Heike Peulen*, Frederick Mantel*, Matthias Guckenberger*2,3, José Belderbos1, Maria Werner-Wasik4, Andrew Hope5, Meredith Giuliani5, Inga Grills6, Jan-Jakob Sonke1

*Both authors contributed equally to this manuscript

1  Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
2 Department of Radiation Oncology, University of Wuerzburg, Wuerzburg, Germany
3 Department of Radiation Oncology, University Hospital Zurich, Zurich, Switzerland
4 Department of Radiation Oncology Thomas Jefferson University Hospital, Philadelphia, PA, USA
5 Department of Radiation Oncology University of Toronto and Princess Margaret Cancer Center, Toronto, Canada
6 Department of Radiation Oncology Beaumont Hospital, Royal Oak, USA
CHAPTER 4

Validation of high-risk CT features for detection of local recurrence after stereotactic body radiotherapy for early stage non-small cell lung cancer

ABSTRACT

Purpose
Fibrotic changes after stereotactic body radiation therapy (SBRT) for stage I non-small cell lung cancer (NSCLC) are difficult to distinguish from local recurrences (LR), hampering proper patient selection for salvage therapy. This study validates previously reported high-risk CT features (HRFs) for detection of LR in an independent patient cohort.

Methods and Materials
From a multicenter database 13 biopsy proven LR were matched 1:2 to 26 non-LR controls based on: 1) dose 2) PTV 3) follow-up time 4) lung lobe. Tested HRFs were: enlarging opacity, sequential enlarging opacity, enlarging opacity after 12 months, bulging margin, linear margin disappearance, loss of air bronchogram and cranio-caudal growth. Additionally two new features were analyzed: the occurrence of new unilateral pleural effusion and growth based on relative volume, assessed by manual delineation.

Results
All HRFs were significantly associated with LR, except for loss of air bronchogram. The best performing HRFs were bulging margin, linear margin disappearance and cranio-caudal growth. ROC analysis of the number of HRFs to detect LR had an area under the curve (AUC) of 0.97 (95%CI 0.9-1.0), which was identical to the performance of the original paper. The best compromise (closest to 100% sensitivity and specificity) was found at ≥ 4 HRFs with a sensitivity of 92% and a specificity of 85%. A model consisting of only two HRFs, bulging margin and cranio-caudal growth, resulted in a sensitivity of 85% and a specificity of 100% with an AUC of 0.96 (95%CI 0.9-1.0) (HRFs ≥ 2). Pleural effusion and relative growth did not significantly improve the model.

Conclusion
We successfully validated CT based HRFs for detection of LR after SBRT for early stage NSCLC. As an alternative to number of HRFs, we propose a simplified model with the combination of the two best HRFs: bulging margin and cranio-caudal growth, although validation is warranted.
INTRODUCTION

Lung cancer is one of the most common cancers and among the leading cause of cancer deaths [1]. The majority of patients diagnosed with non-small cell lung cancer (NSCLC) present with an advanced stage at diagnosis. However, early stage NSCLC is likely to increase as screening of high-risk patients with low-dose computed tomography (CT) is being used more frequently [2,3]. The historical gold standard treatment for early stage NSCLC has been surgery, more precisely lobectomy and hilar and mediastinal lymph node dissection. Stereotactic body radiation therapy (SBRT, also referred to as SABR) has become the treatment of choice for medically inoperable patients with early stage NSCLC and led to a decrease of untreated elderly patients [4]. Several phase II trials have shown excellent local control rates after SBRT comparable to those after surgery in patients that were medically inoperable or had refused surgery [5-8]. Further analyses in the operable patient cohort encouraged the implementation of phase III trials comparing SBRT to surgery in early stage NSCLC. Two independent randomized phase III trials closed early because of slow accrual (STARS and ROSEL), but allowed for a pooled analysis providing the highest evidence so far of non-inferiority of SBRT compared to surgery in stage I NSCLC [9]. With more patients in good general condition becoming eligible for SBRT, accurate follow-up (FU) to detect local failures is crucial as these patients are eligible for salvage treatment [10-13].

The major challenge in detection of local recurrence (LR) after SBRT is the distinction of LR from benign ‘mass like’ fibrosis that occurs in the majority of patients up to 12 months of FU [14-16]. This may lead to unnecessary examinations (FDG-PET), diagnostic interventions (biopsies) and even futile resections [17]. Reliable features to differentiate LR from fibrosis allow for the proper selection of salvage therapy and prevent unnecessary harm to the patient. Only a few studies have identified high-risk CT features for LR, based on a limited number of events [18-20]. Huang et al. proposed seven CT-based high-risk features (HRFs) of LR in a matched control analysis of 12 patients with pathology-proven stage I NSCLC treated with SBRT [21,22]. The presence of ≥ 3 HRFs was highly sensitive and specific for LR (both 92%) and was deemed appropriate to indicate high suspicion of LR and to proceed to biopsy or salvage treatment.

The aim of this study was to validate the CT-based HRFs investigated by Huang et al. in an independent patient cohort. In a second step we tried to further improve the prediction model.

METHODS AND MATERIALS

Patient data was used from a multicenter database of patients treated with SBRT for early stage NSCLC between 2006 and 2013. In total 850 patients were treated at five institutions
internationally (Netherlands Cancer Institute, Amsterdam, The Netherlands; William Beaumont Hospital, Royal Oak, Michigan, USA; Princess Margaret Hospital, University of Toronto, Toronto, Ontario, Canada; University of Wuerzburg, Wuerzburg, Germany and Thomas Jefferson University Hospital, Philadelphia, USA). Target volume definition, treatment planning and delivery have been described in detail before [23,24]. All patients underwent routine FU including CT scans of the chest every 3-6 months for the first 2 years and annually thereafter. In case of suspect findings CT scans were obtained more frequently and in some cases FDG PET-CT was performed. In the database 53 patients were identified as having a LR, of which 13 were biopsy-proven. The biopsy proven LRs were matched to patients without LR in a 1:2 proportion, using the following procedure. Each potential matched control patient from the database was assigned a score. This score was a sum of scores based on several weighted selection criteria. The selection criteria, in order of highest weight with the proportion of maximum points in brackets, were: 1) same prescription dose (29%) 2) FU time (29%): time to LR was matched with FU time of non-LR with at least the same FU 3) same number of fractions (15%) 4) PTV ±5% (12%) 5) PTV ±10% (9%) 6) lung lobe (6%). From the list of matched controls with equal scores, two patients were chosen manually based on the most favorable match of PTV and FU time.

The presence of seven HRFs was evaluated in every FU CT [22]: enlarging opacity, sequential enlarging opacity, enlarging opacity after 12 months, bulging margin (curved, round, convex margin in contrast with linear, straight-lined, concave), linear margin disappearance, loss of air bronchogram and cranio-caudal growth. Two additional features were tested: the occurrence of new unilateral pleural effusion and volume increase [18]. The latter was based on manual delineation by one radiation oncologist of baseline tumor volumes and solid masses post treatment. Relative growth was determined by calculating the FU volume relative to baseline tumor volume. This ‘relative growth’ will be from now on referred to as HRF ‘growth’. Seven patients developed atelectasis during FU and were excluded from this analysis.

All CT images were assessed by two dedicated radiation oncologists blinded for treatment outcome by consensus method. For each patient anonymized planning CT scans, FU CT images and if available FDG PET-CTs, were displayed in sequential order with all window level settings and plane reconstructions available. Enlarging opacity (measured in axial plane) and cranio-caudal growth was assessed based on RECIST 1.1: 20% volume increase and an absolute increase of ≥ 5 mm [25]. Measurements were always compared with those of the prior CT-scan, i.e. FU 3 was compared with FU 2, not baseline or FU1. Maximum diameters and delineations were determined in mediastinal window setting to improve reproducibility. Pleural thickening was included if directly adjacent to the solid mass. Sequential enlarging opacity was only scored as HRF in case of an enlarging opacity two times in a row, i.e. not scored when the enlarging opacity was stable in between two scans.
Differences in occurrence of HRFs between LR and non-LR were calculated with Chi-square and Fisher’s exact test where appropriate. Sensitivity and specificity in predicting LR were assessed for each individual HRF and for each additional cumulative HRF using Fisher’s exact test. Investigation of difference in performance was done using area under the curve (AUC) analysis. Statistical analysis was performed using SPSS software (IBM version 22) and p-values <0.05 were considered statistically significant.

RESULTS

The study cohort consisted of 39 patients, of whom 13 had a biopsy proven LR and 26 matched controls without a LR. Patient, tumor and treatment characteristics are presented in Table 1, which shows a typical lung SBRT population of elderly patients with peripheral tumors with a median diameter of 2.1 cm predominantly located in the upper lobes. LR and non-LRs were well balanced with regard to patient, tumor and treatment characteristics. Median FU was 3.1 years (range 1.2-6.6 years) and median time to LR was 1.5 years (range 1.0-3.8 years). The median number of FU CT scans was five (range 2-11). Median FU of the matched controls was ten months longer than LR patients. Treatment plans were delivered using volumetric image guidance with different techniques: 3D-CRT, IMRT and VMAT [23]. Follow-up scans of two example patients are shown in Figure 1.

FDG PET at LR was performed in 7 patients. Median SUVmax values at baseline and LR were 2.9 (1.1-19) and 7.6 (4.4-17.1). Due to the limited number of events this was not incorporated in the prediction model.

All HRFs were significantly associated with LR, except for loss of air bronchogram (Table 2). Enlarging opacity, cranio-caudal growth and enlarging opacity after 12 months were scored most frequently in 79%, 67% and 64% of all patients respectively. In Figure 2 the ROC curve is displayed of all 7 HRFs with an AUC of 0.97 (95% CI 0.9-1.0). Since loss of air bronchogram was not significant in our analysis, we removed this HRF and repeated the analysis with six HRFs revealing the same outcome as 7 HRFs with an AUC of 0.97 (95% CI 0.9-1.0). To compare our results with Huang et al. we created an ROC curve based on the sensitivity and specificity given in their paper and found the same AUC of 0.97 [22].

To test what number of HRFs succeeded best in detecting LR, we identified the point on the ROC curve closest to the upper left corner of the ROC space representing 100% sensitivity and specificity. Four or more HRFs gave the best result with a sensitivity of 92% and a specificity of 85%.
Since not all HRFs perform equally well, we analyzed whether a combination of fewer HRFs performed similar. The best performing HRFs were: bulging margin and linear margin disappearance with a sensitivity of 85% and a specificity of 100%. These two HRFs were scored exactly identical in every patient, i.e., if there was a bulging margin, there was always linear margin disappearance and vice versa. Therefore, in our search for the best combination with a minimal set of HRFs we selected only bulging margin. Adding cranio-caudal growth to bulging margin gave the best result with an AUC of 0.96 (95%CI 0.9-1.0), a sensitivity of 85% and a specificity of 100% (≥ 2 HRFs). This was not significantly different from 6 HRFs.

In addition we tested two new features. Unilateral pleural effusion was scored in 13% of all patients and had a sensitivity of 31% and specificity of 96% (p=0.035). However, the addition of unilateral effusion to the 6 HRFs (nog significant HRF loss of air bronchogram removed) or the addition of pleural effusion to the two best performing HRFs did not improve the AUC.

<table>
<thead>
<tr>
<th>Table 1. Patient, tumor and treatment characteristics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>Age (y) - median (range)</td>
</tr>
<tr>
<td>Gender – male n (%)</td>
</tr>
<tr>
<td>PTV(cc) - median (range)</td>
</tr>
<tr>
<td>Max tumor diameter (cm) median (range)</td>
</tr>
<tr>
<td>Peripherally located – n (%)</td>
</tr>
<tr>
<td>Lobe – n (%)</td>
</tr>
<tr>
<td>RUL</td>
</tr>
<tr>
<td>RML</td>
</tr>
<tr>
<td>RLL</td>
</tr>
<tr>
<td>LUL</td>
</tr>
<tr>
<td>LLL</td>
</tr>
<tr>
<td>Treatment scheme – n (%)</td>
</tr>
<tr>
<td>3x18 Gy</td>
</tr>
<tr>
<td>5x12 Gy</td>
</tr>
<tr>
<td>4x12 Gy</td>
</tr>
<tr>
<td>8x7.5 Gy</td>
</tr>
<tr>
<td>FU time (y) - median (range)</td>
</tr>
<tr>
<td>No. of FU scans - median (range)</td>
</tr>
</tbody>
</table>

*Central based on RTOG 0236 [8]. PTV=planning target volume; RUL=right upper lobe; RML= right middle lobe; RLL=right lower lobe; LUL=left upper lobe; RLL=right lower lobe; FU=follow up.
A relative growth of 1.9 corresponded to a sensitivity of 85% and a specificity of 65% ($p=0.001$). The combination of growth with the 6 HRFs revealed an AUC of 0.99 (95% CI 0.96-1.00), a sensitivity of 92% and a specificity of 96%. Since four out of seven HRFs are based on growth (enlarging opacity, sequential enlarging opacity, enlarging opacity after 12 months, cranio-caudal growth), we tested whether HRF growth could replace these 4 HRFs. Growth combined with bulging margin and linear margin disappearance gave an AUC of 0.97 (95% CI 0.9-1.0), a sensitivity of 85% and a specificity of 100% (≥2 HRFs). Growth and bulging margin revealed an AUC of 0.96 with a slightly lower sensitivity of 77% and a specificity of 100% (≥2 HRFs). As detection of LR was not improved using relative growth, cranio-caudal growth and bulging margin remained the best combination.

**Table 2.** Sensitivity and specificity of the 7 high risk CT features reported by Huang *et al.* and 2 additionally investigated HRFs.

<table>
<thead>
<tr>
<th>High risk CT feature for local recurrence</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enlarging opacity (≥5mm and ≥20%)</td>
<td>100</td>
<td>31</td>
<td>0.035</td>
</tr>
<tr>
<td>Sequential enlarging opacity</td>
<td>62</td>
<td>77</td>
<td>0.033</td>
</tr>
<tr>
<td>Enlarging opacity after 12 months</td>
<td>92</td>
<td>50</td>
<td>0.013</td>
</tr>
<tr>
<td>Bulging margin</td>
<td>85</td>
<td>100</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Linear margin disappearance</td>
<td>85</td>
<td>100</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>Loss of air bronchogram</td>
<td>15</td>
<td>100</td>
<td>0.105</td>
</tr>
<tr>
<td>Cranio-caudal growth (≥5mm and ≥20%)</td>
<td>100</td>
<td>50</td>
<td>0.001</td>
</tr>
<tr>
<td>Unilateral pleural effusion</td>
<td>31</td>
<td>96</td>
<td>0.035</td>
</tr>
<tr>
<td>Relative growth</td>
<td>85</td>
<td>65</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Figure 1a. Local recurrence

- 14 months: Enlarging opacity, cranio-caudal growth
- 9 months: Bulging margin, linear margin disappearance, loss of air bronchogram, enlarging opacity > 12 months
- 5 months: Enlarging opacity, cranio-caudal growth
- 4 months: Sequential enlarging opacity
- Baseline: Local recurrence
Figure 1b. Non-local recurrence

Figure 1. CT images of patients post SBRT presented in axial and coronal view with scored HRFs. (a) Patient with a biopsy proven local recurrence at 14 months follow-up. After LR diagnosis a salvage pneumonectomy was performed. (b) Patient without a local recurrence. At 7 months follow-up benign mass like fibrosis or HRF enlarging opacity with cranio-caudal growth. From 23 months onwards fibrotic mass with a typical benign wedge linear shape (axial plane) without bulging. Thereafter stabilization of fibrosis.
DISCUSSION

In this independent patient cohort of 13 biopsy proven LR matched with 26 non-LR controls, we successfully validated the clinical applicability of CT based HRFs as predictors for LR reported by Huang et al. [22]. With the same AUC of 0.97, results were comparable, although Huang et al. achieved a sensitivity and specificity of 92% using ≥ 3 HRFs, whereas our results show the same sensitivity of 92%, but a slightly lower specificity of 85% using 1 extra HRF. We tested several other combinations of HRFs, all performing well without significant differences. However, using numbers of HRFs assumes that all HRFs predict LR on average equally well. As displayed in Table 2 this is not the case, whereas in Huang et al. almost all HRFs performed equally well except for linear margin disappearance (p=0.002 instead of p<0.001) [22]. Consequently we selected the best performing HRFs. Moreover, simpler models with fewer variables are easier to apply and in the situation of analyzing FU CT-scans less time consuming and cumbersome. We therefore propose the combination of only two HRFs: cranio-caudal growth and bulging margin, which resulted in a comparable AUC using all HRFs and a sensitivity of 85% and a specificity of 100%.

![ROC curve of number of high-risk CT features to detect local recurrence.](image)

Figure 2. ROC curve of number of high-risk CT features to detect local recurrence. Symbols represent sensitivity and specificity of: 1-7 HRFs (dot) ≥ 4 HRFs (diamond); ≥ 3 HRFs of Huang et al. (triangle), bulging margin and cranio-caudal growth (star).
We tried to improve detection of LR with a new feature ‘growth’ based on relative volume increase, supported by the fact that four out of seven HRFs reported by Huang et al. are based on growth derived from increasing diameters (enlarging opacity, sequential enlarging opacity, enlarging opacity after 12 months, cranio-caudal growth). Another aim was to replace these 4 HRFs by HRF growth. Using our cut-off of 1.9 (near doubling volume) in addition to the six HRFs improved the model slightly. Additionally, growth could replace the four HRFs mentioned above. However, it did not improve the detection of LR. In addition to the risk of overfitting the data and the fact that manual contouring is time consuming, we therefore do not recommend to use relative growth as HRF.

Unilateral pleural effusion has been proposed as HRF by Kato et al., but had not been tested in an independent data cohort thus far [18]. Although it was significant, it did not improve the prediction in addition to the 7 HRFs or the simplified model of bulging margin and cranio-caudal growth. As pleural effusion is also seen as an inflammatory response post SBRT [26], accompanying HRFs are of substantial importance to estimate the risk of a LR. In our cohort unilateral pleural effusion was detected in 5 patients, of which 3 had a LR. Those with a LR also had 5 to 6 other HRFs. Those without a LR had progressive disease with growing nodules in both lungs and the other developed an effusion 13 months prior to LR as the only HRF, which resolved spontaneously and was therefore regarded as reactive. Therefore, based on our data we suggest that pleural effusion without any other HRFs should not result in additional investigations.

Seven patients developed atelectasis during their FU, of which one had a LR. Two cases spontaneously resolved, one having a central tumor. Given the low incidence of central tumors in our cohort, we can only speculate about a causal relationship between central tumor location and atelectasis [27,28]. It is however noteworthy that in 5 patients with peripheral non-LR tumors (20%) atelectasis still developed. This was not correlated with dose or fractionation.

Limitations of this study are its retrospective design and small number of LRs. Numerous papers have described radiological changes over time and have warned for overestimation of tumor recurrence [20,29]. Thus including patients in the analysis that were diagnosed with a LR based on CT imaging only, would have overestimated the sensitivity and specificity of the assessed HRFs. Against the background of very high local control rates after lung SBRT for early stage NSCLC, the available number of events is low: in our large combined multicenter database 53 out of 850 patients were identified as LR of which 13 biopsy-proven. Moreover, the 1:2 proportion used in this paper is obviously not representative for daily practice, but for the feasibility of the study only 26 matches were selected. Still the sample size within this study is comparable to the report by Huang et al. (LR: non LR= 12: 24) that also derived their cases from a large institutional database, and represents the largest case number for...
the assessment of high-risk CT-features associated with LR to date. Another limitation of the study is that there might be an interobserver variability of investigators in determining HRFs, similar to scoring of benign features [30]. Furthermore, patients in the matched control group obviously were not pathologically confirmed. Matched patients were selected based on at least the same FU as the interval between treatment and LR. Although the median FU of the match patients was ten months longer, in individual cases FU was equal to LR or in two patients FU was six months shorter due to lack of a better match. It is possible that few of these patients develop a late LR in the future.

In our cohort enlarging opacity and cranio-caudal growth were always compared with the previous CT-scan, that were performed on a 3-6 monthly interval or more frequently if concerns. Too frequent scanning could potentially lead to miss of gradual growth. We analyzed the growth rate 1 year prior to LR (data not shown) and found that with 6 monthly intervals, growth was correctly identified in 92% and 73% of the aforementioned HRFs, in contrast with 42% and 45% at 3 months respectively. We therefore advise in case of more frequent FU than 6 monthly CT-scans, to not only compare with the previous, but also the preceding CT.

FDG-PET was performed in only seven of the 13 LR patients and was not routinely done in the matched control patients. Therefore performance of FDG-PET in predicting LR could not be investigated. The minimum SUVmax in our seven LRs was 4.4, which is consistent with literature [21]. Nakajima et al. looked at SUVmax values in combination with presence of mass like changes, which may decrease the false positive rate due to inflammatory changes [31]. Especially in case of moderate FDG uptake (SUVmax 2.5-5.0) clinical interpretation remains unclear and other PET tracers might be of use [32].

An alternative readily available in clinic is MRI, albeit patients receiving SBRT often have MR contraindications such as implanted devices. Although currently unexplored because of poor image quality due to moving tissue, technical improvements can (partially) overcome these limitations as shown in the liver [33].

Another strategy to detect LR is by using quantitative information of CT-imaging, e.g. HU-units changes or CT-texture changes over time [34,35]. Or in a more sophisticated way with radiomics, which is the quantification of tumor phenotypes by applying a large number of quantitative image features capturing tumor heterogeneity; tumor image intensity, shape and texture [36,37]. Although the first results are promising, large databases with high numbers of events are needed to create a solid predictive signature, which is unfortunately not yet available for lung SBRT. Given the low incidence of LR, combining of existing large databases seems the way to move forward with the final aim to detect LR at an earlier stage, e.g. during the first FU instead of watchful waiting using serial CTs.
Validation of high-risk CT features

In conclusion, with more patients being treated with SBRT for early stage NSCLC, the need for objective methods to identify LR is rising. Until the role of other imaging modalities and/or more automated image information analyses is clarified, the use of CT imaging based radiographic high-risk features is a helpful option with excellent sensitivity and specificity for the diagnosis of LR. We have validated the 7 CT based HRFs according to Huang et al. in an independent cohort with biopsy proven recurrences with excellent results. As an alternative to $\geq$ 3 HRFs we propose a simplified model with the two best predictors for LR: bulging margin and cranio-caudal growth. Validation of this new model in an external dataset is warranted.

Conflicts of interest
This research was partially supported by Elekta through a research grant with all institutions being members of the Elekta Lung Research Group. This work and these data, however, are the intellectual property of the individual group members and their sponsoring institutions.

Acknowledgements
This work is supported by a research grant from the Dutch Cancer Society (NKI 2009-4568). We thank Wilma Heemsbergen for her support with the statistical analysis.
REFERENCES


Validation of high-risk CT features