Methods and validation of nodule measurement in a lung cancer screening

Ying, W.
Chapter 1

General Introduction
1.1 Lung cancer screening

Lung cancer is the most common and fatal cancer in the world [1]. In Europe, there were 391,000 (12% of the total, third place) new lung cancer cases and 342,000 lung cancer deaths (20% of total, first place) in 2008 [2]. Despite the modest improvements in treatments during the last few decades, the prognosis of lung cancer is still poor and the 5 year survival rate is 15% in the United States, 10% in Europe and 9% in developing countries [1]. The survival of lung cancer is closely related to the stage at the time of diagnosis, ranging from 70% for limited, stage I disease to less than 5% for stage IV disease [3]. Usually, lung cancer does not cause symptoms early in the disease process, and is mostly diagnosed at a late stage in a clinical setting, when the probability of cure is rare.

It is expected that screening can detect lung cancer at an early stage and reduce mortality. A number of clinical trials have been performed trying to prove this hypothesis. The screening trials using chest radiography alone or in combination with sputum cytological examination failed to demonstrate a reduction of mortality due to lack of sensitivity in picking up suspicious lesions [4-8]. With the development of multi-detector spiral computed tomography (CT) and its advantage in the detection of small pulmonary nodules [9-11], the interest in lung cancer screening rekindled. The results of several observational CT screening trials showed that CT is effective in the detection of early stage lung cancer, with a percentage of stage I lung cancers ranged from 68% to 96% of all detected [12-27]. Despite its efficiency on early detection, lung cancer CT screening is still not being recommended by any public health department due to the inherent biases in cohort studies, including lead-time bias, length-time bias and over-diagnosis bias. It is accepted that only a randomized controlled lung cancer screening trial can eliminate these biases to the highest degree and answer the question about mortality reduction by comparing lung-cancer mortality in the screening arm (with CT screening) and the control arm (without CT screening). The Dutch-Belgian lung cancer screening trial (NELSON trial) [28, 29] is an ongoing multicenter randomized controlled low-dose CT lung cancer screening trial in the Netherlands and Belgium. The primary objective of the study is to determine whether screening with chest CT reduces lung cancer mortality in a high-risk cohort compared to a control group receiving no screening. This study recruited 15,530 smokers and former smokers aged 50 to 75 years. The incidence screens take place 1 year, 3 years and 5 years after the prevalence screen. The NELSON trial uses 16-row multi-detector CT and semi-automated 3-dimensional volumetric assessment of nodular growth exclusively.
1.2 Nodule measurement in Lung cancer screening

A large number of pulmonary nodules are being detected in lung cancer screening and only a few of them are malignant. How to accurately pick up the malignant nodule is of vital importance and can potentially influence the outcome of screening. In most screening trials, the nodule management protocols held the same concept: work-ups are recommended for large-size nodules, short-term follow-ups and growth evaluations for median-size nodules and long-term follow-up for small-size nodules. Therefore, an accurate method for the estimation of nodule size is a prerequisite for lung cancer screening, especially in the growth assessment of pulmonary nodules. Furthermore, the volume doubling time (VDT) based on serial CT examinations has been proven to be a valuable tool in the differentiation between benign and malignant pulmonary nodules. Malignant nodules typically present with a VDT between 30 and 400 days [30-35] and thus may be applied as criterion in lung cancer screening.

In most screening trials, the determination of nodule size and its growth were based on two-dimensional (2D) manual diameter measurements. The 2D measurement is subject to substantial measurement variability. This could be the reason that most trials had no quantitative criteria of significant growth (table 1). The recent development of three-dimensional (3D) software-generated volumetric measurement has provided a new method for nodule measurement. Compared to traditional 2D measurement, 3D measurement is inherently more sensitive since the change of volume is more distinct than diameter as a doubling in volume is equal to a 26% increase in diameter in a spherical nodule. For instance, the diameter of a 5 mm nodule will increase to 6.3 mm when its volume doubles. Such a subtle change in size is difficult to be recognized by 2D diameter measurement, especially when the nodule has a non-smooth margin or irregular shape.

The 3D volumetric measurement software is currently commercial available. Although it is more accurate and reproducible than traditional 2D measurement [35-38], it still has some degrees of measurement error. Understanding and quantifying the sources of volumetric measurement error is important to optimize the growth criterion in a screening project.

Software-generated nodule volumetric measurement is typically initiated by defining a region of interest around a nodule or by a user-provided point inside the nodule area. Depending on the application, segmentation algorithms based on attenuation and morphology are then employed to delineate nodules from the surrounding lung parenchyma and neighboring structures such as attached vasculature and pleural surfaces. The estimate of nodule volume is then based on the total number of voxels within the segmented region.
Table 1. The methods for nodule measurement, growth assessment and reading strategy in lung cancer screening trials

<table>
<thead>
<tr>
<th>Name or institution and year started [references]</th>
<th>Nodule size determination</th>
<th>Quantitative criterion for significant growth</th>
<th>Reading strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-ELCAP, 1993, International [18]</td>
<td>2D diameter</td>
<td>No</td>
<td>Consensus double reading</td>
</tr>
<tr>
<td>University of Munster, 1995, Germany [20,21]</td>
<td>2D diameter</td>
<td>No</td>
<td>Single reading</td>
</tr>
<tr>
<td>Hitachi Health Care Center, 2002, Japan [12]</td>
<td>2D diameter</td>
<td>No</td>
<td>NR</td>
</tr>
<tr>
<td>NY-ELCAP, 2000, US [19]</td>
<td>2D diameter</td>
<td>No</td>
<td>Consensus double reading</td>
</tr>
<tr>
<td>ITALUNG, 2004, Italian [75]</td>
<td>2D diameter</td>
<td>Increase of at least 1mm in mean diameter</td>
<td>Consensus double reading</td>
</tr>
<tr>
<td>NELSON, 2003, Netherland [29]</td>
<td>3D volume</td>
<td>Volume change&gt;25% and VDT&lt;400 days</td>
<td>Consensus double reading</td>
</tr>
</tbody>
</table>

ELCAP indicates Early Lung Cancer Action Program; I-ELCAP indicates international ELCAP; NY-ELCAP indicates New York ELCAP; NR, not reported; 2D, two dimensional; 3D, three dimensional
The accuracy of volumetric measurement is firstly affected by partial volume effect, an imaging artifact caused by the limited resolution of CT scanners and the averaging of the linear attenuation coefficients of all materials in a voxel. The problem results when an algorithm counts a percentage of voxels as containing pure nodule tissue or pure parenchyma tissue when, in fact, the voxels contain both [39], which leads to an over- or under-estimation of true volume of nodule tissue. It has been shown that the percentage of partial volume voxels decreases as nodule size increases and as section thickness decreases [40-42]. Secondly, the performance of the segmentation algorithms used to distinguish nodule boundaries from attached structures in most software did not work optimally. Kostis et al. developed a model-based volumetric measurement method which classified nodules into four types on the basis of surrounding structures and found acceptable segmentations (as determined by a radiologist) in approximately 80% and 72% of segmented nodules with vascular and pleural attachments, respectively [43]. Das et al. reported the overall absolute percentage error varied for different nodule attachment categories using LungCare software and was highest for pleural nodules, ranging from 10.3% to 21.2% across CT vendors [44]. Other influencing factors of volumetric measurement accuracy include scan acquisition and reconstruction parameters, such as section thickness, collimation, radiation exposure and the type of CT system [39, 40, 42, 44-47].

The reproducibility of volumetric measurement is another important indicator, representing the consistency of the measured volume. It consists of both the variability presented in the contemporary repeat scans of the same object (inter-scan variability) and the variability caused by the reader (intra- or inter-observer variability). The inter-scan variability can be caused by real change of the objects. For instance, the nodule itself and its margins with surrounding lung parenchyma can vary from scan to scan, physiological changes such as lung volume; furthermore, variation in phase of cardiac cycle, microatelectasis, or patient position on the table may occur [48-52]. In addition, the performance of segmentation could also play an important role [50]. The intra- or inter-observer variability is primarily due to semi-automated software because the observer is responsible for setting a point inside the nodule area to initiate the whole process. In some cases, the observer even needs to modify unsatisfying segmentation by the software. The disagreement between observers mostly occurs in case of some “difficult “nodules, such as those with non-smooth margin, irregular shape or attachment to surrounding structures [53].

There have been a number of studies quantifying the degree of the variability. Wormanns et al. [54] reported an inter-observer variability ranging from -5.5 to 6.6% using LungCare software (Siemens). Revel et al. [53] reported the upper limit of the 95% limits of
acceptability was 6.38% of the previously measured volume using CT Lung Analysis software (GE Medical Systems Europe). Goodman et al. [49] reported the mean inter-observer variability was 0.018% with a standard deviation of 0.73% on 43 successfully segmented nodules. However, multiple interrelated variables are involved in their studies, such as the variance in CT vendor, dose exposition, nodule selection, software vendor. Thus, it is currently difficult to give a general estimate of variability. Validation is thus necessary for the application of certain software in certain circumstances. Moreover, several studies [45,55,56] have found a significant effect of reconstruction settings on volume measurement variability, but these results are derived from normal-dose chest CTs, and no data are available from low-dose chest CT studies so far. Furthermore, all these studies focused on the agreement between the volumes measured with the different reconstruction settings, but none of them addressed the repeatability of each individual reconstruction setting. For reconstruction settings with a lower repeatability, agreement is also expected to be lower.

1.3 Reading strategy in lung cancer screening trial

Achieving a maximum diagnostic yield in lung cancer screening programs is not only dependent on the image quality but also on the appropriate reading of the images. Efforts to improve accuracy and to reduce variability in the interpretation can potentially increase the effectiveness of a screening program.

Most studies in breast cancer screening have shown that double reading increased the cancer detection rate with 6-15% compared to single reading [57-63]. Despite inconsistency in the data as some investigators only found an increase in the cancer detection rate of 2-5% after double reading [64-66], the strategy of double reading is recommended for breast cancer screening today. Taking the costs of double reading into account, double reading also appeared to be more cost-effective than a single reading policy [67, 68]. Double reading can be generally performed by three different methods: independent double reading in which the final result is in favor of the positive diagnosis of either reader; consensus double reading, in which both readers try to reach consensus; and independent double reading arbitrated by a third reader. The first method of double reading has been reported to increase the sensitivity of cancer detection at the expense of a decrease in specificity in screening projects [58, 63,64], while the latter two methods have been reported to achieve an increase in sensitivity without changing the specificity [59,60,62].

In lung cancer screening trials, most studies adopt a double reading strategy (table 1). However, there is few data to date focusing on the reading strategy in lung cancer screening,
especially no information is available for the most important indicator, the change of cancer detection rate due to the double reading. Gierada et al. explored the inter-reader variability in lung screening project and showed that the difference between readers could have occurred in lesion detection, characterization of a lesion as a nodule or non-nodule, and lesion measurement. The inter-observer agreement was moderate to substantial and potential for considerable improvement existed [69]. Similar results were found in other studies in a clinical setting: a relatively high inter-observer variability for the detection and characterization of pulmonary nodules [70-72]. In a lung cancer screening project, image interpretation can be performed in different situations and these situations could result in different outcomes. For instance, displaying the chest CT data on a workstation, reconstructing the data with maximum intensity projection or volume rendering algorithm have been proven to increase the accuracy of the reader [73, 74]. Furthermore, using 3D software-generated volume as the criterion for nodule classification could also reduce the inter-observer variability. Therefore, the value of double reading could be different in different trials due to the difference in methodology.

1.4 Overview of this thesis
The issues of methods and validation of nodule measurement in lung cancer screening are studied in this thesis. In chapter 2, we describe the screening results of the first and second round of the NELSON trial, in which the software-generated volume and volume-doubling time of a non-calcified lung nodule was used as the main criterion for deciding on further action. The inter-observer variability of semi-automated volumetric measurement of pulmonary nodules and the effect of nodule characteristics were assessed in chapter 3 and 4. Furthermore, we explored the effect of reconstruction parameters on the repeatability of semi-automated nodule volumetric measurement and assessed the agreement between three reconstruction settings in chapter 5. The value of consensus double reading in the baseline screening of the NELSON trial was assessed in chapter 6.
References


55. 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 2006; 235:850–856


60. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241-244


65. Denton ER, Field S. Just how valuable is double reporting in screening mammography? Clin Radiol 1997; 52:466-468