortality may be seen in both the screened and non-screened cohorts if the screening intervention is effective.

1.2.3.3 Adverse effects

The potential harms of screening can occur due to both false positive and false negative results. A false positive result can lead to two adverse consequences: morbidity arising from investigation of the detected abnormality, and also undue anxiety. Both of these consequences have been quantified in mammographic screening, and are not insignificant (see section 1.2.4). In the context of overdiagnosis, undue harm to the screened individual may also be caused by true positive results triggering the treatment of a lesion that is relatively indolent.

In contrast, a false negative result may provide false reassurance, and act as license for an individual to continue engaging in risk-taking behaviour [92]. Such an individual may also suffer morbidity in later years when disease manifests clinically, as they may feel that the screening process has failed them [112].

Another potential adverse effect of screening from both an individual and a health resource allocation perspective is opportunity costs. For the screened individual, these include lost wages or expenses incurred by participating in screening [91], while at the population level these would include the diversion of resources away from other healthcare priorities.

1.2.4 Cancer screening in practice

Screening for cancer has had varying degrees of success. The National Health Service currently runs three programmes to screen for breast, cervical and colon cancer [113]. Here the example of mammographic screening is considered, to illustrate the potential problems and adverse effects discussed earlier.

The first large scale assessment of screening for breast cancer was undertaken in the Health Insurance Plan (HIP) randomised control trial in New York, initiated in 1963. This trial randomised 62,000 women aged 40-64 years old to receive either a combination of annual double-view mammography and clinical breast examination, or usual care (no screening). It demonstrated a 25% reduction in breast cancer-specific mortality in the screened cohort after 18 years of follow-up [114]. Subsequently, four trials from Sweden (The Two Counties, Malmö, Stockholm and Gothenburg trials) [115-118], which compared mammography alone to no screening, as well as trials in Edinburgh [119] and Canada [120], provided compelling evidence of reduction in breast cancer-specific mortality. However, 49 years after the initiation of the HIP trial, questions are still asked with respect to the length of follow-up, biases in randomisation, and the benefit of screening in the 40-49 age group [121].

On the basis of the accumulating evidence of benefit at the time, the National Health Service Breast Screening Programme (NHS BSP) was initiated in 1988 [122]. Similar national programmes have been initiated in other countries, such as Sweden [123] and the Netherlands [124]. The NHS BSP initially invited women aged 50-64 for single view mammography, but it has since been continuously modifying its practice. Currently, two-view mammography is being offered to women aged 50-70,

with plans to extend routine three-yearly invitations to those aged 47-49 and 71-73 by 2016 [125].

However, conflicting evidence regarding the mortality benefit from breast screening continues to cast doubt over its effectiveness. For instance, Nystrom et al. quoted a 24% reduction in mortality in an overview of five Swedish trials [126]. In a recent review of breast screening trials for the Cochrane database, Gøtzsche and Nielsen found problems with misclassification of death that led to inaccurate estimates of cancer-specific mortality, and inadequate randomization in four of the eight trials evaluated. They concluded that while mammographic screening is likely to reduce mortality, the effect is probably closer to a 15% relative risk reduction in cancer-specific mortality rather than the higher levels previously described, and could also result in an absolute increase in overdiagnosis of 0.5% [127]. This view of increased overdiagnosis has been disputed by evidence from the Swedish two-county trial and the NHS BSP [128]. However, an independent UK review recently concluded that 43 deaths from breast cancer would be prevented and 129 cases of breast cancer would be overdiagnosed for every 10,000 UK women aged 50 years invited to screening for the next 20 years, i.e. one breast cancer death prevented for about every three overdiagnosed cases [129].

The psychological impact of breast screening has been evaluated in detail. In a systematic review of 54 papers from 13 countries, Brett et al. concluded that women who were recalled for further investigations suffered significant short-term, and possibly long-term anxiety. The increased risk of anxiety was associated with multiple factors, including pain during mammograms and previous false positive rates [130].

From the above discussion, it can be seen that the breast screening experience holds important lessons for other screening modalities that are still building an evidence base, including low-dose CT (LDCT) screening for lung cancer.

1.3 History and practice of lung cancer screening

1.3.1 Screening principles as applied to lung cancer

Lung cancer meets the majority of the principles for screening set forth by Wilson and Jungner (Table 1.2) [85]. Considering each principle in turn:

1) It is an important health problem with a significant human and financial cost (section 1.1.1).

2) and 3) There are acceptable surgical and non-surgical treatments, with facilities for diagnosis and treatment available (section 1.1.6). However, access to some diagnostic tests and treatments may be concentrated in large centres and vary regionally.

4) and 8) Stage I cancers can be recognised in a preclinical phase, and have improved survival (section 1.1.5). There are also international guidelines as to who should be offered treatment [29, 61].

5) and 6) LDCT is a possible test with a low amount of radiation (section 1.4). However, it is unclear if LDCT is acceptable to the population as a whole.

7) The understanding of the natural history of the various histological subtypes of lung cancer continues to improve. However, new advances in

molecular and genetic subtyping are forcing reappraisal and reclassification of some forms of lung cancer, for example adenocarcinoma [35]. New staging information has helped to redefine stage groups for lung cancer which more accurately reflect prognosis [58].

9) and 10) The feasibility and cost-effectiveness of "case-finding", i.e. recruitment to a lung screening programme, are aspects being evaluated in multiple LDCT lung screening trials at present. Cost-effectiveness has been evaluated in observational studies of lung cancer screening [106-109], but not explored fully in randomised control trials.

1.3.2 Lung cancer screening with chest radiography

There have been several trials initiated between the 1950s and 1990s to evaluate lung screening with chest radiography, all but one of which (The Tokyo Metropolitan Government Study) involved exclusively men [90]. Two of the earliest screening studies on chest radiography and sputum cytology screening for lung cancer were performed in the UK, and also exclusively targeted men [131, 132].

Five large randomised trials of varying design have been conducted. Three of these trials, the Memorial Sloan Kettering Lung Project (MSKLP), the Johns Hopkins Lung Project (JHLP) and the Mayo Lung Project (MLP) together constituted the Cooperative Early Lung Detection Program of the National Cancer Institute (NCI) [133], targeting male heavy smokers aged 45 and over. The other two trials were the Czech Study on Lung Cancer Screening and the Prostate, Lung, Colorectal and Ovarian (PLCO) randomised trial.

The MSKLP and the JHLP both randomised participants to a single screen with an annual chest radiograph, or a dual screen consisting of an annual radiograph and sputum cytology examination every 4 months, over 5 years. Both trials in effect had two different interventions and so arguably they did not contain an effective control group. No effect on any outcome parameter (stage distribution, resectability, survival or mortality) was observed in either trial after 5 years [134, 135].

In the third component of the NCI study, the MLP, participants initially underwent a prevalence screen with a single chest radiograph and sputum study; those with no cancer were then randomised to a four-monthly chest radiograph and sputum cytology study, or a control group where annual chest radiograph and sputum cytological examination were recommended but not enforced (i.e. an "advice" group). Again, the control group here may thus not have been sufficiently rigorous. The MLP also suffered from a poor adherence rate. At initial follow-up, there were no differences in mortality, but 5-year survival in the screened group was more than double that of the control group. However, there was an excess of 46 lung cancer cases (i.e. a 29% higher incidence) in the screened group [136]. After 20.5 years of median follow-up, the survival benefit in the screened arm not only persisted but was greater (by about 3 times); again, lung cancer deaths were not statistically significantly different in the two groups, but were in fact higher in the intervention arm (4.4 per 1000 in the screened group and 3.3 per 1000 in the control arm) [137]. Furthermore, the persistence excess cancer rate in the screened group (17% after more than 20 years follow-up) indicates that overdiagnosis (see section 1.2.3.1) may have been a contributory factor [138].

The PLCO trial is the largest randomised control trial (RCT) to date to assess chest radiography screening. It was initiated in 1993 and has only recently completed follow-up and analysis [139]. The PLCO randomised 154,901 participants aged 55-74 (men and women) to receive either an annual chest radiograph or no intervention ("usual care") over 4 years. This trial was population-based, and thus unlike the previous chest radiography trials, did not target a high-risk group. The trial was designed to detect a 10% or greater reduction in lung cancer-specific mortality with at least 90% power. After 13 years, no mortality difference was observed between the groups, with 1213 lung cancer deaths in the screened group, versus 1230 in the control group. Unlike the MLP, no statistically significant difference in incidence was found between the two groups (6% higher incidence in the screened group), but the large majority of lung cancers in the screened group were detected outside of screening.

Thus, the overall conclusion from the trials assessing screening with chest radiography has been that there is no significant reduction in mortality, and overdiagnosis is a significant concern.

1.3.3 Lung cancer screening with low-dose computed tomography (LDCT)

1.3.3.1 Observational trials assessing screening with LDCT

The lack of benefit from screening with chest radiography not only led to recommendations against screening, but arguably also extended caution to screening with alternative strategies, including CT [140].

LDCT had emerged as a potential screening tool by the end of the 1980s [141]. The first trial to assess LDCT was the Early Lung Cancer Action Project (ELCAP) in the 1990s that enrolled 1000 volunteers with a 10 pack-year smoking history to undergo two-view (posteroanterior and lateral) chest radiography and an LDCT scan. Enrollees with between 1 and 6 nodules had each nodule evaluated for size, location, calcification, shape and edge. Patients with more than 6 nodules were considered to have diffuse nodular disease. Nodules measuring up to 20mm that had smooth edges with benign patterns of calcification were classified as benign. For non-calcified nodules, a follow-up algorithm based on diameter was devised [142]. Using this algorithm allowed the detection of 27 (2.7%) lung cancers with LDCT, as opposed to 7 (0.7%) with chest radiography. Also, 23 of the 27 (85.2%) cancers on LDCT as opposed to 4 of 7 (57.1%) on chest radiography were stage I, with 26 of the 27 LDCT-detected cancers being resectable.

The success of ELCAP provided the springboard for the International ELCAP (I-ELCAP) [143], as well as 10 other studies initiated between 1993 and 2000 [144-153]. However, these were all single-arm observational cohort studies, as opposed to RCTs, and therefore their ability to detect a true effect of LDCT on mortality was limited [91]. Consequently, a multitude of recently concluded or ongoing RCTs assessing LDCT either against usual care, or some other intervention were initiated. A selection of these is now discussed.

1.3.3.2 Randomised control trials assessing LDCT screening

Tables 1.6 and 1.7 summarise the nodule management algorithms of recent randomised control trials assessing LDCT. In general, these trials have follow-up algorithms that are based on that of ELCAP.

Diameter (mm)	Volume (mm^3)	NLST [154]	NELSON [155]	DLCST [156]	UKLS [157]
< 2	< 15	None	LDCT at 12 months		None
3	15-50				LDCT at 12 months
4 5 6 7 8 9 10	50-500	LDCT at 3,6,12 or 24 months, depending on lesion size and level of suspicion of malignancy	LDCT at 3 months; assessment of VDT; if VDT < 400 days- refer to pulmonologist VDT=400-600 days- LDCT at 9 months VDT > 600 days- LDCT at 9 months	Combination of VDT assessment and FDG-PET for nodules 5-15mm	LDCT at 3 months; assessment of VDT; if VDT < 400 days- refer to pulmonologist; if VDT > 400 days- LDCT at 9 months
15 > 20	> 500	LDCT at 3,6,12 or 24 months, depending on lesion size and level of suspicion of malignancy; FDG-PET, dynamic contrast-enhanced CT, and/or biopsy	Referral to pulmonologist for work-up		Referral to multi-disciplinary team for work-up

Table 1.6. Management algorithms for solid nodules in the NLST, NELSON, DLCST and UKLS trials.

Diameter (mm)	Volume (mm^3)	DANTE [158]		ITALUNG [159]	MILD [160]
		Smooth margin	Non-smooth margin		
< 2	< 15	LDCT at 3,6 and 12 months; if unchanged, further CT at 24 months	LDCT at 3,6 and 12 months; if unchanged, further CT at 24 months Oral antibiotics, LDCT in 6-8 weeks; if no regression, follow-up CT or invasive procedure	None	None
3	15-50				
4					
5	50-500			LDCT at 3 months; if ≥ 1mm increase in diameter, FDG-PET or invasive procedure	Nodules 60-250mm^3 - LDCT at 3 months; if volume increased by ≥ 25%, suspected malignancy
6					
7					
8				FDG-PET; if PET-positive, FNA recommended. If PET-negative, repeat CT in 3 months.	
9					Nodules > 250mm^3 - FDG-PET or lung biopsy
10		Oral antibiotics, LDCT in 6-8 weeks; if no regression, FDG-PET			
15	> 500				
> 20		Discretionary oral antibiotics, then LDCT or standard contrast-enhanced CT, and FDG-PET		NB: for cases with suspected inflammatory nodules, antibiotics and repeat CT in 4 weeks performed	

Table 1.7. Management algorithms for solid nodules in the DANTE, ITALUNG and MILD trials.

The National Lung Screening Trial (NLST) was a 33-centre prospective RCT in the United States that randomised high-risk participants to LDCT or single-view chest radiography. The NLST was allied to the PLCO trial, and this was part of the reason why the NLST used radiography, and not usual care in the control group [161]. Three rounds of annual screening were performed, and interim analyses were conducted from April 2006 through to 2010, encompassing the last round of screening in 2007. The study was designed to have 90% power to detect a 21% reduction in mortality, independent of the number of rounds of screening [162]. A positive result in the NLST constituted a finding suspicious for lung cancer, defined on LDCT as non-calcified nodule(s) (NCNs) ≥4mm in greatest transverse dimension, or any other suspicious finding. The NLST protocol was not prescriptive, but some recommendations for follow-up of positive screens based on contemporary standard practice were provided [154].

The NLST used outreach methods such as direct mailings and advertisements in the mass media, to enrol 53,454 enrolees between August 2002 and April 2004. The NLST participants were younger, substantially better educated, and were less likely to be current smokers, compared to the general population as assessed by the Tobacco Use Supplement of the US Census Department Continuous Population Survey for 2002-2004 [163], but were otherwise representative of the general population. The interim analysis performed in October 2010 revealed 356 deaths from lung cancer in the LDCT arm of the trial, compared to 443 in the chest radiography arm, corresponding to cumulative lung cancer mortality rates of 247 and 309 deaths per 100,000 person-years, respectively. This 20.0% reduction in lung cancer-specific mortality was statistically significant, exceeding that expected by

chance. A lower but still statistically significant decrease in all-cause mortality of 6.7% was also observed in the LDCT cohort. The majority of patients were diagnosed at an early stage, with 93% of stage I lung cancer patients detected by CT undergoing surgery with curative intent [164].

However, the false positive rate of screens varied between 95% and 98% on LDCT in the three screening rounds, compared to between 93% and 96% on radiography. Also, the proportion of confirmed cancers increased in the final screening round (5.2% versus 2.4% in the second round), although the proportion of cases designated positive decreased (16.8% versus 27.9%). The reasons for this are still unclear [164]. Although the NLST does not directly compare LDCT with usual care, a recent sub-analysis of NLST-matched participants in the PLCO trial has revealed no significant difference in lung cancer mortality between CXR screening and usual care [139]; this has been taken to indirectly imply no significant difference between CT and usual care.

The NELSON (Nederlands-Leuvens Longkanker Screenings Onderzoek) trial in the Netherlands and Belgium has been ongoing since 2003 [165]. In contrast to the NLST, NELSON randomised participants to screening with LDCT or 'usual care', i.e. no screening at all. Rather than pre-defined eligibility criteria, NELSON used a population-based questionnaire to determine what the optimum risk-based selection criteria would need to be to achieve a balance between risk profile, sample size and required rates of participation and retention [165]. This approach to recruitment seemed to provide better concordance of characteristics between the trial and general populations [166]. NELSON offered four rounds of screening: participants underwent screening at baseline, one year, three years and finally 5.5 years later

[167]. Further, participants will be followed up over a longer period of 10 years in total.

NELSON is the first study to incorporate volume doubling time (VDT) calculations using semi-automated volumetry into a prescriptive nodule management algorithm. Nodules are classified into four categories: categories I and II are considered negative [benign or NCN less than category III, respectively], category III indeterminate (a solid NCN, or solid component of a part-solid nodule with a volume of 50-500mm^3; a non-solid nodule or non-solid component of a part-solid nodule with a mean diameter of ≥ 8mm; or a solid pleural-based nodule 5-10mm in minimum diameter i.e. perpendicular to the pleura), and category IV positive (a solid NCN, or solid component of a part-solid nodule with a volume > 500mm^3, or a solid pleural-based nodule > 10mm in minimum diameter) [155]. Participants with indeterminate nodules are invited back for a subsequent low-dose CT at 3 months (baseline screening) or 6-8 weeks (subsequent incidence screening). Cases where the VDT is less than 400 days are then considered positive, and further diagnostic strategies are then pursued. Finally, a VDT of > 600 days was considered negative. The threshold of a VDT < 400 days was chosen as lung cancers with a VDT greater than this may be overdiagnosed cases that do not contribute to mortality [168-170]. The rationale for a shorter initial interval of follow-up CT during incidence screening was that any new malignant nodule that had appeared over the 1 year between baseline and first incidence screen must have a short VDT, and therefore, would be shown to be growing rapidly even after a brief period of follow-up. Combining volumetric assessment with a standardised growth measurement in this way

potentially limited the unnecessary investigation of indeterminate but ultimately stable nodules.

NELSON also differs from the NSLT in that two independent readers initially interpreted LDCT studies, unlike the single reader strategy used in the NLST. On baseline screening, 8623 NCNs were detected in approximately half of all participants, and 98% of these nodules were solid. According to the NELSON nodule definitions, 196 participants had 260 nodules that were positive, in which 70 lung cancers were finally detected. This detection rate of 0.9% was lower than that of other published trials [142, 152, 171-173], but the proportions of Stage I disease in the baseline, second and third screening rounds were similar to that of other trials (64.9%, 70.7% and 62.3% respectively) [167]. Results of the fourth screening round and mortality data are still awaited. Recruitment characteristics and key results from the NLST and NELSON trials are compared in Table 1.8.

	NLST [164]			NELSON [169, 174]	
Parameter assessed	Mean diameter			Volume and diameter	
Screening round	1	2	3	1	2
No. recruited	53454			15822	
No. screened by LDCT	26309	24715	24102	7557	7289
Positivity rate[1](%)	27.3	27.9	16.8	20.8	7.8
No. of lung cancers in LDCT arm	270	168	211	70	54
Lung cancer detection rate (%)	1.0	0.7	0.9	0.9	0.5
Stage I cancer (%)	63.0 (across 3 screening rounds)			64.9	70.7
Invasive procedures (%)[2]	1.7	1.1	1.3	1.2	0.8
% with no lung cancer	29.8 (across 3 screening rounds)			27.2	21.3
Positive predictive value (%)	3.8	2.4	5.2	4.6	9.5

Table 1.8. Characteristics and key results of the NLST and NELSON trials.

[1]Positivity rate is defined as the number of subjects with a positive finding, divided by the total number of subjects screened in the LDCT arm, expressed as a percentage. Note that the nodule management strategy of the NELSON study allowed studies to be called indeterminate and subjected to follow-up scans, but for the purposes of standardization with other trials, indeterminate scans have been considered as positive and included in the calculation of the positivity rate.

[2]The ways in which invasive procedures have been defined and reported so far has varied between trials. For example, the rates of invasive procedures for NLST shown here include bronchoscopy, while those of NELSON do not.

Four smaller European randomised control trials that have recently published results are comparing LDCT with usual care in some form. These are the Danish Lung Cancer Screening Trial (DLCST), The Detection And Screening of early lung cancer by Novel imaging Technology and molecular Essays (DANTE) study, the Italian Lung (ITALUNG) trial, and the Multicentric Italian Lung Detection (MILD) trial. The DLCST was initially allied to the NELSON study, with a similar nodule management protocol [156]. The DANTE study is using an initial baseline screening

chest radiograph to then randomise participants to LDCT screening or usual care. All four trials have incorporated FDG-PET into their algorithms (Tables 1.4 and 1.5). In addition, the DANTE and ITALUNG trial incorporated discretionary antibiotic use and follow-up imaging into their nodule follow-up pathways. The MILD trial is unique in two respects: it is only following a maximum of four NCNs, and it is further randomising those in the screened cohort to repeat LDCT in 1 or 2 years [160].

The DANTE and ITALUNG studies have reported lower percentages of stage I cancers at baseline analysis compared to earlier observational studies [158, 159]. Furthermore, the DANTE, DLCST and MILD trials have revealed no significant difference in mortality between screened and control cohorts, but the short median follow-up duration (less than 5 years) means these results should be interpreted with caution [160, 175, 176].

1.3.3.3 Potential questions to be answered by LDCT trials

The current European randomised screening trials are well-placed to answer questions regarding differences in lung cancer detection and mortality rates in different populations, the optimal nodule measurement and management strategy, the duration and interval of screening, LDCT reading strategies, and cost-effectiveness.

1.3.3.4 The UK Lung Screening Trial

In comparison to some other European countries, the UK has been slow to commence a national LDCT screening trial. An RCT comparing LDCT with usual care has been planned since 2000, by the UK Cancer Coordinating Committee for Research (UKCCR) [177], but faced various delays. However, a pilot study of the

UK lung cancer screening trial (UKLS) is now underway, and aims to randomise 4000 participants to a single LDCT or usual care arm, with a shorter total follow-up period of 18 months. The recruitment and nodule management strategies of the study have been based on the NELSON trial, but with important differences. Individuals aged 50-75 will initially be approached via a questionnaire, and a cohort with a higher risk profile than that of NELSON will be selected based on a five-year lung cancer predictive risk model devised by the Liverpool Lung Project [178]. Unlike NELSON, nodules measuring 15-50mm^3 will be targeted for follow-up, in view of the single screen design.

The participants will undergo only a single screening LDCT unless follow-up is required based on the management algorithm. This 'single-screen design' is felt to be the most cost-effective and rapid method of answering the primary question - does LDCT reduce lung cancer mortality [157]? While it is not reflective of an actual implementable screening programme in which repeated screening would occur, the ideal interval for screening may be theoretically determined. The single-screen design has also been used in other screening studies [179].

1.4 Screening with low-dose computed tomography: technical considerations

1.4.1 Computed tomography systems

The technique of computed tomography was pioneered by Sir Geoffrey Hounsfield in 1973 [180]. Modern CT scanners fundamentally have an x-ray source emitting a fan beam and rotating on a gantry, through a ring composed of a detector array. The patient is moved through the ring during the image acquisition, and the x-rays penetrate the patient and are incident on the detector. The signals from the detector are reconstructed by a computer, and the resulting image is a cross-sectional display that corresponds to the degree of attenuation of the x-rays by the various materials within the scanned volume. The attenuation of a given material is a function of its density - the higher the density, the greater the attenuation. The density of a material on CT is calculated by its relative difference from the attenuation of water (arbitrarily assigned a value of 0), on a logarithmic scale, called a CT number (or Hounsfield number), and expressed in Hounsfield Units (HU).

The newest generations of CT scanners have multiple rows of detectors, and are called multidetector CT (MDCT) or multislice CT. The multiple rows allow thinner collimation and overlapping coverage in the longitudinal (z) axis, and so can perform volumetric scanning (continuous scanning of the whole volume). This has enabled near isotropic resolution - that is, all three dimensions of the volumetric data element (or voxel) are nearly equal [181]. Isotropy has facilitated accurate three-dimensional image reconstructions without distortion along the z-axis.

The lungs already have an inherently high contrast resolution on CT because the attenuation of the lung predominantly consists of two materials of markedly different density (air and the vasculature). The increased z-axis coverage of MDCT, coupled with greatly improved gantry rotation speeds, now allows the thorax to be scanned within one breath-hold. Thus contemporary CT scanning of the lungs has both high spatial and contrast resolution.

1.4.2 Radiation exposure

The dose delivered by a conventional chest CT can be greater than 100 times that of a posteroanterior chest radiograph [182]. As with any medical exposure to ionising radiation, the dose to the patient from CT should always follow the ALARA (as low as reasonably achievable) principle: the delivery of the minimum radiation dose possible to achieve an image of diagnostic quality, so that the hazards of ionising radiation are minimised.

CT dose depends on a number of acquisition parameters and patient factors. The energy of the x-ray beam for any given CT acquisition depends upon the *anode-cathode voltage* (in kilovolts, and may be stated as its peak value, kVp, or now more usually as kV), the *tube current* (in milliamperes, mA), and the *tube current-time product* (in milliampere-seconds, mAs), which is the product of tube current and the exposure time per gantry rotation. Lowering the kV reduces the effective energy and number of photons of the x-ray spectrum, while lowering the mAs reduces the number of photons but does not change the effective energy of the spectrum [183]. A

reduced number of photons will result in a decreased dose, but will also affect image quality due to an increase in image noise.

Scanner geometry can affect absorbed dose because of differences in the distance between the focal spot of the radiation source and the scanner isocentre; according to the inverse square law, radiation intensity is inversely proportional to the square of the distance between the source and the point of measurement. MDCT scanners have a shorter focal spot to isocentre distance compared to single-slice CT scanners, and so may result in increased dose if all other factors are kept constant [184].

Helical *pitch* is defined as the ratio of table feed per gantry rotation to the nominal width of the x-ray beam. Thus, an increased pitch results in a decreased scan time and reduced dose. Pitch, section collimation and table speed are all intertwined: for a given collimation, a faster table speed will increase pitch. However, an increased pitch can result in decreased spatial resolution, since it decreases the duration of radiation exposure for a particular anatomical section scanned [185]. Similarly, a thicker collimation is more dose-efficient; this is because a thicker collimation on MDCT can limit a phenomenon known as "overbeaming", whereby a proportion of the x-ray beam falls beyond the edge of the detector rows and so does not contribute to image quality [185]. However, a thicker collimation can also limit the thickness of the reconstructed image. In this way, trade-offs between dose and image optimization must be carefully balanced.

Patient factors that affect radiation dose include body weight and shape. The attenuation of a given x-ray beam increases with the thickness of the material in its

path, and thus, with increasing body mass index, there is a decrease in the detected x-ray intensity. To compensate, an increased tube current and kilovoltage may be employed for a person with a higher body mass index, thereby increasing the effective dose [186]. However, as body shape varies, the intensity detected can vary along the z-axis according to the thickness of the part of the body scanned [187]. Modern CT scanners have automated exposure control (AEC) functions that allow modulation of the exposure to compensate for some of these patient factors [188].

Common descriptors of CT dose are the CT dose index over the volume scanned ($CTDI_{vol}$), which takes the axial scan spacing and helical pitch into account, and the dose-length product (DLP), which is the product of $CTDI_{vol}$ and the length covered. The DLP is measured in milligray-centimetres (mGy-cm). A conversion factor for the chest can be used to estimate the effective dose, in millisieverts (mSv). [189]. There is variability in the effective dose delivered by a standard dose CT thorax, primarily due to the variation in tube current-exposure time product. For instance, a thoracic CT performed with a 4-slice MDCT at 120 kVp tube voltage, 4×1mm detector configuration, 0.5s rotation time, a pitch of 1.75, and effective mAs of 100 can result in an effective dose of 6.8 mSv [190]. The risk of radiation-induced malignancy from such a standard-dose CT chest examination has been estimated at approximately 1 in 4000 [191].

1.4.3 Low-dose CT technique

The feasibility of low-dose CT (LDCT) in the thorax was reported by Naidich et al. in 1990 [141]. They compared CT images obtained at 120kV, 140 mA, 2 second scan time, and 10mm thick sections reconstructed with a standard algorithm, to two different low-dose protocols: in the first protocol, they altered only the tube current to 10 mA and acquired images at 5 different levels selected from the initial standard dose CT. In the second protocol, they acquired a half-scan at 10 mA that was performed with an acquisition time that was two-thirds that of a full scan. Acceptable visualization was obtained at all chosen levels [141].

Subsequently, other investigators have proven that LDCT is acceptable for viewing normal structures [192, 193] and for a variety of pathological conditions [194, 195], including pulmonary nodule detection (see section 1.6.2.1). The dose reduction using such protocols is significant: For example, Remy-Jardin et al. estimated that using a low-dose protocol (120 kVp, 60-100 mAs depending on body weight) provided a dose of 1.9mSv and 2.4mSv across a 30-cm section of the thorax for male and female patients respectively, compared to 3.4 mSv and 4.4 mSv respectively at standard dose [194]. LDCT thus provides a viable imaging strategy at reduced dose in those subjects requiring serial CT surveillance. It is worth noting that some authors have taken the view that due to a lack of a precise definition, the term "low-dose CT" should be abandoned in favour of accurate reporting of dose parameters [196].

1.4.4 Image reconstruction, post-processing and reading

Thoracic CT images are normally reconstructed using a high-frequency reconstruction algorithm (also called a reconstruction kernel) that maximises the inherent high contrast ratio within the lungs [197]. A volumetric MDCT study can be reconstructed at varying slice thickness - thinner slices provide more resolution but have increased noise, whereas the converse is true for thicker slices. As for all CT studies, the grayscale is optimized for the structures being analysed, by altering the window settings of the image. For the lungs, a frequently used window setting is with the centre level at -600 HU, with a width of 1500 HU [198].

Modern CT images are viewed on softcopy on a Picture Archiving and Communications (PACS) workstation, which improves a radiologist's workflow, given the large number of images generated by MDCT. A variety of reconstructions can be performed using the raw data at the scanner console or from thin slices on most PACS workstations. Multiplanar reconstructions (MPRs) can be performed in any orthogonal and non-orthogonal plane [199]. Maximum intensity projection (MIP) reconstructions are obtained by summing the pixels with the highest CT numbers within a given voxel. In this way, the conspicuity of high attenuation structures is increased (Figure 1.5) (see also section 1.6.2.1). Conversely, minimum intensity projection (minIP) improves visualisation of low attenuation structures, and so for example, makes air within bronchi below the subsegmental level more readily visible [200].

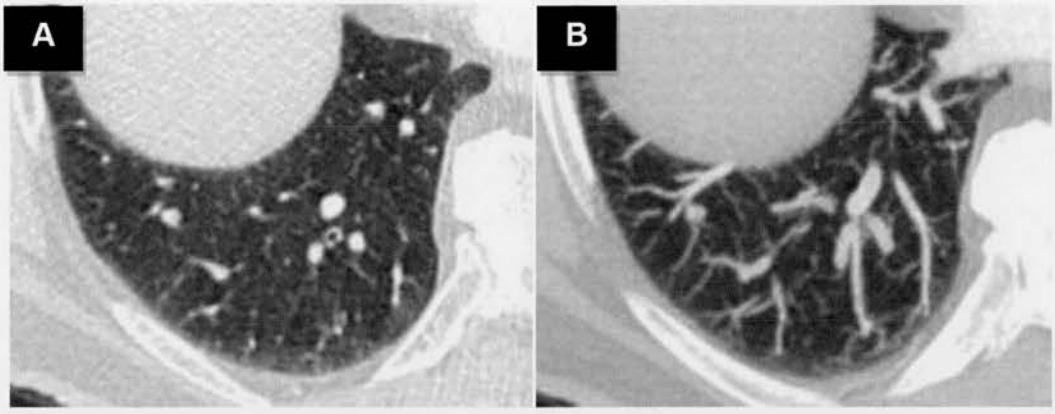

Figure 1.5. Maximum intensity projection (MIP). (A) 1mm-collimation CT image demonstrates a nodule in the right lower lobe adjacent to a vessel, but that is quite easy to miss. (B) 5mm MIP reconstruction depicts the adjacent vessel more clearly and increases nodule conspicuity.

1.4.5 CT measurement methods for lung nodules

The size of a lung nodule can be measured using diameter, area or volume, though traditionally the diameter of a nodule has been the most widely used method in clinical practice. This diameter can either be a uni-dimensional measurement of the maximum diameter of the lesion, or using length and width to provide a mean diameter, as recommended by the Fleischner Society [44]. Bi-dimensional measurements can also be performed to obtain a cross-sectional area; in turn, this cross-sectional area can be multiplied by sectional increment to obtain a volume measurement [201]. Different modifying equations to account for section-to-section variability in nodule shape can be applied for spherical, elliptical or irregularly-shaped nodules [202].

For three-dimensional segmentation of lung nodules, volume rendering is first performed. Volume rendering is a 3D technique that applies a histogram-derived tissue classification based on attenuation to the entire CT dataset. By mapping CT attenuation values to opacity, brightness and colour, structures of different density

can be selectively concealed or revealed [203]. The exact technique and algorithm differs between manufacturers, but in principle segmentation of a nodule can be performed, to define the structure of interest and exclude structures such as vessels and adjacent lung parenchyma (Figure 1.6). A measurement algorithm can be initiated with a mouse click, to calculate the nodule volume [204]. As this requires high computational power, volumetric analysis is not routinely available on PACS workstations, but is available on dedicated workstations provided by CT manufacturers. The volumetric segmentation method used in the UKLS and in this thesis is described in more detail in Chapter 2, section 2.7.2.

Figure 1.6. 3D segmentation and volume calculation of a small lung nodule using LungCARE software (Siemens Medical Solutions, Erlangen, Germany). Segmentation is initiated simply by clicking on a nodule on the CT image.

1.4.6 Overview of computer-aided detection and diagnosis

Investigations into using computer-aided detection (CAD) in thoracic CT began in earnest in the 1990s [205]. Currently, CAD in thoracic CT is used to describe a spectrum of activities encompassing detection, interpretation, decision-making, quantitative analysis and enhanced visualization [206]. The term CADx is sometimes also used to distinguish computer-aided diagnosis from other aspects of CAD. However, for the purposes of this discussion they will be considered as synonymous.

CAD has three basic components: an imaging processing step, a segmentation step, and finally a feature extraction or classification step. Ko and Naidich provided an illustration of how these components may be involved in nodule detection and interpretation [207]. The initial image processing is necessary to separate the thoracic cage from the surrounding lung, using downsampling (a technique whereby pixel width is changed to make the image matrix smaller) and subsequent application of a threshold based on attenuation to remove air from the surrounding tissues within the image. Subsequently, the pleura can be removed, and the lung parenchyma can then be segmented so that the CAD algorithm can analyse candidate regions for the presence of nodules and normal structures (Figure 1.7). Regions of interest (ROIs) are drawn over potential target candidates and interrogated using a method known as feature extraction or classification, which relies on a database of predetermined features such as sphericity [207].

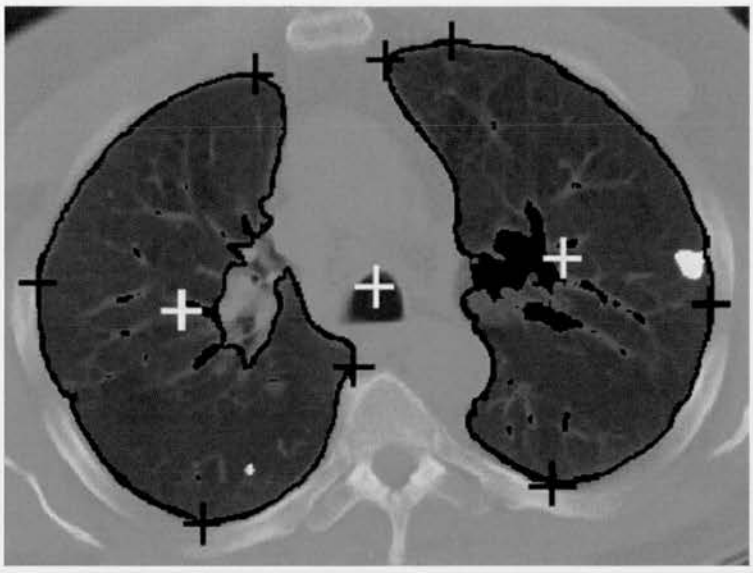

Figure 1.7. Lung segmentation and detection in a CAD algorithm. The lung border has been traced, with the black crosses indicating the most anterior, posterior, lateral, and medial pixels of each lung. The pixels of both lungs are used to calculate the centre of the thorax (white cross over trachea), and the pixels of the individual lungs are used to calculate the respective right and left lung centres (white crosses in the respective lungs). Candidate regions have then been detected by the computer, and colour-coded white for nodules and black for vessels. (From reference [207].)

Multiple feature classification methods are available, and most use a set of features such as size, shape and intensity, to form the inputs into the algorithm, and generate a single output (malignant or benign). These methods include artificial neural networks, support vector machines, linear discriminant analysis and traditional and belief decision trees. The data used to generate the feature classification is derived from expert panels forming a consensus of "truth" as to what constitutes a nodule, except for neural networks, which can adapt and learn. However, approaches using the distribution of interpretations (rather than the final consensus answer) performed by a radiologist are also possible, provided multiple aspects of those interpretations have been adequately captured [208]. Thus, the reference standard used by a particular CAD algorithm is heavily dependent on the robustness of its "truth" dataset. The Lung Image Database Consortium (LIDC) has recently

completed the compilation of a large database of corroborated lesions, where the various features of lesions that have been defined as nodules and non-nodules by multiple radiologists have been recorded in detail, to assist in the development of more robust reference standards [209].

1.5 The pulmonary nodule

1.5.1 Relevant pulmonary anatomy

The normal right lung consists of three lobes: the upper, middle and lower lobes. The normal left lung consists of two lobes, the upper and lower lobes. The trachea divides into right and left bronchi, which in turn divide into the lobar bronchi and then ramify into segmental bronchi. Each segmental bronchus is part of a separate functionally independent unit of lung termed a bronchopulmonary segment. There are usually 10 bronchopulmonary segments in the right lung and eight on the left, as described by Jackson and Huber in 1946, and Boyden in 1955 (Table 1.9) [210].

The segmental bronchi then repeatedly divide until the terminal bronchiole is encountered, which is the most peripheral bronchiole without alveoli. At the end of each terminal bronchiole is the acinus, consisting of respiratory bronchioles that lead into alveolar ducts, which in turn lead into the alveoli.

Jackson and Huber 1946	Boyden 1955
Right upper lobe:	
Apical	S^1
Anterior	S^2
Posterior	S^3
Right middle lobe:	
Lateral	S^4
Medial	S^5
Right lower lobe:	
Superior	S^6
Medial basal	S^7
Anterior basal	S^8
Lateral basal	S^9
Posterior basal	S^{10}
Left upper lobe:	
Upper division:	
Apicoposterior	$S^{1\&3}$
Anterior	S^2
Lower (lingular division):	
Superior lingular	S^4
Inferior lingular	S^5
Left lower lobe:	
Superior	S^6
Anteromedial	$S^{7\&8}$
Lateral basal	S^9
Posterior basal	S^{10}

Table 1.9. Nomenclature of normal bronchopulmonary segmental anatomy. (Modified from [210].)

The term *pulmonary lobule (*previously secondary pulmonary lobule) refers to the smallest unit of lung that is bound by connective tissue septa, as described by Miller (Figure 1.8) ([211].

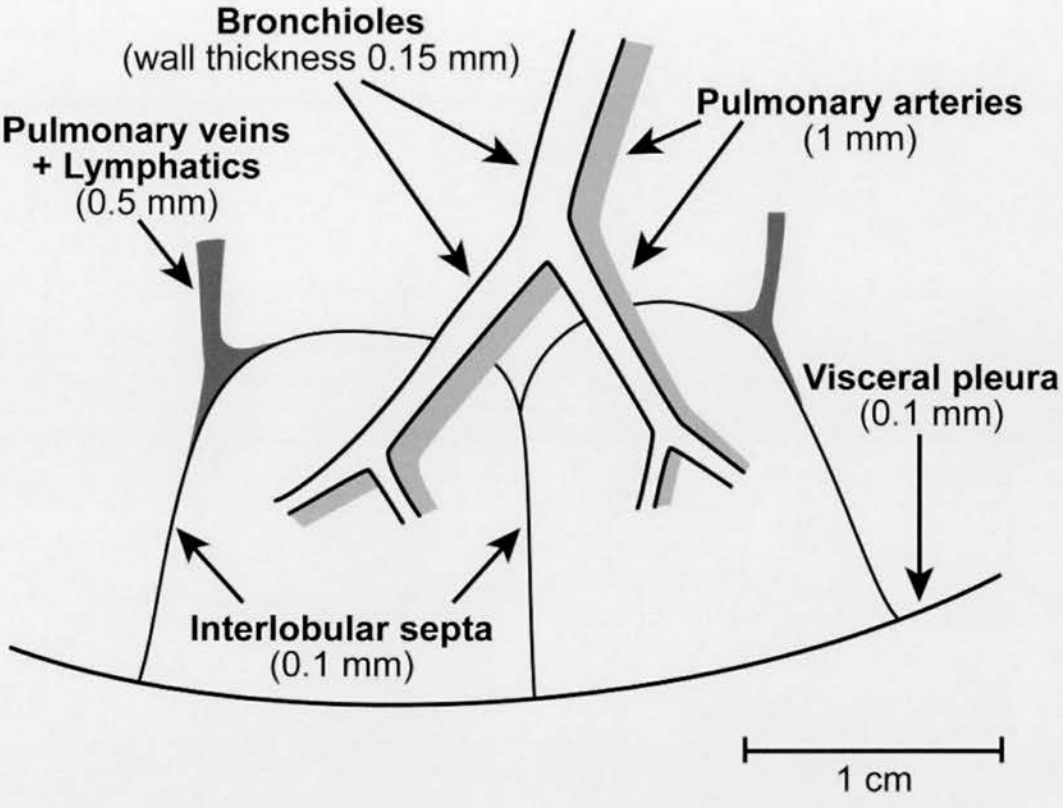

Figure 1.8. The normal pulmonary lobule. (From reference [212].)

Each pulmonary lobule is about 1-2.5cm in diameter and polyhedral in shape [213], and contains between 3 and 24 acini depending on its size [214]. At the centre of each pulmonary lobule is a bronchiole and adjacent artery, while pulmonary venous branches and lymphatics lie within the interlobular septa [213].

In the right lung, the lower lobe is divided from the upper and middle lobes by the right oblique or major fissure. The upper lobe is divided from the middle lobe by the horizontal or minor fissure, which extends anteriorly from the oblique fissure in the axial plane. On the left the upper and lower lobes are also divided by the left oblique fissure, orientated more vertically than the right. The oblique and minor fissures may be incomplete in 12.5-90% [215, 216] of cases, respectively. Some

form of accessory fissure dividing a segment from the remainder of the lobe may be encountered in 22-32% [217, 218].

The connective tissue which supports the lung is termed the interstitium. The interstitium as described by Weibel consists of the peribronchovascular interstitium (a strong sheath surrounding the large bronchi and arteries emanating from the hilum); the centrilobular interstitium (a peripheral continuation of the peribronchovascular interstitium); the subpleural interstitium (from which the interlobular septa that form the boundaries of the pulmonary lobule project); and the intralobular septa (which connects the centrilobular interstitium and interlobular septa) [219].

1.5.2 Pulmonary nodules

1.5.2.1 Historical aspects

Isaac Adler recognised the presentation of a lung cancer as a "single nodule, usually quite small", but viewed it as a rare occurrence [220]. The following year, Wenkebach provided one of the first discourses on radiographic patterns of lung pathology. He emphasised the sharp boundaries and homogeneous nature of tumours of the lung on chest radiography as a distinguishing characteristic from "infiltrations" such as pneumonia [221]. It was not until the mid-20th century, however, that numerous reports begun to describe the difficulties in managing "the solitary pulmonary mass", "nodule" or "coin lesions" [222-224]. By the 1970s, radiographic criteria for the definition of a solitary pulmonary nodule had been established. These included an oval or round shape, a diameter of 4-6cm or less in diameter,

homogeneous density, surrounded by lung, and with delineated margins [225]. The advent of conventional whole lung tomography [226] and subsequently thoracic CT [227] increased the detectability of pulmonary nodules. Subsequent refinements in CT technique with thick-section spiral CT [228] and MDCT, with its thinner sections [229] have made the increased detection of pulmonary nodules inevitable.

1.5.2.2 Definition, types and location of pulmonary nodules

Nodules may be defined by their size, attenuation, location and/or radiologic pattern. The Fleischner Society defines a pulmonary *nodule* as "a rounded opacity, well or poorly defined, measuring up to 3cm in diameter" [230]. A *micronodule* is an opacity measuring less than 3mm in diameter. A solid nodule has homogeneous soft-tissue attenuation, while a *ground-glass nodule* (also termed a *non-solid nodule*) has a hazy attenuation - lower than soft tissue, but not obscuring the bronchial or vascular structures within it. A *part-solid nodule* (also termed a semi-solid nodule) contains both solid and ground-glass density [230, 231]. The term *subsolid nodule* encompasses both pure ground-glass nodules and part-solid nodules (Figure 1.9) [232].

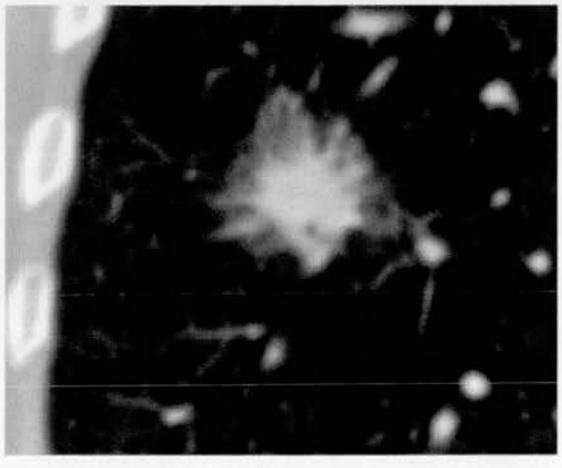

Figure 1.9. A part-solid nodule with a predominantly solid component.

Centrilobular nodules are located a few millimetres away from the pleural surface or interlobular septa as their name implies [233]. Perilymphatic nodules lie within the peribronchovascular, subpleural or centrilobular interstitium, as well as within the interlobular septa [212]. The terms *subpleural*, *perifissural*, *juxtapleural* and *juxtavascular* have also been used to describe nodule location [234, 235] but do not have standardised definitions.

The pattern of nodularity can be a helpful descriptor. Tree-in-bud nodularity refers to branching centrilobular nodularity that is due to either endobronchial impaction by inflammatory material or fluid, or a peribronchial abnormality such as fibrosis. Multiple causes for tree-in-bud nodularity have been described, but most relate to infection and bronchiolitis [236]. Miliary nodularity refers to the random distribution of micronodules throughout the lung, usually seen in disseminated haematogenous spread of tuberculosis or metastases.

1.5.2.3 Causes of pulmonary nodules

Some of the causes of a solitary pulmonary nodule are listed in Table 1.10. These can broadly be categorised into neoplastic and non-neoplastic causes. Non-neoplastic causes can further be divided into inflammatory, vascular, congenital and miscellaneous causes [237, 238].

Aetiology	Disease
Neoplastic:	
Malignant	Primary lung cancer Metastatic disease[1] (including colon, breast, prostate, testicular, renal cell carcinoma, melanoma, and osteosarcoma) Primary carcinoid Primary lymphoma
Benign	Hamartoma
Non-neoplastic:	
Inflammatory:	
Infectious	Granulomatous (mycobacterial, fungal) Abscess Septic embolus[1] Bacterial
Non-infectious	Sarcoidosis[1], Wegener granulomatosis[1], Idiopathic bronchocentric granulomatosis, Rheumatoid arthritis[1],Amyloidosis1 Intrapulmonary lymph node
Vascular	Infarct
Congenital	Intraparenchymal bronchogenic cyst Bronchopulmonary sequestration/ congenital pulmonary airway malformation

Table 1.10. Some causes of pulmonary nodules. Modified from Brandman et al. [238] and Erasmus et al. [237].

[1] These conditions commonly cause multiple nodules.

It is important to note that an opacity that is considered a nodule on chest radiograph may not be a true nodule at all. Instead, it may be due to an external object, due to composite shadows from the overlap of two structures, or "psuedotumour" due to fluid in a fissure, for example. In the case of external objects this can usually be resolved by ensuring the confounding objects are removed, while different projections or repeat films at different inspirations can be obtained to analyse composite shadows. CT can help resolve apparent persistent nodular opacities on a chest radiograph that are in fact caused by fluid or composited bony structures.

The prevalence of solitary pulmonary nodules on LDCT of high-risk patients in screening studies has been estimated at between 8 to 57%, but differences in the definition of a nodule and in reporting methods (e.g. no clear separation of those with multiple versus single nodules, reporting of percentage of patients with nodules, rather than the number of nodules itself) is at least partly responsible for this wide variation [239].

1.5.3 Differentiating benign and malignant nodules on CT

The initial radiological differentiation of benign from malignant nodules relies on morphological evaluation of size, contour, internal characteristics and location, in combination with clinical probability of malignancy. If a nodule remains indeterminate after this evaluation, assessment of growth and sometimes functional characteristics are necessary.

1.5.3.1 Size

In general, a small nodule is more likely to be benign, and indeed 80% of benign nodules are less than 2cm in diameter. However, up to 42% of nodules measuring less than 2cm may be malignant, as shown in a series of 634 solitary pulmonary nodules assessed by CT [240]. A systematic review of eight trials (including ELCAP) has reinforced the notion that nodules measuring under 5mm are almost always benign [239]. The likelihood of malignancy increases as size increases, with odds ratios for malignancy increasing from 0.74 for nodules measuring 1.1-2.0cm, to 3.67 and 5.23 for nodules measuring 2.1-3.0cm and > 3.0cm, respectively [241].

1.5.3.2 Shape and contour

A sharp well-defined margin may be seen in both benign and malignant nodules, and is thus not a particularly helpful discriminator. In a series of 85 malignant and 11 benign nodules, Zwirewich et al. reported that a sharp margin was seen with almost equal frequency (85% of malignant and 82% of benign nodules) [242]. However, a spiculated or irregular margin, especially with distortion of the adjacent vessels (described as a "corona radiata" appearance) is known to correlate well with malignancy. Kuriyama et al. showed that fine spiculations, pleural retraction, and peripheral vessel convergence (analogous to the "corona radiata") were strong indicators of malignancy in peripheral lung cancers [243]. However, these edge characteristics are not sufficiently sensitive or specific for benignity or malignancy in isolation.

1.5.3.3 Internal characteristics

Internal characteristics include calcification, cavitation, attenuation (fat, solid or ground-glass), bubble-like lucencies (sometimes referred to as "pseudocavitation") and air bronchograms. Benign lesions are associated with rather dense and discrete calcification in four characteristic patterns. Central, diffuse, solid or laminated calcification may be seen in calcified granulomas, such as histoplasmosis or tuberculosis, while "popcorn" calcification indicates chondroid calcification within a hamartoma (Figure 1.10) [237]. Hamartomas are benign neoplasms composed of mesenchymal tissues such as fat and connective tissue, typically combined with respiratory epithelium [244]. Absence of such discrete calcification is not indicative of malignancy, as benign lesions often have no calcification. On the other hand, foci of calcification on CT may be seen in malignant lesions, either due to a focus of pre-

existing granuloma that has been subsequently encompassed by tumour, or due to dystrophic calcification. In a series of 353 patients undergoing CT for lung cancer, 20 (6%) showed some evidence of calcification [245].

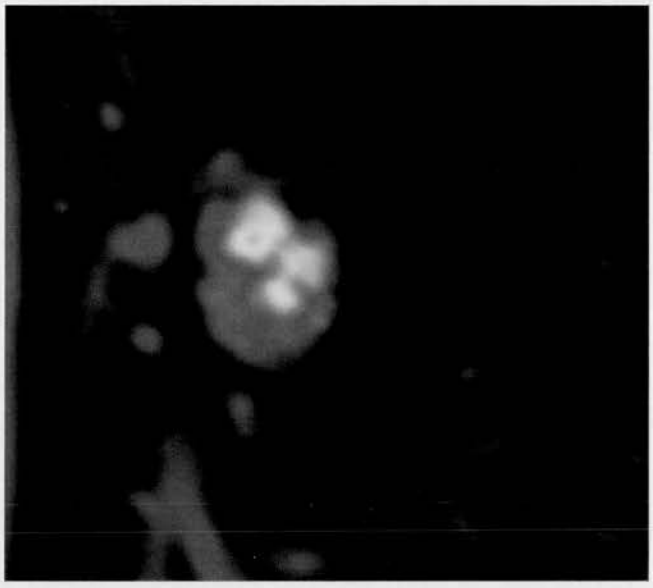

Figure 1.10. Hamartoma demonstrating "popcorn" calcification on a coronal CT image with bone window settings. (From reference [238].)

In addition to "popcorn" calcification, the presence of intralesional fat on thin-section CT points towards a hamartoma. In one series of 47 histologically-proven hamartomas, fat was seen on CT in 18 (38%), and fat and calcium in 10 nodules (21.3%), using a definition of -40 HU to -120 HU on CT. However, only about half of hamartomas contain fat [246].

Both benign and malignant lesions may cavitate. The thickness of the cavity wall can provide a clue to aetiology: cavitary nodules on chest radiograph with a wall thickness less than 4mm were benign in 92% of cases, while those with a wall thickness greater than 16 mm were malignant in 95% of cases in one series [247].

However, cavities with an intermediate thickness of 5 to 15mm were malignant in just over half of cases in that investigation, underscoring the lack of sensitivity and specificity of this sign in the majority of cavitary nodules.

Ground-glass nodules may be seen in both inflammatory and neoplastic conditions, with different radio-pathologic correlations. Multifocal centrilobular ground-glass nodularity may be seen in respiratory bronchiolitis-interstitial lung disease, for example [248]. It is now recognised that the subtype of adenocarcinoma formerly termed bronchioloalveolar carcinoma (BAC) has a range of heterogeneous ground-glass CT appearances. These include ground-glass nodules (with or without associated features such as bubble-like lucencies or psuedocavitation, and air bronchograms), part-solid nodules with varying proportions of solid and ground-glass attenuation, and multifocal consolidation [249]. In a seminal paper, Noguchi proposed six subtypes of "small" (2cm or less in greatest dimension) adenocarcinoma of the lung, labelled A-F. Types A-C were termed replacement adenocarcinomas, as they represented progressive forms of localised BAC with increasing degrees of invasion, while types D-F were subtypes of invasive adenocarcinoma that could arise *de novo* [250]. There is limited evidence suggesting that these subtypes correlate with different but overlapping radiological appearances [251]. The heterogeneity of cancers classified as BAC are partly behind its removal in the recent reclassification of adenocarcinoma [35]. Using the new terms, ground-glass and part-solid nodules that are malignant represent an overlapping pathological and radiological spectrum of adenocarcinoma-in-situ (AIS), minimally-invasive adenocarcinoma (MIA), lepidic predominant adenocarcinoma (LPA) and invasive adenocarcinoma (Figure 1.11) [252].

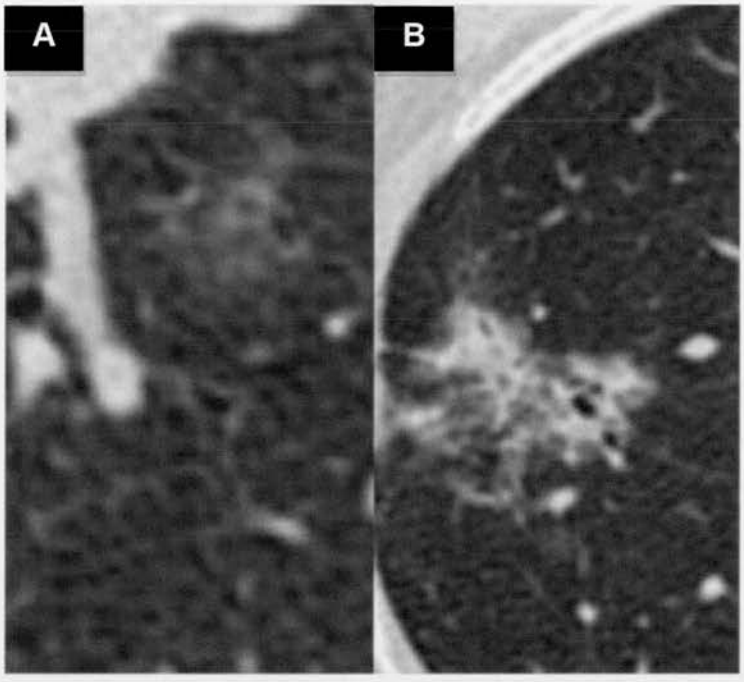

Figure 1.11. Internal characteristics of two different subsolid nodules in the same patient. (A) CT image demonstrates a pure ground-glass nodule, which was revealed to be an adenocarcinoma-in-situ (AIS) on surgical resection. (B) A larger and more dense nodule with some air bronchograms and pleural retraction was also resected and corresponded to a lepidic predominant adenocarcinoma (LPA).

1.5.3.4 Location

As a general rule, malignant nodules are more likely to be found in the upper lobes [253]; however, this feature is not specific for malignancy. Intrapulmonary lymph nodes (IPLNs) are a group of nodules that can be recognised by their location, as long as they meet other morphologic criteria. Some histopathology-corroborated CT features of IPLNs have been described in studies with relatively small samples [235, 254-256]. Typically, an IPLN is up to 12mm in maximum diameter, is polygonal or coffee-bean shaped, lies within 15mm of a pleural surface, has a smooth surface, and often has at least one linear opacity connecting it to the pleural surface, reflecting an interlobular septum [235, 255]. IPLNs are relatively uncommon in the upper lobes. In an analysis of LDCT scans from screening subjects, Ahn et al.

postulated that the lack of malignancy in perifissural nodules seen in their cohort was possibly due to the fact that they were IPLNs, as the majority of these nodules were triangular or ovoid (44% and 42% respectively), and had a septal connection (73%) [234].

1.5.3.5 Growth assessment

The demonstration of a growing nodule remains one of the most useful discriminators of malignancy. Cancers may grow at different rates and in different patterns. The two mathematical models most often used to describe cancer growth are the exponential model and the Gompertzian model.

The exponential model presupposes that the cancer has a constant rate of doubling, most conveniently calculated by the doubling time [257]. The time taken for one volume doubling, the volume doubling time (VDT) can be calculated using the formula:

$$VDT = \frac{\Delta t \times \ln 2}{\ln (V2/V1)}$$

where V1 is the volume of a given nodule at time t1, V2 is the volume at time t2 when it is next measured, and Δt is the interval between t1 and t2 [258].

However, the exponential model does not provide an adequate explanation for the types of tumour growth seen *in vivo*. For example, doubling times are known to exceed cell cycle times [259]. The Gompertzian model allows for an exponential growth phase in the early stage, which then saturates and reaches a plateau with

increasing tumour size, possibly due to some form of negative feedback inhibiting growth [260].

In experimental studies, lung cancers *in vivo* have been found to grow at quite different rates, and in patterns that conform to neither model of growth alone. Lindell et al. studied the growth patterns of 18 screening-diagnosed lung cancers of different subtypes, using long- and short-axis diameter CT measurements to estimate volume [261]. They found that while the majority of lesions grew at a constant rate, there were also cancers that grew very slowly - these were mostly seen in the BAC and adenocarcinoma groups (Figure 1.12). Furthermore, 22% of lesions showed a decrease in volume at some point on both visual estimates and calliper assessments.

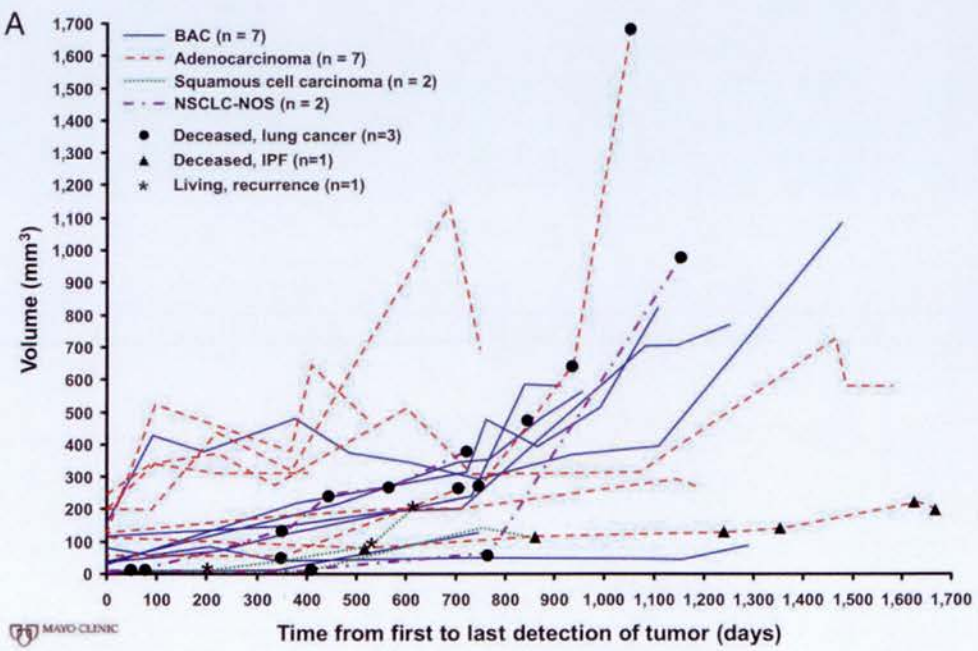

Figure 1.12. Growth curves of 18 non-small cell lung cancers (NSCLC) followed over five years. Many adenocarcinomas demonstrate periods of quite slow growth, followed by a rapid accelerated growth, as well as periods of decreases in size. NSCLC-NOS= non-small cell lung cancer not otherwise specified. (From reference [261].)

Size and growth of a nodule can be measured using diameter or volume. Volumetric analysis is intuitively more sensitive to changes in size as compared to diameter measurement, since volumetry captures the lesion in its entirety, and not just a cross-sectional area [258, 262]. The improved sensitivity of volumetry for growth is highlighted by the fact that for a completely spherical nodule to undergo one volume doubling, there needs to be only a 26% increase in diameter. Volume change is now the main parameter of assessment in the NELSON, DLCST, MILD and UKLS trials [155-157, 160].

The VDT of malignant nodules is highly variable, in the range of 20 and 400 days [202, 258, 261, 263-265], with shorter VDTs noted for small cell lung cancers. A recent analysis of 111 lung cancers detected after an initial negative baseline screen (i.e. incident cancers) in the I-ELCAP trial found that such cancers had a median VDT of 98 days [266]. A solid nodule that has demonstrated stability (i.e. no growth) over 2 years has generally been considered to be benign [44]. However, ground-glass nodules that may represent in-situ or minimally invasive forms of adenocarcinoma (formerly BAC) may have much longer doubling times that could exceed two years (730 days) and so falsely be considered benign. For example, Hasegawa et al. demonstrated that the mean VDT varied between 817 days in pure ground-glass lesions, to 149 days in solid lesions, in a screening-derived cohort of proven lung cancers [263]. Slower growth in such lesions should not be mistaken for benignity [267]. Indeed, a recent analysis of 42 prevalent and 21 incident non-small cell lung cancers from the Pittsburgh Lung Screening Study has demonstrated that prevalent cancers were more likely to have a longer VDT (> 365 days) and contain a higher proportion of cancers in the adenocarcinoma spectrum [268]. On the other

hand, nodules with VDT of greater than 400 days may represent overdiagnosed cases [168].

1.5.3.6 Density change

Because growth may not be an accurate predictor of malignancy in ground-glass or part-solid nodules, the assessment of change in density (with an increasing solid component) may be more useful in these lesions [269, 270]. Henschke et al. reported that the frequency of malignancy in a high-risk population was higher in part-solid nodules compared to pure ground-glass nodules (63% vs 18%) [271]. De Hoop et al. recently proposed that mass (i.e. the product of density, as measured by CT attenuation number and volume) be used as a novel method to determine growth in ground-glass nodules. They found that this parameter could predict growth earlier than either density or size alone, and seemed to have a lower interobserver variability than that of volume [272].

1.5.3.7 Functional characteristics

Functional evaluation can be performed with FDG-PET and PET-CT, or dynamic contrast-enhanced CT. Magnetic resonance (MR) assessment is currently only performed in experimental studies.

FDG-PET can demonstrate increased metabolic activity indicating malignancy in non-calcified nodules > 10mm in diameter with a sensitivity of 96.8% and specificity of 77.8% [273]. Accordingly, the American College of Chest Physicians (ACCP) recommends that nodules measuring at least 8 to 10mm with a low to moderate pre-test probability for malignancy should be referred for FDG-PET for further characterisation, but there is little value to be gained in FDG-PET work-up

for smaller nodules [274]. The role of integrated PET-CT in nodule characterisation has yet to be extensively evaluated [275], but new insights may be gained from the Italian screening studies [DANTE, MILD, ITALUNG and the Continuous Observation of Smoking Subjects (COSMOS) studies] incorporating PET and PET-CT into their algorithms for nodule management. For example, Veronesi et al. found that the diagnostic sensitivity of PET-CT in such cases was 88%, but increased to 100% for solid non-calcified > 10mm, in an analysis of the COSMOS study [276]. In a later analysis they have suggested that decreasing the maximum standardised uptake value (SUV_{max}) (see section 1.1.5) to 1.5 from 2.0 improves sensitivity without compromising specificity in NCNs < 10mm, but this strategy will require further validation [173].

Dynamic contrast-enhanced CT is a technique which uses measurable contrast enhancement above a defined threshold as a surrogate for the increased vascularity within malignant nodules. The threshold for classifying enhancement as malignant can be varied between 15 HU and 30 HU with corresponding sensitivities and specificities of 98% and 58%, and 99% and 54% respectively [275, 277]. This method has shown good correlation with angiogenesis, as measured by vascular endothelial growth factor (VEGF) of nodules on subsequent immunohistochemical staining of pathological samples [278]. In the screening setting, it has only been used thus far in the COSMOS study, where a subset of 54 subjects with intraparenchymal but not perihilar nodules measuring > 8mm underwent dynamic contrast-enhanced CT, resulting in 100% sensitivity but a low 59% specificity for lung cancer [173].

1.5.4 Follow-up strategies for the indeterminate pulmonary nodule

Strategies to follow up the indeterminate pulmonary nodule are necessarily conservative, since the vast majority of small nodules are benign. These strategies must maximize the chances of lung cancer detection, minimize false positive rates, avoid unnecessary anxiety to the patient, and avert undue harm as a result of repeated radiation exposure or unnecessary invasive procedures, while also being cost-effective. The Fleischner guidelines for small (< 8mm) non-calcified solid pulmonary nodule management is the most well-known strategy (Table 1.11) that is applicable to persons 35 years or older, and is based on the mean diameter (length and width) of a nodule [44]. The Fleischner guidelines establish intervals for CT follow-up based on diameter categories.

Diameter (mm)	Low-risk	High-risk
< 4	No follow-up	CT at 12 months; if unchanged, no further follow-up
4-6	CT at 12 months; if unchanged, no further follow-up	CT at 6-12 months, then at 18-24 months if unchanged
6-8	CT at 6-12 months, then at 18-24 months if unchanged	CT at 3-6 months, then 9-12 months, and 24 months if unchanged
> 8	CT at 3, 9 and 24 months if unchanged, FDG-PET, dynamic contrast-enhanced CT, and/or biopsy	CT at 3, 9 and 24 months if unchanged, FDG-PET, dynamic contrast-enhanced CT, and/or biopsy

Table 1.11. Fleischner Society recommendations for the management of small pulmonary nodules. Adapted from MacMahon et al. 2005 [44].

However, the Fleischner guidelines do not distinguish between solitary or multiple pulmonary nodules. The exact prevalence of individuals with multiple pulmonary nodules in the population is unclear; in the high-risk populations involved

in screening studies between 29-32% in those with at least one non-calcified nodule had multiple nodules [142, 159].

Recently, the Fleischner society has also published guidelines for the management of subsolid nodules [232]. These guidelines advocate longer periods of follow-up and greater intervals between follow-up CT scans, with distinct recommendations made for the follow-up of solitary ground-glass nodules, solitary part-solid nodules, and multiple subsolid nodules.

1.5.5 Geographical and ethnic variations

Indeterminate pulmonary nodules do not seem to have a higher prevalence in any one geographic location [239]. A study from Western Australia concluded that the prevalence of small nodules was significantly lower (39%) than an equivalent cohort from the Mayo Clinic Lung Screening Project (51%), but this study was performed on only 49 asymptomatic smokers [279]. It is expected that calcified pulmonary nodules would be seen in areas with endemic granulomatous diseases such as histoplasmosis or tuberculosis, and in populations with a high incidence of pneumoconioses. However, recent data from the NLST has shown no variation in false positive rates between screening centres within and outside the US "histoplasmosis belt", suggesting that such geographical differences may not substantially contribute to variation in nodule frequency [280].

1.6 Reader performance in pulmonary nodule perception on CT

1.6.1 Reference standards and reader variation in nodule perception

A nodule first needs to be perceived, i.e. correctly detected and interpreted, before the strategies to designate it as benign or malignant as outlined in the preceding section can be used. Kundel et al. state that perception has at least three components: search, pattern recognition and decision-making [281]. Applying these components to CT nodule reading, a nodule must:

1. First be detected (a function of *search*);
2. Then be identified as different in contrast to normal structures or opacities (a function of *pattern recognition*); and finally
3. Have a decision made on it, as to whether it constitutes a true "nodule" and needs follow-up, in terms of the likelihood of malignancy and with or without the help of a nodule management strategy *(decision-making)*.

It is well known that radiologists fail to detect lung cancers on chest radiographs that were visible as pulmonary nodules in retrospect, to varying degrees [282, 283]. Although some of the technical factors that would make missing such nodules on chest radiograph (e.g. overlapping vascular structures or obscuring ribs) are not a significant problem on thoracic CT due to its cross-sectional nature and high spatial and contrast resolution, there are still a host of other factors that lead to variation in performance for nodule perception (discussed in the next section). To be able to quantify such differences in perception, a reference standard for the number

of "true" nodules in any given scan is crucial. Ideally, the reference standard would be the "gold" standard with actual histopathological corroboration of each nodule. Indeed some authors have argued that clinical, follow-up and histopathological proof should be the basis of "truth" for such interpretation: in one of the earliest investigations of "truth" in 1978, Revesz et al. highlighted the problem posed by using reference standards that approximate the "truth" [284]. They investigated the accuracy of three different chest radiographic techniques, based on radiologists' interpretations as compared to five different reference standards: majority vote (majority of expert panel agree), consensus opinion (all experts agree), expert judgment (a further expert arbitrating), feedback review, and clinical/pathologic proof. They found that each of the radiographic techniques could be proven to be superior to the other depending on which definition of truth was used, the implication being that any diagnostic test could be falsely thought to be accurate if measured against a particular truth standard, i.e. the truth is subjective.

However, it would not only be impractical but also unethical to seek histopathological corroboration for multiple nodules found on a CT scan, the vast majority of which are likely to be benign [239]. Instead, reference standards using a majority or consensus of opinions are probably the most popular form of "truth". It could be argued that if a test measures favourably against different surrogate reference standards, it would serve as a more robust validation of that test in the absence of a definitive truth.

Two studies from the Lung Image Database Consortium (LIDC) have elegantly illustrated the above concept. Armato et al. asked four radiologists to analyse 25 thoracic CT studies for nodules $\geq$ 3mm in diameter, and created 24 different expert

"truth" sets based on combinations of pairs and triplets of these radiologists, against which the sensitivity of each radiologist was measured [285]. They showed that sensitivity varied widely between 51-83.2%, depending on the definition of "truth". Another study used a similar design of different combinations of four radiologists as a reference standard for CAD performance assessment [286]. Between the most liberal (any 1 of 4 readers) and strictest (all 4 readers in consensus) definitions of reader agreement as the reference standard, the number of "true" nodules $\geq$ 3 mm decreased 48% (from 174 to 90) and CAD sensitivity for nodules $\geq$ 3 mm increased from 70% to 79%, while a much larger decrease (84%) was seen in the number of nodules < 3mm. These studies also highlight two further issues: first, the sensitivity of a reader (whether human or CAD) increases as the truth panel used becomes more restrictive, since the denominator of agreed nodules becomes smaller. Secondly, by using the same dataset to both define the reference standard and test readers, a level of uncertainty as to the validity of the reference standard may be introduced, i.e. there may be "chance" agreement between test and reference standard reading [287]. This uncertainty may be reduced by using different datasets to establish the reference standard and to test readers, or by resampling the expert panel with a different dataset to ensure consistency at a different point in time.

Reader performance or CAD studies generally report the establishment of their own reference standard [288-294]. In light of the above discussion it is imperative to view all reported sensitivity data in the context of the "truth" standard used. All CAD systems should also ideally report the type of truth panels on which their algorithms were based [295].

1.6.2 Factors influencing reader performance

1.6.2.1 Influence of technical factors

Several investigators have already demonstrated the feasibility of low-dose CT for the detection of pulmonary nodules, predominantly by lowering tube current-time product [296-303]. The feasibility of pulmonary nodule detection is probably a consequence of the inherent high contrast within the lung, which enables it to tolerate reductions in signal-to-noise ratio. A recent study showed a decrease in sensitivity for only one reader out of three, when tube current-time product was lowered to 5mAs compared to 300mAs [304]. Despite this, there can be decreased detection if tube current-time products are decreased to lower than 20mAs, particularly for nodules less than 5mm in experimental settings [299].

The measurement of diameter with electronic callipers is subject to considerable intra- and inter-reader variability [305-307]. Revel et al. showed that, when measuring nodules 2cm or less in a non-screening cohort, the limits of intra- and inter-observer variability were 1.32mm and 1.73mm respectively. This means that a nodule could confidently be said to have grown only if its diameter had increased beyond these limits [307]. For example, because only a 26% increase in diameter of a spherical nodule is required to represent a single volume doubling [308], it could be falsely concluded that a nodule measuring 5mm at baseline and then 6.3mm at 3 month follow-up has doubled in volume, when it is in fact stable and the difference is due to measurement error. This has implications for the confidence assigned to the interpretation of manual nodule measurements made by readers.

Volumetric analysis, especially semi-automated volumetry, also demonstrates a degree of interobserver variability, but less so compared to that of diameter measurements. Volume measurements can be expected to be more accurate than uni- or bi-dimensional measurements alone, since a volumetric measurement evaluates the entirety of a nodule [201, 258, 309]. In a study of 322 synthetic nodules implanted in porcine lungs, Bolte and colleagues demonstrated that volumetry could estimate true lesion size to within a mean deviation of –9.2% for semi-automated and –0.3% for manual-corrected volumetry [310]. Volumetric software-derived measurements also demonstrate good reproducibility [258, 311, 312] and high inter-observer agreement [311, 313, 314]. The measurement error between volumetric measurements of the same nodule is up to about 27%, with the majority of volumetric measurements demonstrating variability of less than 10% [311, 313]. Many studies have evaluated different technical parameters that may be responsible for variation in volume reproducibility. The conclusions of these studies can be summarised as follows:

- There is less variability in volumetry as the size of measured nodules increases [315-318];
- Volumetric segmentation is less accurate when nodules are non-spherical and have irregular margins (Figure 1.13) [317, 319];
- Measurement reproducibility is increased with decreasing section thickness [316, 320]; and
- Using different segmentation algorithms within the same software package can increase inter-observer variability [321].

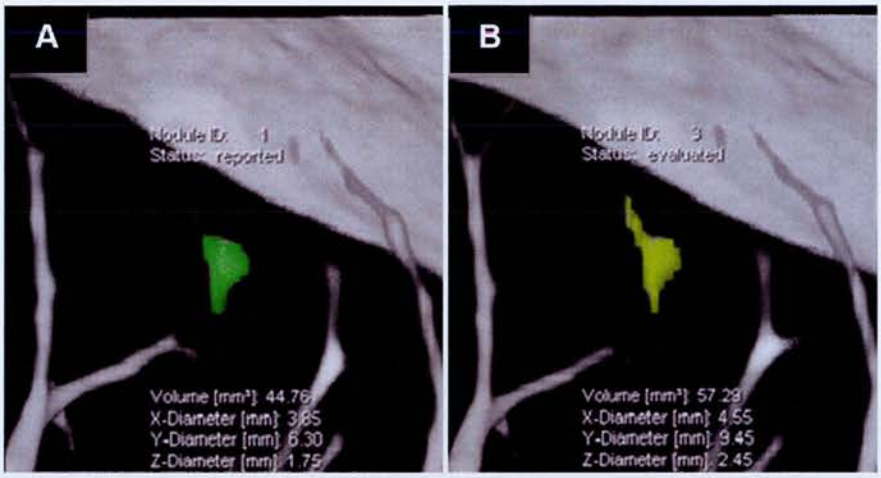

Figure 1.13. 3D volumetric segmentation of a non-spherical nodule. Segmentation performed on the same nodule yields different volumes due to its irregular shape.

The volumetric analysis of non-solid nodules or the ground-glass component of subsolid nodules is still suboptimal, but improvements in segmentation algorithms for this purpose are ongoing [322].

Nodule detection also varies with slice thickness; reduced slice thickness with MDCT has led to an increased ability to detect small pulmonary nodules. Fischbach et al. used two readers to evaluate 100 CT thoracic examinations that had been reconstructed at 1.25mm and 5mm slice thickness [229]. They found that although nodules greater than 10mm in diameter were equally well depicted at both slice thicknesses, there was a significant improvement in detection of nodules < 5mm. Also, interobserver agreement was high for nodules detected at 1.25mm but only moderate for those detected at 5mm.

The use of maximum intensity projections (MIP) can also contribute to increased detection. This was first demonstrated by Gruden et al., in a study of 25 patients with pulmonary metastases (all 3-9mm in diameter), whose scans were reviewed by 5 radiologists of varying experience [323]. They showed that the use of

10mm thick, 8mm interval MIP cine reconstructions significantly improved the detection of central nodules that can potentially be obscured by vasculature for senior radiologists, while the detection of peripheral nodules was improved for the junior radiologists, hence mitigating the effect of reader experience in nodule detection. Valencia et al. subsequently demonstrated that the use of 10mm thick axial and coronal MIP images resulted in increased detection of nodules less than 5mm, as compared to 1mm and 10mm axial images, whereas the detection of nodules greater than 5mm was not statistically significantly different between all reconstruction types [324]. An investigation performed on LDCT screening images has suggested that MIP review may also be the least time-consuming reading technique as compared to axial 1mm review and to CAD as a second reader, while simultaneously providing the highest sensitivity for detection of all nodules among the three methods [325].

A technique of volume rendering that uses different densities to depict vessels and nodules has been suggested as superior to MIP [326]. However, such a volume rendering technique has not been used widely, whereas a MIP function is now quite easily integrated into 3D reconstruction packages within a PACS system such that a radiologist can view MIPs as part of his or her reporting workflow.

CT reporting using cine review of softcopy images on a workstation ("scrolling") is now the norm. Thus, intuitively it can be expected that an increased cine frame rate would result in a decreased sensitivity for pulmonary nodule detection. Copley et al. investigated four different cine speeds (1, 5, 10 and 15 frames per second) for a single craniocaudal evaluation of the lungs for pulmonary nodules. They found that there was a trend towards reduced detection of pulmonary

nodules with increasing cine speeds that was independent of observer experience (expert radiologists or junior residents formed their reading panel); however, this trend was not statistically significant and no MIPs were used during the reading process [327].

The presence of co-existent pulmonary disease also affects perception of nodules, but leads to errors of interpretation rather than detection. In a study of 32 missed lung cancers in a Japanese LDCT screening programme, 13 cancers that had mean diameters of 15.9mm (range 6-26mm) were reported but misinterpreted as findings other than lung cancer on 16 CT scans (i.e. 16 lesions wrongly interpreted). 14 out of these 16 lesions had features that mimicked benign disease, and were associated with underlying tuberculosis, emphysema, fibrosis, silicosis and asbestosis [328].

1.6.2.2 Influence of psychophysical factors

Psychophysics is the exploration of the relationship between physical image quality and diagnostic performance [329]. Much of the literature informing this field has involved plain film radiography, and not CT. However, a brief overview of psychophysical factors is useful in understanding factors affecting reader performance.

The fundamental unit of visual processing is a fixation, during which the image projected onto the retina is processed. Various models to explain how such visual processing occurs have been devised, but none comprehensively explains visual processing. As discussed earlier, Kundel et al. took the view that perception involves *search*, *pattern recognition*, and *decision-making*. Intertwined with these perceptual

components are *object knowledge* and *background knowledge* [281]. *Object knowledge* influences pattern recognition: the observer must be aware of the features of the target lesion before a search for this can begin. *Background knowledge* refers to understanding of the features of the normal structures, such as the expected anatomy and course of bronchi and arteries, or the location of fissures. In this way, any object that is not part of the "background" normal lung would be perceived as abnormal.

A similar model put forward by Lesgold et al. describes *cognitive* and *perceptual* components of search [330]. The *perceptual component* controls where the eye moves, but depends on the knowledge base, which is informed by the *cognitive component*. This is analogous to object and background knowledge, because the cognitive component creates a prototype of "normal" such that deviations from that prototype can be analysed to see if it fits a target.

Visual search can also be understood using the "global-focal" model, which hypothesizes that the two processes of *global analysis* and *focal feature analysis* interact during visual search [331]. *Global analysis* uses sensory input from the whole retina to give a general overview of a scene, while *focal feature analysis* is concentrated on the detailed interpretation of the sensory input in the central retinal fields, along the axis of gaze [329]. These two processes can occur during scanning an image either during a single fixation, or a fixation cluster (a number of fixations at a particular location of the image). Pattern analysis occurs during global analysis, while integration of all features occurs to give a detailed analysis during focal feature analysis.

The "global-focal" model can be integrated into another decision-making model of visual perception, the Gregory-Rock model (Figure 1.14). In this model, global and focal feature analyses inform a rapid phase of "literal" or "bottom-up" perception, while a slower phase of "preferred" or "top-down" perception is informed by knowledge of the outside world, similar to the object knowledge described above [329].

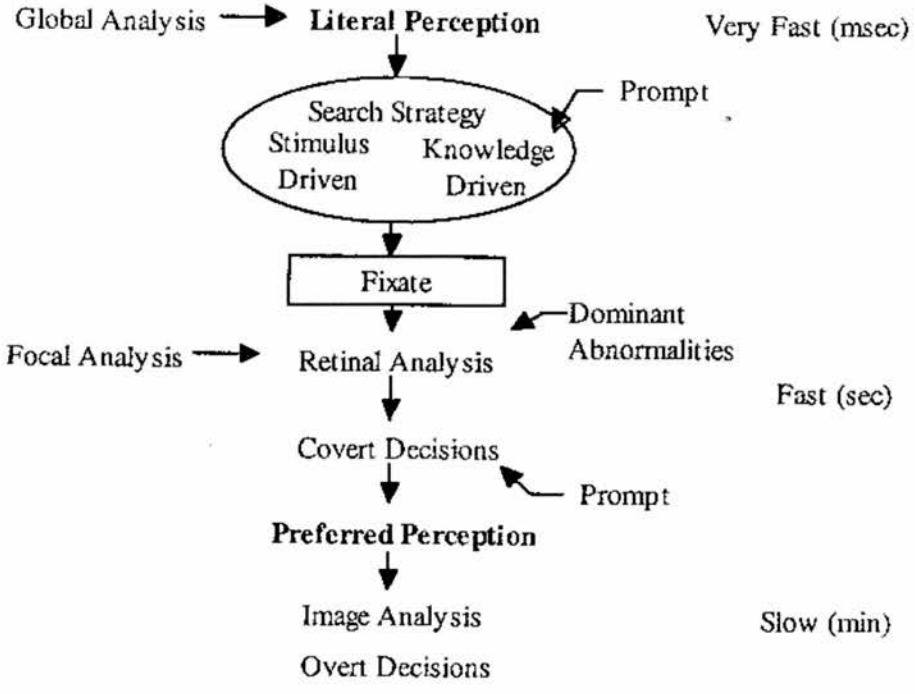

Figure 1.14. A model of visual perception that integrates the "global-focal" and Gregory-Rock models. (From reference [329].)

Satisfaction of search is an important source of errors of detection. Satisfaction of search describes the phenomenon where in the absence of a specific target, a general free-search (i.e. undirected) visual task is performed, with no defined expectation as to the outcome of the task [332]. As such, the observer has no

reference point from which he or she can decide to call a study normal, and may selectively divert attention to different fixation clusters, thereby missing some targets. If there are an unknown number of targets, the detection of one target will make the detection of subsequent ones less likely [333]. Of course, CAD readers do not suffer from satisfaction of search.

1.6.2.3 Influence of the number of readers

Only a handful of studies have evaluated the influence of double- versus single-reading on pulmonary nodule detection, and with the exception of one recent study, these have looked at non-screening populations. Wormanns et al. evaluated the performance of three readers in reading scans of nine patients with pulmonary metastases [334]. They simulated double-reading on both standard-dose (SDCT) and low-dose (LDCT) by using all possible pair combinations of the three readers. Sensitivity for nodule detection (the majority of which were less than 5mm) increased from 63% to 74% on SDCT and 64% to 79% on LDCT for single and double readers respectively. However, they concluded that 20% of nodules still remained undetected.

A number of European randomized screening trials are using double-reading with consensus to evaluate pulmonary nodules, but only the NELSON trial has provided an analysis thus far. Wang et al. retrospectively looked at 74 lung cancers that were detectable at baseline (prevalence) screening [174]. At that stage of screening the NELSON trial used two readers: a local radiologist and a second unblinded central radiologist. In discordant cases consensus was attempted, and an expert third radiologist's opinion was used in the event of non-consensus. They

compared the detection of nodules using this double-reading process to the findings of the initial single reader, and found that although the detection rate of nodules increased by 19%, the cancer detection or recall rates did not increase significantly, with 2.7% additional subjects with lung cancer found. They concluded that there was no benefit in using consensus double-reading, because the use of volumetry in the NELSON nodule management algorithm is more objective. However, this conclusion was probably premature, due to their small sample size (as the authors concede) and also because in order for volumetry to be performed, all important nodules must already have been detected. In other words, volumetry helps with interpretation, but not with detection itself.

1.6.2.4 Influence of reader experience and education

The visual processes described in section 1.6.2.2 occur differently in an untrained observer and an experienced radiologist. By studying eye positions, Kundel and La Follette found that the scan paths of first year medical students were localised to central parts, whereas trained radiologists used a circumferential pattern of scanning while simultaneously moving their eyes to a target lesion faster [335]. As the majority of missed lung nodules on chest radiographs are fixated in the central visual field, this difference in perception between expert and novice may have implications for nodule detection rates [281].

However, the effect of the level of reader experience in nodule detection on CT is still debatable. For instance, Marten et al. found that two experienced radiologists outperformed two inexperienced radiologists in nodule detection [336]. In contrast, Awai et al. found no significant difference in nodule detection sensitivity between

five board-certified radiologists and five radiology residents with only two years of training [337]. They argued that the residents would have accumulated sufficient cross-sectional anatomical CT knowledge to perform a nodule detection task as well as an experienced radiologist. In two other studies, the most inexperienced reader in fact had the highest sensitivity on axial (1-1.25mm slice thickness) image review [325, 338].

Thus far, there are no studies evaluating the role of radiographers in pulmonary nodule reading on LDCT. The lack of radiographer studies is not surprising, given that radiographers do not undertake CT reporting. However, as LDCT lung cancer screening is actually a very specific task, it is useful to look at the feasibility of radiographers in LDCT screening, especially when considering the limited evidence for improved detection with increased experience.

A precedent for radiographer reading exists in mammographic screening, where several countries, including the UK and Netherlands, involve trained radiographers in reading screening studies [339-341]. These radiographers (or advanced practitioners) undergo continuous self- and external-assessment. A study from the Netherlands compared the effect of two periods of screening with different reading methodologies. The first period involved double-reading by two radiologists who read simultaneously to reach consensus. In the second period, independent double-reading by radiographers (reading simultaneously) followed by an independent double-reading by two radiologists (who were blinded to the radiographers' opinions) was performed, and results were compared with the first period [341]. In the event that no consensus could be reached by the radiologists, a third radiologist would provide arbitration. Cases were referred for workup if any single reader (either

radiographer or radiologist) remained adamant that workup was necessary. The radiographer reading resulted in a higher referral rate (1.43% vs 1.02%) and a non-statistically significant higher cancer detection rate per 1000 women screened (5.25 vs 4.86), but decreased the positive predictive value due to the higher number of false positives. Thus, an independent double-reading radiographer approach can increase the sensitivity of a breast cancer detection programme, but at the expense of diminished specificity.

1.6.2.5 Influence of methods of consensus and arbitration

Traditionally discrepancies in observer readings have been dealt with either by:

- combining radiologists' readings (which assumes all discordant readings are due to differences in detection) [334, 342, 343];
- a process of consensus where readers discuss individual discrepancies [159]. The latter method deals with both discrepancies due to detection and interpretation, and is commonly accepted both as a valid method of resolving discrepancy, and of deriving the reference standard in radiological studies [208, 344]. However, consensus interpretation is itself subject to inherent limitations such as "groupthink" [344];
- using an independent "expert" arbiter (where one [334], two [289, 290, 293, 294, 336, 337] or even three [345, 346] experts have been used);
- using "internal" arbitration where a reader is independently shown and asked to form an opinion on nodules identified by other readers and not by themselves; or

- using a combination of these methods.

Thus far, few studies have evaluated the impact of different methods of consensus and arbitration on observer performance.

1.6.3 Comparisons between radiologists and computer-aided detection (CAD)

Multiple studies have been performed assessing the role of CAD in combination with radiologists in pulmonary nodule detection (Table 1.12). While the results of the individual studies vary, there are four important conclusions that can be drawn from these studies:

1. CAD is not effective as a first reader (whereby a CAD reading is performed and then presented to a radiologist for acceptance or rejection of CAD findings, without the radiologist also independently reading the study);

2. CAD is effective as a second reader (whereby a CAD reading is performed after an initial radiologist's independent reading, so that the radiologist can then accept or reject any additional CAD findings);

3. The beneficial effect of CAD holds for both inexperienced and experienced readers, but may be greater when used in combination with an experienced reader; and

4. The sensitivity of CAD increases with increasing nodule diameter.

Author	Journal	No. Of Nodules	Sensitivity of CAD
Armato et al. [347]	Medical Physics 2001 28(8):1552-1561	171	70%
Ko et al. [348]	Radiology 2001 218(1):267-273	370	81%
Wormanns et al. [288]	European Radiology 2002 12(5):1052-1057	153	38%
Rubin et al. [289]	Radiology 2005 234(1):274-283	195	86%
Bae et al. [349]	Radiology 2005 236(1):286-293	164	95%
Kim et al. [350]	Radiology 2005 236(1):295-299	126	95%
Marten et al. [351]	Clinical Radiology 2005 60(2):196-206	135	76%
Yuan et al. [352]	American Journal of Roentgenology 2006 186(5):1280-1287	628	73%
Das et al. [291]	Radiology 2006 241(2):564-571	116	74%
Sahiner et al. [353]	Academic Radiology 2009 16(12):1518-1530	241	54-76%[1]

Table 1.12. A selection of trials investigating the sensitivity of computer-aided detection.

[1]Sensitivities in this study were 54%, 64%, 68% and 76% for nodules of 3, 4, 5 and 6mm, respectively.

CAD can potentially be calibrated to a particular specificity and sensitivity, depending on the desired goal (increased detection or minimization of false positives). For example, in a study by Rubin et al., independent double-reading by radiologists was simulated and compared to CAD as a second reader [289]. Simulated double-reading by two radiologists resulted in a mean of 2.8-3.0 false positive detections per patient (depending on the reader pair), and improved sensitivity to 63% on average, from 50% for a single reading. In comparison, the

mean sensitivity for a radiologist-CAD first reader-second reader modelled combination, assuming all true positive CAD detections were accepted by the radiologist, would be 76%, 79%, 84% and 85% for maximum false positivity thresholds of 3,5,10 and 15 false positive detections allowed, respectively. Thus, even at the strict threshold of 3 false positive detections allowed for CAD, the sensitivity of the reader-CAD combination was higher than that of a radiologist alone.

There are also conflicting reports of the varying benefit of CAD according to reader experience. Marten et al. investigated four radiologists, two experienced (8 and 6 years experience) and two inexperienced (residents with 6 months experience) who read 18 MDCT scans containing a total of 96 nodules according to their reference standard [336]. Thereafter, a CAD evaluation was performed. CAD significantly outperformed the inexperienced radiologists, but did not significantly outperform the experienced radiologists. Combinations of the readers and CAD revealed that an experienced radiologist-CAD combination significantly outperformed a single reader and an experienced-inexperienced reader combination, while an inexperienced reader-CAD combination did not outperform an experienced reader or CAD. The authors concluded that an inexperienced reader-CAD combination was thus probably inadequate as an alternative to a single experienced reader. However, the readers in their study were not allowed to take advantage of image manipulation tools such as MIPs, and so may have been at a disadvantage. The study sample was also small, and studies were performed at standard and not low-dose. Another study with a much larger population has illustrated the utility of CAD as a second reader for both experienced and inexperienced radiologists [293].

Recently, there has also been interest in using CAD as a concurrent rather than second reader. In concurrent reading, the first round of radiologist reading is removed; instead, the study is read by the CAD, the radiologist then reviews the CAD marks, and confirms or deletes them accordingly, followed by an independent reading by the radiologist. This method has undergone only limited evaluation, with conflicting reports with regards to its effect on reading times and sensitivity [294, 354].

There is thus scope for further investigation of the optimum reader paradigm where an assisted reader, be it CAD or a radiographer, is involved.

1.6.4 Economic aspects

The question of the most cost-effective reading strategy goes hand-in-hand with analyses of reader performance, both in screening and regular clinical practice. Some evidence for the cost-effectiveness of double-reading again comes from mammographic screening. For example, in an analysis from the Netherlands, Groenewaud et al. suggest that while a double-reading with a "referral if any reader suggests" strategy could result in four times as high referral rates and an increase of biopsies or other invasive procedures, the cost-effectiveness of €4,190 per life-year gained may still be economically acceptable [105]. Interestingly, a recent analysis from the UK concluded that at current false positive rates prompting recall, CAD in mammographic screening would not be a cost-effective alternative to double-reading [355]. The issue of cost-effective reading strategies remains largely unexplored in

LDCT screening but needs addressing before a national lung screening programme can be recommended.

1.7 Aims of thesis

This thesis aims to identify an optimum and pragmatic reading strategy in lung cancer screening using computed tomography, by investigating:

1.The training of radiographers as readers;

2.The performance of radiographers in direct comparison with radiologists;

3. The use of radiographers as concurrent readers to assist radiologists;

4. The performance of radiographers compared to a CAD system; and

5.The impact of double- and triple- reading strategies, and of different methods of arbitration for discordant findings, on radiologists' performance.

CHAPTER 2: GENERAL METHODOLOGY

2.1 Introduction

This chapter describes the methods used to construct the training and evaluation datasets used in Chapters 3 and 7, the recruitment and selection criteria for the United Kingdom Lung Cancer Screening (UKLS) pilot trial participants that formed the study groups for Chapters 4 to 6, the image acquisition and post-processing techniques used, nodule reading protocols, and statistical methods. Methodology that was specific to a chapter will be described in that chapter itself.

2.2 Ethical approval for the UKLS

Ethical approval for the UKLS trial was granted by the National Research Ethics Service Committee, NHS R&D, the National Information Governance Board for Health and Social Care, and the Administration of Radioactive Substances Advisory Committee (ARSAC) (application no. 0521, reference ECC 2-02(a)/2011, approved 15.03.2011). In addition, all participating sites had undergone site specific assessment conducted by their local R&D department.

Informed consent had been obtained from all participants. In consenting to the trial, participants also consented to trial CT screening, investigations, treatment, follow-up and data collection. The investigations performed for this thesis formed part of the investigations into CT screening undertaken within the UKLS pilot trial.

2.3 Construction of a training dataset from the NELSON study

Two-hundred and sixty-three consecutive LDCT screening studies were initially procured from the NELSON study following approval from the NELSON investigators (see Appendix 1). These studies had been performed at Utrecht Hospital, Netherlands. Two-hundred and two of these studies were reported by NELSON radiologists to contain a total of 701 nodules (696 solid and five sub-solid nodules).

To ensure adequate representation of part-solid and nodules greater than $500mm^3$, this initial dataset was enriched by a further 10 cases containing 20 ground-glass nodules and 20 Category 4 nodules, also obtained from the NELSON study.

The selected NELSON cases were all pre-anonymised: each study had been given a unique reference number, and no patient identifiable data was visible. Data pertaining to image reconstruction parameters, and metadata, such as scanner type, were retained.

Volumetric datasets containing contiguous images were obtained via encrypted hard disk, in Digital Imaging and Communications (DICOM) format and in accordance with the data protection protocols of both the NELSON and UKLS studies. All studies were then transferred from hard disk to the internal hard drive of the digital workstation used at the three participating trial sites (Royal Brompton Hospital, London; Liverpool Heart & Chest Hospital, Liverpool; and Papworth Hospital, Cambridge).

2.4 Participants in the UKLS pilot study

2.4.1 Recruitment methods

Participants (male and female) aged 50 to 75 years of age were selected randomly from National Health Service (NHS) or Strategic Health Authority (SHA) records, and approached with an invitation letter and a questionnaire to assess their risk for lung cancer. The responses to the questionnaire were entered into the Liverpool Lung Project (LLP) Risk Prediction Model for each patient to calculate their risk of developing lung cancer [157, 157]. The LLP Risk Prediction Model calculates the absolute risk of lung cancer over a defined period, based on age, sex, smoking duration, family history of lung cancer, history of non-pulmonary malignant tumour, history of pneumonia, and occupational exposure to asbestos; it has been internally and externally validated [356].

2.4.2 Selection criteria

Inclusion criteria were those patients who have a 5% risk of developing lung cancer in 5 years, based on the LLP Risk Prediction Model. Exclusion criteria were those subjects who: were unable to give consent; had a co-morbidity that would unequivocally contraindicate either screening or treatment if lung cancer were detected; had a CT scan performed within one year of the invitation to be screened; were unable to lie flat; weighed greater than 200 kg (too large for the CT scanner).

2.5 UKLS Low-dose CT screening

2.5.1 Participating sites

LDCT scans were performed and read at two participating local sites (Liverpool Heart & Chest Hospital, Liverpool and Papworth Hospital, Cambridge). Scans were then transmitted to a central site (Royal Brompton Hospital, London - see section 2.6.2) for a second reading and arbitration in the event of discrepant findings.

2.5.2 Scanning technique

Participants' weight and height were ascertained prior to scanning to allow selection of appropriate exposure factors. Imaging was performed during suspended maximal inspiration. No intravenous contrast material was administered. The lung parenchyma (lung apices to bases) was scanned in its entirety in a single craniocaudal acquisition. The field of view (FOV) selected was the smallest diameter as measured from the widest point of the outer rib to outer rib large enough to accommodate the entire lung parenchyma (usually no more than 35cm). Thin detector collimation (0.5 – 0.625mm) was used with a pitch of 0.9-1.1. Exposure factors were tailored to patient height and weight (Table 2.1). CT dose index (CTDIvol) was kept below 4 mGy, with the effective radiation dose below 2 mSv. All appropriate dose modulation was used according to manufacturer's guidelines and local practice.

	Body weight (kg)		
	<50	**50-90**	**>90**
Voltage (kVp)	100	120	140
Tube current-time product (mAs)	Adjusted depending on scanner type to achieve target CTDIvol		
CTDIvol (mGy)	0.8	1.6	3.2
Effective dose taking into account scout view (mSv)	< 0.7	< 1.0	< 1.6

Table 2.1. Exposure factors used in the UKLS trial, modified from Baldwin et al. [157].

2.5.3 Image reconstruction

Images were reconstructed at 1mm thickness with 0.7 increment, using both moderate and high spatial frequency kernels.

2.5.4 Quality control

Ten randomly selected cases from both sites were reviewed every month at the central site with respect to adequacy of craniocaudal coverage, field of view, degree of inspiration, motion artefacts, radiation exposure parameters, radiation dose, and reconstruction algorithms. Feedback was provided to local sites.

2.6 Data collection

2.6.1 UKLS database

The UKLS database was managed and maintained by the UKLS project team. It was a web-based database that was built on the Nelson Management System

(NMS) design [155], and accessible upon registration with the UKLS project manager. Access rights were configurable such that readers in the study were not able to view the recordings of other readers unless they were involved in consensus or arbitration. The database held all trial information for the participants. The participant's surname and demographic details were visible to the reader to ensure accurate identification.

2.6.2 Data storage and retrieval

All scans performed at the local sites were transferred to the reading workstation at the local site (Syngo, Siemens Medical Solutions, Erlangen, Germany) as well as to the local Picture Archiving and Communications System (PACS). At the same time, the scans were also transferred via encrypted network transfer to the Image Exchange Portal (Burnbank Systems, Ipswich, England) and then downloaded onto the Syngo workstation at the central site, as well as to the central site PACS server, ensuring backup of data at both local and central sites.

A UKLS reader opened the study in "LungCARE", the volumetric segmentation package used for reading (LungCARE, version Somaris/5 VB 10A, Siemens Medical Solutions). Each detected nodule was evaluated and recorded in the UKLS database using a non-commercial database electronic soft-copy entry proforma (Artex Nodule input for UKLS version 4.4, Logiton, Netherlands) (Figure 2.1). Options for nodule categorisation and segment location were available from drop-down menus, while the slice position of the nodule was entered using free-text. Once a reading for a particular nodule had been completed, the information from the

proforma was copied and pasted into a structured DICOM report in an Extensible Markup Language (XML) format. The XML file was then transferred to a local network drive, so that the file could be uploaded to the patient's record on the UKLS database via a PC with web access. The XML file contained information regarding a nodule's size, table position location, lung and segment location, category and volume.

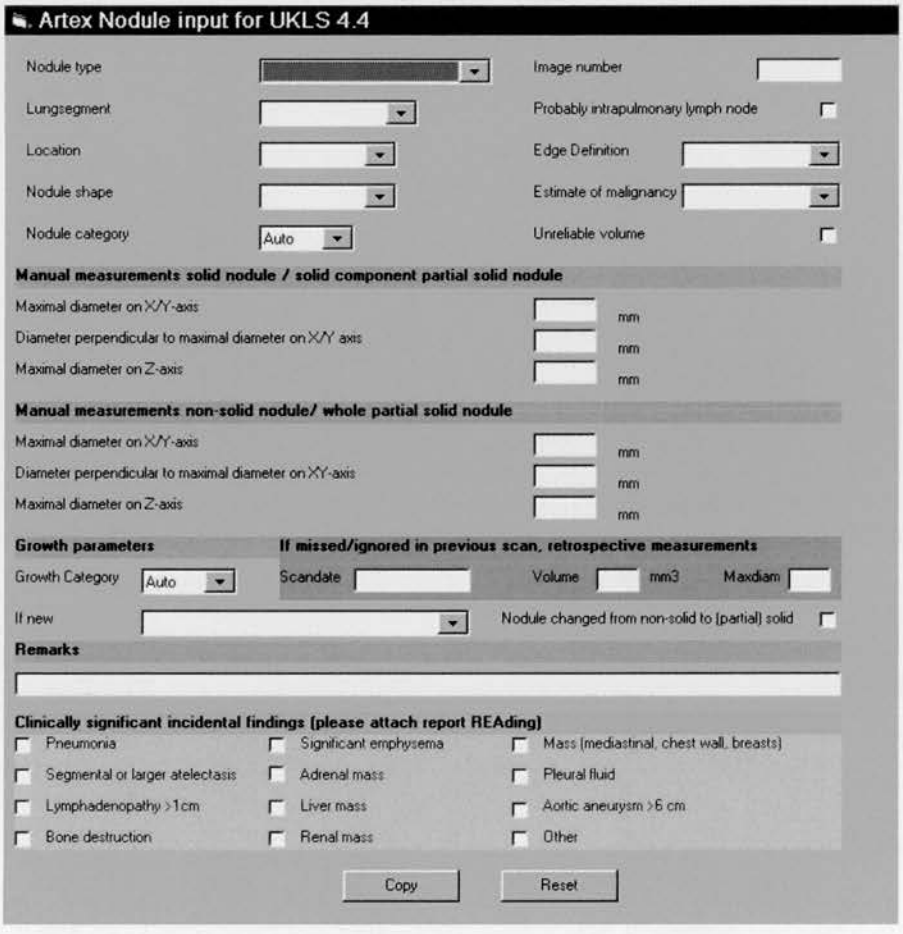

Figure 2.1. The database electronic soft-copy entry proforma used for UKLS nodule recording.

2.7 CT interpretation

2.7.1 Nodule definitions

Nodules were defined as follows:

Category 1 Benign nodules: Nodules fulfilling one of the following criteria; a benign pattern of calcification, fat, measuring less than 3 mm in diameter or volume < 15 mm^3. Intrapulmonary lymph nodes (IPLNs) if the following criteria were fulfilled: lie within 5 mm of the pleura, < 8 mm in diameter, smooth border, ovoid or non-spherical, and at least one interlobular septum radiating from surface of nodules.

Category 2 If solid and intraparenchymal with a maximum diameter of 3.1-4.9 mm or a volume of 15-49 mm^3; if solid and pleural or juxtapleural with a maximum diameter of 3.1-4.9 mm; if non-solid or part-solid with a maximum non-solid component diameter of 3.1-4.9 mm, and, where there was a solid component, this had a diameter of < 3 mm and/or volume of < 15 mm^3.

Category 3 If solid and intraparenchymal with a volume of 50-500 mm^3; if solid and pleural or juxtapleural with a diameter 5.0-9.9 mm; if non-solid with a maximum diameter of $\geq$ 5mm; if part-solid with a maximum non-solid component diameter of $\geq$ 5mm, and with a solid component volume of 15-500 mm^3 or solid component diameter of 3.0– 9.9 mm.

Category 4 If solid and intraparenchymal with a volume > 500 mm^3; if solid and pleural or juxtapleural with a diameter of $\geq$ 10 mm; if part-solid and the solid component had a diameter of $\geq$ 10 mm or had a volume > 500 mm^3.

2.7.2 Semi-automated volumetric nodule segmentation method

The volumetric segmentation method was performed using LungCARE (Siemens Medical Solutions, Erlangen, Germany) and is similar to the technique described in other studies using this software [313]. First, the reader marked a candidate nodule with a mouse click. Then the programme automatically defined a volume of interest around the candidate nodule, which could be further analyzed by using volume-rendered displays or a coronal reformation. The candidate nodule could then be either approved or discarded. The evaluation of a nodule with a second mouse click initiated an automated volume measurement programme. The programme used a fixed-attenuation threshold of -400 HU to extract a three-dimensional connected "structure of interest". This structure of interest consisted of the nodule and, if present, connected structures such as vessels or parts of the chest wall. Subsequently, a small spherical three-dimensional template that originated from the click point was gradually expanded; its cross-correlation with the segmented nodule was computed for each step. The peak value of the cross-correlation curve was determined, and an empirical cut-off value close to the peak value was used to separate the nodule from potential adjacent structures. In this manner, an optimum three-dimensional template was generated. Finally, the nodule was segmented by fusing the optimum three-dimensional template and the structure of interest; this was followed by spatial reasoning to remove adjacent structures. The segmented nodule was then shown in yellow on the volume-rendered display of the volume of interest (Figure 2.2).

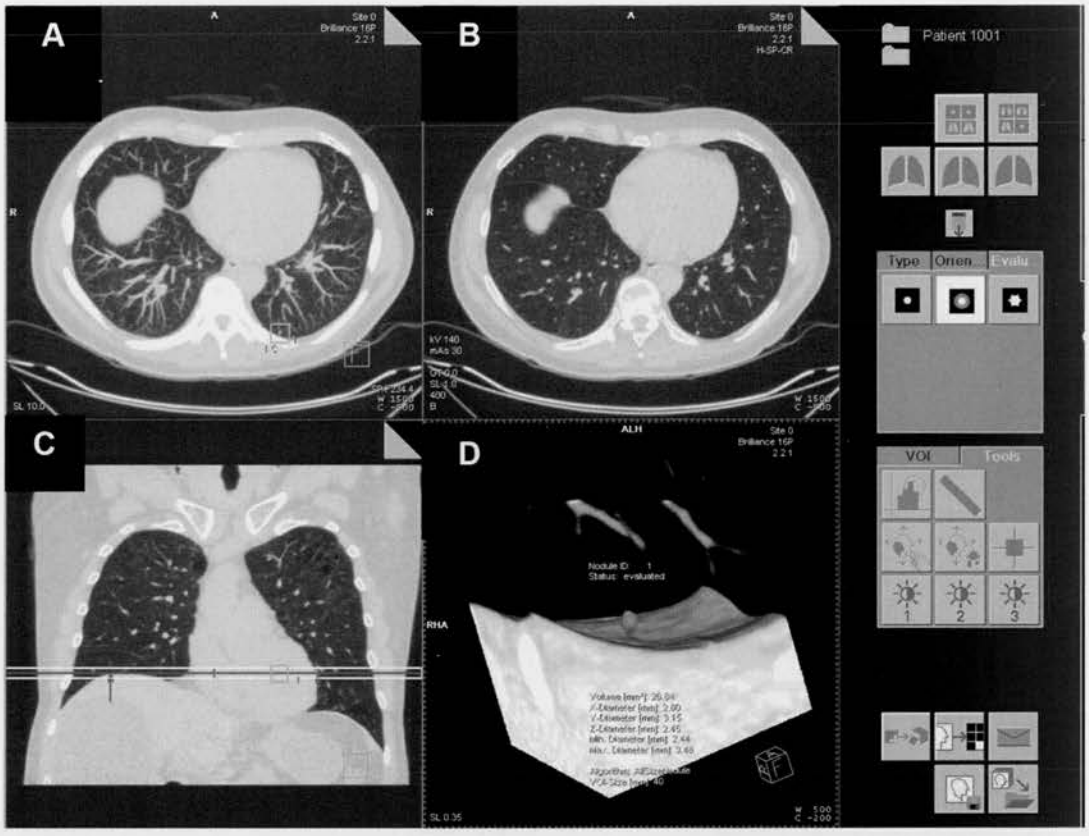

Figure 2.2. Example of marking and annotation of a nodule within LungCARE by a radiologist on an anonymised LDCT study. The CT is presented in a 2 x 2 viewing partition with a default window setting level -500 HU, width 1500 HU, and arranged in the following manner: (A) top left panel, maximum intensity projections (MIPs) with a default setting of 10mm thickness, showing a nodule marked with a yellow outline box in the left lower lobe; (B) top right panel, 1mm-collimation axial images; (c) bottom left panel, 0.7mm-collimation coronal images; and (D) bottom right panel, displaying volumetric segmentation of a nodule once that nodule is selected with a mouse click.

2.7.3 Follow-up protocols

The follow-up protocols for the nodules were based on initial volume and volume doubling time (VDT), as shown in Figure 2.3.

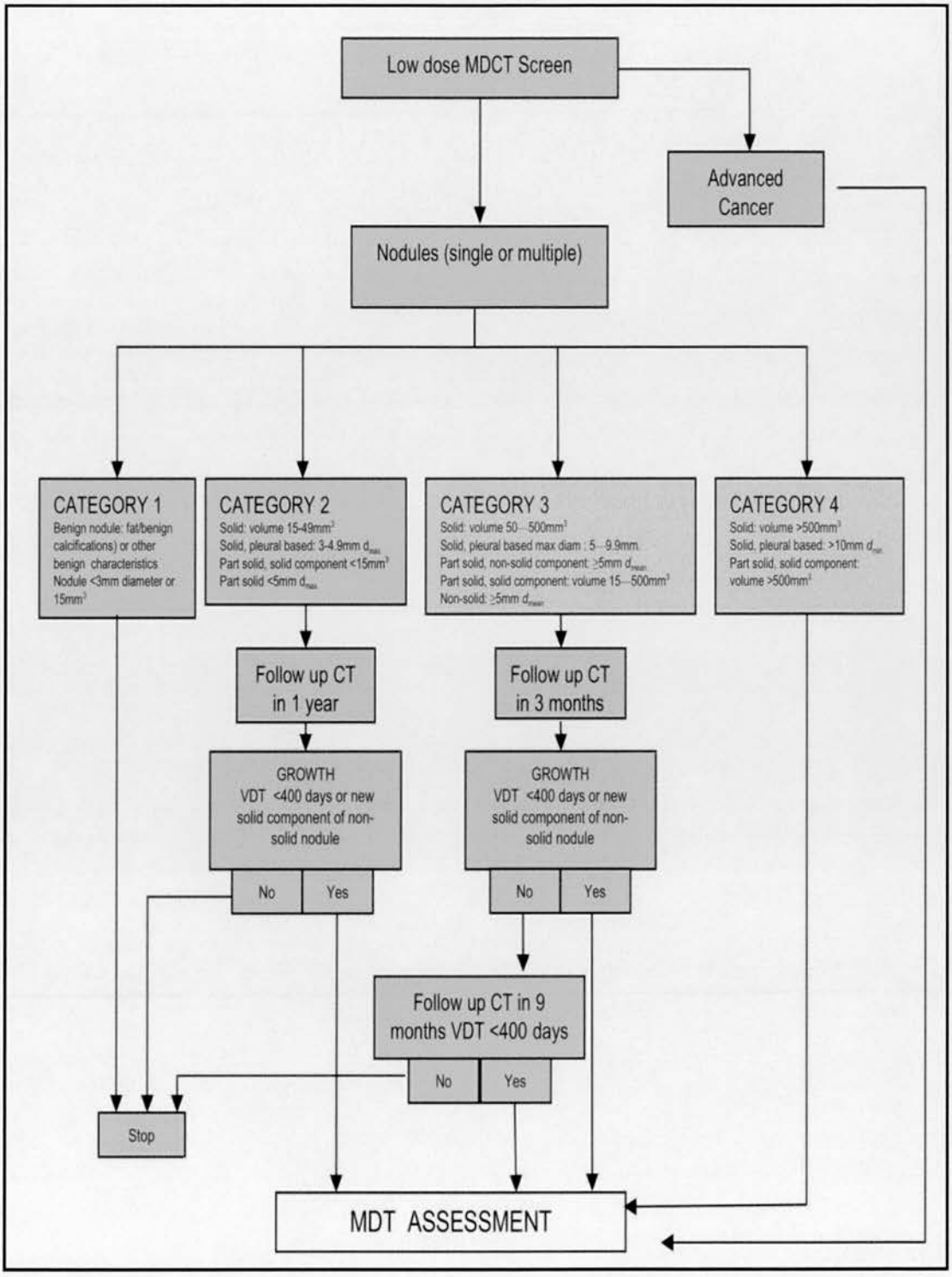

Figure 2.3. Follow-up algorithm of the UKLS trial, from Baldwin et al. [157]. MDT= multidisciplinary team.

2.8 Statistical methods

2.8.1 Measures of observer performance

In investigations of image interpretation, the observer, rather than the imaging equipment itself, may be considered as the diagnostic test under scrutiny [357].

2.8.1.1 Sensitivity

Sensitivity is the proportion of subjects with a condition (for example, a pulmonary nodule) who have a positive test [97]. It is calculated by the formula:

$$\frac{\text{(true positives)}}{\text{(true positives)} + \text{(false negatives)}} \times 100$$

The higher the sensitivity, the lower the proportion of false negative cases. As such, a highly sensitive test can be considered a good "rule out" test.

2.8.1.2 Specificity

Specificity is the proportion of people without a disease who have a negative test [97]. It is calculated by the formula:

$$\frac{\text{(true negatives)}}{\text{(false positives)} + \text{(true negatives)}} \times 100$$

The higher the specificity, the lower the proportion of false positive cases A highly specific test can be considered a good "rule in" test.

Sensitivity and specificity are measures of a diagnostic test which are unaffected by disease prevalence.

2.8.1.3 Positive predictive value

The positive predictive value of a test is the probability of disease given a positive test [358]. It is calculated by the formula:

$$\frac{\text{(true positives)}}{\text{(true positives)} + \text{(false positives)}} \text{x}100$$

2.8.1.4 Negative predictive value

The negative predictive value of a test is the probability of absence of disease given a negative test [358]. It is calculated by the formula:

$$\frac{\text{(true negatives)}}{\text{(true negatives)} + \text{(false negatives)}} \text{x}100$$

Unlike sensitivity and specificity, both positive and negative predictive values are affected by the prevalence of disease.

2.8.1.5 Average false positive detections per case

The average number of false positive detections (FPs) per case for a given reader is calculated by dividing the total number of FPs by the total number of cases read. The average FPs per case is a parameter that is often quoted for computer-aided

detection (CAD) systems, when it is usually described as the "false positive rate" [359]. However, as explained in section 1.2.2.2, the false positive rate is actually defined as the probability of obtaining a false positive result, and is given by (1-specificity) [97]; use of the term "false positive rate" in the CAD literature is therefore potentially confusing. As such, the terms "average number of false positive detections per case" or "average FPs per case" will be used in this thesis with respect to false positive detections, for the purposes of clarity.

2.8.2 Comparison between groups

The type of test used to compare the differences between groups depends on whether the data are:

- independent or paired (i.e. related);
- categorical or continuous; and
- normally or non-normally distributed.

2.8.2.1 Categorical data

The *chi-square test* of homogeneity is used to compare two groups of subjects that have been sampled from two independent populations and a binary outcome (e.g. positive or negative) is used for classification [360]. It tests the null hypothesis that there is no difference between the observed and expected frequencies of a result. The expected frequency is calculated according to the null hypothesis in each cell of a 2×2 contingency table for the binary outcomes. If the expected

frequencies are close to the observed frequencies, the model according to the null hypothesis fits the data well; thus, the null hypothesis should not be rejected.

However, the chi-square test may not be accurate when sample sizes are small [360]. In such instances, *Fisher's exact test* is preferred. Although there is no definite cut-off defined for when the Fisher's exact test should be used, a useful rule is to apply this test when the total sample size is less than 30 [360] or the expected cell count within any cell in the 2 x 2 contingency table is less than five [361].

McNemar's test is used to compare paired observations of categorical data, for example the comparison of the observations of different readers on the same subject. The data are not independent (as they are being performed on the same subject) and thus the chi-square test would not be appropriate. Unlike the chi-square test, the McNemar's test only considers those pairs of observations which are discordant (e.g. true positive and false negative, false positive and true negative), while ignoring the concordant pairs. For such situations, the null hypothesis is that the proportions of positive results are the same for two observations by two different readers, versus the alternative hypothesis that they are not the same [360].

2.8.2.2 Continuous data

The *independent samples t-test* is a parametric test for comparing the means between observations from independent samples [362]. The non-parametric equivalent is the *Mann Whitney U test* [363]. When observations are paired, the *paired student's t-test* and the *Wilcoxon signed rank test* are the parametric and non-parametric equivalents for comparing the mean difference, respectively [364].

2.8.3 Measures of agreement

Observer variation of categorical data, such as in the classification of nodules, can be measured by the kappa coefficient (multirater κ), which represents the agreement obtained between two readers beyond that expected by chance [365]. The kappa coefficient is calculated by

$$\kappa = P_A - P_E / 1-P_E$$

where P_A is the proportion of cases for which agreement exists, and P_E is the proportion of cases in which raters would agree by chance. The *weighted kappa coefficient* adjusts for the extent to which disagreement occurs when a grading system consisting of more than two categories is used [365].

CHAPTER 3: TRAINING RADIOGRAPHERS IN LUNG NODULE DETECTION

3.1 Introduction

Radiologists are arguably best suited to the task of lung nodule identification on CT because they have medical knowledge, an understanding of CT anatomy, and reading experience. However, it could be reasoned that these attributes help mainly with the interpretation and the assignment of some level of clinical significance to a particular finding, rather than to the detection of that finding itself. As discussed in section 1.6.1, the detection of a lung nodule is mainly a matter of perception. If the task of detection could be reliably and consistently performed by a suitably trained non-radiologist, there would in theory be a greater pool of readers available for CT lung cancer screening. Radiologists could then focus on the important tasks of interpretation of findings, recommendations for follow-up and arbitration, rather than on nodule detection alone.

An alternative reader to a radiologist is a radiographer (or technologist as they are referred to in the USA). Radiographers are an obvious first choice for this task, as they have a basic understanding of both technical and anatomical aspects of thoracic CT. Radiographers have already been assessed as readers in screening mammography [339, 366] and screening CT colonography [367, 368]. After a period of training, key performance indicators can be identified for such tasks [369]. However, no precedent for radiographers reading thoracic CTs currently exists, and

therefore the optimum type and duration of training they require to develop the skills necessary for this task is unknown.

It is generally accepted that feedback provided by tutors is integral to medical training [370]. Feedback is a process that allows two-way interaction between a learner and tutor, so that learning takes place in a non-evaluative and non-judgemental environment, allowing both the learner and tutor to identify areas for modification and reinforcement in their learning.

The aim of the current investigation was to evaluate the effect of continuous feedback as a means of training radiographers in the task of lung nodule detection.

3.2 Materials and Methods

3.2.1 Training dataset and subsets

The methods of CT acquisition, procurement of scans from the NELSON study, and data anonymisation and storage have been described in Chapter 2. From the 202 CT studies initially procured for the training dataset (section 2.3), 100 studies containing a variable number of nodules were chosen at random. The training dataset was arbitrarily divided into 10 training subsets, each containing 10 CT scans. Each of these scans had been read by one NELSON reader at the local scanning site, a NELSON reader at a central site, and in cases where there had been disagreement, by a third NELSON reader. The final consensus answer for the nodules identified by the NELSON readings in these studies had all been recorded on a spreadsheet that accompanied the CTs from the NELSON database. Each nodule identified by the

NELSON reading was visually confirmed by one of two thoracic radiologists [Dr A. Devaraj (AD) or me (AN)]. As the NELSON nodule protocol did not specifically discriminate between intrapulmonary lymph nodes (IPLNs) and nodules, these radiologists (AD or AN) also recorded IPLNs.

3.2.2 Selection of radiographers

Four radiographers who were able to commit at least four hours a week over a five-month period were selected as readers. All four radiographers had experience in thoracic CT scan acquisition and thus had a basic understanding of acquisition parameters and thoracic CT anatomy. Readers 1 to 3 had 10 or more years' radiography experience, while reader 4 had 4 years.

3.2.3 Introductory tutorial

The radiographers were first given a PC-based presentation (prepared by AD) covering the basic principles of thoracic CT anatomy and low-dose CT acquisition. They were then shown 20 CTs containing examples of different types of lung nodules as well as focal opacities such as pleural thickening and atelectasis that can mimic lung nodules. They were provided with the UKLS definitions of lung nodules, as described in section 2.7.1. The images of these various opacities were obtained from 20 CT studies from the original pool of 202 LDCT studies that had been obtained from the NELSON study, but that had not been included in the 100 scans forming the training dataset for the current study.

The radiographers were also tutored in the use of the LungCARE software on the Syngo workstation (Siemens Medical Solutions, Erlangen, Germany), and were trained to use the software's measurement tools as well as image manipulation tools such as magnification and maximum intensity projection (MIP), to ensure they were sufficiently able to optimise their nodule detection technique.

3.2.4 Radiographer reading

Each radiographer was provided with the unique identifier numbers of the 10 anonymised CT studies comprising each training subset. Radiographers were only given one subset at a time to analyse, and were all given the CTs and the subsets in the same order.

Radiographers were asked to identify all category 2, 3 and 4 nodules, and only to identify category 1 nodules if they met all criteria for intrapulmonary lymph nodes (see section 2.7.1). For each nodule, radiographers were asked to record the slice position, location (right or left lung), and category of the nodule. The radiographers performed semi-automated volumetric analysis on any segmentable solid nodule or solid component of a part-solid nodule on the LungCARE software. For non-solid nodules, the radiographers categorised the nodules based on the maximum diameter. All recordings were entered on a spreadsheet created using Microsoft Excel (version 2007, Microsoft Corp., Redmond, CA).

The time taken for each radiographer to complete each subset of 10 scans was recorded to the nearest 0.25 hours.

3.2.5 Evaluation of radiographer answers and feedback to radiographers

After each radiographer had completed a subset of 10 scans, one of the two radiologists (AD or AN) reviewed each opacity that had been identified by the radiographers. All opacities were classified by the radiologist into one of four levels (Table 3.1). This process was performed in the presence of that radiographer, and so acted as feedback for the radiographer, providing him or her with the opportunity to ask questions on a case-by-case basis. The radiographers could only move on to the next subset of 10 scans after completion of this feedback process.

Classification level	Description	Explanatory notes
1	Agreement with NELSON readings	If a nodule recorded by the radiographer matched the slice position, location and size (volume or diameter) of the NELSON reading, the nodule was recorded as being in agreement.
2	Extra nodule	If a genuine nodule that was not part of the NELSON reading was identified by the radiographer, it was considered an extra positive finding.
3	Missed nodule	A verified nodule (whether discovered on the NELSON reading or by at least one other radiographer) that had not been identified by the radiographer was considered a missed nodule.
4	"Overcall"	An opacity not thought to represent a nodule or a nodule that was too small (i.e. category 1 and not matching criteria for an IPLN) was considered an 'overcall'. The nature of the opacity was recorded by the radiologist (see Table 3.5).

Table 3.1. Classification of opacities identified by radiographers.

3.2.6 Reference standard

After four radiographers had completed the analysis of all 100 cases, all extra findings identified by at least one radiographer (but not by the NELSON readers) were reviewed by two radiologists in consensus. Thus in summary, the reference standard for this study consisted of:

1. An initial evaluation by at least 2 NELSON readers;
2. A validation of that evaluation by one of two participating radiologists (AD or AN) in this study;
3. A two-stage refinement process consisting of radiographer identification of extra findings and consensus interpretation of those extra findings by two radiologists.

3.2.7 Statistical analysis

The sensitivity, specificity, average number of false positive detections per case, and the ratio of true positive to false positive detections for each radiographer as compared to the reference standard was calculated for the 100 scans in total. The ratio of true positive to false positive detections per unit hour for each radiographer was also calculated. The sensitivity and specificity of each radiographer was also calculated for each subset of 10 scans. A two-tailed Fisher's exact test was used to compare the sensitivities and specificities from one subset to the next for a given radiographer. All analysis was performed using Medcalc (version 12.5.0.0, MedCalc

Software, Mariakerke, Belgium). A P value of less than 0.05 was assumed to be statistically significant.

3.3 Results

3.3.1 General data

All 100 subjects in the CT test cohort were male, aged 53-79 years old. A total of 417 opacities were identified in 91 out of the 100 CT studies. A range of 1 to 23 opacities were identified per scan. The range of opacities per subset of 10 studies varied between 14 and 79. Figure 3.1 illustrates the distribution of opacities by each set. In the remaining 9/100 studies, no opacities were identified by radiographers.

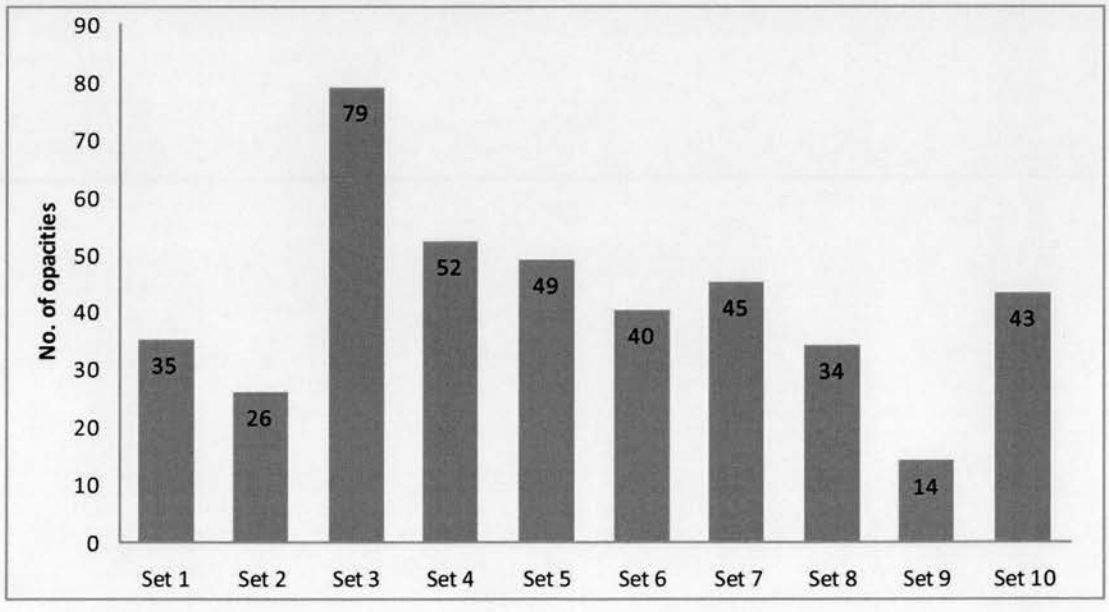

Figure 3.1. Number of opacities identified per subset of 10 CT studies.

The mean time taken for each subset by radiographers 1,2,3 and 4 was 3.1, 2.2, 2.4 and 2.3 hours respectively, corresponding to a mean time per subject of 18, 13, 14 and 14 minutes respectively.

3.3.2 Reference standard

A total of 282 opacities were considered identifiable nodules (IPLNs or category 2 to 4 nodules) in 83/100 CT studies. 76 (27.0%) were designated IPLNs, 150 (53.2%) category 2, 51 (18.1%) category 3, and 5 (1.8%) category 4 nodules. In total, 17/100 CT studies had no opacities designated as positive nodules in the reference standard; these included the 9 CT studies with no positive findings identified by any radiologist or radiographer, and 8 CT studies with at least one false positive finding by a radiographer.

3.3.3 Overall performance of individual radiographers

The overall performance of each radiographer over the 100 scans is shown in Table 3.2. The mean sensitivity and specificity were 60.5% and 68.4% respectively. The average number of false positive detections per case for all four readers was 0.46.

	Radiographer			
	1	2	3	4
Sensitivity (%)	78.0	75.5	39.7	48.6
Specificity (%)	74.3	65.3	70.1	63.9
Average no. of FP detections per case	0.37	0.50	0.43	0.52
Ratio of TP:FP detections	5.95	4.26	2.60	2.63

Table 3.2. Overall performance of individual radiographers over the 100 scans. TP= true positive; FP= false positive.

When the amount of time for task performance was considered, the ratios of true positive (TP): false positive (FP) detections per unit hour for radiographers 1, 2, 3 and 4 were 1.9, 1.9, 1.1 and 1.1 respectively.

3.3.4 Effect of feedback on radiographer performance

The sensitivities of each radiographer for each subset of 10 scans are illustrated in Table 3.3. Statistically significant increases in sensitivity were observed for radiographers 3 and 4 between subsets 4 and 5 only (increases of 38.3% and 40.3% respectively). Nevertheless, radiographers 3 and 4 showed a slight trend towards improvement in the later subsets (Figure 3.2).

	Radiographer							
	1		2		3		4	
Subset	Sens (%)	*P*	Sens (%)	*P*	Sens (%)	*P*	Sens (%)	*P*
1	86.4	NA	81.8	NA	22.7	NA	54.5	NA
2	73.7	0.44	68.4	0.47	42.1	0.31	31.6	0.21
3	84.8	0.31	73.9	0.76	34.8	0.59	34.8	1.00
4	65.9	0.05	75.0	1.00	20.5	0.16	27.3	0.50
5	82.4	0.13	67.6	0.61	58.8	**0.0008**	67.6	**0.0005**
6	82.1	1.00	85.7	0.14	35.7	0.08	60.7	0.60
7	75.9	0.75	72.4	0.33	48.3	0.42	69.0	0.59
8	79.2	1.00	75.0	1.00	58.3	0.58	54.2	0.39
9	81.8	1.00	63.6	0.69	36.4	0.29	27.3	0.17
10	72.0	0.69	88.0	0.17	48.0	0.72	60.0	0.15

Table 3.3. Sensitivities of radiographers for each subset. *P* values are for the difference in sensitivity between the particular subset and the one preceding it, using the two-tailed Fisher's exact test. *P* values in bold indicate statistically significant differences.
Sens = sensitivity. *P* = *P* value. NA = not applicable.

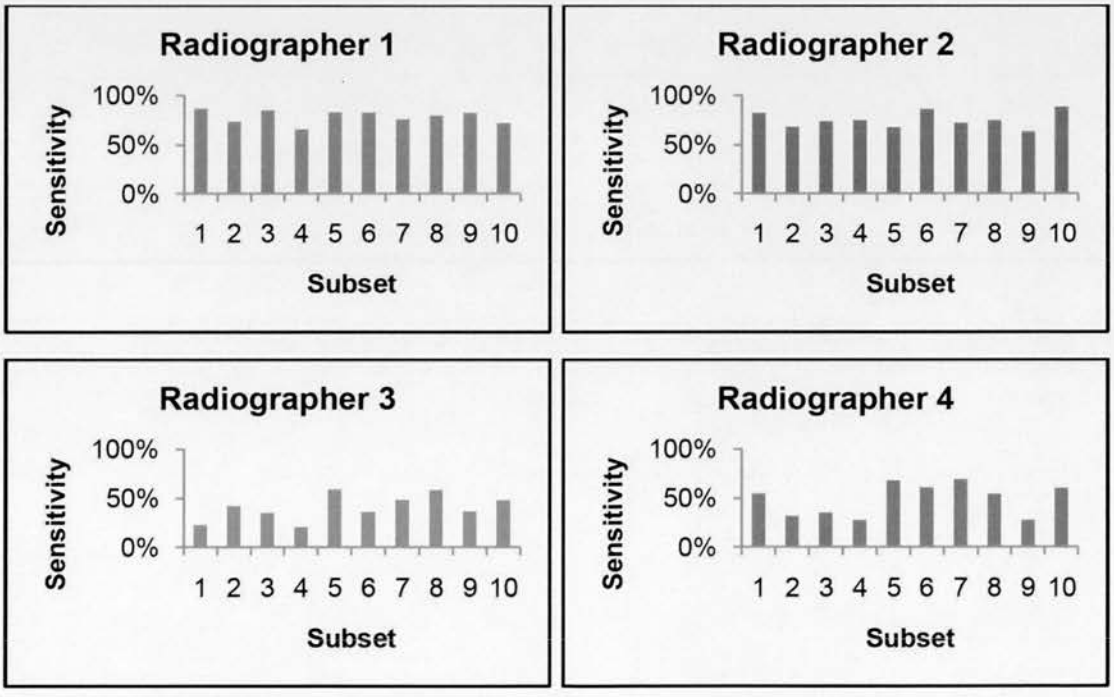

Figure 3.2. Variation in sensitivity between subsets for radiographers 1 to 4.

The specificities of each radiographer for each subset of 10 scans are illustrated in Table 3.4. Statistically significant decreases in specificity were observed for Radiographer 2 between subsets 3 and 4, 5 and 6, and 9 and 10 (47.7%, 33.3% and 65.0% respectively), and for Radiographer 4 between subsets 4 and 5 (73.3%). Statistically significant increases in specificity were seen for Radiographer 2 between subsets 4 and 5, and for Radiographer 3 between subsets 3 and 4 (75.0% and 42.4% respectively).

No particular trend was observed for the changes in specificity for any radiographer (Figure 3.3).

	Radiographer							
	1		2		3		4	
Subset	Spec (%)	*P*	Spec (%)	*P*	Spec (%)	*P*	Spec (%)	*P*
1	64.3	NA	64.3	NA	92.9	NA	50.0	NA
2	57.1	1.00	71.4	1.00	85.7	1.00	42.9	1.00
3	72.7	0.41	72.7	1.00	57.6	0.22	78.8	0.08
4	75.0	1.00	25.0	**0.03**	100.0	**0.04**	100.0	0.31
5	80.0	1.00	100.0	**0.0003**	66.7	0.12	26.7	**0.001**
6	50.0	0.13	66.7	**0.03**	75.0	0.69	58.3	0.13
7	75.0	0.24	62.5	1.00	56.3	0.43	50.0	0.72
8	100	0.11	58.3	1.00	58.3	1.00	75.0	0.25
9	85.7	0.37	100.0	0.11	85.7	0.33	85.7	1.00
10	80.0	1.00	35.0	**0.006**	70.0	0.63	70.0	0.63

Table 3.4. Specificities of radiographers for each subset. *P* values are for the difference in specificity between the particular subset and the one preceding it, using the two-tailed Fisher's exact test. *P* values in bold indicate statistically significant differences.
Spec= specificity. *P* = *P* value. NA = not applicable.

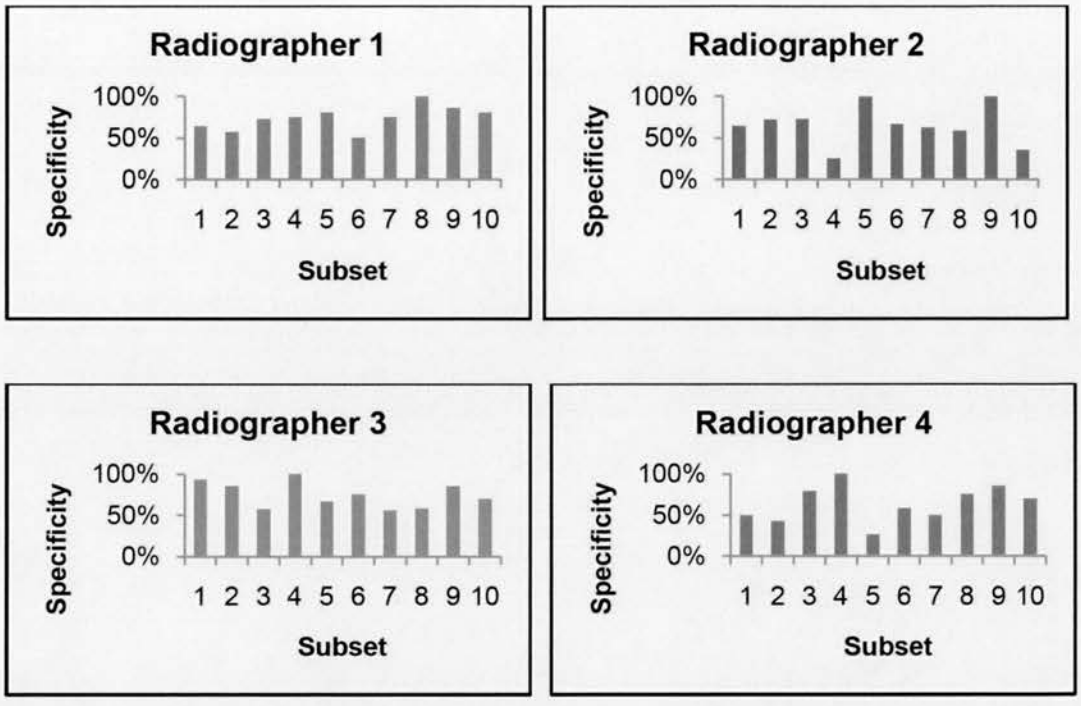

Figure 3.3. Variation in specificity between subsets for radiographers 1 to 4.

3.3.5 Characteristics of false positive findings

The types of false positive findings identified by each radiographer varied widely between each radiographer (Table 3.5). In general, false positive findings were most frequently due to a combination of pleural or fissural thickening, atelactasis, presumed scarring and overcalling vascular structures.

Type of false positive finding	Radiographer 1	2	3	4
	No. (%)			
Pleural/fissural thickening, pleural plaques	9 (24.3)	11 (22.0)	4 (9.3)	11 (21.2)
Atelectasis	6 (16.2)	7 (14.0)	9 (20.9)	9 (17.3)
Presumed scarring	7 (18.9)	7 (14.0)	10 (23.3)	7 (13.5)
Linear opacity	0 (0.0)	3 (6.0)	2 (4.7)	1 (1.9)
Calcified nodule	3 (8.1)	0 (0.0)	3 (7.0)	2 (3.8)
Nodule < 3mm in diameter	0 (0.0)	8 (16.0)	4 (9.3)	7 (13.5)
Vascular structure	7 (18.9)	10 (20.0)	9 (20.9)	8 (15.4)
Diaphragm/diaphgramatic fat	1 (2.7)	2 (4.0)	0 (0.0)	1 (1.9)
Mucus plug	2 (5.4)	2 (4.0)	1 (2.3)	3 (5.8)
Others (motion artefact / osteophyte-related opacity)	2 (5.4)	0 (0.0)	1 (2.3)	3 (5.8)
Total	**37**	**50**	**43**	**52**

Table 3.5. Types of false positive findings. Numbers in parentheses indicate the proportion of false positive findings in that category as a percentage of the total false positive findings for a particular radiographer.

3.3.6 Characteristics of false negative findings

Table 3.6 illustrates the wide variation in the types of nodules missed between the 4 radiographers. The majority of nodules missed by all radiographers were small (category 2) or IPLNs.

	Radiographer			
	1	2	3	4
Nodule category	No. (%)			
1 (i.e. IPLN)	18 (29.0)	17 (24.6)	45 (26.5)	38 (26.2)
2	41 (66.1)	39 (56.5)	94 (55.3)	84 (57.9)
3	3 (4.8)	11 (14.5)	28 (16.5)	22 (15.2)
4	0 (0.0)	2 (2.9)	3 (1.8)	1 (0.7)
Total	**62**	**69**	**170**	**145**

Table 3.6. False negative findings. Numbers in parentheses indicate the proportion of false negative findings in that category as a percentage of the total false negative findings for a particular radiographer.

3.4 Discussion

The mean and range of sensitivities for pulmonary nodule detection by radiographers in this study were largely comparable to the wide range (between 21% and 70%) reported for radiologists in the literature [289, 290, 293, 334, 336, 338, 371]. However, a learning effect on sensitivity and specificity over the sequential 10 subsets was not demonstrated. Training would intuitively be expected to enhance performance. In an article examining this issue, Wood stated that "with practice…the novice will build a mental library of patterns of normality and abnormality, along with a library of information. Frequent feedback and affirmation will ensure that this experience grows" [372]. The lack of a conspicuous learning curve in the current study may be due to the small number of scans in each subset, as well as varying levels of difficulty that may have existed within each subset. However, this level of difficulty is not something that is easy to quantify or standardise.

The fact that the first two readers had a higher sensitivity than radiographers 3 and 4 could be due to differences in inherent ability. Over time it might have been expected that the poorer sensitivity of readers 3 and 4 in comparison to readers 1 and 2 would have improved. Indeed, a small but inconsistent trend in this direction was observed for radiographers 3 and 4 in the later subsets, but their sensitivity still did not match that of the two most sensitive radiographers.

The lack of a learning curve has been shown in other detection performance studies. Burling et al. did not manage to find a learning effect for the selection of the correct management strategy or for lesion detection in an evaluation of five radiographers performing CT colonography interpretation, and similarly concluded that this may have been due to the underpowered nature of their study [373]. Kinnunen et al. found no improvement effect of training for four groups of radiologists in the radiographic diagnosis of midfacial trauma [374].

As the current study is the first, to my knowledge, to evaluate radiographers in a potential LDCT lung cancer screening setting, comparative evidence for radiographer performance in this field is unavailable. However, there is some evidence to support a hypothesis of innate perceptual skill from studies involving radiographer or "naïve" readers in mammography and CT colonography settings. Davies et al. evaluated expertise in categorising mammograms, showing that a group of 3 "naïve" readers with no radiographic experience performed as well as three radiographers (who had exposure to mammograms but no formal training in reading them) and that their performance exceeded that expected by chance, after only a few hours of training [375]. Studies evaluating various training methods for

radiographers in CT colonography have demonstrated a similar training effect for detecting polyps [367, 368, 373, 376].

In the current study, the length of radiographers' "experience" did not seem to affect sensitivity; indeed, the radiographer with the longest duration of "experience" (Radiographer 3) had the lowest sensitivity. This is not unexpected since any radiographer, independent of his or her seniority, would be assumed to be starting from the same baseline as far as pulmonary nodule detection on CT is concerned.

Studies using radiologists with different levels of experience have shown conflicting results in terms of the level of reader experience on nodule detection ability. For instance, Awai et al. demonstrated that the area under the free-response receiver-operator characteristic (ROC) curve was not significantly different when comparing five board-certified radiologists (who did not specialise in thoracic radiology) against five radiology residents [337]. Brochu et al. found there was a significant difference in the sensitivities of 3 radiologists with different levels of experience for the detection of pulmonary nodules in an LDCT screening setting, but that it was the radiologist with the least experience that had the greatest sensitivity (70%) compared to two other thoracic radiologists with 15 and 20 years' experience respectively (sensitivity 54% and 38% respectively) [338]. Brown et al. found that when six non-radiologists, four non-thoracic radiologists and six thoracic radiologists read 8 LDCT cases for pulmonary nodules, the mean detection rates were 62%, 62% and 72% for the non-radiologists, non-thoracic radiologists and thoracic radiologists, respectively, indicating that experience did not have a substantial impact on nodule

detection [345]. Again, these findings reinforce the idea that nodule detection ability is more closely related to an inherent perceptual ability unique to each individual.

An attempt was made to understand the types of opacities that cause difficulty in classification leading to false positives, as well as the characteristics of nodules that were missed by radiographers. It appears that radiographers, perhaps predictably, had difficulty in detecting smaller nodules, and were more likely to "overcall" opacities due to pleural thickening, pleural plaques, atelectasis, or prominent vascular structures.

It is important to recognise that the mean times taken by the radiographers per CT study (range 13-18 minutes) in this study were longer than that taken by the radiologists in more or less comparable studies, for instance by Rubin et al. (range 4.7-9.8 minutes) and Roos et al. (range 4.9-12.3 minutes); the longer reading times could have helped increase radiographer sensitivity. In a CT colonography setting, it has been demonstrated that radiographers (but not radiologists) have an increased accuracy with longer reading times [377], and vice-versa [373]. This has also been shown to a certain extent in the present study: Reader 1, who took the longest time, had the highest sensitivity. However, the ratio of true positive to false positive findings is more instructive when determining the likelihood that a nodule of true significance versus one of less or no significance will be detected, and in this regard, readers 1 and 2 had similar ratios per unit time, indicating that overall Reader 2 made the most efficient use of time. An increase in reading time may have cost-effectiveness implications if the use of radiographers is contemplated in a lung cancer screening programme, as the volume of CTs read by a radiographer could be substantially less than that of an experienced radiologist. However, the fact that the

additional time spent appears to increase the number of true positive detections (i.e. sensitivity) rather than the number of false positive detections is still reassuring in the specific context of a screening programme.

This study has a few limitations. First, like all nodule detection studies without histopathological proof, the determination of a nodule as a true positive necessarily depends on a less than "golden" reference standard. In this case, a reference standard was used that was a composite of the radiographers' readings, as well as a total of four radiologists' readings. In terms of bias considerations, two of the four radiologists determining the reference standard were also training the radiographers and providing feedback; inevitably the radiologists' own interpretations of findings would have influenced that of the radiographers and of the reference standard. However, as has been concluded by Armato et al., variation in opinion as to what constitutes a "true nodule" in a reference standard is extensive [378, 379], and as such this variation would be present in any reference standard that was derived from radiologists' interpretations.

As discussed earlier, the subsets may have been too small to demonstrate a learning effect. However, as the CT studies in the training set were chosen at random, the number of nodules within each subset was not predetermined, inevitably leading to some subsets with smaller number of nodules. Such variation in the number of nodules would of course also be encountered in clinical practice.

Although the length of time taken for the radiographers to read each subset was recorded, this was not done for the radiologists who originally read these studies, as those evaluations were performed as part of the original screening trial and time

data was not collected in the NELSON trial. It is therefore not possible to make a direct observation regarding the time-effectiveness of the radiographers as compared to radiologists.

3.5 Summary

- Radiographers' performance in lung nodule detection, after a short period of training and using continuous feedback, was variable but for some individuals was similar to that reported for radiologists.

- A learning effect could not be demonstrated, indicating that an innate perceptual ability is likely to be a significant determinant of performance.

- The performance of radiographers as readers in a lung cancer screening programme thus needs to be formally evaluated against that of radiologists; this will be the subject of the next investigation.

CHAPTER 4: PROSPECTIVE COMPARISON OF TRAINED RADIOGRAPHERS WITH EXPERIENCED THORACIC RADIOLOGISTS FOR CT LUNG NODULE DETECTION

4.1 Introduction

A CT lung screening programme requires a reading radiologist to dedicate a significant amount of time to the task of lung nodule detection. Should a national lung screening programme be launched, the number of radiologists reading lung screening CTs would, in terms of current working practices, need to increase, especially if the reading protocol involved more than one reader with consensus and arbitration.

As discussed in Chapter 3, one possible method of circumventing the increase in radiologist-hours required for screening is to use radiographers as part of the CT reading process. Having provided radiographers with a basic level of training, it is imperative that their performance is assessed against the established methodology of various screening trials, in which radiologists perform all reading. The aim of this investigation was therefore to evaluate the performance of radiographers in lung nodule detection on CT, compared with radiologists in the setting of the UK Lung Cancer Screening (UKLS) pilot trial.

4.2 Materials and Methods

4.2.1 Study design and case selection

This study was performed prospectively. Two-hundred and ninety consecutive CT studies performed for the UKLS pilot trial were read for this study between November 2011 and April 2012. All CT studies were performed at one of two participating sites, according to trial protocols as specified in section 2.5.

4.2.2 Classification of nodules

Nodules were classified according to the UKLS definitions described in section 2.7.1. The electronic database entry proforma (Artex VOF, Logiton, Netherlands) used for recording nodules as described in section 2.6.2 also allowed for intrapulmonary lymph nodes (IPLNs) to have an additional unique designation so that these could be identified separately.

4.2.3 CT evaluation by radiologists

Each CT examination was read by a single radiologist at each of the two participating sites (Radiologist A at Local Site 1 and Radiologist B at Local Site 2, both with more than 10 years of specialist thoracic imaging experience). The CTs were then transmitted to the central reading site on the same day via the Image Exchange Portal (Burnbank Systems, Ipswich, England) described in section 2.6.2 for a second independent reading (by Radiologist C, 10 years' experience), and read as described in section 2.7.2.

The access rights to the UKLS database of the two reading radiologists at the participating sites were configured such that they were not able to view the recordings of other readers, but the central radiologist (Radiologist C) could access these readings as he had to identify any discrepant findings that required arbitration. The central radiologist only viewed the readings of the local site radiologists once he had completed his own reading.

4.2.4 Selection of reading radiographers

Four radiographers who were able to commit at least four hours a week over the study period were selected as readers. Radiographer 1 read CTs at Local Site 1, and Radiographer 2 read CTs at Local Site 2. Two radiographers (Radiographers 3 and 4) read CTs at the central site. As such, each CT study was read by two radiologists and at least one radiographer, with a maximum of two radiographers (one local site radiographer and one central site radiographer).

All four radiographers had experience in thoracic CT scan acquisition. Radiographers 3 and 4 had participated in the investigation described in Chapter 3 and had thus already undergone training, while Radiographers 1 and 2 were trained using a method similar to that described in Chapter 3. However, as no useful learning effect had been noted in Chapter 3 using feedback after every 10 training cases read, the radiographers instead read 80 training cases (the same training cases used in Chapter 3) - 20 cases where reading was directly supervised by the local site reading radiologist, followed by 60 cases of self-directed training by radiographers with indirect supervision from the local site reading radiologist.

Prior to commencing reading, all four radiographers and all three radiologists also achieved at least 80% sensitivity (as compared to nodules identified by NELSON radiologists) on a test set of 25 different cases that were gleaned from the NELSON training dataset described in section 2.3. This training target was set at a higher level than that achieved by most radiologists in the literature (between 21%-70%, as described in section 3.4) because the CT studies were being read prospectively for clinical, and not solely experimental, purposes. As such, the ability of a reader to detect a high proportion of potentially clinically relevant nodules had to be ensured.

4.2.5 Reference standard

After the second reading, the radiologist at the central site (Radiologist C) reviewed, weekly, all nodule candidates identified by radiologists for each subject on the database to identify any discrepancies. Arbitration on discrepancies was provided at the central site by a thoracic radiologist with more than 20 years of experience, and the final consensus view was recorded on the database. All agreed category 2 to 4 nodules and intrapulmonary lymph nodes were included in the reference standard. In addition, the maximum diameter of each nodule comprising the reference standard as recorded on the UKLS database was recorded.

4.2.6 Classification of discrepancies

For each CT, a list of each reader's reading was generated and compared against the reference standard. A nodule was considered to have been missed by a

reader if it was included in the reference standard but not recorded by that reader. In addition note was made of nodules classified as IPLNs by the reference standard but recorded as category 2 to 4 by the reader, and vice versa.

4.2.7 Statistical analysis

The sensitivity (the percentage of reference standard nodules identified) and the average false positive detections (FPs) per case (expressed as mean and standard deviation) were calculated:

- for each reader, for all cases read by him or her;
- for each radiographer and radiologist within a particular radiographer-radiologist combination (10 combinations in total), taking into account only cases read by that combination, to enable direct comparison between that radiographer and radiologist (comparisons of sensitivity and average FPs were performed using McNemar's test and paired student's t-test, respectively);
- for each reader in the first 10 weeks (P1), and compared to that in the second 10 weeks (P2) of the study (comparisons of sensitivity and average FPs per case were performed using the Chi-square test and independent samples student's t-test, respectively).

A post-hoc analysis of differences between radiographer and radiologist sensitivity when considering only reference standard nodules at two higher diameter thresholds ($\geq$ 5mm and $\geq$ 6mm) was subsequently conducted , to investigate the effect of increasing the threshold for nodule positivity on radiographer performance.

The agreement between each radiographer and the reference standard for the classification of IPLNs and category 2 to 4 nodules was assessed using the weighted kappa statistic (multirater κ), as detailed in Chapter 2. Kappa values were defined as follows: a κ less than or equal to 0.2, poor agreement; 0.21-0.40 fair agreement; 0.41-0.6 moderate agreement; 0.61-0.80, good agreement; and 0.81-1.00, very good agreement [380]. All analysis was performed using Medcalc (version 12.5.0.0, MedCalc Software, Mariakerke, Belgium). A *P* value of less than 0.05 was assumed to be statistically significant.

4.3 Results

4.3.1 Reference standard

Eighty-one (27.9%) of the 290 CT studies did not contain any nodules. The reference standard thus consisted of 599 nodules in the remaining 209 (72.1%) CT studies. The majority of CTs had 1 (35.4%), 2 (23.0%) or 3(21.5%) nodules per CT (Figure 4.1). The median number of nodules per scan was 2, with a range of 1 to 18 nodules. The mean diameter of reference standard nodules was 5.2 ± 2.9mm (median 4.4mm, range 1.6-30.0mm).

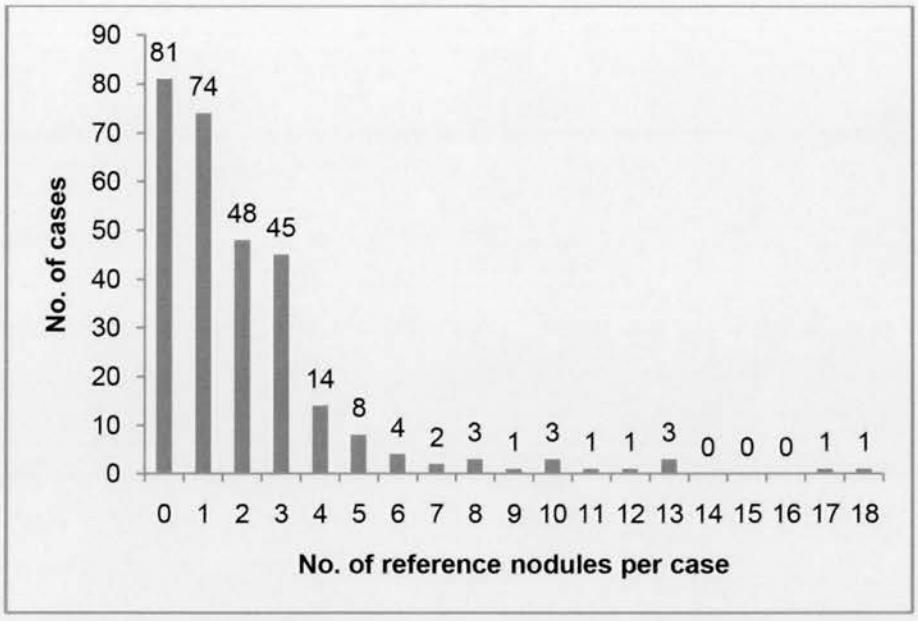

Figure 4.1. Distribution of the number of nodules per CT scan.

The breakdown of nodules in the reference standard according to UKLS size category and type is illustrated in Table 4.1. The majority were solid nodules (62.1%), with just under a third (32.6%) classified as IPLNs. 98.2% of nodules were category 1 to 3.

	Nodule type				
Size Category	**Solid (non-IPLN)**	**Part-solid**	**Non-solid**	**IPLN[1]**	**Total**
1	NA	NA	NA	195 (32.6)	**195 (32.6)**
2	247 (41.2)	2 (0.3)	7 (1.2)	n/a	**256 (42.7)**
3	115 (19.2)	4 (0.7)	18 (3.0)	n/a	**137 (22.9)**
4	10 (1.7)	1 (0.2)	0 (0)	n/a	**11 (1.8)**
Total	**372 (62.1)**	**7 (1.2)**	**25 (4.2)**	**195 (32.6)**	**599 (100.0)**

Table 4.1. Distribution of nodules according to UKLS size category and nodule type. Figures in parentheses are percentages of the total number of nodules. Minor inconsistencies in summation of the percentages is due to rounding of percentages to 1 significant figure. IPLN= intrapulmonary lymph node. NA= not applicable.
[1]IPLNs are category 1 by definition in the UKLS.

4.3.2 Overall performance of radiographers and radiologists

Radiographers 1, 2, 3 and 4 had sensitivities of 67.6%, 77.8%, 79.4% and 61.6% respectively (mean sensitivity 71.6 ± 8.5%). Radiologists A, B and C had sensitivities of 88.9%, 87.0% and 74.0% respectively (mean sensitivity 83.3 ± 8.1%).

The average FPs per case for radiographers 1, 2, 3 and 4 were 1.2 ± 2.1, 2.9 ± 2.8, 0.6 ± 1.0 and 1.1 ± 1.3 respectively, while that of radiologists A, B and C were 0.5 ± 0.8, 0.7 ± 1.0 and 0.2 ± 0.5 respectively.

4.3.3 Comparison of radiographer and radiologist performance

The sensitivities of each radiographer compared to those of the corresponding radiologists within a particular radiographer-radiologist combination are illustrated in Table 4.2. Radiographers 1 and 2 could only be compared with their corresponding local site radiologists (i.e. radiologists A and B respectively) and the central site radiologist (radiologist C). Radiographer sensitivity was significantly lower than radiologist sensitivity in 7 of 10 radiographer-radiologist combinations (range of difference, 9.7%-32.8%, $P < 0.05$), and not significantly different in 3/10 combinations.

Radiographer-radiologist combination	No. of CTs read	Sensitivity			
		Radiographer (%)	Radiologist (%)	Difference (%)	*P* value
1-A	130	67.6	88.0	20.4	**0.0008**
1-C	130	67.6	74.5	6.9	0.30
2-B	139	77.8	87.4	9.6	**<0.0001**
2-C	139	77.8	74.5	-3.3	0.20
3-A	68	81.0	92.2	11.2	**0.01**
3-B	87	78.5	88.2	9.7	**0.0087**
3-C	155	79.4	76.2	-3.2	0.32
4-A	64	53.8	86.6	32.8	**<0.0001**
4-B	49	68.7	85.5	16.8	**0.0051**
4-C	113	61.6	72.0	10.4	**0.0119**

Table 4.2. Comparison of radiographer and radiologist sensitivity for the 10 radiographer-radiologist combinations. A negative value for the difference in sensitivity indicates a lower radiologist sensitivity compared to a radiographer. *P* values are those derived from the McNemar's test. *P* values in bold are statistically significant.

Radiographers had significantly higher average FPs per case than radiologists in 8/10 combinations (range of difference, 0.4-2.6, $P < 0.05$), and there was no significant difference in the remaining two combinations (Table 4.3).

Radiographer-radiologist combination	Average FPs per case			
	Radiographer	Radiologist	Difference	*P* value
1-A	1.1 ± 1.3	0.5 ± 0.8	-0.6	**<0.0001**
1-C	1.1 ± 1.3	0.1 ± 0.5	-1.0	**<0.0001**
2-B	2.8 ± 2.8	0.7 ± 1.1	-2.1	**<0.0001**
2-C	2.8 ± 2.8	0.2 ± 0.5	-2.6	**<0.0001**
3-A	0.9 ± 1.4	0.5 ± 0.8	-0.4	**0.0176**
3-B	1.4 ± 2.5	0.6 ± 1.0	-0.8	**0.0015**
3-C	1.2 ± 2.1	0.1± 0.5	-1.1	**<0.0001**
4-A	0.4 ± 0.8	0.5 ± 0.8	0.1	0.2009
4-B	0.8 ± 1.2	0.8 ± 1.3	0	0.71
4-C	0.6 ± 1.0	0.2 ± 0.5	-0.4	**0.0001**

Table 4.3. Comparison of radiographer and radiologist average FPs per case for the 10 radiographer-radiologist combinations. A negative value for the difference indicates a lower radiologist average FPs per case compared to a radiographer. *P* values are those derived from the paired student's t-test. *P* values in bold are statistically significant.

4.3.4 Reader performance in the first 10 weeks versus second 10 weeks

The two radiographers with the lowest overall sensitivity (Radiographers 1 and 4) showed a significant improvement in sensitivity between the first and second 10-week period (sensitivity 50.0% in P1 versus 74.1% in P2 for Radiographer 1, 41.8% in P1 versus 67.2% in P2 for Radiographer 4, $P < 0.005$), but their sensitivity in P2 still did not reach the level of Radiographers 2 and 3, who showed no significant difference in their sensitivity between the two periods. Radiologists' sensitivity did not significantly differ between the two periods. No radiographer or radiologist demonstrated a significant difference in average FPs per case between the

two periods. As such, the improved sensitivity of Radiographers 1 and 4 in the second 10 weeks did not come at the expense of increased average FPs per case.

4.3.5 Characterisation of intrapulmonary lymph nodes and nodules

Radiographers showed moderate to good agreement with the reference standard for the classification of IPLNs and category 2 to 4 nodules (Table 4.4).

Radiographer	Weighted kappa statistic	Level of agreement
A	0.634	good
B	0.554	moderate
C	0.571	moderate
D	0.603	moderate

Table 4.4. Agreement between radiographer and the reference standard for IPLN and nodule classification.

4.3.6 Comparisons of sensitivity using alternate diameter thresholds

There were 236 reference standard nodules that were 5mm or greater in diameter. When considering only these nodules, the number of radiographer-radiologist combinations with significantly lower radiographer sensitivity decreased to 4/10 combinations (range of difference, 17.2%-37.3%, $P < 0.05$) (Table 4.5). No significant difference was seen in the remaining 6/10 combinations.

Radiographer-radiologist combination	No. of nodules ≥ 5 mm diameter	Sensitivity			*P* value
		Radiographer (%)	Radiologist (%)	Difference (%)	
1-A	87	63.2	80.5	17.2	**0.0041**
1-C	87	63.2	82.8	19.5	**0.0053**
2-B	128	82.0	89.1	7.0	0.1508
2-C	128	82.0	83.6	1.6	0.8501
3-A	45	80.0	88.9	8.9	0.3438
3-B	82	79.3	89.0	9.8	0.1153
3-C	127	79.5	85.0	5.5	0.2482
4-A	51	41.2	78.4	37.3	**0.0005**
4-B	42	78.6	88.1	9.5	0.4240
4-C	93	58.1	79.6	21.5	**0.0011**

Table 4.5. Comparison of radiographer and radiologist sensitivity for the 10 radiographer-radiologist combinations for nodules ≥ 5mm diameter. *P* values are those derived from the McNemar's test. *P* values in bold are statistically significant.

There were 165 reference standard nodules that were 6mm or greater in diameter. As with nodules that were 5mm or greater in diameter, the number of radiographer-radiologist combinations with significantly lower radiographer sensitivity also decreased to 4/10 combinations (range of difference, 18.3%-40.6%, $P < 0.05$) (Table 4.6). No significant difference was seen in the remaining 6/10 combinations.

Radiographer -radiologist combination	No. of nodules ≥6mm diameter	Sensitivity			*P* value
		Radiographer (%)	Radiologist (%)	Difference (%)	
1-A	60	65.0	83.3	18.3	**0.0127**
1-C	60	65.0	85.0	20.0	**0.0169**
2-B	91	79.1	90.1	11.0	0.05
2-C	91	79.1	82.4	3.3	0.66
3-A	30	83.3	90.0	6.7	0.63
3-B	60	75.0	88.3	13.3	0.08
3-C	90	77.8	85.6	7.8	0.14
4-A	37	40.5	81.1	40.6	**0.0015**
4-B	29	75.9	89.7	13.8	0.34
4-C	66	56.1	80.3	24.2	**0.0009**

Table 4.6. Comparison of radiographer and radiologist sensitivity for the 10 radiographer-radiologist combinations for nodules ≥ 6mm diameter. *P* values are those derived from the McNemar's test. *P* values in bold are statistically significant.

4.4 Discussion

The mean and range of sensitivities of radiographers and radiologists in this study group were comparable to figures reported for radiologists (Table 4.7) [289, 290, 293, 334, 336, 338, 371].

Authors	Year	No. and type of readers	Sensitivity (%)	
			Mean	Range
Marten et al.	2004	4 radiologists	40	21-57
Brochu et al.	2007	3 radiologists	54	38-70
Rubin et al.	2005	3 radiologists	50	41-60
Roos et al.	2010	3 radiologists	53	44-59
Wormanns et al.	2005	3 radiologists	64	NR
Beigelman-Aubry et al.	2007	2 radiologists	52	46-58
Fraioli et al.	2007	3 radiologists	57	46-68
Current study	2012	3 radiologists 4 radiographers	83 72	74-89 62-79

Table 4.7. Sensitivities of radiologists in a selection of nodule detection studies. NR= not reported.

Caution should always be exercised when comparing sensitivities between nodule detection studies, as differences in the derivation and stringency of the reference standard [285], and in the types of patients undergoing CT examinations (e.g. patients with multiple metastases versus lung screening studies) may profoundly affect sensitivity. Nevertheless, it is reassuring that not only did the mean sensitivity of radiologists in the current study exceed those of previous studies, but radiographers had sensitivity comparable to radiologists in reported previous studies.

The importance of a high rate of nodule detection- i.e. high sensitivity - is underscored by the fact that most failures in lung cancer diagnosis are due to errors of detection rather than interpretation [381, 382]. Considering this in isolation, the desideratum of any CT screening programme is to have readers with the highest possible sensitivity, and in this context radiographers cannot be considered ideal first

readers for lung screening, since their overall performance in the majority of cases was statistically significantly lower than that of the radiologists reading the same CTs.

However, initial experience from the UKLS pilot trial has indicated that the vast majority of nodules smaller than 5mm in diameter (i.e. category 2 nodules) demonstrate no growth on volumetry within a 12-month period. As such, it may be prudent to raise the threshold for considering a nodule positive, to avoid unnecessary follow-up. A recent retrospective analysis from the I-ELCAP trial has reinforced this notion, suggesting that using diameter thresholds of 6mm, 7mm and 8mm could decrease further work-up by 36%, 56% and 68% respectively, while resulting in a maximum delay in lung cancer diagnosis of 9 months in 0%, 5% and 5.9% of cases respectively [383]. It could thus be argued that the sensitivity of radiographers in a screening programme need not be as high as that of radiologists across the whole spectrum of nodules, but should be optimised for larger nodules. This notion prompted the post-hoc analysis of radiographer sensitivity confined to reference standard nodules that were 5mm and greater (the current definition of a positive result in I-ELCAP), and 6mm and greater in diameter. Radiographer sensitivity became more closely aligned with that of radiologists using these higher diameter thresholds in the majority of combinations, but there were still four combinations where radiographers were less sensitive at both diameter thresholds. Nevertheless, such comparability between radiographer and radiologist sensitivity, when viewed in conjunction with their satisfactory ability to accurately categorise nodules according to precise definitions, suggests that a radiographer could function adequately as an

aid to nodule detection rather than a first reader, in the same way that computer-aided detection (CAD) does.

The average false positive detections per case of the radiographers in this study were significantly higher than those of radiologists. However, it is reassuring that radiographers in general did not exceed 3 average FPs per case, and are thus comparable to CAD systems, where average FPs per case between 0.3 and 15 per case have been reported [384]. When viewed in this context, the higher average FPs per case of radiographers compared to radiologists in this study does not disqualify them from being used as aids to reading.

In contrast to the lack of a learning effect observed in Chapter 3, an improvement in sensitivity of the two least sensitive radiographers between the first and second 10-week periods in the present study was noted, and occurred despite the lack of a formalised programme of feedback. It is thus likely that exposure to more nodules over a longer period may help to improve the sensitivity of less sensitive readers. Importantly, the improved sensitivity did not come with the penalty of increased average false positive detections per case.

This study has a few limitations. First, the reference standard presumed that any nodule detected by a radiographer but not by a radiologist was not a true positive. This was a necessary condition within the UKLS trial because radiographers are not currently validated as readers in lung screening, and so it could therefore be argued that it would have been unethical for any radiographer findings to direct clinical care. However, this does potentially mean that there were less true positives and more false positives within the radiographer readings, thereby

potentially exaggerating their false positive detection rate. Even so, this reference standard still serves as a valid relative standard against which to measure radiographer performance. Also, radiographers were not provided with continuous feedback in this study, which might have been useful to assess the effect of feedback over a larger set of nodules than that used in Chapter 3. Such an assessment was not the primary objective of this study, but it is worth noting that radiographers were comparing their own readings with those in the consensus, once consensus had been achieved, and so were enacting a form of self-directed learning. However, this effect cannot be quantified.

4.5 Summary

- Radiographers' in this study displayed, on average, a lower sensitivity for nodule detection and higher false positive detections per case compared to radiologists reading the same CTs.
- Radiographers are thus probably not sensitive enough to be used as first readers.
- However, the fact that the performance of some radiographers compared favourably with radiologists, especially when larger nodules were considered, suggests that radiographers could fulfil the role of an assistant reader.

CHAPTER 5: THE EFFECT OF RADIOGRAPHERS AS CONCURRENT READERS ON THE PERFORMANCE AND READING TIME OF EXPERIENCED RADIOLOGISTS IN LUNG CANCER SCREENING

5.1 Introduction

The use of radiographers as human aids to radiologists for the task of lung nodule detection remains untested. In contrast, computer-aided detection (CAD) software as a second reader for this task has been extensively evaluated and shown to improve sensitivity, as discussed in detail in section 1.6.3.

Despite this extensive evaluation, CAD has not been universally adopted in lung nodule detection, and has also not been prospectively evaluated in any of the randomised control trials in CT lung cancer screening. The reticence to use CAD may be due to the practicalities of integrating it with CT reporting workflow, but these difficulties are not insurmountable. A greater encumbrance may be the need for two rounds of reading by a radiologist when using CAD as a second reader - the radiologist has to first independently read the study, present it to the CAD system, and then re-evaluate the study with the specific aim of assessing the CAD marks, in order to arrive at a final set of agreed findings. This has led to interest in using CAD as a concurrent reader. In concurrent reading, the first round of radiologist reading is removed; instead, the study is processed by the CAD and presented to the radiologist, who accepts or rejects the CAD marks, and finally performs an independent search for any missed nodules. Although this method has undergone only limited

evaluation, it has been shown to decrease reading times without compromising sensitivity [294, 354].

The aim of the present investigation was thus to prospectively compare the performance of radiologists reading CTs independently (i.e. unaided) with their performance when using radiographers as concurrent readers, with respect to sensitivity, false positive detection, and reading times, in an actual lung screening setting.

5.2 Materials and Methods

5.2.1 Study design and case selection

This study was performed prospectively. Between June and October 2012, the baseline CT studies of 369 consecutive participants in the LDCT arm of the UKLS trial were read for this study. All CT studies were performed at one of two participating sites, according to trial protocols as specified in section 2.5.

5.2.2 CT evaluation by radiologists

Each CT study was read by a single radiologist at one of the two participating sites, namely Radiologist A at Local Site 1 and Radiologist B at Local Site 2, both with more than 10 years of specialist thoracic imaging experience. The studies were then transmitted to a central reading site as described in section 2.6.2 for a second independent reading by Radiologist C, 10 years' experience, and read as described in section 2.7.2. Radiologists A, B and C were the same radiologists as in Chapter 4. In

addition, a second central reading radiologist with 7 years' experience, Radiologist D, also participated in this study.

Nodule classification and recording was exactly the same as described in sections 2.6.2 and 4.2.2.

5.2.3 Selection of reading radiographers

The same radiographers as in Chapter 4 participated in this study: Radiographer 1 read CTs at Local Site 1, Radiographer 2 read CTs at Local Site 2, and Radiographers 3 and 4 read CTs at the central site.

5.2.4 Concurrent reading workflow

In concurrent reading, the following steps were performed:

1. Each CT was first read by a radiographer on the LungCARE workstation, who uploaded his or her report to the UKLS database under a personal login, as explained in section 2.6.2.
2. The radiographer's stored nodule recordings (in the form of a DICOM structured report, or DICOM SR) were then made available to the reading radiologist. For each recording, the radiologist had one of three options. He could **accept** a particular finding if he agreed with it, and leave the radiographer recording unmodified. He could **reject** a finding if he disagreed with it (i.e. he thought the finding did not represent a nodule), in which case he would delete the recording. Finally, he could **amend** the recording if he

agreed that the finding represented a nodule, but disagreed with its categorisation (for example, he thought it was a subsolid nodule where a radiographer had classified it as solid), in which case he would re-report the nodule using the electronic soft-copy entry proforma (Artex VOF, Logiton, Netherlands) (see section 2.6.2) and copy and paste his own report into the nodule report.

3. After reviewing the radiographer's recordings, the radiologist performed another search to identify any additional nodules missed by the radiographer.
4. The radiologist then saved his recordings as a new DICOM SR, and uploaded this report to the UKLS database.

In this way, the radiologist still only performed a single reading of the CT. Figure 5.1 summarises the concurrent reading process as compared to first reading and second reading.

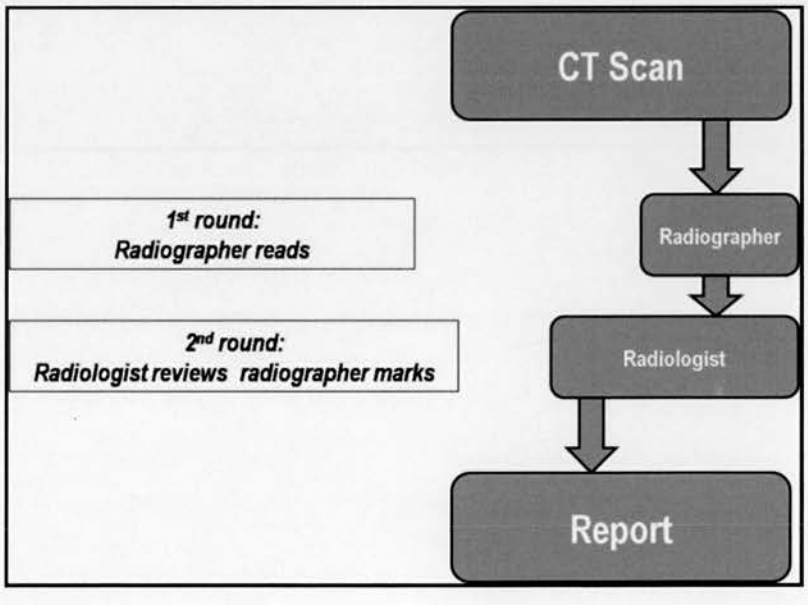

Figure 5.1A.

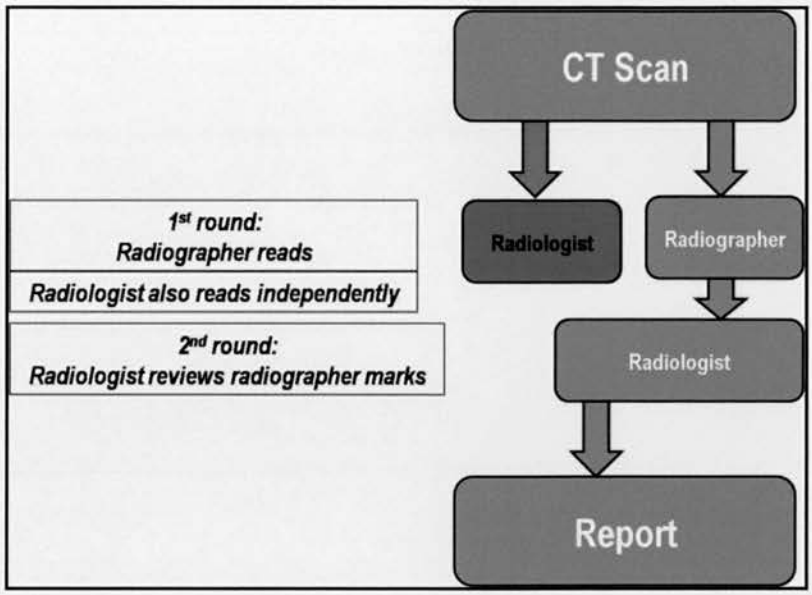

Figure 5.1B.

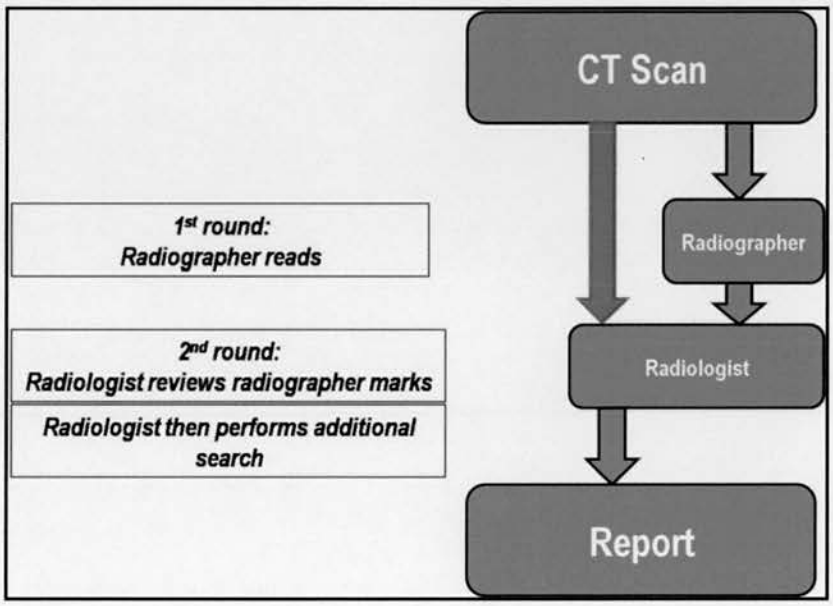

Figure 5.1C.

Figure 5.1. The process of reading in first (5.1A), second (5.1B) and concurrent (5.1C) reading with radiographers.

5.2.5 Selection of CT studies for independent single reading versus concurrent reading

It was decided that the most practical strategy of comparing the concurrent reading and independent single (i.e. unaided) reading performance for each radiologist would be for him to read a number of CTs using one method, and then to switch to the other method after two months. In this way, the radiologist would be reading in actual screening conditions, potential variability in CT studies (such as the number and type of nodules per CT) would be adjusted for, and central and local site radiologists could all be assessed. As such, for the first two months (June-August), the local radiologists (Radiologists A and B) were instructed to read independently, while the central radiologists (Radiologists C and D) read concurrently, and the reading methods were switched between the sites over the second two months (August-October).

5.2.6 Reading times

Radiologists and radiographers were instructed to record the method and the times, to the nearest minute, on the CTs they read. The decision to record to the nearest minute, rather than second, was made so that a meaningful change in reading time (i.e. measurable in minutes, not seconds) that would translate into a potential real saving in an actual screening setting could be detected. Recording to the nearest minute was also less disruptive to reading workflow.

Reading time recordings were saved together with the radiologists' reports on the UKLS database. Recognising that recording reading times for every CT when

reading large numbers of screening studies during the UKLS could be difficult, the aim was to ensure at least two-thirds of CTs read had a recorded reading time.

5.2.7 Reference standard

After both local and central readings had been performed, the radiologists at the central site (Radiologist C or D) reviewed all identified nodule candidates for each subject on the database to identify any discrepancies. Arbitration on discrepancies was provided at the central site by a thoracic radiologist with more than 20 years of experience, and the final consensus answer was recorded on the database. All agreed category 2 to 4 nodules and intrapulmonary lymph nodes were considered positive in the reference standard. The reference standard was thus composed of any nodules which had been identified and agreed on by both radiologists, as well as any nodule that had been identified by at least one radiologist, and subsequently ratified by the expert arbiter, including those radiographer-identified nodules that had been agreed or amended by a radiologist during concurrent reading.

5.2.8 Statistical analysis

The numbers of concurrently and independently read CTs were calculated for each radiologist. Comparisons of the number of reference standard nodules between concurrent and independent reading datasets for each radiologist were performed using the two-tailed Mann-Whitney Test (a non-parametric distribution of nodules was assumed).

Sensitivity, the absolute number of false positive detections (FPs) and average FPs per case (expressed as mean and standard deviation) were separately determined for the cohorts of CTs read independently and concurrently per radiologist. The sensitivity of each radiologist was calculated by dividing the number of true positive nodules detected by the total number of nodules in the reference standard for the cases read by that radiologist. Average FPs per case was calculated by dividing the total number of FPs by the total number of cases read by that radiologist. Differences in proportions were compared using the Chi-squared test or the Fisher's exact test in the case of smaller sample sizes (see section 2.8.2.2) [360, 361] as appropriate.

Differences in reading times between concurrent and independent reading for each reader were compared using the independent samples t-test. A post-hoc analysis to determine the correlation between numbers of nodules and reading time was subsequently performed using Spearman's rank correlation, and differences in correlation coefficients analysed for statistical significance.

All analysis was performed using Medcalc (version 12.5.0.0, MedCalc Software, Mariakerke, Belgium). A P value of less than 0.05 was assumed to be statistically significant.

5.3 Results

5.3.1 Reference standard

A total of 369 LDCT studies were read during the study period. 123 (33.3%) of the 369 CT studies did not contain any nodules. The reference standard thus consisted of 694 nodules in the remaining 246 (67.7%) CT studies. Figure 5.2 illustrates the distribution of the number of nodules per CT study. The majority of CTs had 1 (23.6%), 2 (16.3%), 3 (8.9%) or 4 (6.8%) nodules. The median number of nodules per CT was 1, with a range of 1 to 17 nodules. The majority of reference standard nodules were solid nodules (51.3%) and intrapulmonary lymph nodes (42.1%) (Table 5.1). Ninety-six point eight percent of nodules were category 1 to 3.

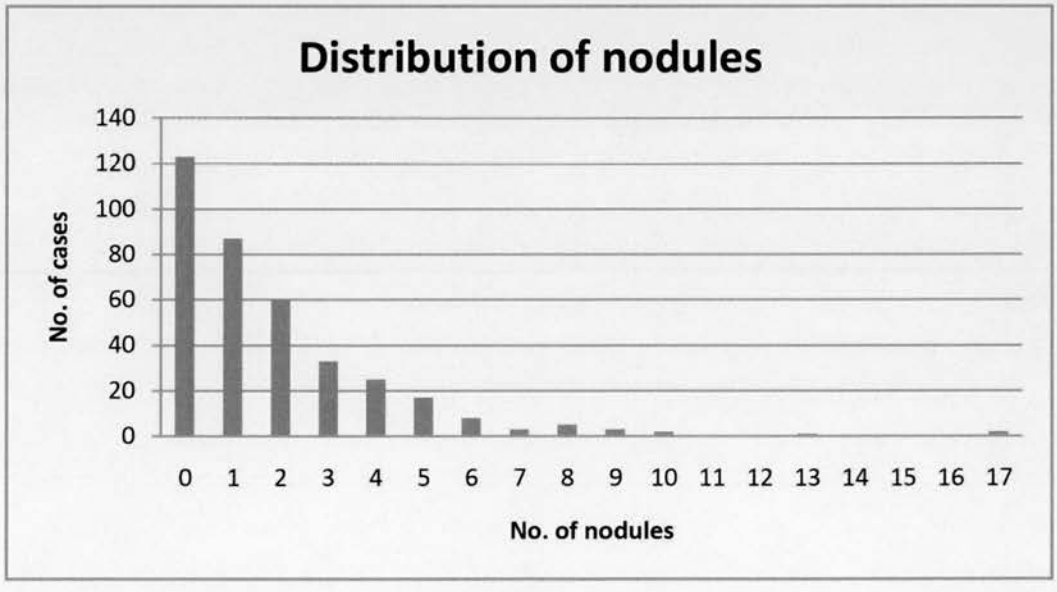

Figure 5.2. Distribution of the number of nodules per CT study.

		Nodule type				
		Solid (non-IPLN)	Part-solid	Non-solid	IPLN[1]	Total
Size category	1	NA	NA	NA	292 (42.1)	292 (42.1)
	2	219 (31.6)	2 (0.3)	11 (1.6)	NA	232 (33.4)
	3	119 (17.1)	8 (1.2)	21 (3.0)	NA	148 (21.3)
	4	18 (2.6)	1 (0.1)	0 (0.0)	NA	19 (2.7)
	U[2]	0 (0.0)	0 (0.0)	3 (0.4)	NA	3 (0.4)
	Total	356 (51.3)	11 (1.6)	35 (5.0)	292 (42.1)	694 (100.0)

Table 5.1. Distribution of nodules according to UKLS size category and nodule type. Figures in parentheses are percentages of the total number of nodules.
[1]Intrapulmonary lymph nodes (IPLNs) are category 1 by definition in the UKLS.
[2]U=uncategorised. Three nodules were categorised as non-solid nodules (Category C) but could not have their sizes calculated due to a database error.
NA= not applicable.

5.3.2 Number of cases and number of nodules by reading method

The number of cases read by each radiologist varied from 83 to 119 using independent reading and from 69 to 122 using concurrent reading (Table 5.2). However, there was no significant difference between the numbers of reference standard nodules per case in the independent versus concurrent reading cohorts for any of the four radiologists (Figure 5.3).

		Radiologist			
		A	B	C	D
Reading Method	Independent	83	119	94	84
	Concurrent	88	79	122	69
	P value	0.83	0.18	0.41	0.74

Table 5.2. Numbers of cases read by each radiologist using each reading method. *P* values are for differences between the number of reference standard nodules per case between independent and concurrent reading, derived from the Mann-Whitney test.

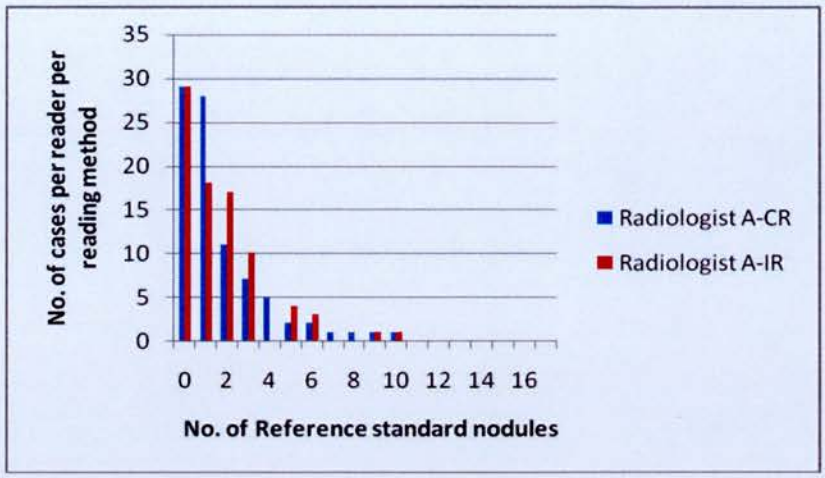

Figure 5.3A.

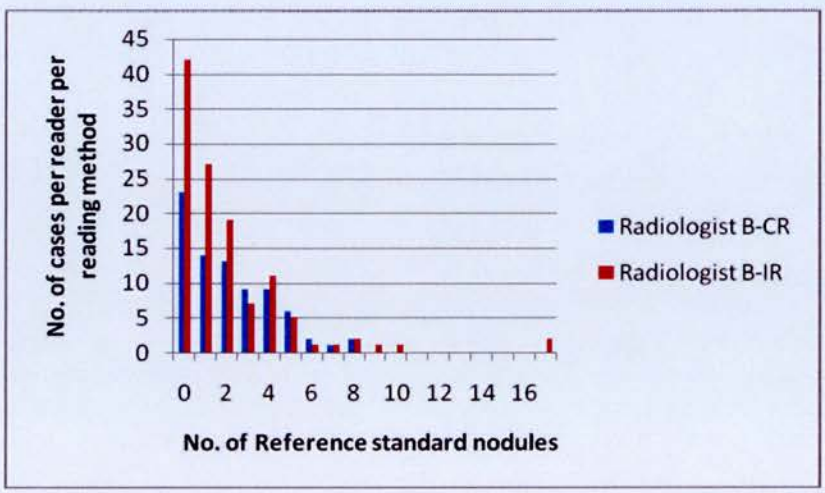

Figure 5.3B.

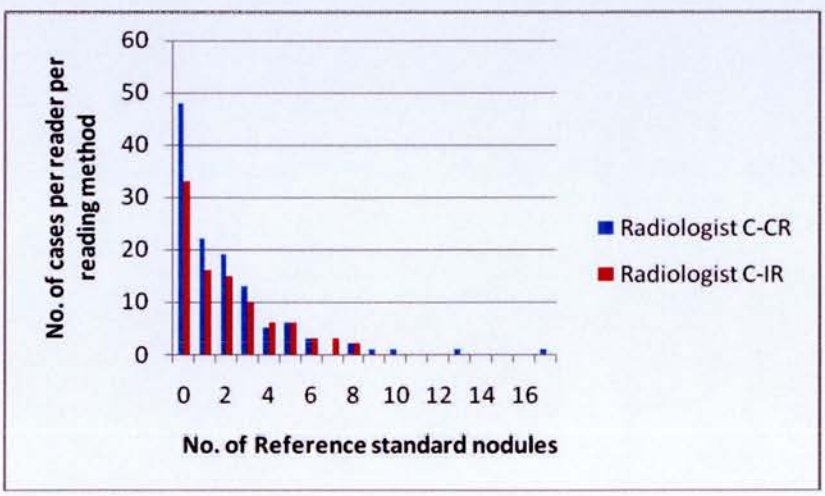

Figure 5.3C.

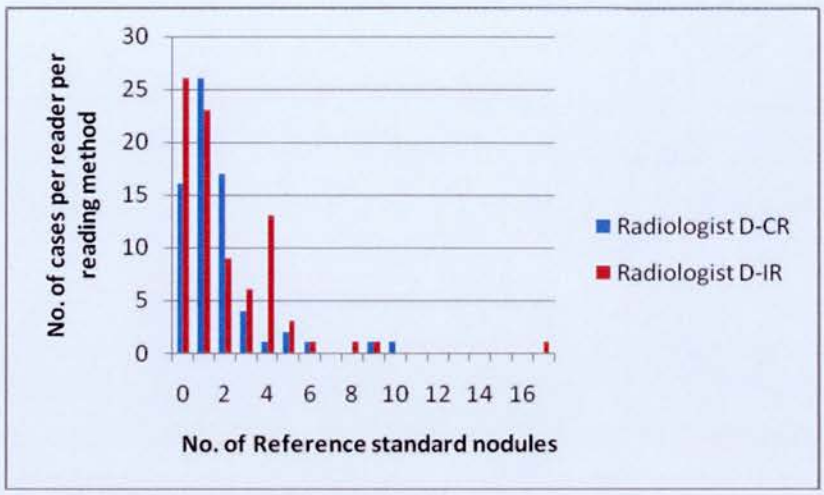

Figure 5.3D.

Figure 5.3. No. of reference standard nodules per case for each reading method for Radiologists A-D.
CR= concurrent reading, IR= independent reading.

5.3.3 Sensitivity of radiologists

The overall sensitivity of each radiologist for the different reading methods is detailed in Table 5.3. The mean sensitivity for radiologists reading independently was 77.5 ± 11.2%, increasing to 90.8 ± 5.6 % with the use of concurrent reading. For all but one radiologist (Radiologist D), statistically significant higher sensitivity was achieved with concurrent reading compared to independent reading.

		Radiologist			
		A	B	C	D
Reading Method	Independent	78.9	79.8	62.2	89.2
	Concurrent	90.4	98.2	84.5	90.1
	Difference	11.5	18.4	22.3	0.9
	***P* value**	**0.01**	**<0.0001**	**<0.0001**	0.97

Table 5.3. Sensitivity of radiologists for each reading method. Except for *P* values, figures shown are percentages. *P* values are those derived from the Chi-square test. *P* values in bold indicate statistically significant results.

5.3.4 False positive detections (FPs)

There was a wide variation in the average FPs per case. While the overall mean of average FPs per case increased from 0.33 ± 0.20 with independent reading to 0.60 ± 0.53 with concurrent reading, average FPs per case ranged between 0.06 and 1.38, increasing with concurrent reading for Radiologists A, B and C (and statistically significant for Radiologists B and C), but decreasing for Radiologist D (Table 5.4).

		Radiologist			
		A	B	C	D
Reading Method	Independent	0.31 ± 0.75	0.47± 1.10	0.06 ± 0.25	0.48 ± 0.96
	Concurrent	0.37± 0.65	1.38 ± 1.46	0.21 ± 0.61	0.42 ± 0.76
	Difference	0.06	0.91	0.15	-0.06
	***P* value**	0.56	**<0.001**	**0.03**	0.69

Table 5.4. Average FPs per case for each reading method. A negative difference indicates a lower average FP per case with concurrent compared to independent reading. *P* values are those derived from the independent samples t-test. *P* values in bold indicate statistically significant results.

5.3.5 Reading times

Radiologists A, B, C and D recorded their reading times for 100%, 77%, 75%, and 90% of their concurrently read cases, respectively, and for 83%, 71%, 85% and 89% of their independently read cases, respectively. The mean reading times per case for concurrent reading ranged from 6.2 minutes to 8.6 minutes, compared to 7.0 to 12.4 minutes for independent reading (Table 5.5).

Concurrent reading was faster than independent reading for all radiologists, but this increase in reading speed was not statistically significant for Radiologist C. Furthermore, the maximum decrease in mean reading time was just under 4 minutes (Radiologist A).

Reading method	Radiologist			
	A	B	C	D
Independent	12.4 (11.1,13.5)	8.8 (7.8, 9.8)	7.0 (6.5, 7.5)	8.3 (7.4, 9.3)
Concurrent	8.6 (7.9, 9.3)	6.2 (5.5, 7.0)	6.9 (6.4, 7.4)	7.0 (6.2, 7.8)
Difference	-3.8	-2.6	-0.1	-1.3
***P* value**	**<0.0001**	**0.0001**	0.65	**0.03**

Table 5.5. Mean reading times of radiologists for each reading method. Numbers shown are time in minutes, except for *P* values. Numbers in parentheses are 95% confidence intervals for the mean. A negative difference indicates a shorter time with concurrent compared to independent reading. *P* values are those derived from the independent samples t-test. *P* values in bold indicate statistically significant results.

5.3.6 Relationship between number of nodules per CT and reading time

The finding that concurrent reading was quicker despite there being no significant difference between the number of reference standard nodules per CT led to a post-hoc analysis to investigate the relationship between the number of nodules per CT and reading time. As could be expected, there was a significant correlation between the number of reference nodules per subject and the time taken for each CT, and for Radiologists A, C and D, this correlation was stronger for independently read CTs (Table 5.6). However, the strengthening of this correlation was only statistically significant for Radiologist D.

Reading Method	Radiologist			
	A	B	C	D
Independent	0.612 (**<0.0001**)	0.420 (**0.0001**)	0.432 (**0.0001**)	0.766 (**<0.0001**)
Concurrent	0.572 (**<0.0001**)	0.546 (**<0.0001**)	0.422 (**<0.0001**)	0.523 (**<0.0001**)
***P* value**	0.71	0.34	0.94	**0.01**

Table 5.6. Rank correlation between number of nodules per patient and time taken. Figures shown are Spearman's coefficient of rank correlation (Spearman's rho), and figures in parentheses are *P* values for significance level. *P* values in the final row are for significance levels of the difference between correlation coefficients for concurrent and independent reading per radiologist. Bold font indicates a statistically significant *P* value.

5.4 Discussion

The current investigation is, to my knowledge, the first evaluation of radiographers as concurrent readers for pulmonary nodule detection in CT lung cancer screening. It has shown that the sensitivity of the majority of radiologists in this study group improved by using radiographers as concurrent readers, with a statistically significant reduction in mean reading time, but accompanied by a simultaneous increase in false positive detections.

The improvement in sensitivity in this study was seen for all except the most sensitive radiologist, but the range of sensitivities of radiologists in this study was reassuringly comparable to that reported for radiologists (between 40% to 83%), as discussed earlier in sections 3.4 and 4.4. As such, there is every reason to expect that concurrent reading with radiographers could improve the sensitivity of the majority of expert thoracic radiologists.

The mean sensitivity achieved by the radiologists reading independently in this study (77.5 ± 11.2%) was lower than that in Chapter 4 (82.5 ± 7.7%), although a statistical comparison of these means is difficult due to the different numbers of cases read. However, the reference standards in the two investigations were also different: the reference standard in the current investigation included all radiographer-identified nodules that had been agreed by at least one radiologist and an arbiter, or both radiologists, while radiographers' readings were not considered by radiologists and so not included in the reference standard in Chapter 4. As such, the denominator of the total number of positive nodules in the current standard may be larger, so apparently lowering sensitivity during independent reading.

Of course, the increased average false positive detections per case that accompanied the increases in sensitivity with concurrent reading is once again a salutary reminder of the trade-off between sensitivity and "overcalling" of nodules, especially for the two readers with the greatest sensitivity improvements (Radiologists B and C). Given the design of the present study, it is not possible to ascertain whether a radiologist may have designated the same false positives as a radiographer in the absence of a radiographer mark. However, the fact that the same radiographers in the current study demonstrated in general higher average FPs per case compared to radiologists in Chapter 4 implies that radiographers are probably the main drivers of false positive detections in the concurrent reading scenario. Intuitively, it could be expected that the increase in FP detection is particularly due to radiographers identifying a large number of small opacities as nodules, and subsequent reticence of the radiologists to reject such marks once presented with them. Certainly, a previous study with CAD has shown that a radiologist's confidence in designating a small nodule as positive was enhanced when presented with a CAD mark, even if it did not improve overall accuracy [385].

It is important to put the increase in average FPs per case into perspective when considering concurrent reading with radiographers as an alternate strategy to independent reading, as compared to CAD. CAD results in average FPs per case of 3.7 to 4.15 when including nodules as small as 3mm, the size-cut off used for inclusion in the UKLS and the present study [354, 386]. This rate is much higher than the highest rate with concurrent reading in the present study (Radiologist B, 1.38). Nevertheless, the consequences that can be triggered by false positives (such

as increased number of CTs requiring arbitration, follow-up, and increased anxiety to patients) must be acknowledged.

Concurrent reading also proved statistically significantly faster than independent reading for three of the four radiologists. However, the statistical significance of the decrease in time may not translate into a clinically relevant time-saving benefit for every radiologist; in this study, Radiologist A would save just under 4 minutes per subject with concurrent reading, whereas Radiologist C was hardly affected by it. Nevertheless, when extrapolated to the 187,500 or so CTs per year that would need to be read in a UK national screening programme (Dr A Devaraj, written communication, 22nd February 2013), a time saving of 4 minutes per CT for each radiologist becomes significant. More importantly, concurrent reading did not increase the mean reading time.

The reasons for the variation in time-saving are probably very much radiologist-dependent. Much of the time that may be saved when using radiographers in concurrent reading may be attributable to the time taken to fill in the nodule input proforma for each opacity. Although the time taken to report each nodule for each reader was not recorded (as this would not be practical), it takes approximately 0.5 minutes to complete the reporting of a single nodule, from volumetric segmentation to filling in and electronically pasting the proforma into the nodule report, whereas the review of a radiographer mark takes only a few seconds. However, this time-saving could be nullified by the need to review a large number of identified opacities, especially if modifications to the interpretation need to be made. This may explain why the correlation between the number of nodules per CT and reading time was poorer for concurrent reading. Whereas an increase in nodules per CT can be

expected to cause an increase in reading time when reading independently, the interaction of numbers of nodules and the number of marks requiring changes or deletion may all contribute to a less linear relationship between nodule number and reading time in concurrent reading.

In contrast to concurrent reading using radiographers, there is some evidence for concurrent reading using CAD in lung nodule detection, with evaluations in two studies [294, 354]. Both of these studies assessed CAD as a second reader and as a concurrent reader, using a concurrent reading sequence similar to the current investigation, and performing concurrent and independent reading on the same CTs after an interval of 12-16 weeks and 2 months respectively. However, both studies also have important methodological differences.

Beyer et al. evaluated four radiologists reading 50 CTs performed for varying clinical indications under experimental conditions [294]. Independent and concurrent reading were performed on the same CTs 12 to 16 weeks apart. An individual who was not part of the reading process timed the radiologists. They found that while concurrent reading reduced reading time (by a mean of only 19.8 ± 14.5 seconds), in contrast to the current investigation, pooled sensitivity using concurrent reading with CAD resulted in either lower (for nodules < 1.75mm) or equivalent (for nodules $\geq$ 1.75mm) sensitivity compared to independent reading. They hypothesised that this apparently paradoxical result could have occurred due to the interaction of two effects: (1) a decreased vigilance, as a result of a shortened reading time, and (2) an increase in sensitivity due to additional CAD-identified nodules. Their reference standard was established by including all CAD marks accepted by all radiologists, and review by a consensus panel of two radiologists of all opacities detected by at

least one radiologist (whether or not detected by the CAD system). All CAD marks not detected by any radiologists did not form part of the reference standard. As such, their reference standard was not dissimilar to the current study. However, as explained earlier, the findings of the present study do not support their hypothesis: radiologists' sensitivity increased with concurrent reading, and there were more false positives, possibly signalling an increased, not decreased vigilance. If anything, the increased stringency of their reference standard as result of using two arbiters in consensus would have made the denominator of total positive nodules smaller, and so should have resulted in overall increased sensitivity compared to the readers in the present study, as has been previously illustrated by Armato et al. [285] and discussed in Chapter 1.

In contrast, Foti et al. found that concurrent reading with CAD increased sensitivity for nodule detection in 100 patients with pulmonary metastases, but not to a level reaching statistical significance [354], and with an increased mean reading time (by 60 seconds). However, the two radiologists performing concurrent reading were different to the two radiologists who performed the initial independent reading. They have thus compared two different pairs of radiologists with no statement regarding the baseline independent sensitivity of the radiologists performing concurrent reading, and so their conclusions should be interpreted with caution. It was also unclear from their study whether the reading time stated included the time taken for CAD to process the study, but given the small increase in reading time, it is unlikely that the CAD processing time was taken into account.

The present study has some limitations. The concurrent and independent reading cohorts consisted of different patients, unlike the previous studies mentioned

above [294, 354]. However, this was the most practical way of performing this study prospectively, with real lung screening cases being read under actual reporting rather than experimental conditions, and as such has practical applicability. To minimise potential variations in the number of nodules per CT (in other words, the "level of difficulty" of each CT) affecting the cohorts, a strategy of switching the reading methods after 2 months was adopted. It is reassuring that the number of reference standard nodules per subject between reading methods was not different for any of the four radiologists. The strategy of switching between reading methods would also have mitigated variations in each radiologist's performance during one or other period.

Timing was performed by each individual, to the nearest minute, and not second. A greater statistically significant reduction in time could potentially have been seen if reading times had been recorded to the second, but it was important in this study to detect time reductions that would translate into clinically meaningful reductions, i.e. minutes, not seconds. There was no way to blind each radiologist to the reading method he was using, by definition, and so there could have been a selection bias of the radiologists in choosing which cases they recorded times for. Again, this was the most practical way to record timing given the large volumes of CTs to be reported.

5.5 Summary

- This study demonstrated that radiologists' sensitivity in lung nodule detection could be improved with the use of radiographers as concurrent readers.

- An increase in false positive detections with radiographer-assisted concurrent reading occurred, but this increase was still below that reported for CAD systems.

- Concurrent reading with radiographers was also faster than single reading, but on a per case basis the time saved was relatively modest.

CHAPTER 6: PERFORMANCE OF RADIOGRAPHERS COMPARED TO COMPUTER-AIDED DETECTION (CAD) IN LUNG NODULE DETECTION FOR LUNG CANCER SCREENING

6.1 Introduction

In the preceding chapters, it has been inferred that radiographers have similar sensitivities and lower average false positive detection rates per case than some computer-aided detection (CAD) systems in the published literature. The aim of this investigation was to directly compare radiographer performance with that of a CAD system for screening studies performed in the UKLS trial.

6.2 Materials and Methods

6.2.1 Study design and case selection

Between April and June 2012, the baseline CT examinations of 108 consecutive participants in the LDCT arm of the UKLS trial were read. This time period was chosen to ensure that these examinations had not been included in the investigations performed for Chapters 4 and 5.

6.2.2 LDCT evaluation, arbitration and consensus

The same radiographers in Chapters 4 and 5 performed CT reading prospectively: Radiographer 1 at Local Site 1, Radiographer 2 at Local Site 2, and Radiographers 3 and 4 at the central site. In addition, the same UKLS radiologists

who had read examinations in Chapters 5 and 6 had also read these examinations. Nodules were classified according to the UKLS definitions described in section 2.7.1, and the procedure for nodule recording was exactly the same as described in sections 2.6.2 and 4.2.2.

Radiographers and radiologists performed evaluations independently and uploaded these to the UKLS database, and radiologists did not refer to radiographers' recordings during the course of their reading. To create a database of all recorded nodules, the category and location of opacities identified by each radiographer and radiologist (available on the UKLS database) were separately recorded on a Microsoft Excel spreadsheet (version 2007, Microsoft Corp., Redmond, CA).

6.2.3 Computer-aided detection (CAD) software

The CAD system used for this study (Visia CT Lung System version 3.1, Mevis Medical Solutions, Bremen, Germany; formally known as the ImageChecker CT Lung system, R2 Technology, California) is approved by the US Food and Drug Administration (F.D.A.) for commercial use in the detection of lung nodules.

The CAD algorithm detected and analysed focal opacities by first isolating the lung parenchyma using threshold-based segmentation methods. The software then used techniques such as 3D region-growing and attenuation thresholding to identify groups of voxels with attenuation numbers above a pre-specified threshold, and created regions of interest (ROIs) in each image corresponding to these groups. These ROIs represented potential candidate lesions for the software to analyse. Most of the candidate lesions represented blood vessels and airway walls; thus, the

software needed to distinguish these from true abnormalities. It did this by calculating geometric parameters for each ROI including size, shape, density and location. These features were used to classify each lesion by comparison with a reference database of expected appearances of nodules and vessels. The software then used a decision-making tree to assign a likelihood rating to a lesion, as to whether it represented a true lesion. The CAD system would only highlight a lesion to the reader with a mark (Figure 6.1) if the likelihood rating exceeded a pre-defined threshold.

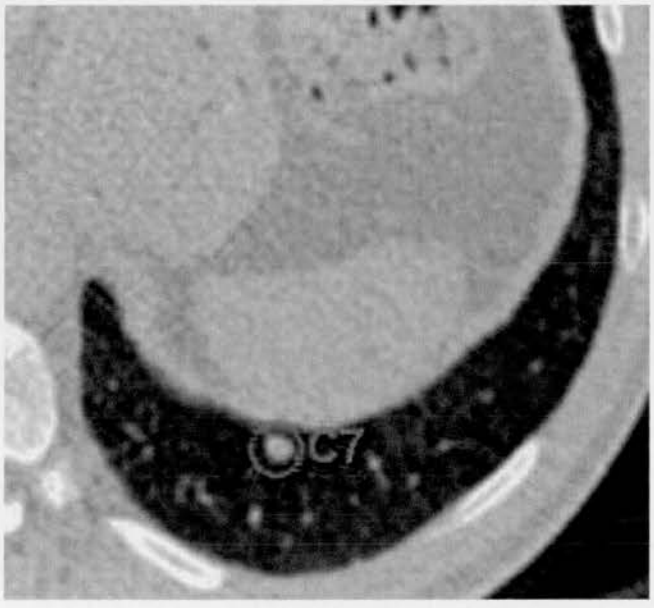

Figure 6.1. Nodule mark in the left lower lobe by the Visia CAD system.

The Visia CT Lung System was configured to detect solid nodules if greater than 4mm and less than 30mm in diameter, but it could detect smaller and larger nodules if they were dense, well-defined and completely surrounded by aerated lung. It was also calibrated to have an average false positive detection (FP) per case of 2.0. It was not designed to detect part-solid or non-solid lesions. It was able to generate volumetric measurements of each lesion. A user could also click on (and annotate)

other potential nodule candidates not highlighted by the CAD. If the system accepted such a candidate, it would generate volumetric measurements of this candidate as well. Even if it was not able to accept the candidate and generate volumetric measurements, the system still allowed the additional nodule to be annotated and measured using callipers. A final report was then generated which indicated the slice number, whether it was detected by CAD or the user, and size and density parameters.

The CAD system had two components: a server and a workstation. The server ran on a 64-bit Windows XP Professional operating system (Microsoft systems, Richmond, Virginia) that was based on a network server at the Royal Brompton Hospital. The workstation ran on a normal Windows-based PC.

6.2.4 CAD evaluation and comparison with radiographers and radiologists

The selected 108 CTs were imported from PACS in Digital Imaging and Communications (DICOM) format for processing on the workstation. Once processed, the CAD results were sent to the server, and CAD marks were available for review on the workstation.

Readings from radiographers and radiologists were visually matched to the CAD readings on the CAD workstation. Opacities that CAD had identified were designated CAD-positive. In order to accurately compare CAD readings against those of radiographers and radiologists, only nodules that the CAD should be expected to identify were included for comparison. However, readers in the UKLS

would designate any nodule greater than 3mm as positive according to UKLS definitions, while the CAD, with its minimum diameter threshold of 4mm, would not necessarily do so. Opacities that had been identified by at least one radiographer or radiologist but that had been missed by CAD were thus measured. If they exceeded the size thresholds or they were subsolid, they were excluded. Any cases of diffuse infection thought secondary to infection (by the arbitrating radiologist) were also excluded. Thus, only nodules that were solid and met size criteria were considered CAD-negative, and radiographer - or radiologist-positive as appropriate. The exclusion of nodules that the CAD could not be expected to identify, based on its operating parameters, ensured that the final dataset of nodules considered positive was not biased towards the radiographers and against the CAD system.

6.2.5 Reference standard

Any opacity agreed by both local and central radiologists was designated a reference standard nodule and required no arbitration. Opacities were presented for arbitration if they had been identified by only one radiologist, or by only a radiographer or CAD and not by at least one radiologist. Arbitration of these discrepancies was provided at the central site by a thoracic radiologist with more than 20 years of experience (the same arbiter as for Chapters 4 and 5).

6.2.6 Statistical analysis

As this study evaluated the performance of radiographers against CAD, the readings of radiologists were excluded from comparison. The number of cases read

by each radiographer was calculated. For each radiographer and for CAD, sensitivity was calculated by dividing the number of true positive nodules identified by the total number of true positive nodules in the reference standard. Sensitivity for subgroups of nodules measuring < 5mm & ≥ 5mm respectively was also calculated. The average FPs per case (expressed as mean and standard deviation) was calculated by determining the total number of FP detections for each case, summing this number and dividing it by the total number of cases read. Each radiographer was compared against CAD only for the cases read by both. Sensitivity was compared using McNemar's test, and average FPs per case were compared using the paired samples t-test.

All analysis was performed using Medcalc (version 12.5.0.0, MedCalc Software, Mariakerke, Belgium). A *P* value of less than 0.05 was assumed to be statistically significant.

6.3 Results

6.3.1 Included studies

Of the 108 baseline CT examinations performed during this study period, 9 cases had to be excluded. Four cases were excluded due to technical difficulties that led to the CAD server being unable to process the CTs. Three cases were excluded because they contained no nodules suitable for CAD analysis (i.e. no solid nodules between 4mm and 30mm in diameter). Two cases were excluded because they were deemed to contain infection-related diffuse nodularity by the arbitrating radiologist.

As such, the final number of cases that were eligible for CAD analysis was 99 examinations. Radiographers 1, 2, 3 and 4 read 40, 36, 47 and 43 cases respectively.

6.3.2 Reference standard

Thirty-two (32.3%) of the 99 CT studies did not contain any nodules. The reference standard thus consisted of 180 nodules in the remaining 67 CT studies. The majority of CTs had 1 (29.3%), 2 (14.1%), 3 (5.1%) or 4 (6.1%) nodules per study (Figure 6.2). The median number of nodules per CT was 1, with a range of 1 to 8 nodules.

The mean diameter of reference standard nodules was 6.6 ± 3.3mm (median 5.7mm, range 2.8-25.2mm).

Of the 180 reference standard nodules, 125 (69.4%) were identified by at least one radiologist, 24 (13.3%) by at least one radiographer but not by any radiologist or CAD, 21 (11.7%) by CAD only, and 10 (5.6%) by both CAD and at least one radiographer but not by any radiologist.

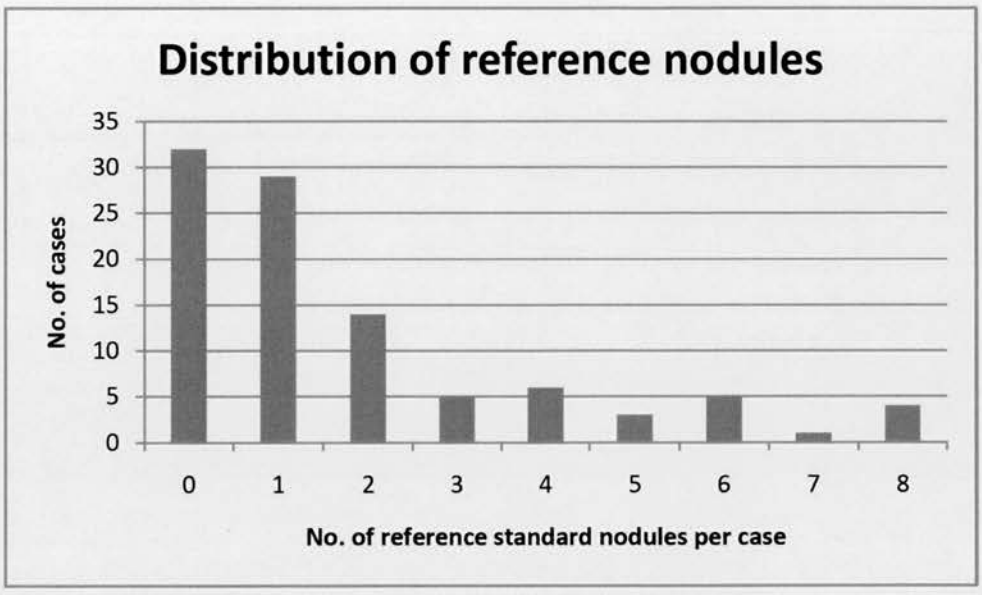

Figure 6.2. Frequency distribution of the number of nodules per CT study.

6.3.3 Comparison of sensitivity

The mean sensitivity of radiographers was 70.1 ± 13.0 % (range, 53.3%-85.0%) compared to an overall CAD sensitivity of 58.9%. Only Radiographer 2 had a significantly higher sensitivity compared to CAD (85.0% versus 63.3% respectively, $P < 0.05$) for the same cases read, while the sensitivities of the remaining three radiographers did not differ significantly from CAD (Table 6.1).

	Radiographer			
	1	2	3	4
No. of cases read	40	36	47	43
Sensitivity of radiographer (%)	71.8	85.0	70.4	53.3
Sensitivity of CAD (%)	59.2	63.3	57.1	63.3
Difference	12.6	21.7	13.3	-10.0
***P* value**	0.19	**0.02**	0.07	0.26

Table 6.1. Sensitivity of radiographers compared to CAD. *P* values are derived from McNemar's test. *P* values in bold indicate statistically significant results.

When subgrouped according to reference standard nodules with diameter < 5mm and ≥ 5mm respectively, no statistically significant difference was found between radiographer and CAD sensitivity (Table 6.2).

	Radiographer							
	1		2		3		4	
Size (mm)	**<5**	**≥ 5**	**<5**	**≥ 5**	**<5**	**≥ 5**	**<5**	**≥ 5**
Radiographer sensitivity (%)	77.3	69.4	87.0	83.8	71.0	70.1	52.0	54.3
CAD sensitivity (%)	54.5	61.2	52.2	70.3	48.4	61.2	48.0	74.3
Difference (%)	22.7	8.2	34.8	13.5	22.6	8.9	4.0	20.0
***P* value**	0.23	0.56	0.06	0.27	0.14	0.35	1.00	0.12

Table 6.2. Sensitivity of radiographers compared to CAD for nodules <5mm and ≥ 5mm. *P* values are derived from McNemar's test.

6.3.4 Comparison of average false positive detections per case

The average FPs per case by the CAD system across all 99 CT examinations was 2.19. When comparing only the same cases read by both CAD and a particular radiographer, the average FPs per case was lower than that of CAD for all four radiographers, and was statistically significantly lower for three of the four radiographers (Table 6.3). No radiographer exceeded more than 2 FPs per case, on average.

		Radiographer			
		1	**2**	**3**	**4**
Average FPs per case	**Radiographer**	0.65 ± 1.25	1.58 ± 2.03	1.15 ± 1.41	0.16 ± 0.43
	CAD	2.10 ± 2.28	2.19 ± 2.82	2.36 ± 2.67	2.09 ± 2.20
	Mean difference	1.45	0.61	1.21	1.93
	***P* value**	**0.0002**	0.10	**0.0004**	**<0.0001**

Table 6.3. Average FPs per case for each radiographer as compared to CAD. *P* values are those derived from the paired samples t-test. *P* values in bold indicate statistically significant results.

6.4 Discussion

The radiographers in this study demonstrated sensitivities that were comparable, and possibly superior, to that of a commercially available CAD system reading the same CTs for the detection of pulmonary nodules in baseline lung cancer screening CT examinations. This comparable sensitivity was maintained even when nodules were subgrouped according to a higher diameter threshold of 5mm or greater - that is, the same higher threshold diameter at which radiographers had already demonstrated greater alignment with radiologists' sensitivity in Chapter 4. In

addition, the majority of radiographers demonstrated a significantly lower average number of FP detections per case than the CAD system.

CAD systems have been extensively investigated as second readers in double-reading scenarios to aid radiologists and shorten reading times. Various studies have demonstrated the ability of CAD to either improve the performance of experienced radiologists, or to augment the performance of junior radiologists or residents such that it becomes comparable to accredited radiologists [288, 289, 291, 293, 336, 337, 346-348, 353, 371, 387]. However, an acknowledged trade-off in the increased sensitivity of a CAD system is the penalty of increased false positive detections (FPs).

CAD systems vary widely in their reported baseline sensitivities. A literature review by Li et al., for example, showed that the variation in sensitivity between nine systems was as much as 21.4% (between 69.8% and 91.2%), with average FPs of between 0.3 and 15 per case reported [384]. This variation between systems can be expected due to differences in detection techniques, CT acquisition and reconstruction parameters, and the databases of "ground truth" that have been used in the establishment of thresholds for lesion identification. A further source of variability is introduced by the use of different reference standards in different studies, which can have a profound impact on sensitivity [285]. Thus, comparison between different studies using different CAD systems is extremely difficult, and efforts to combine these algorithms and encourage more open assessment are underway [295].

However, even when the same CAD system is evaluated in different studies, markedly variable results have been reported. This is best illustrated by three evaluations of predecessor versions of the CAD system (previously known as ImageChecker) used in the current study. Lee et al. investigated the ImageChecker CT LN-1000 CT and reported a CAD sensitivity of 60%, with an average of 1.56 FPs per subject [387]. However, Das et al. compared the ImageChecker system to the Nodule Enhanced Viewing (NEV) system (Siemens Medical Solutions, Forchheim, Germany) and quoted a sensitivity of 73% and an average of 6 FPs per case [291]. Godoy et al. assessed the performance of ImageChecker CT V2.0 for pathologically proven lung cancers that manifested as nodules greater than 4mm, finding a sensitivity of 87.7% and average FPs of 0.9 per case [388]. The current results (sensitivity 58.9%, average FP detections of 2.15 per subject) thus most closely approximate those of Lee et al., but such variation again can be attributable to differences in reference standards. For example, Lee et al. excluded CT examinations with multiple nodules (more than 6 nodules) from analysis while this has not been done in the current study, unless such nodularity was in keeping with infection.

There are also conflicting reports of the impact of CAD-identified FPs on readers' confidence in nodule designation. In a mammography study, a CAD system operating at high average FPs per case lowered confidence in the system, because readers had to review a large number of CAD marks, and so were reluctant to accept these marks [389]. As such, there may be an element of the CAD "crying wolf" on too many occasions, resulting in more CAD rejections, whether true or false positive. However, Nietert et al. recently demonstrated that when radiologists had less than 100% "confidence" in designating a particular opacity as a nodule, their

"confidence" was significantly enhanced by the presence of a CAD mark, although this did not help improve reader accuracy [385]. Indeed, readers' FP rates per case seem to increase with CAD as a second reader [387, 390]. The pattern of increase is also interesting: if an increasing number of CAD detections is sequentially presented to a reader, an initial sharp increase in true positive detections is subsequently attenuated, and FP detections increase, suggesting that there is an optimum FP operating point beyond which the benefit of CAD is negated [290]. Of course, this operating point would again differ between systems.

Given the many wide variations that can occur with a CAD system, the alternative of using a radiographer as a human aid to detection should thus not be discounted. As demonstrated in this study, the sensitivity of radiographers for detecting solid nodules that would warrant follow-up was comparable to that of CAD. Furthermore, the number of nodules identified by only radiographers or only CAD but not by any radiologist was almost equal, suggesting that the additive effect of either method on the sensitivity of a radiologist would probably be equal (although each method would result in different nodules being called and missed).

The argument for using CAD rather than radiographers as aids in pulmonary nodule detection are the same as against using any human reader, in that fatigue may interfere with the nodule reading task, there may be a high turnover of personnel, more person-hours are required overall, and greater demands may be placed on infrastructure (e.g. more reporting workstations needed). However, the radiographers in the current study read the CT examinations prospectively in the course of the UKLS pilot trial, and not under experimental conditions. Despite this their sensitivity at the very least matched the CAD system, suggesting that a radiographer's vigilance

can be maintained without adversely affecting performance. Importantly, the maintained sensitivity of radiographers did not come at the expense of higher average FP detections per case.

The notion that a CAD system can be recalibrated, by adjusting the average FPs per case and sensitivity such that it is more aligned with a particular group of readers, has been explored under experimental conditions [289, 290]. However, these parameters are not adjustable in commercial software such as the CAD system in this study, and so such re-alignment is presently a fallacy. Also, the reference standards for CAD nodule identification - the databases of "ground truth" that have been used in the establishment of thresholds and feature classification - are not known to the reader, and cannot be re-adjusted. In contrast, findings can be discussed with radiographers, and the reasons for rejections of nodule marks explained. All radiographers in a particular screening programme can be evaluated on the same training and testing datasets, and re-training is possible. In this way, direct "calibration" by a radiologist is more feasible for radiographers than it presently is for CAD. A radiologist is arguably more likely to have confidence in an assisted reader when that reader can be aligned with the radiologist in this manner. Of course, such "calibration" by a radiologist with inherently low sensitivity could also have a detrimental effect on radiographer's sensitivity, underscoring the need to ensure that reading radiologists in a screening programme meet a minimum acceptable level of sensitivity at the outset.

Another important advantage of using radiographers is that their time-saving role extends not only to nodule detection, but to logistical tasks such as matching nodules between readers on a database, as well as nodule matching on baseline and

follow-up subjects. Although nodule matching software for comparing studies is available and accurate [391], human confirmation of matching accuracy will still be required. This time-saving should not be underestimated in a screening programme, as experience from the UKLS pilot trial suggests considerable effort needs to be expended on such tasks.

Radiographers can also alert the radiologist to findings (such as a large nodule) that need urgent referral. CAD systems such as the one used in this study do not currently have the functionality to generate alerts for urgent findings.

This study has some limitations. Only solid nodules were included, as the CAD algorithm was not optimized to detect non- or part-solid nodules. The impact of CAD (whether as a concurrent or second reader) on radiologists' sensitivity in the setting of an actual lung screening trial cannot be directly evaluated, as there was no precedent or ethical approval for CAD to be used in such a decision-making capacity in the UKLS pilot trial. As with all studies on nodule detection in the absence of histopathological proof, the reference standard was determined by radiologists' agreement.

6.5 Summary

- Sensitivity of radiographers for CT pulmonary nodule detection was comparable, and possibly superior, to that of the CAD system in this study.
- The comparable sensitivity of radiographers was maintained both for nodules smaller than 5mm, and nodules equal to or greater than 5mm.
- The majority of radiographers had significantly lower average false positive detections per case compared to CAD.
- Radiographers may thus be able to act as aids for nodule detection that are at least as sensitive as CAD, and with a potentially lower false positive penalty.

CHAPTER 7: THE IMPACT OF THE NUMBER OF READERS AND METHODS OF ARBITRATION ON READER PERFORMANCE IN LUNG NODULE DETECTION

7.1 Introduction

The numbers of readers used in lung cancer screening studies with low dose CT have varied. Some have used single reading [151, 164, 392] whereas the majority have used double-reading strategies [142, 146, 150, 156, 158-160, 174, 393]. The rationale behind using two readers is that lung nodule detection rates are improved [334, 336] and hence potentially fewer lung cancer are missed. However few studies have evaluated the impact of reader numbers on lung nodule identification, and these have been limited in certain ways.

First, the reference standards in previous studies of nodule detection performance have varied in their derivation. The reference standard usually incorporates the readings of observers, as well as an expert opinion. The expert opinion may be that of a single radiologist only [334], or a consensus opinion of two [289, 290, 293, 294, 336, 337] or sometimes even three [345, 346] experienced radiologists. The variation in experience of these experts may in turn influence the reference standard [394]. Additionally, the expert or experts may either perform an initial independent search ("free-search") for nodules themselves, followed by a review of all positive findings identified by the observers (directed search), or they may only perform the latter.

In addition the issue of resolving discordant readings has not been fully addressed in previous studies. Discordant readings arise because one radiologist has missed (a discrepancy in detection) or deliberately chosen to ignore (a discrepancy in interpretation) a finding. Discrepancies between radiologists can be resolved in a variety of ways, including:

1. Combining radiologists' readings (which assumes all discordant readings are positive findings, regardless of whether the discordance is due to differences in detection or interpretation) [334, 342, 343];
2. A process of consensus where readers discuss individual discrepancies [159, 393] to determine if these findings should be designated positive, thus facilitating debate when there are discordances in both detection and interpretation. This method is commonly accepted both as a valid method of resolving discrepancy [208, 344]. However, consensus interpretation is itself subject to inherent limitations such as "groupthink" [344];
3. Using an independent arbiter;
4. Using a form of "internal" arbitration where a reader is independently shown and asked to form an opinion on nodules identified by other readers and not by themselves [285, 395]; and
5. A combination of these methods [142, 156, 174].

The previous chapters have shown that radiographers can assist radiologists in lung nodule detection. However, regardless of whether radiologists read independently or concurrently, the final interpretation of an opacity rests with

radiologists, and thus more needs to be known about the impact of the numbers of readers and of methods of resolving discrepancy.

Therefore, the purpose of the current investigation was to evaluate the impact of double- and triple-reading on lung nodule detection accuracy, and to assess the effect of different methods of arbitration on this detection accuracy.

7.2 Materials and methods

7.2.1 Construction of the reading dataset

The CT studies used in this study were all obtained from the NELSON study, as described in Chapter 2. From the 202 procured scans, 100 CT studies were randomly chosen that contained a variable mix of nodules (as identified by NELSON readers). Fifteen of these studies could not be processed by the LungCARE application and were excluded. Thus, 85 CT scans formed the total dataset for this investigation, hereon referred to as the *reading dataset.*

7.2.2 CT reading

7.2.2.1 Reader training and image manipulation

Five expert thoracic radiologists who work in different institutions were invited to take part in this study. The radiologists had different levels of radiology experience: Reader A, 19 years; Reader B, 13 years; Reader C, 7 years; Reader D, 21 years; and Reader E, 9 years.

All radiologists were given a tutorial on usage of the LungCARE software on the Syngo workstation, including image manipulation and measurement tools. The radiologists were also given examples of the nodules they were being asked to mark based on the nodule definitions below. These examples were taken from NELSON CT examinations which had been obtained during the procurement exercise but which were not part of the reading dataset, to avoid recall bias.

Studies loaded into LungCARE were presented in a 2 x 2 viewing partition with a default window setting level -500 HU, width 1500 HU, and arranged in the following manner: top left panel, maximum intensity projections (MIPs) with a default setting of 10mm thickness; top right panel, 1mm-collimation axial images; and bottom left panel, 0.7mm-collimation coronal images. The bottom right panel was initially blank, and was used to display volumetric segmentation of a selected nodule (Figure 7.1).

Figure 7.1. Example of marking and annotation of a nodule within LungCARE by a radiologist on an anonymised LDCT study.

Readers were free to alter MIP thickness and window settings. Readers were not asked to record volumetric measurements of each nodule, nor any nodule characteristics, so that there was no interruption to their workflow.

Nodules were marked and annotated by a reader using the "set marker" and "annotate" options respectively, both accessible on a right mouse click. The "annotate" feature provided a free text box where readers inserted their categorisation of an opacity according to the definitions in Table 7.1 below. Each reader saved their readings as a DICOM structured report (DICOM SR) with their initials as a unique identifier.

7.2.2.2 Nodule characterisation and reading timeframe

The radiologists performed two rounds of reading each.

First reading (R1)

During the first reading (R1), each radiologist performed a free search of the entire CT dataset, and classified all opacities as positive or negative (Table 7.1). A nodule was defined according to the Fleischner Society Glossary of Terms for Thoracic Imaging - that is, a rounded or irregular opacity, well or poorly defined, measuring up to 3cm in diameter [230].

Category	***Definition***
Positive	Definite non-calcified solid nodule > 3mm in maximum transverse diameter; **OR** any non-solid or part-solid nodule; **OR** an intrapulmonary lymph node, that satisfied all 5 criteria of: -a smooth margin -an ovoid or triangular (but not round or spherical) shape - < 8mm in maximum transverse diameter - within 5mm of the pleura (or lies within an interlobar or accessory fissure) - at least one interlobular septum radiating from its surface
Negative	A non-nodular opacity, that did not have all the characteristics of a nodule or intrapulmonary lymph node (e.g. a linear or curvilinear opacity); **OR** A very small opacity, that met the definition of a nodule but was < 3mm (NB: Radiologists were asked to ignore opacities that they judged to be an obvious benign opacity (e.g. scar-like opacities or nodules with a benign pattern of calcification)

Table 7.1. Definition of opacities.

The purpose of requiring radiologists to mark opacities that were considered negative was to subsequently allow the distinction to be made between opacities that were missed and those that had been deliberately ignored.

Following the completion of R1, the DICOM SR of each reader was reviewed in LungCARE to identify each recorded opacity. As LungCARE provides only z-axis and not x- and y-axis coordinates, each opacity was visually matched to the marks on LungCARE using a separate open-source DICOM software package (Osirix version 4.1.1, Osirix Foundation, Geneva, Switzerland) capable of recording all 3 spatial coordinates, so that each nodule had a unique set of identifying coordinates. In this way, two nodules within the same axial slice could be spatially distinguished and not mistaken for each other.

The designation of each opacity as positive or negative was also recorded for each reader on Osirix. The spatial coordinates and designations of each opacity were tabulated for each reader using a commercially available PC database (Microsoft Excel version 2007, Microsoft Corp., Redmond, CA, USA). Once all detected opacities were matched and tabulated in this manner, it was possible to generate, for each radiologist, a list of opacities that had not been recorded by him or her, but had been identified by at least one other radiologist.

Second reading (R2)

In the second reading (R2), each radiologist was directed to opacities that had not been recorded by them but had been identified as a positive nodule by at least one other radiologist, and asked to categorise them as positive or negative. Each radiologist was not shown the number of other radiologists who had identified the opacity. This was done to prevent a "forced" consensus, whereby a radiologist's opinion might inadvertently be influenced by the opinion of the majority.

7.2.3 Derivation of the reference standard

It was decided at the outset that for a panel of five radiologists, agreement between any four radiologists after R2 as to the positivity or negativity of a nodule would constitute the reference standard. A reference standard requiring all five to be in agreement was regarded as being too stringent, and would have resulted in a small number of nodules within the reference standard. In turn, this could have spuriously elevated sensitivity [285] and could also have excluded potentially important nodules. A reference standard requiring only three radiologists to be in agreement, on the other hand, was regarded as unsatisfactory, as it would have required only a small majority to be in agreement.

7.2.4 Simulation of reader combinations and methods of arbitration

With five radiologists it was possible to simulate 10 pairs of double-readers and 20 combinations of triple readers. The complex permutations and combinations that arose from these pair and triplet-combinations are illustrated in section 7.3.

7.2.4.1 Double-reading

In double-reading, there are a variety of methods available to deal with discordant readings. As indicated in the introduction, a method commonly used is "consensus" whereby readers discuss (often alongside one another) discrepancies and decide upon a final answer. However, this method was not used because of its inherent limitations [344]. Instead four possible ways of dealing with discrepancies in double-reading were evaluated:

1. **Double-reading with no arbitration:**

   In this scenario, any finding deemed positive in the first round of reading by one OR other radiologist in the pair was considered positive for that pair, without any need for arbitration. In other words readers' positive nodules were combined.

2. **Double-reading with only internal arbitration:**

   In this scenario, a reader's interpretation of nodules identified by another radiologist but not by him or her was taken into account. After the second round of reading (R2), opacities which were considered positive by both parties were deemed as positive. Opacities that still had discrepant designations were considered negative.

3. **Double-reading with only external arbitration:**

   In this scenario, each radiologist within a pair was assumed to only have performed a single round of reading (R1), and would not have had the opportunity to view the other radiologists' detections. An opacity identified by both radiologists was accepted as positive without any arbitration. Any opacity designated a nodule by only one radiologist was referred for external arbitration by an independent thoracic radiologist with 10 years of experience.

4. **Double-reading with internal arbitration followed by external arbitration:**

   In this scenario, readers' interpretation after both R1 and R2 was taken into account (akin to method 2 above), but this time with external arbitration of outstanding discrepancies by an independent reader (the same independent thoracic radiologist as in method 3 above).

7.2.4.2 Triple reading

Triple reading was modelled by combining each pair after R2 with one of the remaining three readers. For each pair, there were 3 possible triplet combinations, giving rise to 30 triplet combinations in total. For a particular triplet, an opacity was regarded as positive if it had been identified by at least two of the three radiologists.

7.2.5 Measurement of opacity size

The maximum and minimum transverse diameters of each opacity were recorded, and the average diameter calculated. These measurements were performed separately by a radiology resident with three years of general radiology experience, to ensure that this task did not distract any of the five radiologists in performing their primary functions of detection and interpretation during the two rounds of reading.

7.2.6 Statistical analysis

The total number, median and range of all opacities and of reference standard nodules identified in the 85 CT studies was calculated. Average nodule diameter for both reference standard nodules and negative opacities was expressed as mean, standard deviations, median and range.

Sensitivity and specificity were calculated as explained in Chapter 2. The following comparisons of the sensitivities and specificities of the different methods of reading and arbitration were made:

1. Double-reading with no arbitration versus single reading
2. Double-reading with only internal arbitration versus single reading
3. Double-reading with only external arbitration versus single reading
4. Double-reading with only external arbitration versus double-reading with only internal arbitration

5. Double-reading with internal arbitration followed by external arbitration versus single reading
6. Double-reading with internal arbitration followed by external arbitration versus double-reading with only internal arbitration only
7. Triple reading versus double-reading with internal arbitration.

All comparisons were made using McNemar's test. All analysis was performed using Medcalc (version 12.5.0.0, MedCalc Software, Mariakerke, Belgium). Results were considered statistically significant if the *P* value was less than 0.05.

7.3 Results

7.3.1 General data

All subjects were male, aged 53-79 years old (mean 62.6 years, median 63). Three of the 85 CTs contained no opacities identified by any of the five radiologists. A total of 528 opacities were identified by all radiologists in the remaining 82 CTs, corresponding to 6.44 nodules per subject. A range of 0 to 22 opacities (median 5) were identified per subject. The distribution of subjects according to the number of opacities is shown in Figure 7.2.

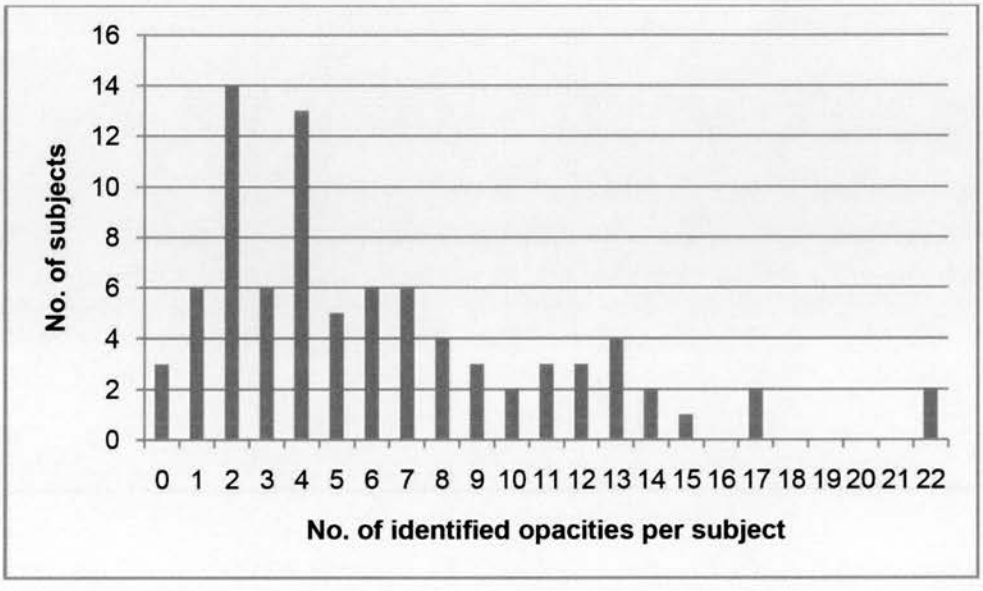

Figure 7.2. The number of identified opacities per subject.

7.3.2 Reference standard

Of the 528 opacities, 186 (35.2%) met the reference standard criteria (i.e. agreed as a nodule by any four of five radiologists) in 64/85 patients (75.3%). The nodules comprising the reference standard had a mean diameter of 4.6 ± 1.6mm (median 4.4mm, range 2.1mm to 15.6mm). The negative opacities had a mean diameter of 3.5 ± 1.3mm (median 3.4mm, range 1.5 to 13.3mm).

7.3.3 Performance of individual radiologists

The sensitivity and specificity of the individual radiologists is shown in Table 7.2. The mean and median sensitivity were 64.5% and 58.6% respectively; the mean and median specificity were 84.6% and 89.3% respectively.

Radiologist	*A*	*B*	*C*	*D*	*E*
Sensitivity (%)	86.6	74.7	52.2	50.5	58.6
Specificity (%)	66.7	85.2	89.3	91.6	90.4

Table 7.2. Performance of individual radiologists.

7.3.4 Summary of comparisons

The effects of double- and triple reading and methods of arbitration are summarised in Table 7.3.

Comparison	*Effect on sensitivity*			*Effect on specificity*		
	Increase	**Decrease**	**Unchanged**	**Increase**	**Decrease**	**Unchanged**
Double-reading						
No arbitration vs single reader	20/20	0/20	0/20	0/20	20/20	0/20
Internal arbitration vs single reader	8/20	2/20	10/20	16/20	1/20	3/20
External arbitration vs single reader	11/20	4/20	5/20	4/20	6/20	10/20
External arbitration vs internal arbitration	4/10	2/10	4/10	0/10	10/10	0/10
Internal and external arbitration vs single reader	18/20	0/20	2/20	4/20	10/20	6/20
Internal and external arbitration vs internal arbitration	9/10	0/10	1/10	0/10	10/10	0/10
Triple reading vs internal arbitration	30/30	0/30	0/30	0/0	30/30	0/30

Table 7.3. Summary of effects of double- and triple reading methods on sensitivity and specificity. Figures are proportions of pairs (for double-reading) or triplets (for triple reading).

7.3.5 Performance of double-reading and arbitration

The effect of double-reading using the four scenarios is presented in Tables 7.4-7.15. For each reader, four different pair combinations were possible; thus, a total of 20 pair-combinations were assessed.

7.3.5.1 Double-reading without arbitration

Sensitivity significantly improved for all readers across all pair-combinations (Table 7.4). The median sensitivity improved from 58.6% (range 50.5%-86.6%) to 83.6% (range 66.7%-92.5%). There was a corresponding significant decrease in specificity for all readers, with median specificity decreasing from 89.3% (range 66.7%-91.2%) to 77.1% (range 57.7%-84.9%) (Table 7.5).

7.3.5.2 Double-reading with only internal arbitration

Double-reading with internal arbitration had a variable effect on sensitivity as compared to a single reader (Table 7.6). Sensitivity significantly increased for 8 of the 20 pair-combinations (median 16.4%, range 9.1%-32.3%), decreased in 2 pair-combinations (mean 20.7%, range 16.1%-25.3%), and did not change in the remaining 10.

The improvement in specificity was more consistent: specificity increased significantly for 16 pairs (median 7.7%, range 4.6%-29.3%), decreased in 1 pair (4.6%), and did not change in 3 pairs (Table 7.7).

Before double-reading without arbitration		*After double-reading without arbitration*				
		Reader				
		A	**B**	**C**	**D**	**E**
Reader	**A**		5.4 **(0.002)**	5.9 **(0.0010)**	4.8 **(0.039)**	4.3 **(0.008)**
	B	17.2 **(<0.0001)**		9.1 **(<0.0001)**	5.9 **(0.001)**	8.6 **(<0.0001)**
	C	40.3 **(<0.0001)**	31.7 **(<0.0001)**		17.7 **(<0.0001)**	23.1 **(<0.0001)**
	D	40.9 **(<0.0001)**	30.1 **(<0.0001)**	19.4 **(<0.0001)**		16.1 **(<0.0001)**
	E	32.2 **(<0.0001)**	24.7 **(<0.0001)**	16.7 **(<0.0001)**	8.1 **(0.0001)**	

Table 7.4. Changes in sensitivity using double-reading without arbitration. For a given pair, each figure represents the percentage change in sensitivity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant change.

Before double-reading without arbitration		*After double-reading without arbitration*				
		Reader				
		A	**B**	**C**	**D**	**E**
Reader	**A**		-6.1 **(<0.0001)**	-9.0 **(<0.0001)**	-5.8 **(<0.0001)**	-5.5 **(<0.0001)**
	B	-24.6 **(<0.0001)**		-9.3 **(<0.0001)**	-7.0 **(<0.0001)**	-6.7 **(<0.0001)**
	C	-31.6 **(<0.0001)**	13.3 **(<0.0001)**		-7.0 **(<0.0001)**	-8.1 **(<0.0001)**
	D	-30.7 **(<0.0001)**	-13.3 **(<0.0001)**	-9.3 **(<0.0001)**		-6.7 **(<0.0001)**
	E	-29.3 **(<0.0001)**	-11.9 **(<0.0001)**	-9.3 **(<0.0001)**	-5.5 **(0.0001)**	

Table 7.5. Changes in specificity using double-reading without arbitration. For a given pair, each figure represents the percentage change in specificity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

Before double-reading with internal arbitration		*After double-reading with internal arbitration*				
		Reader				
		A	**B**	**C**	**D**	**E**
Reader	**A**		1.6 (0.5465)	-25.3 **(<0.0001)**	-3.8 (0.1904)	-0.5 (1.000)
	B	13.4 **(<0.0001)**		-16.1 **(<0.0001)**	-3.8 (0.2109)	3. 2 (0.2636)
	C	9.1 **(0.0002)**	6.5 **(0.0139)**		-2.7 (0.3588)	-1.8 (0.7893)
	D	32.3 **(<0.0001)**	20.4 **(<0.0001)**	-1.1 (0.8875)		9.1 **(0.0008)**
	E	27.4 **(<0.0001)**	19.4 **(<0.0001)**	-5.4 (0.2453)	1.1 (0.8312)	

Table 7.6. Changes in sensitivity using double-reading with only internal arbitration. For a given pair, each figure represents the percentage change in sensitivity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

Before double-reading with internal arbitration		*After double-reading with internal arbitration*				
		Reader				
		A	**B**	**C**	**D**	**E**
Reader	**A**		17.4 **(<0.0001)**	29.3 **(<0.0001)**	26.4 **(<0.0001)**	19.1 **(<0.0001)**
	B	-1.1 (0.6171)		11.9 **(<0.0001)**	11.6 **(<0.0001)**	5.8 **(0.0008)**
	C	6.7 **(<0.0001)**	7.8 **(<0.0001)**		8.4 **(<0.0001)**	7.5 **(<0.0001)**
	D	1.4 (0.4414)	5.3 **(0.0013)**	6.1 **(<0.0001)**		4.6 **(0.0014)**
	E	-4.6 **(0.027)**	0.6 (0.8597)	6.4 **(0.0003)**	5.8 **(<0.0001)**	

Table 7.7. Changes in specificity using double-reading with only internal arbitration. For a given pair, each figure represents the percentage change in specificity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

7.3.5.3 Double-reading with only external arbitration

Double-reading using only external arbitration also had a variable effect on sensitivity as compared to a single reader (Table 7.8). Sensitivity significantly increased for 11 of the 20 pair-combinations (median 14.0%, range 5.9%-26.9%), decreased in 4 pair-combinations (median 7.8%, range 4.8%-10.2%), and did not change in the remaining 5 pair-combinations.

However, specificity increased significantly for only 4 pairs (median 14.3%, range 12.8%-15.7%), decreased in 6 pairs (median 6.4%, range 4.9%-10.4%), and did not change in 10 pairs (Table 7.9).

When compared to double-reading with only internal arbitration, double-reading with only external arbitration improved sensitivity in 4 of 10 pairs (median 12.9%, range 9.1%-17.7%), decreased sensitivity in 2 of 10 pairs (median 7.0%, range 6.5%-7.5%), and did not change sensitivity in the remaining 4 pairs (Table 7.10). However, significantly decreased specificity was seen for all 10 pairs (median 6.4%, range 4.6%-13.6%) (Table 7.11).

Before double-reading with external arbitration — Reader	*After double-reading with external arbitration* — Reader A	B	C	D	E
A		-4.8 **(0.0490)**	-7.5 **(0.0013)**	-10.2 **(0.0005)**	-8.1 **(0.0041)**
B	7.0 **(0.0146)**		-1.6 (0.6291)	-5.4 (0.0525)	-2.2 (0.5563)
C	26.9 **(<0.0001)**	21.0 **(<0.0001)**		6.5 (0.0973)	12.4 **(0.0025)**
D	25.8 **(<0.0001)**	18.8 **(<0.0001)**	8.1 **(0.0041)**		11.3 **(0.0002)**
E	19.9 **(<0.0001)**	14.0 **(<0.0001)**	5.9 **(0.0266)**	3.2 (0.2101)	

Table 7.8. Changes in sensitivity using double-reading with only external arbitration. For a given pair, each figure represents the percentage change in sensitivity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

Before double-reading with external arbitration — Reader	*After double-reading with external arbitration* — Reader A	B	C	D	E
A		12.8 **(<0.0001)**	15.7 **(<0.0001)**	14.5 **(<0.0001)**	14.2 **(<0.0001)**
B	-5.8 **(0.0021)**		1.7 (0.2632)	1.4 (0.4414)	0 (0.8501)
C	-7.0 **(0.0075)**	-2.3 (0.3580)		2.6 (0.2002)	0.9 (0.7656)
D	-10.4 **(<0.0001)**	-4.9 **(0.0147)**	0.3 (1.0000)		-1.4 (0.4244)
E	-9.6 **(<0.0001)**	-5.2 **(0.0046)**	-0.3 (1.0000)	-0.3 (1.0000)	

Table 7.9. Changes in specificity using double-reading with only external arbitration. For a given pair, each figure represents the percentage change in specificity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

		Sensitivity		
Reader Pair	**Only internal arbitration (%)**	**Only external arbitration (%)**	**Absolute percentage change**	***P* value (McNemar's test)**
AB	88.2	81.7	-6.5	**0.019**
AC	61.3	79.0	17.7	**<0.0001**
AD	82.8	76.3	-6.5	0.082
AE	86.0	78.5	-7.5	**0.0108**
BC	58.6	73.1	14.5	**0.0003**
BD	71.0	69.4	-1.6	0.7277
BE	78.0	72.6	-5.4	0.0776
CD	49.5	58.6	9.1	**0.0171**
CE	53.2	64.5	11.3	**0.0043**
DE	59.7	61.8	2.1	0.4807

Table 7.10. Comparison of the sensitivity of double-reading with only external arbitration to double-reading with only internal arbitration. *P* values in bold font indicate statistically significant changes.

		Specificity		
Reader Pair	**Only internal arbitration (%)**	**Only external arbitration (%)**	**Absolute percentage change**	***P* value (McNemar's test)**
AB	84.1	79.4	-4.7	**<0.0001**
AC	95.9	82.3	-13.6	**<0.0001**
AD	93.0	81.2	-11.8	**<0.0001**
AE	85.8	80.9	-4.9	**<0.0001**
BC	97.1	87.0	-10.1	**<0.0001**
BD	96.8	86.7	-10.1	**<0.0001**
BE	91.0	85.2	-5.8	**<0.0001**
CD	97.7	91.9	-5.8	**<0.0001**
CE	96.8	90.1	-6.7	**<0.0001**
DE	96.2	90.1	-6.1	**<0.0001**

Table 7.11. Comparison of the specificity of double-reading with only external arbitration to double-reading with only internal arbitration. *P* values in bold font indicate statistically significant changes.

7.3.5.4 Double-reading with internal arbitration followed by external arbitration

Introducing external arbitration on the pair combinations only when there were outstanding discrepancies after internal arbitration demonstrated increased sensitivity for 18 out of 20 pairs over a wide range (median 18.5%, range 4.3%-40.9%), and did not change in the remaining 2 pairs. Sensitivity did not decrease for any reader (Table 7.12).

This method increased specificity in only 4 pair-combinations (median 9.1%, range 5.5%-12.8%), and decreased specificity in 10 pair-combinations (median 8.4%, range 4.1%-17.1%), with no change in the remaining 6 pairs (Table 7.13).

When compared to double-reading with only internal arbitration, double-reading with internal followed by external arbitration significantly improved sensitivity in 9 out of 10 pairs (median 8.6%, range 3.8%-26.3%) (Table 7.14), but at the expense of diminished specificity in all 10 pairs (median 11.4%, range 7.0%-16.5%) (Table 7.15).

Before double-reading with internal and external arbitration		After double-reading with internal and external arbitration				
		Reader				
		A	B	C	D	E
Reader	A		4.3 **(0.0215)**	1.1 (0.8445)	4.8 **(0.0117)**	3.2 (0.1460)
	B	16.1 **(<0.0001)**		7.0 **(0.0146)**	4.8 **(0.0352)**	7.5 **(0.0013)**
	C	35.5 **(<0.0001)**	29.6 **(<0.0001)**		18.8 **(<0.0001)**	24.7 **(<0.0001)**
	D	40.9 **(<0.0001)**	29.0 **(<0.0001)**	20.4 **(<0.0001)**		17.2 **(<0.0001)**
	E	31.2 **(<0.0001)**	23.7 **(<0.0001)**	18.3 **(<0.0001)**	9.1 **(0.0001)**	

Table 7.12. Changes in sensitivity using double-reading with internal followed by external arbitration. For a given pair, each figure represents the percentage change in sensitivity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

Before double-reading with internal and external arbitration		After double-reading with internal and external arbitration				
		Reader				
		A	B	C	D	E
Reader	A		5.5 **(0.0279)**	12.8 **(<0.0001)**	11.6 **(<0.0001)**	6.7 **(0.0042)**
	B	-13.0 **(<0.0001)**		0.9 (0.6676)	-0.6 (0.8638)	-5.2 **(0.0019)**
	C	-9.9 **(0.0002)**	-3.2 (0.1930)		1.4 (0.5322)	-2.9 (0.1939)
	D	-13.3 **(<0.0001)**	-7.0 **(0.0011)**	-0.9 (0.6476)		-5.2 **(0.0058)**
	E	-17.1 **(<0.0001)**	-10.4 **(<0.0001)**	-4.1 **(0.0140)**	-4.1 **(0.0140)**	

Table 7.13. Changes in specificity using double-reading with internal followed by external arbitration. For a given pair, each figure represents the percentage change in specificity as compared to the single reader, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

		Sensitivity		
Reader Pair	**Only internal arbitration (%)**	**Internal and external arbitration (%)**	**Absolute percentage change**	***P* value (McNemar's test)**
AB	88.2	90.9	2.7	0.4984
AC	61.3	87.6	26.3	**<0.0001**
AD	82.8	91.4	8.6	**0.0002**
AE	86.0	89.8	3.8	**0.0233**
BC	58.6	81.7	23.1	**<0.0001**
BD	71.0	79.6	8.6	**0.0002**
BE	78.0	82.3	4.3	**0.0133**
CD	49.5	71.0	21.5	**<0.0001**
CE	53.2	76.9	23.7	**<0.0001**
DE	59.7	67.7	8.0	**0.0003**

Table 7.14 Comparison of the sensitivity of double-reading with internal followed by external arbitration, to double-reading with only internal arbitration. *P* values in bold font indicate statistically significant changes.

		Specificity		
Reader Pair	**Only internal arbitration (%)**	**Internal and external arbitration (%)**	**Absolute percentage change**	***P* value (McNemar's test)**
AB	84.1	72.2	-11.9	**<0.0001**
AC	95.9	79.4	-16.5	**<0.0001**
AD	93.0	78.3	-14.7	**<0.0001**
AE	85.8	73.3	-12.5	**<0.0001**
BC	97.1	86.1	-11.0	**<0.0001**
BD	96.8	84.6	-12.2	**<0.0001**
BE	91.0	80.0	-11.0	**<0.0001**
CD	97.7	90.7	-7.0	**<0.0001**
CE	96.8	86.4	-10.4	**<0.0001**
DE	96.2	86.4	-9.8	**<0.0001**

Table 7.15 Comparison of the specificity of double-reading with internal followed by external arbitration, to double-reading with only internal arbitration. *P* values in bold font indicate statistically significant changes.

7.3.6 Performance of triple reading

With triple reading, a universal improvement in sensitivity occurred for each of the 10 pairs, across each of its 3 possible combinations (median 23.9%, range 6.5%-47.8%) (Table 7.16). However, as with double-reading with external arbitration (either alone or following internal arbitration), a universal decrease in specificity for all combinations was also observed, with a mean reduction of 10.5% (Table 7.17).

			After triple reading				
			Additional reader forming triplet				
			A	**B**	**C**	**D**	**E**
Before triple reading	**Reader pair**	**AB**			9.2 **(<0.0001)**	6.5 **(0.0015)**	7.0 **(0.0009)**
		AC		36.0 **(<0.0001)**		36.0 **(<0.0001)**	36.5 **(<0.0001)**
		AD		11.8 **(<0.0001)**	14.5 **(<0.0001)**		11.3 **(<0.0001)**
		AE		9.1 **(<0.0001)**	11.8 **(<0.0001)**	8.1 **(0.0003)**	
		BC	38.7 **(<0.0001)**			30.1 **(<0.0001)**	31.7 **(<0.0001)**
		BD	23.7 **(<0.0001)**		17.7 **(<0.0001)**		14.5 **(<0.0001)**
		BE	17.2 **(<0.0001)**		12.4 **(<0.0001)**	7.5 **(0.0005)**	
		CD	47.8 **(<0.0001)**	39.2 **(<0.0001)**			34.4 **(<0.0001)**
		CE	44.6 **(<0.0001)**	37.1 **(<0.0001)**		30.6 **(<0.0001)**	
		DE	34.4 **(<0.0001)**	34.9 **(<0.0001)**	24.9 **(<0.0001)**		

Table 7.16. Changes in sensitivity using triple reading, compared to double-reading with only internal arbitration. Figures are percentage changes in sensitivity, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

			After triple reading				
			Additional reader forming triplet				
			A	**B**	**C**	**D**	**E**
Before triple reading	**Reader pair**	**AB**			-5.2 **(<0.0001)**	-6.1 **(<0.0001)**	-9.6 **(<0.0001)**
		AC		-17.1 **(<0.0001)**		-8.4 **(<0.0001)**	-15.7 **(<0.0001)**
		AD		-15.1 **(<0.0001)**	-5.5 **(<0.0001)**		-15. 7 **(<0.0001)**
		AE		-11.3 **(<0.0001)**	-5.5 **(<0.0001)**	-8.4 **(<0.0001)**	
		BC	-18.3 **(<0.0001)**			-5.5 **(<0.0001)**	-10.4 **(<0.0001)**
		BD	-18.8 **(<0.0001)**		-5.2 **(<0.0001)**		-12.2 **(<0.0001)**
		BE	-16.5 **(<0.0001)**		-4.3 **(0.0003)**	-6.4 **(<0.0001)**	
		CD	-10.1 **(<0.0001)**	-6.1 **(<0.0001)**			-5.5 **(<0.0001)**
		CE	-16.5 **(<0.0001)**	-10.1 **(<0.0001)**		-4.6 **(0.0002)**	
		DE	-18.8 **(<0.0001)**	-18.3 **(<0.0001)**	-4.1 **(0.0005)**		

Table 7.17. Changes in specificity using triple reading, compared to double-reading with only internal arbitration. Figures are percentage changes in sensitivity, with *P* values (McNemar's test) in parentheses. *P* values in bold font indicate statistically significant changes.

7.4 Discussion

The results of this investigation show that the effect of double-reading on the performance of individual experienced radiologists is immensely variable, depending on the method used to deal with discordant readings.

Double-reading without arbitration was shown to universally increase sensitivity for nodule detection. The impetus for improving nodule detection alone -

without necessarily an interpretation component - is strong when one considers that most failures of lung cancer diagnosis are due to failures of detection [381, 382]. The findings of the present investigation support the notion that increasing the number of readers in detection tasks improves the sensitivity of detection, simply by combining the evaluation of both readers. This observation is consistent with those of Wormanns et al., in one of the few studies of double-reading in pulmonary nodule detection. That study, performed in a group of patients with pulmonary metastases rather than a lung cancer screening population, found that mean sensitivity for detection using independent double-reading without any arbitration improved sensitivity from 64% for a single reader to 79% on LDCT [334].

A similar effect of double-reading on sensitivity when using this independent double-reading strategy has been observed in mammographic screening, with increased cancer detection rates [342, 343]. However, the penalty for an increased sensitivity is usually an increase in false positive detections, resulting in a decreased specificity and positive predictive value [341]. This decrease in specificity has been replicated in the current study, and has particular relevance to lung cancer screening with CT, which is known to generate large of numbers of false positive detections even with single readers [164].

For this reason, lung cancer screening studies and other evaluations of double-reading in nodule detection have also incorporated some form of arbitration or consensus, so that there is a mechanism to decide what should be designated a positive or an "actionable" finding. Double-reading in lung cancer screening has been in use since the Early Lung Cancer Action Project (ELCAP) [142] and has also been used in multiple European randomised control trials, including the NELSON

trial. In both the ELCAP and NELSON trials, two readers initially tried to achieve consensus through discussion. The NELSON study used two readers, one at a peripheral reading site and one at a central site, and in cases of discrepancy the central reader would attempt to achieve consensus with the first through discussion. If no consensus could be reached, arbitration by a third reader of 20 years' experience was used, with the majority decision of the three readers used as the final result [169, 174]. In these studies attempting consensus, the use of discussion could potentially result in forced or "coerced" consensus, where the opinions of more experienced radiologists supersede those of less experienced ones [344]. As noted by Bankier et al., participation in a group consensus effort may have a modifying effect on an individual radiologist's opinion, diminishing the value of a consensus agreement (which they termed "groupthink"). In any case, the merit of consensus discussion on the nature of a small opacity is questionable; such opacities have very few defining characteristics, and so do not lend themselves to discussion.

Thus, instead of using consensus by discussion, three different methods of resolving discrepancies using double-reading with arbitration were modelled. First, with double-reading using only internal arbitration, radiologists were shown nodules that they had missed during the first round of reading, providing them with the opportunity to record nodules that they may have failed to detect during the initial read. This method had the advantage of allowing "internal" arbitration regarding an opacity without requiring all radiologists to be present simultaneously, and also without any particular radiologist forming a dominant opinion that may inadvertently coerce the agreement of other radiologists. It also prevented radiologists from altering the categorisation of nodules which they had initially detected on free search.

Using this method had a variable effect on sensitivity. The variability of the results of this method of arbitration was understandably a reflection of the baseline sensitivity of the individual readers comprising the pair. A very sensitive reader could significantly improve the sensitivity of a pair, as demonstrated by the 7 pairs involving readers A or B using this method. However, the effect of less sensitive radiologists on the very sensitive radiologists is not easy to predict: in the present investigation, reader C significantly weakened the sensitivity of readers A and B (the two most sensitive readers) while readers D and E had no effect.

These findings thus underscore the importance of establishing the baseline performance characteristics of any reader in a multiple reader setting. Notably, the sensitivities of the individual readers (50.5% to 86.6%) in the current study were similar to those in previous studies [289, 293, 325, 334], and so there is no reason to suspect that these findings are unique to the expert readers in the present study.

Despite this variability in performance, there are two trends revealed by this method of double-reading with internal arbitration that are complementary and potentially advantageous to a lung cancer screening programme. First, sensitivity was either unchanged or improved in the majority of pair-combinations. Second, specificity was improved in the majority of combinations.

Next double-reading with external arbitration was simulated, where the opinion of the third independent reader was sought in all cases that were discordant after round 1. This had a markedly variable effect on sensitivity and specificity as well, compared to a single reader. Furthermore, this method was not superior to simply using internal arbitration alone. In the third method, double-reading with

internal followed by external arbitration was simulated. On the whole this led to increased sensitivity, compared to a single reader, and compared to internal arbitration alone. However, this increased sensitivity came at the expense of either unchanged or diminished specificity, when compared to a single reader in the majority of combinations.

Finally, triple reading was investigated. Triple reading seemed to be universally reliable in improving sensitivity, but again at the expense of specificity, as compared to double-reading with internal arbitration only.

There are two implications of these results for lung screening in practice. First, if external arbitration does not consistently improve sensitivity, while leaving specificity diminished or unchanged, reliance on this method is questionable. Second, the highly labour- and time-intensive natures of this strategy are additional reasons for caution when applying these methods to lung screening practice. It is worth noting here that "double"-reading in cases where some form of external arbitration is used is in fact fallacious; such cases invariably will have to involve the services of a third reader, and so are really also forms of triple reading. The identification of discrepant nodules requiring arbitration and the recording of final readings during external arbitration are administrative tasks that can lengthen the total time taken for this process.

The potential lack of benefit of double-reading has been demonstrated by Ying Wang et al. In a retrospective analysis of 74 proven lung cancer cases from the prevalence screening round of the NELSON study, they concluded that double-reading detected only 2 (2.7%) more lung cancers with a 0.2% reduction in

specificity [174]. This conclusion is probably premature given the small sample size. It may also not be applicable to the task of lung nodule detection in general, given that the main criteria for assessment of nodules in the NELSON study is growth as demonstrated by semi-automated volumetry, a relatively reproducible parameter [311, 313]; as previously discussed in section 1.6.2.3, volumetry helps with interpretation, but not with detection itself. Nevertheless, this adds to the evidence regarding variable effects of multiple reading strategies shown by the current study.

The findings in the present investigation also have implications for the widely varying reference standards used in studies on lung nodule identification. A single expert radiologist acting as a reference standard would be questionable; it is arguable whether any one radiologist can have the "final word" in defining such a standard, since sensitivity (as demonstrated by previous studies and reinforced by the results of the present investigation) and agreement among radiologists with respect to nodule characteristics are highly variable [142, 396, 397]. In studies on computer-aided diagnosis (CAD) as compared to radiologists, the reference standards were established by two expert radiologists in consensus who reviewed the marks made by CAD software and readers, and were free to add their own detection marks [289, 293, 336]. In defining these standards, no methods of arbitration have been described; it is conceivable that differing reference standards could be produced by different methods of arbitration, with consequently different performance characteristics demonstrated. It is thus important that all studies reporting nodule detection accuracy outline in detail methods used to resolve discrepancies.

This study has a few limitations. The reference dataset has only been established through consensus, without histological data; thus, the true significance

of all nodules is not known. However, in screening practice histological corroboration of the vast majority of nodules would not be available, and as such a consensus standard is the only initial surrogate standard available to enable decision-making. Nevertheless, it should be noted that the sensitivity and specificity described here are with reference to such a standard.

Also, the reference dataset was created through the involvement of the same radiologists that were under investigation, a potential source of error which is difficult to correct for [287]. Although it has been suggested that the reference standard should be created by another group of radiologists who are not being tested [287] this also has its limitations because there is no reason to suppose that another group of independent expert radiologists will perform any better than the group under evaluation and so provide a more accurate estimate of ground "truth". Also, the readers were from different institutions and so no "institutional" bias in terms of attitude towards over- or undercalling nodules should be expected. Furthermore, each radiologist was aware that they were involved in a nodule detection exercise, and hence the potential for heightened vigilance existed. However, such vigilance also exists in the setting of lung cancer screening. Finally, the exact values for double-reading with external arbitration do depend to a certain extent on the sensitivity of the external arbiter itself. The importance of this factor in turn would depend on whether the external arbiter had higher or lower sensitivity specifically for the task of interpretation as opposed to detection, since the process of external arbitration requires that he or she be directed only to discordant findings and asked to provide an interpretation, rather than performing the task of detection as well. However, the

magnitude of this effect in this study cannot be measured because the sensitivity of the external arbiter was not directly assessed.

It was not possible to analyse the level of disagreement between readers due to differences in nodule measurement. It is known that interobserver agreement improves with increasing nodule size [396], but size-based analysis was not the primary aim of this study. Readers were instructed to classify opacities with characteristics of a nodule that were greater than 3mm as positive, but were not asked to record their measurements for each nodule, because (a) this would have been too time-consuming and detract from the primary task of nodule detection and evaluation, and (b) the interobserver variability inherent to both diameter [307] and volumetric [314] measurements of nodules would still have existed as a limitation of any size-based analysis.

7.5 Summary

- The performance of experienced thoracic radiologists for the task of lung nodule detection is not invariably improved by increasing the number of readers, and is significantly affected by the method of arbitration used.

- If double-reading is to be practised, perhaps the most effective and least labour-intensive method is double-reading with only internal arbitration, as it maintained or increased nodule detection with an unchanged or improved specificity, without requiring the services of a third reader.

- The effects of double-reading and arbitration methods must be recognised when devising reading strategies, defining reference standards that rely on consensus, and assessing reader performance against these standards.

CHAPTER 8: CONCLUSION

Over the past fifteen years, lung cancer screening trials using CT have convincingly shown that CT can detect more lung cancers than chest radiography [142, 143, 150, 164, 398]; furthermore, that these cancers are more likely to be at an early stage and, most recently, that a reduction in mortality can be achieved [164]. While further results from different randomised control trials are awaited to confirm such a mortality reduction, the pessimism arising from the lack of benefit seen in earlier screening trials using chest radiography has slowly given way to enthusiasm for screening with CT. However, as yet there has been no detailed consideration of the pragmatic reading strategies required when extrapolating CT lung cancer screening from the trial setting to a national screening programme. The investigations contained in this thesis have provided a number of insights into the implications of different CT reading strategies for nodule detection accuracy and for radiologists' workload should such a programme be implemented.

The evaluation of radiographers as readers in CT lung cancer screening has, to my knowledge, not previously been reported. In Chapter 4, following a short period of standardised training, **radiographers were able to achieve sensitivities for nodule detection that were comparable to radiologists in literature, but were inferior to radiologists reading the same studies**. Importantly, however, **the range of differences seen between radiographers and radiologists was not dissimilar to that seen between radiologists in previous studies.** The implication of these findings is that while radiographers on the whole are not sensitive enough to be used as first readers (a characteristic they share with CAD systems), some radiographers'

detection abilities compare favourably with those of radiologists, potentially allowing them to act as assistant readers. Furthermore, radiographers were clearly able to follow strict size (predominantly volumetry-based) and morphological criteria to designate opacities as nodules and intrapulmonary lymph nodes.

In order to determine whether radiographers could fulfil the role of assistant readers, their effect as concurrent readers on radiologists' performance was subsequently assessed in Chapter 5. Concurrent reading, rather than second reading, was chosen because it is a more pragmatic method of implementing radiographer-assisted reading in a screening programme - a radiologist could perform his or her review of the radiographer's marks in the same sitting as a "free search" for additional nodules. Concurrent reading has also recently been evaluated using CAD, but with conflicting effects on sensitivity and reading time reported [294, 354]. Two important conclusions arose from this investigation. Firstly, **concurrent reading conferred an increase in sensitivity for all but the most sensitive radiologist.** Secondly, **concurrent reading was able to reduce reading time, by as much as 4 minutes**. Such a time-saving, coming as it did with the benefit of increased sensitivity, has positive implications for a national screening programme; the number of radiologists required for such a programme could be substantially reduced by using radiographer-assisted concurrent reading, or alternatively, a radiologist would require less time to read an equivalent number of studies in a session. The importance of these effects in encouraging both radiographers and radiologists to participate in a screening programme should not be underestimated. Radiographers would participate in the knowledge that their contribution is simultaneously enhancing sensitivity and screening throughput. Radiologists, meanwhile, would

arguably be more willing if they could focus their expertise on the tasks of interpretation and decision-making rather than simply detection in screening, while at the same time being able to read faster.

It was also inferred from the investigations in Chapters 4 and 5 that the performance characteristics of radiographers may be similar to those of a CAD system. To address this question directly, the performance of radiographers was directly compared against a commercially approved CAD system in Chapter 6. **Radiographers demonstrated that they were at least as sensitive as CAD for nodule detection**. Arguments against a radiographer rather than computer assistant often centre on the fact that a human is expensive, prone to fatigue and inconsistent, but such arguments are at least partly one-sided; as previously discussed in Chapter 6, CAD systems, including the one evaluated in this thesis, are also prone to marked differences in sensitivity and average false positive detections per case [359].

Ultimately, the strength of a screening programme rests on its ability to maximise early lung cancer detection while minimising false positive detection, and this in turn depends both on the performance of its readers and its mechanism for arriving at a robust consensus on nodules that require follow-up. This aspect of lung nodule detection has not been sufficiently addressed in previous studies on nodule detection or CT lung screening; the advantages of double-reading have been taken at face value, and methods of arbitration have differed substantially without acknowledging their potential impact [159, 169, 289, 336, 337, 345, 346]. The modelling of different methods of arbitration in Chapter 7 exposed the stark variation in performance characteristics that may be encountered when different methods of resolving discrepancies and different numbers of radiologists are used for nodule

detection and interpretation. Significantly, the comparison of models revealed that **the addition of an external arbiter does not invariably improve sensitivity for nodule detection, and may leave specificity decreased or unchanged, making reliance on such a mechanism of arbitration questionable. In contrast, a mechanism of "internal arbitration" maintained or improved both sensitivity and specificity.** Such a mechanism of arbitration provides a second round of reading in which radiologists within a pair independently decide on the significance of nodules that they had missed, without discussion and the potential for a "forced" agreement, and without the services of a third reader. Using such a method, only nodules agreed by both radiologists in a pair would be included in a consensus. However, the logistics of applying such a method of arbitration to lung screening practice are challenging, as each radiologist would be required to read the same case twice, at separate sittings. As such, this investigation provides insights into the delicate balance that must be achieved by the reading strategy of a lung cancer screening programme. In the pursuit of improved sensitivity, the variation in different methods of double-reading has to be borne in mind, in addition to an awareness that "double"-reading when using some form of arbitration is really a misnomer: external arbitration requires a third reader and a total of three rounds of reading, while internal arbitration requires two readers but a total of four rounds of reading.

A recurring theme through all the investigations in this thesis is the accepted trade-off between sensitivity and false positive detection, whether the reader is a radiologist (reading alone, in combination with other radiologists, or with a concurrently reading assistant radiographer), a radiographer, or a CAD system. An

important finding is that **although radiographers reading alone, and radiologists performing radiographer-assisted concurrent reading, had higher average false positive detections per case than independently reading radiologists alone, these average false positive detections per case were still lower than those of CAD systems in the literature, and of the CAD system evaluated in Chapter 6.** A recent investigation has demonstrated the potential for CAD's benefit as a second reader to be negated once a certain false positive detection threshold is reached [290]. When viewed in this light, the relatively lower false positive detection with radiographer-assisted concurrent reading is welcome, as it suggests that the false positive detections arising from such a reading paradigm will not compromise the benefit accrued from its increased sensitivity.

In demonstrating the variation in nodule detection performance between readers, the investigations in Chapters 3 through to 7 have also reinforced notions of an inherent perceptual ability unique to each individual, even if the visual search patterns they use may differ depending on their level of training [335]. However, Chapters 3 and 4 have also captured the potential importance of exposure to a task over a longer period of time, for readers with a lower baseline perceptual ability. While a significant learning effect across the 10 small subsets read was not demonstrated overall for the radiographers in Chapter 3, **the two less sensitive radiographers in Chapter 4 were able to improve their sensitivity while keeping their average false positive detections per case static, when undertaking a 10-week period of reading more CTs.** It is important to note that while these radiographers in the UKLS pilot study were not given direct feedback, they were comparing their own readings with those in the consensus, once it had been achieved,

and so were enacting a form of self-directed learning that would be crucial to the success of assisted reading in a screening programme. Also, the fact that less sensitive radiographers could improve distinguishes them from a CAD system, where, even if an upgrade is performed, the dataset and algorithms for derivation of an improved performance remain impervious to the standard radiologist.

The investigations in Chapter 4 and 6 also provide some insight into what to expect with radiographer reading should higher size thresholds be used for nodule positivity in future screening programmes, as recently suggested [383]. If such an approach is adopted, not only would radiographer sensitivity be maintained when compared to radiologists or to CAD, but radiographer and radiologist sensitivity would probably become more aligned.

The limitations of each investigation have been highlighted in individual chapters, the most obvious, but insurmountable, limitation being the lack of histopathological corroboration of nodules to serve as the standard of reference. However, as histopathological information would not be available for the vast majority of nodules in screening practice, a consensus standard is the only surrogate standard available to enable decision-making, and serves to illustrate the comparisons between the various reading strategies that have been explored.

In summary, the investigations in this thesis have described the performance of different CT reading strategies for nodule detection in lung cancer screening, including the previously unreported role of radiographers as readers. In the future, some of these observations could be used to inform the choice of optimal reading strategy for a national screening programme in the UK.

REFERENCES

1. Osler W. New Growths in the Lung. 2nd ed. New York: D. Appleton and Company, 1895; 590-591

2. Adler I. Forward. Primary malignant growths of the lungs and bronchi: a pathological and clinical study. New York: Longmans, Green and Co, 1912; 3

3. Miller YE. Pathogenesis of lung cancer: 100 year report. Am J Respir Cell Mol Biol 2005; 33:216-223

4. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010; 127:2893-2917

5. Cancer Research UK. Lung Cancer-UK mortality statistics. Last updated on 29-06-2011. Available from http://info.cancerresearchuk.org/cancerstats/types/lung/mortality/ Accessed on 29-12-2011.

6. Holmberg L, Sandin F, Bray F, *et al.* National comparisons of lung cancer survival in England, Norway and Sweden 2001-2004: differences occur early in follow-up. Thorax 2010; 65:436-441

7. Office for National Statistics. Cancer Survival in England, patients diagnosed 2005-2009 and followed up to 2010. Last updated on 15-11-2011. Available from http://www.ons.gov.uk/ons/rel/cancer-unit/cancer-survival/2005-2009--followed-up-to-2010/tbl-cancer-survival.xls Accessed on 12-04-2012.

8. Imperatori A, Harrison RN, Leitch DN, *et al.* Lung cancer in Teesside (UK) and Varese (Italy): a comparison of management and survival. Thorax 2006; 61:232-239

9. Hubbard RB, Baldwin DR. Diagnosing lung cancer earlier in the UK. Thorax 2010; 65:756-758

10. Health and Social Care Information Centre. National Lung Cancer Audit Report 2005. Last updated on 12-2006. Available from https://catalogue.ic.nhs.uk/publications/clinical/lung/nati-clin-audi-supp-prog-lung-canc-nlca-2005/clin-audi-supp-prog-lung-canc-nlca-2005-rep1.pdf Accessed on 01-02-2012.

11. Parkin DM. Trends in lung cancer incidence worldwide. Chest 1989; 96:5S-8S

12. Cancer Research UK. Lung Cancer-UK incidence statistics. Last updated on 20-03-2014. Available from http://info.cancerresearchuk.org/cancerstats/types/lung/incidence/#source2 Accessed on 21-03-2014.

13. Office for National Statistics. Cancer Registrations in England, 2009. London, England.: Office for National Statistics, 2011.

14. Toh CK, Gao F, Lim WT, *et al.* Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity. J Clin Oncol 2006; 24:2245-2251

15. Wakelee HA, Chang ET, Gomez SL, *et al.* Lung cancer incidence in never smokers. J Clin Oncol 2007; 25:472-478

16. Thun MJ, Hannan LM, Adams-Campbell LL, *et al.* Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. PLoS Med 2008; 5:e185

17. U.S.Department of Health and Human Services. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta,GA,USA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2010.

18. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. Br Med J 1950; 2:739-748

19. Levin ML, Goldstein H, Gerhardt PR. Cancer and tobacco smoking; a preliminary report. J Am Med Assoc 1950; 143:336-338

20. Mills CA, Porter MM. Tobacco smoking habits and cancer of the mouth and respiratory system. Cancer Res 1950; 10:539-542

21. Schrek R, Baker LA, Ballard GP, Dolgoff S. Tobacco smoking as an etiologic factor in disease; cancer. Cancer Res 1950; 10:49-58

22. Wynder EL, Graham EA. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. J Am Med Assoc 1950; 143:329-336

23. Parkin DM. 2. Tobacco-attributable cancer burden in the UK in 2010. Br J Cancer 2011; 105 Suppl 2:S6-S13

24. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005; 55:74-108

25. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. BMJ 2000; 321:323-329

26. Doll R, Peto R, Boreham J, Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. Br J Cancer 2005; 92:426-429

27. Lubin JH, Caporaso NE. Cigarette smoking and lung cancer: modeling total exposure and intensity. Cancer Epidemiol Biomarkers Prev 2006; 15:517-523

28. Howlader, N, Noone, AM, Krapcho, M *et al.* Cancer of the Lung and Bronchus (Invasive): Percent Distribution and Counts by Histology among Histologically Confirmed Cases, 2004-2008 Both Sexes by Race. Last updated on 28-01-2010. Available from http://seer.cancer.gov/csr/1975_2008/browse_csr.php?section=15&page=sect_15_table.28.html Accessed on 23-12-2011.

29. National Institute for Health and Care Excellence. Clinical Guideline 121: The Diagnosis and Treatment of Lung Cancer. London: **National Institute for Health and Care Excellence**, 2011.

30. Janssen-Heijnen ML, Coebergh JW. The changing epidemiology of lung cancer in Europe. Lung Cancer 2003; 41:245-258

31. Wynder EL, Muscat JE. The changing epidemiology of smoking and lung cancer histology. Environ Health Perspect 1995; 103 Suppl 8:143-148

32. World Health Organization. Histological Typing of Lung Tumours. Geneva: World Health Organization, 1967.

33. Travis WD, Brambilla E, Muller-Hermelink HK, Harris CC. Tumours of the Lung. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart (IARC/World Health Organization Classification of Tumours). Lyon: IARC Press, 2004;

34. Travis WD, Garg K, Franklin WA, *et al.* Bronchioloalveolar carcinoma and lung adenocarcinoma: the clinical importance and research relevance of the 2004 World Health Organization pathologic criteria. J Thorac Oncol 2006; 1:S13-S19

35. Travis WD, Brambilla E, Noguchi M, *et al.* International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011; 6:244-285

36. Spiro SG, Gould MK, Colice GL. Initial evaluation of the patient with lung cancer: symptoms, signs, laboratory tests, and paraneoplastic syndromes: ACCP evidenced-based clinical practice guidelines (2nd edition). Chest 2007; 132:149S-160S

37. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK. Lung cancer risk prediction: a tool for early detection. Int J Cancer 2007; 120:1-6

38. Koyi H, Hillerdal G, Branden E. A prospective study of a total material of lung cancer from a county in Sweden 1997-1999: gender, symptoms, type, stage, and smoking habits. Lung Cancer 2002; 36:9-14

39. Buccheri G, Ferrigno D. Lung cancer: clinical presentation and specialist referral time. Eur Respir J 2004; 24:898-904

40. Carbone PP, Frost JK, Feinstein AR, Higgins GA, Selawry O. Lung cancer: perspectives and prospects. Ann Intern Med 1970; 73:1003-1024

41. Filderman AE, Shaw C, Matthay RA. Lung cancer. Part I: Etiology, pathology, natural history, manifestations, and diagnostic techniques. Invest Radiol 1986; 21:80-90

42. Sorenson JA, Mitchell CR, Armstrong JD, *et al.* Effects of improved contrast on lung-nodule detection. A clinical ROC study. Invest Radiol 1987; 22:772-780

43. Farr RF, Allisy-Roberts PJ. Fluoroscopy, Digital Imaging and Computed Tomography. Physics for Medical Imaging. 1st ed. London: Saunders (W.B.) Co. Ltd, 1998; 81-117

44. MacMahon H, Austin JH, Gamsu G, *et al.* Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 2005; 237:395-400

45. Naidich DP, Rusinek H, McGuinness G, Leitman B, McCauley DI, Henschke CI. Variables affecting pulmonary nodule detection with computed tomography: evaluation with three-dimensional computer simulation. J Thorac Imaging 1993; 8:291-299

46. Haas AR, Vachani A, Sterman DH. Advances in diagnostic bronchoscopy. Am J Respir Crit Care Med 2010; 182:589-597

47. Ikeda S. Flexible bronchofiberscope. Ann Otol Rhinol Laryngol 1970; 79:916-923

48. Wang KP. Transbronchial needle aspiration and percutaneous needle aspiration for staging and diagnosis of lung cancer. Clin Chest Med 1995; 16:535-552

49. Detterbeck FC, Jantz MA, Wallace M, Vansteenkiste J, Silvestri GA. Invasive mediastinal staging of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; 132:202S-220S

50. Kostakoglu L, Agress H, Jr., Goldsmith SJ. Clinical role of FDG PET in evaluation of cancer patients. RadioGraphics 2003; 23:315-340

51. Kapoor V, McCook BM, Torok FS. An introduction to PET-CT imaging. RadioGraphics 2004; 24:523-543

52. Silvestri GA, Gould MK, Margolis ML, *et al.* Noninvasive staging of non-small cell lung cancer: ACCP evidenced-based clinical practice guidelines (2nd edition). Chest 2007; 132:178S-201S

53. van Tinteren H, Hoekstra OS, Smit EF, *et al.* Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial. Lancet 2002; 359:1388-1393

54. Nomori H, Horio H, Fuyuno G, Kobayashi R, Morinaga S, Suemasu K. Lung adenocarcinomas diagnosed by open lung or thoracoscopic vs bronchoscopic biopsy. Chest 1998; 114:40-44

55. Denoix PF. Enquete permanent dans les centres anticancereux. Bull Inst Nat Hyg 1946; 1:70-75

56. Mountain CF, Carr DT, Anderson WAD. A system for the clinical staging of lung cancer. Am J Roentgenol 1974; 120:130-138

57. UICC International Union Against Cancer. Lung and Pleural Tumours. In:Sobin LH, Gospodarowicz M K, Wittekind C, eds. TNM Classification of Malignant Tumours. 7th ed. n.p.: Wiley Blackwell, 2009; 138-146

58. Goldstraw P, Crowley J, Chansky K, *et al.* The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. J Thorac Oncol 2007; 2:706-714

59. Shepherd FA, Crowley J, Van HP, *et al.* The International Association for the Study of Lung Cancer lung cancer staging project: proposals regarding the clinical staging of small cell lung cancer in the forthcoming (seventh) edition of the tumor, node, metastasis classification for lung cancer. J Thorac Oncol 2007; 2:1067-1077

60. Rami-Porta R, Ball D, Crowley J, *et al.* The IASLC Lung Cancer Staging Project: proposals for the revision of the T descriptors in the forthcoming (seventh) edition of the TNM classification for lung cancer. J Thorac Oncol 2007; 2:593-602

61. Alberts WM. Diagnosis and management of lung cancer executive summary: ACCP evidence-based clinical practice guidelines (2nd Edition). Chest 2007; 132:1S-19S

62. Varlotto JM, Recht A, Flickinger JC, Medford-Davis LN, Dyer AM, DeCamp MM. Lobectomy leads to optimal survival in early-stage small cell lung cancer: a retrospective analysis. J Thorac Cardiovasc Surg 2011; 142:538-546

63. Asamura H, Goya T, Koshiishi Y, Sohara Y, Tsuchiya R, Miyaoka E. How should the TNM staging system for lung cancer be revised? A simulation based on the Japanese Lung Cancer Registry populations. J Thorac Cardiovasc Surg 2006; 132:316-319

64. Birim O, Kappetein AP, Takkenberg JJ, van Klaveren RJ, Bogers AJ. Survival after pathological stage IA nonsmall cell lung cancer: tumor size matters. Ann Thorac Surg 2005; 79:1137-1141

65. Carr SR, Schuchert MJ, Pennathur A, *et al.* Impact of tumor size on outcomes after anatomic lung resection for stage 1A non-small cell lung cancer based on the current staging system. J Thorac Cardiovasc Surg 2012; 143:390-397

66. Flieder DB, Port JL, Korst RJ, *et al.* Tumor size is a determinant of stage distribution in t1 non-small cell lung cancer. Chest 2005; 128:2304-2308

67. Okada M, Nishio W, Sakamoto T, *et al.* Effect of tumor size on prognosis in patients with non-small cell lung cancer: the role of segmentectomy as a type of lesser resection. J Thorac Cardiovasc Surg 2005; 129:87-93

68. Port JL, Kent MS, Korst RJ, Libby D, Pasmantier M, Altorki NK. Tumor size predicts survival within stage IA non-small cell lung cancer. Chest 2003; 124:1828-1833

69. Paoletti L, Pastis NJ, Denlinger CE, Silvestri GA. A decade of advances in treatment of early-stage lung cancer. Clin Chest Med 2011; 32:827-838

70. Hadziahmetovic M, Loo BW, Timmerman RD, *et al.* Stereotactic body radiation therapy (stereotactic ablative radiotherapy) for stage I non-small cell lung cancer--updates of radiobiology, techniques, and clinical outcomes. Discov Med 2010; 9:411-417

71. Timmerman R, Paulus R, Galvin J, *et al.* Stereotactic body radiation therapy for inoperable early stage lung cancer. JAMA 2010; 303:1070-1076

72. Grills IS, Mangona VS, Welsh R, *et al.* Outcomes after stereotactic lung radiotherapy or wedge resection for stage I non-small-cell lung cancer. J Clin Oncol 2010; 28:928-935

73. Howington JA, Blum MG, Chang AC, Balekian AA, Murthy SC. Treatment of stage I and II non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013; 143:e278S-e313S

74. Strauss GM, Herndon JE, Maddaus MA, *et al.* Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the Cancer and Leukemia Group B, Radiation Therapy Oncology Group, and North Central Cancer Treatment Group Study Groups. J Clin Oncol 2008; 26:5043-5051

75. Zemlyak A, Moore WH, Bilfinger TV. Comparison of survival after sublobar resections and ablative therapies for stage I non-small cell lung cancer. J Am Coll Surg 2010; 211:68-72

76. Donington J, Ferguson M, Mazzone P, *et al.* American College of Chest Physicians and Society of Thoracic Surgeons consensus statement for evaluation and management for high-risk patients with stage I non-small cell lung cancer. Chest 2012; 142:1620-1635

77. Markos J, Mullan BP, Hillman DR, *et al.* Preoperative assessment as a predictor of mortality and morbidity after lung resection. Am Rev Respir Dis 1989; 139:902-910

78. Ferguson MK, Little L, Rizzo L, *et al.* Diffusing capacity predicts morbidity and mortality after pulmonary resection. J Thorac Cardiovasc Surg 1988; 96:894-900

79. Brunelli A, Refai M, Salati M, Xiume F, Sabbatini A. Predicted versus observed FEV1 and DLCO after major lung resection: a prospective evaluation at different postoperative periods. Ann Thorac Surg 2007; 83:1134-1139

80. Ferguson MK, Vigneswaran WT. Diffusing capacity predicts morbidity after lung resection in patients without obstructive lung disease. Ann Thorac Surg 2008; 85:1158-1164

81. Lim E, Baldwin D, Beckles M, *et al.* Guidelines on the radical management of patients with lung cancer. Thorax 2010; 65 Suppl 3:iii1-27

82. Brunelli A, Kim AW, Berger KI, Addrizzo-Harris DJ. Physiologic evaluation of the patient with lung cancer being considered for resectional surgery: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013; 143:e166S-e190S

83. Commission on Chronic Illness. Chronic illness in the United States.Vol I. Prevention of chronic illness. Cambridge, MA: Harvard University Press, 1957.

84. Moskowitz M. Screening is not diagnosis. Radiology 1979; 133:265-268

85. Wilson, J. M. and Jungner, Y. G. Principles and practice of mass screening for disease. Geneva: World Health Organization, 1968.

86. Gillon R. Medical ethics: four principles plus attention to scope. BMJ 1994; 309:184-188

87. Andermann A, Blancquaert I, Beauchamp S, Dery V. Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. Bull World Health Organ 2008; 86:317-319

88. Harris R, Sawaya GF, Moyer VA, Calonge N. Reconsidering the criteria for evaluating proposed screening programs: reflections from 4 current and former members of the U.S. Preventive services task force. Epidemiol Rev 2011; 33:20-35

89. UK National Screening Committee. Programme appraisal criteria: Criteria for appraising the viability, effectiveness and appropriateness of a screening programme. Last updated on 2012. Available from http://www.screening.nhs.uk/criteria Accessed on 22-01-2012.

90. Strauss GM, Gleason RE, Sugarbaker DJ. Screening for lung cancer. Another look; a different view. Chest 1997; 111:754-768

91. Black WC. Computed tomography screening for lung cancer: review of screening principles and update on current status. Cancer 2007; 110:2370-2384

92. Steele RJ, Brewster DH. Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? No. BMJ 2011; 343:d6397

93. Penston J. Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? Yes. BMJ 2011; 343:d6395

94. Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? Lancet 2000; 355:129-134

95. Henschke CI, Shaham D, Yankelevitz DF, Altorki NK. CT screening for lung cancer: past and ongoing studies. Semin Thorac Cardiovasc Surg 2005; 17:99-106

96. Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. BMJ 2002; 324:827-830

97. Machin D, Campbell MJ, Walters SJ. Diagnostic Tests. Medical Statistics: A Textbook for the Health Sciences. Chichester, England: John Wiley and Sons, 2007; 49-50

98. Doubilet P, Weinstein MC, McNeil BJ. Use and misuse of the term "cost effective" in medicine. N Engl J Med 1986; 314:253-256

99. Donaldson C, Currie G, Mitton C. Cost effectiveness analysis in health care: contraindications. BMJ 2002; 325:891-894

100. Petitti DB. Advanced cost-effectiveness analysis. Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis. 1st ed. New York: Oxford University Press, 2000; 182-201

101. Torrance GW. Utility approach to measuring health-related quality of life. J Chronic Dis 1987; 40:593-603

102. Petitti DB. Measuring Preferences for Health States. Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis. 1st ed. New York: Oxford University Press, 2000; 169-181

103. National Institute for Health and Clinical Excellence. Methods for the Development of NICE public health guidance. London: **National Institute for Health and Clinical Excellence**, 2009.

104. Leivo T, Salminen T, Sintonen H, *et al.* Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat 1999; 54:261-267

105. Groenewoud JH, Otten JD, Fracheboud J, *et al.* Cost-effectiveness of different reading and referral strategies in mammography screening in the Netherlands. Breast Cancer Res Treat 2007; 102:211-218

106. Wisnivesky JP, Mushlin AI, Sicherman N, Henschke C. The cost-effectiveness of low-dose CT screening for lung cancer: preliminary results of baseline screening. Chest 2003; 124:614-621

107. Marshall D, Simpson KN, Earle CC, Chu C. Potential cost-effectiveness of one-time screening for lung cancer (LC) in a high risk cohort. Lung Cancer 2001; 32:227-236

108. Mahadevia PJ, Fleisher LA, Frick KD, Eng J, Goodman SN, Powe NR. Lung cancer screening with helical computed tomography in older adult smokers: a decision and cost-effectiveness analysis. JAMA 2003; 289:313-322

109. Manser R, Dalton A, Carter R, Byrnes G, Elwood M, Campbell DA. Cost-effectiveness analysis of screening for lung cancer with low dose spiral CT (computed tomography) in the Australian setting. Lung Cancer 2005; 48:171-185

110. Welch HG, Black WC. Overdiagnosis in cancer. J Natl Cancer Inst 2010; 102:605-613

111. Baecke E, de Koning HJ, Otto SJ, van Iersel CA, van Klaveren RJ. Limited contamination in the Dutch-Belgian randomized lung cancer screening trial (NELSON). Lung Cancer 2010; 69:66-70

112. Raffle AE, Muir Gray JA. What Screening Does. Screening: Evidence and Practice. 1st ed. Oxford: Oxford University Press, 2007; 69

113. NHS Cancer Screening Programme. NHS Cancer Screening Programme. Last updated on 2011. Available from http://www.cancerscreening.nhs.uk/index.html Accessed on 08-04-2011.

114. Shapiro S. Periodic screening for breast cancer: the HIP Randomized Controlled Trial. Health Insurance Plan. J Natl Cancer Inst Monogr 1997;27-30

115. Andersson I, Janzon L. Reduced breast cancer mortality in women under age 50: updated results from the Malmo Mammographic Screening Program. J Natl Cancer Inst Monogr 1997;63-67

116. Tabar L, Vitak B, Chen TH, *et al.* Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology 2011; 260:658-663

117. Frisell J, Lidbrink E, Hellstrom L, Rutqvist LE. Followup after 11 years--update of mortality results in the Stockholm mammographic screening trial. Breast Cancer Res Treat 1997; 45:263-270

118. Miller AB, Baines CJ, To T. The Gothenburg breast screening trial: first results on mortality, incidence, and mode of detection for women ages 39-49 years at randomization. Cancer 1998; 83:186-190

119. Alexander FE, Anderson TJ, Brown HK, *et al.* 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. Lancet 1999; 353:1903-1908

120. Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up. A randomized screening trial of mammography in women age 40 to 49 years. Ann Intern Med 2002; 137:305-312

121. Gotzsche PC. Mammography screening: truth, lies, and controversy. Lancet 2012; 380:218

122. The NHS Breast Screening Programme. When was the NHS Breast Screening Programme set up? Last updated on 2012. Available from http://www.cancerscreening.nhs.uk/breastscreen/programme-commencement.html Accessed on 04-01-2012.

123. Olsson S, Andersson I, Karlberg I, Bjurstam N, Frodis E, Hakansson S. Implementation of service screening with mammography in Sweden: from pilot study to nationwide programme. J Med Screen 2000; 7:14-18

124. Fracheboud J, de Koning HJ, Boer R, *et al.* Nationwide breast cancer screening programme fully implemented in The Netherlands. Breast 2001; 10:6-11

125. The NHS Breast Screening Programme. Why are women under 50 not routinely invited for breast screening? Last updated on 2012. Available from http://www.cancerscreening.nhs.uk/breastscreen/under-50.html Accessed on 04-01-2012.

126. Nystrom L, Rutqvist LE, Wall S, *et al.* Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993; 341:973-978

127. Gotzsche PC, Nielsen M. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2011;CD001877

128. Duffy SW, Tabar L, Olsen AH, *et al.* Absolute numbers of lives saved and overdiagnosis in breast cancer screening, from a randomized trial and from the Breast Screening Programme in England. J Med Screen 2010; 17:25-30

129. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. Lancet 2012; 380:1778-1786

130. Brett J, Bankhead C, Henderson B, Watson E, Austoker J. The psychological impact of mammographic screening. A systematic review. Psychooncology 2005; 14:917-938

131. Nash FA, Morgan JM, Tomkins JG. South London Lung Cancer Study. Br Med J 1968; 2:715-721

132. Brett GZ. The value of lung cancer detection by six-monthly chest radiographs. Thorax 1968; 23:414-420

133. Berlin NI, Buncher CR, Fontana RS, Frost JK, Melamed MR. The National Cancer Institute Cooperative Early Lung Cancer Detection Program. Results of the initial screen (prevalence). Early lung cancer detection: Introduction. Am Rev Respir Dis 1984; 130:545-549

134. Flehinger BJ, Melamed MR, Zaman MB, Heelan RT, Perchick WB, Martini N. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Memorial Sloan-Kettering study. Am Rev Respir Dis 1984; 130:555-560

135. Frost JK, Ball WC, Jr., Tockman MS, *et al.* Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. Am Rev Respir Dis 1984; 130:549-554

136. Fontana RS, Sanderson DR, Woolner LB, *et al.* Screening for lung cancer. A critique of the Mayo Lung Project. Cancer 1991; 67:1155-1164

137. Marcus PM, Bergstralh EJ, Fagerstrom RM, *et al.* Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up. J Natl Cancer Inst 2000; 92:1308-1316

138. Marcus PM, Bergstralh EJ, Zweig MH, Harris A, Offord KP, Fontana RS. Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. J Natl Cancer Inst 2006; 98:748-756

139. Oken MM, Hocking WG, Kvale PA, *et al.* Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA 2011; 306:1865-1873

140. Miettinen OS, Henschke CI. CT screening for lung cancer: coping with nihilistic recommendations. Radiology 2001; 221:592-596

141. Naidich DP, Marshall CH, Gribbin C, Arams RS, McCauley DI. Low-dose CT of the lungs: preliminary observations. Radiology 1990; 175:729-731

142. Henschke CI, McCauley DI, Yankelevitz DF, *et al.* Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 1999; 354:99-105

143. Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, Miettinen OS. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 2006; 355:1763-1771

144. Bach PB, Jett JR, Pastorino U, Tockman MS, Swensen SJ, Begg CB. Computed tomography screening and lung cancer outcomes. JAMA 2007; 297:953-961

145. Diederich S, Thomas M, Semik M, *et al.* Screening for early lung cancer with low-dose spiral computed tomography: results of annual follow-up examinations in asymptomatic smokers. Eur Radiol 2004; 14:691-702

146. MacRedmond R, Logan PM, Lee M, Kenny D, Foley C, Costello RW. Screening for lung cancer using low dose CT scanning. Thorax 2004; 59:237-241

147. Miller A, Markowitz S, Manowitz A, Miller JA. Lung cancer screening using low-dose high-resolution CT scanning in a high-risk workforce: 3500 nuclear fuel workers in three US states. Chest 2004; 125:152S-153S

148. Nawa T, Nakagawa T, Kusano S, Kawasaki Y, Sugawara Y, Nakata H. Lung cancer screening using low-dose spiral CT: results of baseline and 1-year follow-up studies. Chest 2002; 122:15-20

149. Pastorino U, Bellomi M, Landoni C, *et al.* Early lung-cancer detection with spiral CT and positron emission tomography in heavy smokers: 2-year results. Lancet 2003; 362:593-597

150. Sobue T, Moriyama N, Kaneko M, *et al.* Screening for lung cancer with low-dose helical computed tomography: anti-lung cancer association project. J Clin Oncol 2002; 20:911-920

151. Sone S, Li F, Yang ZG, *et al.* Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner. Br J Cancer 2001; 84:25-32

152. Swensen SJ, Jett JR, Hartman TE, *et al.* CT screening for lung cancer: five-year prospective experience. Radiology 2005; 235:259-265

153. Tiitola M, Kivisaari L, Huuskonen MS, *et al.* Computed tomography screening for lung cancer in asbestos-exposed workers. Lung Cancer 2002; 35:17-22

154. Aberle, D. R., Black, W. C., Goldin, J., and Patz, E. Gareen I. Gatsonis C. American College of Radiology Imaging Network ACRIN No. 6654 Contemporary Screening for the detection of lung cancer Amendment 10. Last updated on 2004. Available from Available at http://www.acrin.org/6654_protocol.aspx Accessed on 08-04-2011.

155. Xu DM, Gietema H, de KH, *et al.* Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 2006; 54:177-184

156. Pedersen JH, Ashraf H, Dirksen A, *et al.* The Danish randomized lung cancer CT screening trial--overall design and results of the prevalence round. J Thorac Oncol 2009; 4:608-614

157. Baldwin DR, Duffy SW, Wald NJ, Page R, Hansell DM, Field JK. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. Thorax 2011; 66:308-313

158. Infante M, Lutman FR, Cavuto S, *et al.* Lung cancer screening with spiral CT: baseline results of the randomized DANTE trial. Lung Cancer 2008; 59:355-363

159. Pegna AL, Picozzi G, Mascalchi M, *et al.* Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung Cancer 2009; 64:34-40

160. Pastorino U, Rossi M, Rosato V, *et al.* Annual or biennial CT screening versus observation in heavy smokers: 5-year results of the MILD trial. Eur J Cancer Prev 2012; 21:308-315

161. Church TR. Chest radiography as the comparison for spiral CT in the National Lung Screening Trial. Acad Radiol 2003; 10:713-715

162. Aberle DR, Berg CD, Black WC, *et al.* The National Lung Screening Trial: overview and study design. Radiology 2011; 258:243-253

163. Aberle DR, Adams AM, Berg CD, *et al.* Baseline characteristics of participants in the randomized national lung screening trial. J Natl Cancer Inst 2010; 102:1771-1779

164. Aberle DR, Adams AM, Berg CD, *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011; 365:395-409

165. 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 randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007; 120:868-874

166. van der Aalst CM, van Iersel CA, van Klaveren RJ, *et al.* Generalisability of the results of the Dutch-Belgian randomised controlled lung cancer CT screening trial (NELSON): does self-selection play a role? Lung Cancer 2012; 77:51-57

167. Horeweg N, van der Aalst CM, Thunnissen E, *et al.* Characteristics of lung cancers detected by computer tomography screening in the randomized NELSON trial. Am J Respir Crit Care Med 2013; 187:848-854

168. Lindell RM, Hartman TE, Swensen SJ, *et al.* Five-year lung cancer screening experience: CT appearance, growth rate, location, and histologic features of 61 lung cancers. Radiology 2007; 242:555-562

169. van Klaveren RJ, Oudkerk M, Prokop M, *et al.* Management of lung nodules detected by volume CT scanning. N Engl J Med 2009; 361:2221-2229

170. Yankelevitz DF, Kostis WJ, Henschke CI, *et al.* Overdiagnosis in chest radiographic screening for lung carcinoma: frequency. Cancer 2003; 97:1271-1275

171. Diederich S, Wormanns D, Semik M, *et al.* Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers. Radiology 2002; 222:773-781

172. Henschke CI, Naidich DP, Yankelevitz DF, *et al.* Early lung cancer action project: initial findings on repeat screenings. Cancer 2001; 92:153-159

173. Veronesi G, Bellomi M, Mulshine JL, *et al.* Lung cancer screening with low-dose computed tomography: a non-invasive diagnostic protocol for baseline lung nodules. Lung Cancer 2008; 61:340-349

174. Wang Y, van Klaveren RJ, de Bock GH, *et al.* No benefit for consensus double reading at baseline screening for lung cancer with the use of semiautomated volumetry software. Radiology 2012; 262:320-326

175. Infante M, Cavuto S, Lutman FR, *et al.* A randomized study of lung cancer screening with spiral computed tomography: three-year results from the DANTE trial. Am J Respir Crit Care Med 2009; 180:445-453

176. Saghir Z, Dirksen A, Ashraf H, *et al.* CT screening for lung cancer brings forward early disease. The randomised Danish Lung Cancer Screening Trial: status after five annual screening rounds with low-dose CT. Thorax 2012; 67:296-301

177. Ellis SM, Husband JE, Armstrong P, Hansell DM. Computed tomography screening for lung cancer: back to basics. Clin Radiol 2001; 56:691-699

178. Cassidy A, Myles JP, Liloglou T, Duffy SW, Field JK. Defining high-risk individuals in a population-based molecular-epidemiological study of lung cancer. Int J Oncol 2006; 28:1295-1301

179. Ng EH, Ng FC, Tan PH, *et al.* Results of intermediate measures from a population-based, randomized trial of mammographic screening prevalence and detection of breast carcinoma among Asian women: the Singapore Breast Screening Project. Cancer 1998; 82:1521-1528

180. Hounsfield GN. Computerized transverse axial scanning (tomography). 1. Description of system. Br J Radiol 1973; 46:1016-1022

181. Flohr TG, Schaller S, Stierstorfer K, Bruder H, Ohnesorge BM, Schoepf UJ. Multi-detector row CT systems and image-reconstruction techniques. Radiology 2005; 235:756-773

182. Mayo JR, Aldrich J, Muller NL. Radiation exposure at chest CT: a statement of the Fleischner Society. Radiology 2003; 228:15-21

183. Farr RF, Allisy-Roberts PJ. Radiation Physics. Physics for Medical Imaging. 1st ed. London: Saunders (W.B.) Co. Ltd, 1998; 1-33

184. Hamberg LM, Rhea JT, Hunter GJ, Thrall JH. Multi-detector row CT: radiation dose characteristics. Radiology 2003; 226:762-772

185. Kalra MK, Maher MM, Toth TL, *et al.* Strategies for CT radiation dose optimization. Radiology 2004; 230:619-628

186. Modica MJ, Kanal KM, Gunn ML. The obese emergency patient: imaging challenges and solutions. RadioGraphics 2011; 31:811-823

187. Lewis, M. Radiation dose issues in multi-slice CT scanning. ImPACT group, 2005. Available from http://www.impactscan.org/download/msctdose.pdf Accessed on 12-1-2011.

188. Lee CH, Goo JM, Ye HJ, *et al.* Radiation dose modulation techniques in the multidetector CT era: from basics to practice. RadioGraphics 2008; 28:1451-1459

189. McNitt-Gray MF. AAPM/RSNA Physics Tutorial for Residents: Topics in CT. Radiation dose in CT. RadioGraphics 2002; 22:1541-1553

190. International Commission on Radiation Protection (IRCP). Managing Patient Dose in Multi-Detector Computed Tomography (MDCT). ICRP Publication 102. Ann ICRP 2007; 37:

191. National Council on Radiation Protection and Measurements. Risk Estimates for Radiation Protection. Bethesda, MD, USA: NCRP Publications, 1993.

192. Mayo JR, Hartman TE, Lee KS, Primack SL, Vedal S, Muller NL. CT of the chest: minimal tube current required for good image quality with the least radiation dose. AJR Am J Roentgenol 1995; 164:603-607

193. Prasad SR, Wittram C, Shepard JA, McLoud T, Rhea J. Standard-dose and 50%-reduced-dose chest CT: comparing the effect on image quality. AJR Am J Roentgenol 2002; 179:461-465

194. Remy-Jardin M, Sobaszek A, Duhamel A, Mastora I, Zanetti C, Remy J. Asbestos-related pleuropulmonary diseases: evaluation with low-dose four-detector row spiral CT. Radiology 2004; 233:182-190

195. Tack D, de M, V, Petit W, *et al.* Multi-detector row CT pulmonary angiography: comparison of standard-dose and simulated low-dose techniques. Radiology 2005; 236:318-325

196. Bankier AA, Kressel HY. Through the Looking Glass revisited: the need for more meaning and less drama in the reporting of dose and dose reduction in CT. Radiology 2012; 265:4-8

197. Webb WR, Müller NL, Naidich DP. Technical Aspects of High-Resolution Computed Tomography. High-Resolution CT of The Lung. Fourth ed. Lippincott Williams and Wilkins, 2008; 1-41

198. Stern EJ, Frank MS, Godwin JD. Chest computed tomography display preferences. Survey of thoracic radiologists. Invest Radiol 1995; 30:517-521

199. Calhoun PS, Kuszyk BS, Heath DG, Carley JC, Fishman EK. Three-dimensional volume rendering of spiral CT data: theory and method. RadioGraphics 1999; 19:745-764

200. Ravenel JG, McAdams HP. Multiplanar and three-dimensional imaging of the thorax. Radiol Clin North Am 2003; 41:475-489

201. Jennings SG, Winer-Muram HT, Tarver RD, Farber MO. Lung tumor growth: assessment with CT--comparison of diameter and cross-sectional area with volume measurements. Radiology 2004; 231:866-871

202. Winer-Muram HT, Jennings SG, Tarver RD, *et al.* Volumetric growth rate of stage I lung cancer prior to treatment: serial CT scanning. Radiology 2002; 223:798-805

203. Dalrymple NC, Prasad SR, Freckleton MW, Chintapalli KN. Informatics in radiology (infoRAD): introduction to the language of three-dimensional imaging with multidetector CT. RadioGraphics 2005; 25:1409-1428

204. Gavrielides MA, Kinnard LM, Myers KJ, Petrick N. Noncalcified lung nodules: volumetric assessment with thoracic CT. Radiology 2009; 251:26-37

205. Giger ML, Bae KT, MacMahon H. Computerized detection of pulmonary nodules in computed tomography images. Invest Radiol 1994; 29:459-465

206. Armato, S. G., III. CAD in Lung Imaging. Presented at the RSNA Annual Meeting 2011, 2011.

207. Ko JP, Naidich DP. Computer-aided diagnosis and the evaluation of lung disease. J Thorac Imaging 2004; 19:136-155

208. Zinovev D, Duo Y, Raicu DS, Furst J, Armato SG. Consensus versus disagreement in imaging research: a case study using the LIDC database. J Digit Imaging 2012; 25:423-436

209. Armato SG, III, McLennan G, Bidaut L, *et al.* The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011; 38:915-931

210. Osborne D, Vock P, Godwin JD, Silverman PM. CT identification of bronchopulmonary segments: 50 normal subjects. AJR Am J Roentgenol 1984; 142:47-52

211. Webb WR, Muller NL, Naidich DP. Normal Lung Anatomy. High Resolution CT of the Lung. Fourth ed. Lippincott Williams & Wilkins, 2008; 42-64

212. Raoof S, Amchentsev A, Vlahos I, Goud A, Naidich DP. Pictorial essay: multinodular disease: a high-resolution CT scan diagnostic algorithm. Chest 2006; 129:805-815

213. Heitzman ER, Markarian B, Berger I, Dailey E. The secondary pulmonary lobule: a practical concept for interpretation of chest radiographs. I. Roentgen anatomy of the normal secondary pulmonary lobule. Radiology 1969; 93:507-512

214. Itoh H, Murata K, Konishi J, Nishimura K, Kitaichi M, Izumi T. Diffuse lung disease: pathologic basis for the high-resolution computed tomography findings. J Thorac Imaging 1993; 8:176-188

215. Raasch BN, Carsky EW, Lane EJ, O'Callaghan JP, Heitzman ER. Radiographic anatomy of the interlobar fissures: a study of 100 specimens. AJR Am J Roentgenol 1982; 138:1043-1049

216. Medlar EM. Variations in interlobar fissures. AJR Am J Roentgenol 1947; 57:723-725

217. Godwin JD, Tarver RD. Accessory fissures of the lung. AJR Am J Roentgenol 1985; 144:39-47

218. Ariyurek OM, Gulsun M, Demirkazik FB. Accessory fissures of the lung: evaluation by high-resolution computed tomography. Eur Radiol 2001; 11:2449-2453

219. Weibel ER. Fleischner Lecture. Looking into the lung: what can it tell us? AJR Am J Roentgenol 1979; 133:1021-1031

220. Adler I. Primary malignant growths of the lungs and bronchi: a pathological and clinical study. New York: Longmans, Green, and Co., 1912;

221. Wenckebach KF. The radiology of the chest. Arch Roentgen Ray 1913; 18:169-182

222. Good CA, Hood RT, Jr., McDonald JR. Significance of a solitary mass in the lung. Am J Roentgenol Radium Ther Nucl Med 1953; 70:543-554

223. Taylor RR, Rivkin LN, Salyer JM. The solitary pulmonary nodule; a review of 236 consecutive cases, 1944 to 1956. Ann Surg 1958; 147:197-202

224. Burdette WJ, Evans C. Management of Coin Lesions and Carcinoma of the Lung. Ann Surg 1965; 161:649-673

225. Nathan H. Management of solitary pulmonary nodules. An organized approach based on growth rate and statistics. JAMA 1974; 227:1141-1144

226. Schaner EG, Head GL, Kalman MA, Dunnick NR, Doppman JL. Whole body computed tomography in the diagnosis of abdominal and thoracic malignancy: review of 600 cases. Cancer Treat Rep 1977; 61:1537-1560

227. Muhm JR, Brown LR, Crowe JK. Use of computed tomography in the detection of pulmonary nodules. Mayo Clin Proc 1977; 52:345-348

228. Remy-Jardin M, Remy J, Giraud F, Marquette CH. Pulmonary nodules: detection with thick-section spiral CT versus conventional CT. Radiology 1993; 187:513-520

229. Fischbach F, Knollmann F, Griesshaber V, Freund T, Akkol E, Felix R. Detection of pulmonary nodules by multislice computed tomography: improved detection rate with reduced slice thickness. Eur Radiol 2003; 13:2378-2383

230. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. Radiology 2008; 246:697-722

231. Godoy MC, Naidich DP. Subsolid pulmonary nodules and the spectrum of peripheral adenocarcinomas of the lung: recommended interim guidelines for assessment and management. Radiology 2009; 253:606-622

232. Naidich DP, Bankier AA, MacMahon H, *et al.* Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. Radiology 2013; 266:304-317

233. Webb WR. Thin-section CT of the secondary pulmonary lobule: anatomy and the image--the 2004 Fleischner lecture. Radiology 2006; 239:322-338

234. Ahn MI, Gleeson TG, Chan IH, *et al.* Perifissural nodules seen at CT screening for lung cancer. Radiology 2010; 254:949-956

235. Hyodo T, Kanazawa S, Dendo S, *et al.* Intrapulmonary lymph nodes: thin-section CT findings, pathological findings, and CT differential diagnosis from pulmonary metastatic nodules. Acta Med Okayama 2004; 58:235-240

236. Gosset N, Bankier AA, Eisenberg RL. Tree-in-bud pattern. AJR Am J Roentgenol 2009; 193:W472-W477

237. Erasmus JJ, Connolly JE, McAdams HP, Roggli VL. Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. RadioGraphics 2000; 20:43-58

238. Brandman S, Ko JP. Pulmonary nodule detection, characterization, and management with multidetector computed tomography. J Thorac Imaging 2011; 26:90-105

239. Wahidi MM, Govert JA, Goudar RK, Gould MK, McCrory DC. Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; 132:94S-107S

240. Siegelman SS, Khouri NF, Leo FP, Fishman EK, Braverman RM, Zerhouni EA. Solitary pulmonary nodules: CT assessment. Radiology 1986; 160:307-312

241. Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. Radiology 1993; 186:405-413

242. Zwirewich CV, Vedal S, Miller RR, Muller NL. Solitary pulmonary nodule: high-resolution CT and radiologic-pathologic correlation. Radiology 1991; 179:469-476

243. Kuriyama K, Tateishi R, Doi O, *et al.* CT-pathologic correlation in small peripheral lung cancers. AJR Am J Roentgenol 1987; 149:1139-1143

244. Travis WD, Brambilla E, Muller-Hermelink HK, Harris CC. Mesenchymal Tumours: Hamartoma. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart (IARC/World Health Organization Classification of Tumours). Lyon: IARC Press, 2004; 113

245. Mahoney MC, Shipley RT, Corcoran HL, Dickson BA. CT demonstration of calcification in carcinoma of the lung. AJR Am J Roentgenol 1990; 154:255-258

246. Siegelman SS, Khouri NF, Scott WW, Jr., *et al.* Pulmonary hamartoma: CT findings. Radiology 1986; 160:313-317

247. Woodring JH, Fried AM, Chuang VP. Solitary cavities of the lung: diagnostic implications of cavity wall thickness. AJR Am J Roentgenol 1980; 135:1269-1271

248. Holt RM, Schmidt RA, Godwin JD, Raghu G. High resolution CT in respiratory bronchiolitis-associated interstitial lung disease. J Comput Assist Tomogr 1993; 17:46-50

249. Patsios D, Roberts HC, Paul NS, *et al.* Pictorial review of the many faces of bronchioloalveolar cell carcinoma. Br J Radiol 2007; 80:1015-1023

250. Noguchi M, Morikawa A, Kawasaki M, *et al.* Small adenocarcinoma of the lung. Histologic characteristics and prognosis. Cancer 1995; 75:2844-2852

251. Yang ZG, Sone S, Takashima S, *et al.* High-resolution CT analysis of small peripheral lung adenocarcinomas revealed on screening helical CT. AJR Am J Roentgenol 2001; 176:1399-1407

252. Austin JH, Garg K, Aberle D, *et al.* Radiologic implications of the 2011 classification of adenocarcinoma of the lung. Radiology 2013; 266:62-71

253. Byers TE, Vena JE, Rzepka TF. Predilection of lung cancer for the upper lobes: an epidemiologic inquiry. J Natl Cancer Inst 1984; 72:1271-1275

254. Bankoff MS, McEniff NJ, Bhadelia RA, Garcia-Moliner M, Daly BD. Prevalence of pathologically proven intrapulmonary lymph nodes and their appearance on CT. AJR Am J Roentgenol 1996; 167:629-630

255. Ishikawa H, Koizumi N, Morita T, Tsuchida M, Umezu H, Sasai K. Ultrasmall intrapulmonary lymph node: usual high-resolution computed tomographic findings with histopathologic correlation. J Comput Assist Tomogr 2007; 31:409-413

256. Shaham D, Vazquez M, Bogot NR, Henschke CI, Yankelevitz DF. CT features of intrapulmonary lymph nodes confirmed by cytology. Clin Imaging 2010; 34:185-190

257. Schwartz M. A biomathematical approach to clinical tumor growth. Cancer 1961; 14:1272-1294

258. Yankelevitz DF, Reeves AP, Kostis WJ, Zhao B, Henschke CI. Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. Radiology 2000; 217:251-256

259. Bru A, Albertos S, Luis SJ, Garcia-Asenjo JL, Bru I. The universal dynamics of tumor growth. Biophys J 2003; 85:2948-2961

260. Castro MA, Klamt F, Grieneisen VA, Grivicich I, Moreira JC. Gompertzian growth pattern correlated with phenotypic organization of colon carcinoma, malignant glioma and non-small cell lung carcinoma cell lines. Cell Prolif 2003; 36:65-73

261. Lindell RM, Hartman TE, Swensen SJ, Jett JR, Midthun DE, Mandrekar JN. 5-year lung cancer screening experience: growth curves of 18 lung cancers compared to histologic type, CT attenuation, stage, survival, and size. Chest 2009; 136:1586-1595

262. Kostis WJ, Reeves AP, Yankelevitz DF, Henschke CI. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. IEEE Trans Med Imaging 2003; 22:1259-1274

263. Hasegawa M, Sone S, Takashima S, *et al.* Growth rate of small lung cancers detected on mass CT screening. Br J Radiol 2000; 73:1252-1259

264. Revel MP, Merlin A, Peyrard S, *et al.* Software volumetric evaluation of doubling times for differentiating benign versus malignant pulmonary nodules. AJR Am J Roentgenol 2006; 187:135-142

265. Nathan MH, Collins VP, Adams RA. Differentiation of benign and malignant pulmonary nodules by growth rate. Radiology 1962; 79:221-232

266. Henschke CI, Yankelevitz DF, Yip R, *et al.* Lung Cancers Diagnosed at Annual CT Screening: Volume Doubling Times. Radiology 2012; 263:578-583

267. Yankelevitz DF, Henschke CI. Does 2-year stability imply that pulmonary nodules are benign? AJR Am J Roentgenol 1997; 168:325-328

268. Wilson DO, Ryan A, Fuhrman C, *et al.* Doubling times and CT screen-detected lung cancers in the Pittsburgh Lung Screening Study. Am J Respir Crit Care Med 2012; 185:85-89

269. Kakinuma R, Ohmatsu H, Kaneko M, *et al.* Progression of focal pure ground-glass opacity detected by low-dose helical computed tomography screening for lung cancer. J Comput Assist Tomogr 2004; 28:17-23

270. Kodama K, Higashiyama M, Yokouchi H, *et al.* Natural history of pure ground-glass opacity after long-term follow-up of more than 2 years. Ann Thorac Surg 2002; 73:386-392

271. Henschke CI, Yankelevitz DF, Mirtcheva R, McGuinness G, McCauley D, Miettinen OS. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. AJR Am J Roentgenol 2002; 178:1053-1057

272. de Hoop BJ, Gietema H, van d, V, Murphy K, van Klaveren RJ, Prokop M. Pulmonary ground-glass nodules: increase in mass as an early indicator of growth. Radiology 2010; 255:199-206

273. Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. JAMA 2001; 285:914-924

274. Gould MK, Fletcher J, Iannettoni MD, *et al.* Evaluation of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; 132:108S-130S

275. Yi CA, Lee KS, Kim BT, *et al.* Tissue characterization of solitary pulmonary nodule: comparative study between helical dynamic CT and integrated PET/CT. J Nucl Med 2006; 47:443-450

276. Veronesi G, Bellomi M, Veronesi U, *et al.* Role of positron emission tomography scanning in the management of lung nodules detected at baseline computed tomography screening. Ann Thorac Surg 2007; 84:959-965

277. Swensen SJ, Viggiano RW, Midthun DE, *et al.* Lung nodule enhancement at CT: multicenter study. Radiology 2000; 214:73-80

278. Yi CA, Lee KS, Kim EA, *et al.* Solitary pulmonary nodules: dynamic enhanced multi-detector row CT study and comparison with vascular endothelial growth factor and microvessel density. Radiology 2004; 233:191-199

279. Murray CP, Wong PM, Louw J, Waterer GW. Western Australian cigarette smokers have fewer small lung nodules than North Americans on CT screening for lung cancer. J Med Imaging Radiat Oncol 2009; 53:339-344

280. Pinsky PF, Gierada DS, Nath PH, Kazerooni E, Amorosa J. National lung screening trial: variability in nodule detection rates in chest CT studies. Radiology 2013; 268:865-873

281. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Invest Radiol 1978; 13:175-181

282. Quekel LG, Kessels AG, Goei R, van Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. Chest 1999; 115:720-724

283. Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect. Radiology 1992; 182:115-122

284. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Invest Radiol 1983; 18:194-198

285. Armato SG, III, Roberts RY, Kocherginsky M, *et al.* Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". Acad Radiol 2009; 16:28-38

286. Ochs RH, Angel E, Panknin C, McNitt-Gray M, Brown M. Forming a reference standard from LIDC data: impact of reader agreement on reported CAD performance. SPIE Proceedings 2007; 6514:

287. Dodd LE, Wagner RF, Armato SG, III, *et al.* Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium. Acad Radiol 2004; 11:462-475

288. Wormanns D, Fiebich M, Saidi M, Diederich S, Heindel W. Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system. Eur Radiol 2002; 12:1052-1057

289. Rubin GD, Lyo JK, Paik DS, *et al.* Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology 2005; 234:274-283

290. Roos JE, Paik D, Olsen D, *et al.* Computer-aided detection (CAD) of lung nodules in CT scans: radiologist performance and reading time with incremental CAD assistance. Eur Radiol 2010; 20:549-557

291. Das M, Muhlenbruch G, Mahnken AH, *et al.* Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance. Radiology 2006; 241:564-571

292. Marten K, Grillhosl A, Seyfarth T, Obenauer S, Rummeny EJ, Engelke C. Computer-assisted detection of pulmonary nodules: evaluation of diagnostic performance using an expert knowledge-based detection system with variable reconstruction slice thickness settings. Eur Radiol 2005; 15:203-212

293. Fraioli F, Bertoletti L, Napoli A, *et al.* Computer-aided detection (CAD) in lung cancer screening at chest MDCT: ROC analysis of CAD versus radiologist performance. J Thorac Imaging 2007; 22:241-246

294. Beyer F, Zierott L, Fallenberg EM, *et al.* Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader. Eur Radiol 2007; 17:2941-2947

295. van Ginneken B, Armato SG, III, de HB, *et al.* Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study. Med Image Anal 2010; 14:707-722

296. Diederich S, Lenzen H, Windmann R, *et al.* Pulmonary nodules: experimental and clinical studies at low-dose CT. Radiology 1999; 213:289-298

297. Gartenschlager M, Schweden F, Gast K, *et al.* Pulmonary nodules: detection with low-dose vs conventional-dose spiral CT. Eur Radiol 1998; 8:609-614

298. Gergely I, Neumann C, Reiger F, Dorffner R. [Lung nodule detection with ultra-low-dose CT in routine follow-up of cancer patients]. Rofo 2005; 177:1077-1083

299. Hetmaniak Y, Bard JJ, Albuisson E, *et al.* [Pulmonary nodules: dosimetric and clinical studies at low dose multidetector CT]. J Radiol 2003; 84:399-404

300. Itoh S, Ikeda M, Arahata S, *et al.* Lung cancer screening: minimum tube current required for helical CT. Radiology 2000; 215:175-183

301. Karabulut N, Toru M, Gelebek V, Gulsun M, Ariyurek OM. Comparison of low-dose and standard-dose helical CT in the evaluation of pulmonary nodules. Eur Radiol 2002; 12:2764-2769

302. Nitta N, Takahashi M, Murata K, Morita R. Ultra low-dose helical CT of the chest: evaluation in clinical cases. Radiat Med 1999; 17:1-7

303. Rusinek H, Naidich DP, McGuinness G, *et al.* Pulmonary nodule detection: low-dose versus conventional CT. Radiology 1998; 209:243-249

304. Christe A, Torrente JC, Lin M, *et al.* CT screening and follow-up of lung nodules: effects of tube current-time setting and nodule size and density on detectability and of tube current-time setting on apparent size. AJR Am J Roentgenol 2011; 197:623-630

305. Bogot NR, Kazerooni EA, Kelly AM, Quint LE, Desjardins B, Nan B. Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. Acad Radiol 2005; 12:948-956

306. Reeves AP, Biancardi AM, Apanasovich TV, *et al.* The Lung Image Database Consortium (LIDC): a comparison of different size metrics for pulmonary nodule measurements. Acad Radiol 2007; 14:1475-1485

307. Revel MP, Bissery A, Bienvenu M, Aycard L, Lefort C, Frija G. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 2004; 231:453-458

308. Lillington GA, Caskey CI. Evaluation and management of solitary and multiple pulmonary nodules. Clin Chest Med 1993; 14:111-119

309. Marten K, Auer F, Schmidt S, Kohl G, Rummeny EJ, Engelke C. Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. Eur Radiol 2006; 16:781-790

310. Bolte H, Riedel C, Muller-Hulsbeck S, *et al.* Precision of computer-aided volumetry of artificial small solid pulmonary nodules in ex vivo porcine lungs. Br J Radiol 2007; 80:414-421

311. Marchiano A, Calabro E, Civelli E, *et al.* Pulmonary nodules: volume repeatability at multidetector CT lung cancer screening. Radiology 2009; 251:919-925

312. Wormanns D, Kohl G, Klotz E, *et al.* Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. Eur Radiol 2004; 14:86-92

313. Gietema HA, Wang Y, Xu D, *et al.* Pulmonary nodules detected at lung cancer screening: interobserver variability of semiautomated volume measurements. Radiology 2006; 241:251-257

314. Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL. Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. AJR Am J Roentgenol 2006; 186:989-994

315. Goo JM, Tongdee T, Tongdee R, Yeo K, Hildebolt CF, Bae KT. Volumetric measurement of synthetic lung nodules with multi-detector row CT: effect of various image reconstruction parameters and segmentation thresholds on measurement accuracy. Radiology 2005; 235:850-856

316. Ravenel JG, Leue WM, Nietert PJ, Miller JV, Taylor KK, Silvestri GA. Pulmonary nodule volume: effects of reconstruction parameters on automated measurements--a phantom study. Radiology 2008; 247:400-408

317. Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, *et al.* Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. Radiology 2008; 248:625-631

318. Ko JP, Rusinek H, Jacobs EL, *et al.* Small pulmonary nodules: volume measurement at chest CT--phantom study. Radiology 2003; 228:864-870

319. Gietema HA, Schaefer-Prokop CM, Mali WP, Groenewegen G, Prokop M. Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT-- influence of inspiration level, nodule size, and segmentation performance. Radiology 2007; 245:888-894

320. Wang Y, de Bock GH, van Klaveren RJ, *et al.* Volumetric measurement of pulmonary nodules at low-dose chest CT: effect of reconstruction setting on measurement variability. Eur Radiol 2010; 20:1180-1187

321. Ashraf H, de HB, Shaker SB, *et al.* Lung nodule volumetry: segmentation algorithms within the same software package cannot be used interchangeably. Eur Radiol 2010; 20:1878-1885

322. Park CM, Goo JM, Lee HJ, Kim KG, Kang MJ, Shin YH. Persistent pure ground-glass nodules in the lung: interscan variability of semiautomated volume and attenuation measurements. AJR Am J Roentgenol 2010; 195:W408-W414

323. Gruden JF, Ouanounou S, Tigges S, Norris SD, Klausner TS. Incremental benefit of maximum-intensity-projection images on observer detection of small pulmonary nodules revealed by multidetector CT. AJR Am J Roentgenol 2002; 179:149-157

324. Valencia R, Denecke T, Lehmkuhl L, Fischbach F, Felix R, Knollmann F. Value of axial and coronal maximum intensity projection (MIP) images in the detection of pulmonary nodules by multislice spiral CT: comparison with axial 1-mm and 5-mm slices. Eur Radiol 2006; 16:325-332

325. Jankowski A, Martinelli T, Timsit JF, *et al.* Pulmonary nodule detection on MDCT images: evaluation of diagnostic performance using thin axial images, maximum intensity projections, and computer-assisted detection. Eur Radiol 2007; 17:3148-3156

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

327. Copley SJ, Bryant TH, Chambers AA, *et al.* Observer accuracy in the detection of pulmonary nodules on CT: effect of cine frame rate. Clin Radiol 2010; 65:133-136

328. Li F, Sone S, Abe H, MacMahon H, Armato SG, III, Doi K. Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings. Radiology 2002; 225:673-683

329. Beutel J, Kundel HL, van Metter RL. Part II: Psychophysics. In:Beutel J, Kundel H L, van Metter R L, eds. Handbook of Medical Imaging Volume I: Physics and Psychophysics. SPIE, 2000; 557-920

330. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996; 3:1000-1006

331. Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. RadioGraphics 1987; 7:1241-1250

332. Samuel S, Kundel HL, Nodine CF, Toto LC. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. Radiology 1995; 194:895-902

333. Nodine CF, Krupinski EA, Kundel HL, Toto L, Herman GT. Satisfaction of search (SOS). Invest Radiol 1992; 27:571-573

334. Wormanns D, Ludwig K, Beyer F, Heindel W, Diederich S. Detection of pulmonary nodules at multirow-detector CT: effectiveness of double reading to improve sensitivity at standard-dose and low-dose chest CT. Eur Radiol 2005; 15:14-22

335. Kundel HL, La Follette PSJ. Visual search patterns and experience with radiological images. Radiology 1972; 103:523-528

336. Marten K, Seyfarth T, Auer F, *et al.* Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. Eur Radiol 2004; 14:1930-1938

337. Awai K, Murao K, Ozawa A, *et al.* Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance. Radiology 2004; 230:347-352

338. Brochu B, Beigelman-Aubry C, Goldmard JL, Raffy P, Grenier PA, Lucidarme O. [Computer-aided detection of lung nodules on thin collimation MDCT: impact on radiologists' performance]. J Radiol 2007; 88:573-578

339. Bennett RL, Sellars SJ, Blanks RG, Moss SM. An observational study to evaluate the performance of units using two radiographers to read screening mammograms. Clin Radiol 2012; 67:114-121

340. Pauli R, Hammond S, Cooke J, Ansell J. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. J Med Screen 1996; 3:18-22

341. Duijm LE, Groenewoud JH, Fracheboud J, van Ineveld BM, Roumen RM, de Koning HJ. Introduction of additional double reading of mammograms by radiographers: effects on a biennial screening programme outcome. Eur J Cancer 2008; 44:1223-1228

342. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241-244

343. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol 2003; 180:1461-1467

344. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology 2010; 257:14-17

345. Brown MS, Goldin JG, Rogers S, *et al.* Computer-aided lung nodule detection in CT: results of large-scale observer test. Acad Radiol 2005; 12:681-686

346. White CS, Pugatch R, Koonce T, Rust SW, Dharaiya E. Lung nodule CAD software as a second reader: a multicenter study. Acad Radiol 2008; 15:326-333

347. Armato SG, III, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. Med Phys 2001; 28:1552-1561

348. Ko JP, Betke M. Chest CT: automated nodule detection and assessment of change over time--preliminary experience. Radiology 2001; 218:267-273

349. Bae KT, Kim JS, Na YH, Kim KG, Kim JH. Pulmonary nodules: automated detection on CT images with morphologic matching algorithm--preliminary results. Radiology 2005; 236:286-293

350. Kim JS, Kim JH, Cho G, Bae KT. Automated detection of pulmonary nodules on CT images: effect of section thickness and reconstruction interval--initial results. Radiology 2005; 236:295-299

351. Marten K, Engelke C, Seyfarth T, Grillhosl A, Obenauer S, Rummeny EJ. Computer-aided detection of pulmonary nodules: influence of nodule characteristics on detection performance. Clin Radiol 2005; 60:196-206

352. Yuan R, Vos PM, Cooperberg PL. Computer-aided detection in screening CT for pulmonary nodules. AJR Am J Roentgenol 2006; 186:1280-1287

353. Sahiner B, Chan HP, Hadjiiski LM, *et al.* Effect of CAD on radiologists' detection of lung nodules on thoracic CT scans: analysis of an observer performance study by nodule size. Acad Radiol 2009; 16:1518-1530

354. Foti G, Faccioli N, D'Onofrio M, Contro A, Milazzo T, Pozzi MR. Evaluation of a method of computer-aided detection (CAD) of pulmonary nodules with computed tomography. Radiol Med 2010; 115:950-961

355. Guerriero C, Gillan MG, Cairns J, Wallis MG, Gilbert FJ. Is computer aided detection (CAD) cost effective in screening mammography? A model based on the CADET II study. BMC Health Serv Res 2011; 11:11

356. Cassidy A, Myles JP, van TM, *et al.* The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 2008; 98:270-276

357. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology. AJR Am J Roentgenol 1996; 166:517-521

358. Machin D, Campbell MJ, Walters SJ. Bayes Theorem. Medical Statistics: A Textbook for the Health Sciences. Chichester, England: John Wiley and Sons, 2007; 51-57

359. Goo JM. A computer-aided diagnosis for evaluating lung nodules on chest CT: the current status and perspective. Korean J Radiol 2011; 12:145-155

360. Applegate KE, Tello R, Ying J. Hypothesis testing III: counts and medians. Radiology 2003; 228:603-608

361. Machin D, Campbell MJ, Walters SJ. Comparison of two independent groups- categorical outcomes. Medical Statistics: A Textbook for the Health Sciences. Chichester, England: John Wiley and Sons, 2007; 132-136

362. Tello R, Crewson PE. Hypothesis testing II: means. Radiology 2003; 227:1-4

363. Machin D, Campbell MJ, Walters SJ. Comparison of two independent groups- continuous outcomes. Medical Statistics: A Textbook for the Health Sciences. Chichester, England: John Wiley and Sons, 2007; 125-131

364. Machin D, Campbell MJ, Walters SJ. Comparison of two groups of paired observations- continuous outcomes. Medical Statistics: A Textbook for the Health Sciences. Chichester, England: John Wiley and Sons, 2007; 119-125

365. Crewson PE. Reader agreement studies. AJR Am J Roentgenol 2005; 184:1391-1397

366. Pauli R, Hammond S, Cooke J, Ansell J. Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions. Br J Radiol 1996; 69:10-14

367. Lauridsen C, Lefere P, Gerke O, Gryspeerdt S. Effect of a tele-training programme on radiographers in the interpretation of CT colonography. Eur J Radiol 2012; 81:851-856

368. European Society of Gastrointestinal and Abdominal Radiology CT Colonography Group Investigators. Effect of directed training on reader performance for CT colonography: multicenter study. Radiology 2007; 242:152-161

369. Scott HJ, Gale AG. Breast screening: PERFORMS identifies key mammographic training needs. Br J Radiol 2006; 79 Spec No 2:S127-S133

370. Wood BP. Feedback: a key feature of medical training. Radiology 2000; 215:17-19

371. Beigelman-Aubry C, Raffy P, Yang W, Castellino RA, Grenier PA. Computer-aided detection of solid lung nodules on follow-up MDCT screening: evaluation of detection, tracking, and reading time. AJR Am J Roentgenol 2007; 189:948-955

372. Wood BP. Visual expertise. Radiology 1999; 211:1-3

373. Burling D, Wylie P, Gupta A, *et al.* CT colonography: accuracy of initial interpretation by radiographers in routine clinical practice. Clin Radiol 2010; 65:126-132

374. Kinnunen J, Gothlin JH, Totterman S. Effect of training and experience on radiologic diagnostic performance in midfacial trauma. Acta Radiol 1988; 29:83-87

375. Davies IRL, Sowden PT, Hammond SM, Ansell J. Expertise in categorizing mammograms: a perceptual or conceptual skill? Proc SPIE 1994; 2166:86-95

376. Jensch S, van Gelder RE, Florie J, *et al.* Performance of radiographers in the evaluation of CT colonographic images. AJR Am J Roentgenol 2007; 188:W249-W255

377. Burling D, Halligan S, Altman DG, *et al.* CT colonography interpretation times: effect of reader experience, fatigue, and scan findings in a multi-centre setting. Eur Radiol 2006; 16:1745-1749

378. Armato SG, III, McNitt-Gray MF, Reeves AP, *et al.* The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Acad Radiol 2007; 14:1409-1421

379. Armato SG, III, Roberts RY, McNitt-Gray MF, *et al.* The Lung Image Database Consortium (LIDC): ensuring the integrity of expert-defined "truth". Acad Radiol 2007; 14:1455-1463

380. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992; 304:1491-1494

381. Kakinuma R, Ohmatsu H, Kaneko M, *et al.* Detection failures in spiral CT screening for lung cancer: analysis of CT findings. Radiology 1999; 212:61-66

382. White CS, Romney BM, Mason AC, Austin JH, Miller BH, Protopapas Z. Primary carcinoma of the lung overlooked at CT: analysis of findings in 14 patients. Radiology 1996; 199:109-115

383. Henschke CI, Yip R, Yankelevitz DF, Smith JP. Definition of a positive test result in computed tomography screening for lung cancer: a cohort study. Ann Intern Med 2013; 158:246-252

384. Li Q. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Comput Med Imaging Graph 2007; 31:248-257

385. Nietert PJ, Ravenel JG, Taylor KK, Silvestri GA. Influence of nodule detection software on radiologists' confidence in identifying pulmonary nodules with computed tomography. J Thorac Imaging 2011; 26:48-53

386. Zhao Y, de Bock GH, Vliegenthart R, *et al.* Performance of computer-aided detection of pulmonary nodules in low-dose CT: comparison with double reading by nodule volume. Eur Radiol 2012; 22:2076-2084

387. Lee IJ, Gamsu G, Czum J, Wu N, Johnson R, Chakrapani S. Lung nodule detection on chest CT: evaluation of a computer-aided detection (CAD) system. Korean J Radiol 2005; 6:89-93

388. Godoy MC, Cooperberg PL, Maizlin ZV, *et al.* Detection sensitivity of a commercial lung nodule CAD system in a series of pathologically proven lung cancers. J Thorac Imaging 2008; 23:1-6

389. Zheng B, Swensson RG, Golla S, *et al.* Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments. Acad Radiol 2004; 11:398-406

390. Teague SD, Trilikis G, Dharaiya E. Lung nodule computer-aided detection as a second reader: influence on radiology residents. J Comput Assist Tomogr 2010; 34:35-39

391. Koo CW, Anand V, Girvin F, *et al.* Improved efficiency of CT interpretation using an automated lung nodule matching program. AJR Am J Roentgenol 2012; 199:91-95

392. Wilson DO, Weissfeld JL, Fuhrman CR, *et al.* The Pittsburgh Lung Screening Study (PLuSS): outcomes within 3 years of a first computed tomography scan. Am J Respir Crit Care Med 2008; 178:956-961

393. Blanchon T, Brechot JM, Grenier PA, *et al.* Baseline results of the Depiscan study: a French randomized pilot trial of lung cancer screening comparing low dose CT scan (LDCT) and chest X-ray (CXR). Lung Cancer 2007; 58:50-58

394. Petrick, N., Gallas, B. D., Samuelson, F. W., Wagner, R. F., and Myers, K. J. Influence of panel size and expert skill on truth panel performance when combining expert ratings. Proc.SPIE 2005; 5749:49-57

395. New York Early Lung Cancer Action Project Investigators. CT Screening for lung cancer: diagnoses resulting from the New York Early Lung Cancer Action Project. Radiology 2007; 243:239-249

396. Leader JK, Warfel TE, Fuhrman CR, *et al.* Pulmonary nodule detection with low-dose CT of the lung: agreement among radiologists. AJR Am J Roentgenol 2005; 185:973-978

397. Gierada DS, Pilgram TK, Ford M, *et al.* Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. Radiology 2008; 246:265-272

398. Sone S, Takashima S, Li F, *et al.* Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet 1998; 351:1242-1245

Appendix 1: Permission to obtain NELSON nodules

Email correspondence October 2011

RE: NELSON nodules/board

21/10/2011

Hansell David <davidhansell@rbht.nhs.uk>

to Mali, W., Pim, Anand, me

Dear Dr Mali,
So very many thanks for your reply. That is all very clear. Acknowledgement of "Nelson nodules" was taken as read (in fact, the provenance of the cases would have to be stated in the Methods section, at the very least).
As you have suggested, we will route any further questions we have about permission/acknowledgement/authorship thro' the excellent Pim.
Thanks again,
DavidMH

From: Mali, W. [W.Mali@umcutrecht.nl]
Sent: 20 October 2011 10:57
To: Hansell David
Cc: Pim de Jong
Subject: RE: NELSON nodules/board

Dear dr Hansell, thank you for your mail.
I am glad the cases were helpful for you.
I think it is fine when these cases are used for the thesis too.
I don t think we have to be co authors but perhaps you can acknowledge
The Nelson study for supplying you with the cases.
Whenever you have more questions about the Nelson study
Or want to contact us please feel free to approach Pim de Jong
He is my trusted collaborator at the UMCUtrecht and can help you.

Best regards Willem Mali

Van: Hansell David [mailto:davidhansell@rbht.nhs.uk]
Verzonden: donderdag 13 oktober 2011 8:55

Aan: Mali, W.
Onderwerp: NELSON nodules/board

Dear Dr Mali,

Please forgive this out of the blue approach. We have not met but Mathias Prokop suggested I contact you.

I am involved with the UK lung cancer screening pilot study (Prof John Field is the Priniciple Investigator) and as you may be aware Dr Pim de Jong has been hugely helpful in providing our Dr Arjun Nair with some CTs from Utrecht which have been put to good effect in training our radiographers.

Dr Nair is with us for the next year or two and, with myself and a co-supervisor Dr Anand Devaraj, we are putting together a plan for his thesis – which will revolve around the successful, or otherwise, training of non-radiologists for reading screening CTs (approximately 7 chapters, i.e. 7 different investigations). Because recruiting of UKLS cases only starts in the next 3 weeks time it will be a few months before we have sufficient UK studies to make use of, so it would be useful but not absolutely crucial, if we could use some of the Nelson cases for studies in the first two of his chapters.

We understood clearly from the outset, from Pim de Jong and Mathias Prokop, that we would need the sanction of the Nelson Board to use Nelson CTs in any study which is why I am writing to you. I understand that you are on the Nelson Board and since the CTs are Utrecht cases I thought it would be appropriate to route any such request through you. Would that be okay? Of course, if one or other of the investigations using "Nelson nodules" is fit for publication we would be pleased to have relevant coauthors as suggested by you/the Board. I think the plan for Dr Arjun's thesis will be completed in the next couple of weeks so at that point we would have the outline of each investigation for your/Nelson Board's approval.

Again, my apologies for this sudden onslaught. I look forward to hearing from you.

Kind regards,
David Hansell
Royal Brompton Hospital
London UK

---

verzocht de inhoud niet te gebruiken en de afzender direct te informeren door het bericht te retourneren. Het Universitair Medisch Centrum Utrecht is een publiekrechtelijke rechtspersoon in de zin van de W.H.W. (Wet Hoger Onderwijs en Wetenschappelijk Onderzoek) en staat geregistreerd bij de Kamer van Koophandel voor Midden-Nederland onder nr. 30244197.

Denk s.v.p aan het milieu voor u deze e-mail afdrukt.

This message may contain confidential information and is intended exclusively for the addressee. If you receive this message unintentionally, please do not use the contents but notify the sender immediately by return e-mail. University Medical Center Utrecht is a legal person by public law and is registered at the Chamber of Commerce for Midden-Nederland under no. 30244197.

Please consider the environment before printing this e-mail.

DISCLAIMER:

The information contained in this email may be subject to public disclosure under the NHS Code of Openness or the Freedom of Information Act 2000. Unless the information is legally exempt from disclosure, the confidentiality of this email, and your reply cannot be guaranteed.

The information and material in this email is intended for the use of the intended addressee or the person responsible for delivering it to the intended addressee. It may contain privileged or confidential information and/or copyright material.

If you receive this email by mistake please advise the sender immediately by using the reply facility in your email software or notify Royal Brompton & Harefield NHS Trust Help Desk on +44(0) 20 7351 8696

Communication is not sent through a secure server; Royal Brompton & Harefield NHS Trust cannot accept responsibility for the accuracy of outgoing electronic mail. Any views or opinions expressed are solely those of the author and do not represent the view of Royal Brompton & Harefield NHS Trust unless specifically stated.

Appendix 2: Cost quote for Visia CT Lung System version 3.1 (Mevis Medical Solutions, Bremen, Germany)

MeVis
MEDICAL SOLUTIONS

SALES QUOTATION

QUOTE NUMBER	QUOTE DATE
Q-090611-0	09/06/11

QUOTE TO:
DR. ANAND DEVARAJ
ST GEORGE'S HOSPITAL
BLACKSHAW ROAD
LONDON SW17 0QT
UK

SHIP TO:
ST GEORGE'S HOSPITAL
BLACKSHAW ROAD
LONDON SW17 0QT
UK

LINE	PART NO.	DESCRIPTION	QTY.	UNIT LIST PRICE	DISCOUNT	AMOUNT
1	VCT-3100-N	**Visia™ CT Lung System with Nodule CAD (Version 3.1) with AutoPoint™ Temporal Comparison and Pulmonary Artery Patency Exam (PE™) Features** Description: The Visia™ CT Lung System (v3.1) provides exclusive FDA-cleared lung nodule computer-aided-detection (CAD) and automatic measurement tools for chest multi-slice CT (MSCT). Includes Visia™ server (v3.1) and workstation (v3.0) software, AutoPoint™ temporal comparison feature, Pulmonary Artery Patency Exam (PE™) feature, remote installation and applications training, and 1-year warranty. Workstation Client Licenses: 3 Included *(NOTE: Visia™ CT Lung System is a software-only solution. Customer must provide server and workstation hardware compliant with system requirements.)*	1	€52,000	(€13,000) 25% Clinical Partner Discount	€39,000
2	VCT-3100-IT	**Onsite installation and applications training**		€ 1,800		€ 1,800
3	VCT-3100-SA	**Visia™ CT Lung System Service Agreement** Description: The Visia CT Lung System (v3.1) – System Support Agreement includes telephone technical support, remote diagnostics and system maintenance, and software updates and upgrades to existing functionality. Service Agreement Term: 1 Year		€ 5,500		€ 5,500
					SALES TAX:	See Below

Terms: Net 30 days; FOB Factory Freight prepaid and added Quote Valid for 30 days	**Warranty:** MeVis Medical Solutions, Inc. warrants these products to be free from defects in materials and workmanship for a period of one (1) year from date of delivery.
Special Conditions: Sales tax will be added if the Buyer is not tax exempt. Please include your tax exemption number or local tax rate on your purchase order, as applicable.	**Quote Prepared By:** **Signature:** _____
To place an order for the products and/or services included in this quotation, please sign and date this document in the space provided or submit a purchase order via mail or fax to: MeVis Medical Solutions, Inc. N27 W24075 Paul Ct. Ste. 100 Pewaukee, WI USA 53072 Phone: 262-691-9530 Fax: 262-691-9531	**Quote Accepted By:** _____ **Signature:** _____ **Date:** _____