Respiration

Interventional Pulmonology

Respiration 2021;100:1088–1096
DOI: 10.1159/000516505

Received: November 16, 2020
Accepted: April 6, 2021
Published online: June 24, 2021

Accuracy and Reproducibility of Endoscopic Ultrasound B-Mode Features for Observer-Based Lymph Nodal Malignancy Prediction

Roel L.J. Verhoeven[a, b] Fausto Leoncini[c] Jorik Slotman[a, b] Chris de Korte[b] Rocco Trisolini[c] Erik H.F.M. van der Heijden[a] on behalf of the E-predict Study Group

[a]Department of Pulmonary Diseases, Radboud University Medical Center, Nijmegen, The Netherlands; [b]Department of Radiology, Medical Ultrasound Imaging Center (MUSIC), Radboud University Medical Center, Nijmegen, The Netherlands; [c]Interventional Pulmonology, Fondazione Policlinico Universitario A. Gemelli, Rome, Italy

Keywords

Bronchoscopy · Cancer · Endobronchial ultrasound · Endobronchial ultrasound-transbronchial needle aspiration · Endoscopic ultrasound · Esophageal ultrasound · Lung cancer staging · Nonsmall-cell lung cancer · Observer variability · Lymph node staging

Abstract

Background: Endoscopic ultrasound routinely guides lymph node evaluation for the staging of a known or suspected lung cancer. Characteristics seen on B-mode imaging might help the observer decide on the lymph nodes of risk. The influence of nodal size on the predictivity of these characteristics and the agreement with which operators can combine these for malignancy risk prediction is to be determined. ***Objectives:*** We evaluated (1) if prospectively scored individual B-mode ultrasound features predict malignancy when further divided by size and (2) assessed if observers were able to reproducibly agree on still lymph node image malignancy risk. ***Methods:*** Lymph nodes as visualized by EBUS were prospectively scored for B-mode characteristics. Still B-mode images were furthermore collected. After collection, a repeated scoring of a subset of lymph nodes was retrospectively performed (n = 11 observers). ***Results:*** Analysis of 490 lymph nodes revealed the short axis size is an objective measure for stratifying risk of malignancy (ROC area under the curve 0.78). With ≥8-mm size, 210/237 malignant lymph nodes were correctly identified (89% sensitivity, 46% specificity, 61% PPV, and 81% NPV). Secondary addition of B-mode features in <8-mm nodes had limited value. Retrospective analysis of intra- and interobserver scoring furthermore revealed significant disagreement. ***Conclusions:*** Lymph nodes of ≥8-mm size and preferably even smaller should be aspirated regardless of other B-mode features. Observer disagreement in scoring both small and large lymph nodes suggests it is infeasible to include subjective features for stratification. Future research should focus on (integrating) other (semi)quantitative values for improving prediction.

Introduction

Once a suspected or proven lung cancer with abnormal mediastinal findings has been found through CT and/or PET-CT imaging, guidelines recommend systematic

karger@karger.com
www.karger.com/res

Correspondence to:
Roel L.J. Verhoeven, roel.lj.verhoeven@radboudumc.nl

lymph nodal assessment and aspiration through endobronchial ultrasound (EBUS)-transbronchial needle aspiration and preferably combined with esophageal ultrasound-fine-needle aspiration [1]. Whilst a minimum of 3–4 lymph node aspirations is required by endoscopic examination if CT and/or PET-CT has shown abnormal findings (>10-mm short axis size on CT and/or FDG-PET avidity), also less-suspected findings on endoscopic imaging (>5-mm short axis findings on ultrasound) might subsequently incur ultrasound-guided aspiration [1–4]. In cases where the decision-making is not clear-cut, ultrasound imaging is often used to help decide on which lymph nodes to aspirate. Several studies have assessed which routinely available endoscopic ultrasound B-mode features help differentiate malignant and benign lymph nodes. The 5 identified and studied B-mode features include size, nodal heterogeneity, margin distinctiveness, presence of a central necrosis sign, and a central hilar structure [5]. As found by multiple studies, individual features or a combination thereof might have predictive value [5–13]. However, the results in these predominantly single-center studies differ. Consequently, no formal recommendation on the widespread use of these features has been given [14].

Possible reasons for the found differences in the prospective value of B-mode features might be differences in lymph node disease burden under study and/or a lack of consistent information in the features for enabling accurate prediction of malignancy. In clinical practice, one of the first determinants on the need of sampling therefore might remain to be the CT- and US-based lymph node short axis size. The influence of the lymph node size on scoring performance of the other identified B-mode features is however not well established. Some single-center studies report aspirating all lymph nodes of >5-mm short axis size [2, 5], whereas several multicentric studies report a ≥8-mm node as a lower margin for enabling aspiration regardless of the presence of any other features [15, 16]. It is however especially subcentimeter lymph nodes which are subjected to B-mode feature evaluation, with the endoscopist deciding intraprocedurally to proceed with sampling or not. The predictive value in doing so is however complicated not only by differences in reported performance across studies, but also a disagreement in interpretation of B-mode imaging and features could be a potential pitfall [8, 17]. As we have moved toward systematic sampling in endoscopic staging in order to prevent the need of more invasive cervical mediastinoscopy staging, one can question if subjective scoring of ≤8-mm lymph nodes remains desirable or if all lymph node regions should be aspirated regardless.

The aim of the present study is to assess the performance characteristics of the reported ultrasound features evaluated during endosonography in a multicenter international study when further specified by lymph node size. In this study, we therein specifically assess if the often used and clinically feasible 8-mm size cutoff could be further helped by B-mode features in deciding upon aspiration. To assess the reproducibility (inter- and intraobserver variability) of compounded endoscopic B-mode feature scoring for predicting lymph nodal malignancy, we furthermore performed a multiobserver retrospective scoring of a subset of lymph node images.

Materials and Methods

Study Subjects

This study was performed using prospectively collected and scored ultrasound B-mode images as gathered during the E-predict multicenter international trial, a study evaluating the value of ultrasound strain elastography for endosonographic prediction of lymph node malignancy (clinicaltrials.gov identifier: NCT02488928) [18]. All images were acquired and prospectively scored during endoscopic ultrasound-guided fine-needle aspiration procedures (EBUS/esophageal ultrasound) for a known or suspected lung cancer. For retrospective assessment of the reproducibility (inter- and intraobserver variability) of compounded endoscopic B-mode feature scoring in a subset of these prospectively collected images, several observers from different centers with varying experience were recruited on a voluntary basis.

Study Design

This study is 2-fold: (1) We prospectively evaluated performance characteristics of the endoscopist-reported ultrasound B-mode features in a multicenter multiobserver fashion. The performance characteristics were evaluated on all lymph nodes and in the subsets of <8-mm and ≥8-mm lymph nodes. (2) We investigated the observer variability in scoring the combination of B-mode features for predicting lymph node malignancy. By retrospective scoring of a subset of still lymph node images through multiple observers, intra- and interobserver variability is assessed. Knowing pathology outcome, accuracy of endoscopist malignancy scoring (as based on compounded B-mode features) is exploratively assessed.

Methods

1. Prospective collection of B-mode feature scoring was performed as part of the E-predict study protocol, which included 525 lymph nodes [18]. The investigators of the 5 recruiting centers prospectively scored features after careful examination on dynamic imaging: echogenicity (heterogeneous vs. homogeneous), shape (round vs. oval), margins (distinct vs. indistinct), coagulation necrosis sign (present vs. absent), and central hilar structure (present vs. absent) [18, 19]. Lymph node short axis size was determined by intraprocedural caliper measurements. During analysis, combined feature scores as described by Hylton et al. [8] were furthermore retrospectively computed for

Table 1. Performance of prospectively scored ultrasound features for predicting lymph node malignancy on the overall dataset of 490 lymph nodes (disease prevalence 0.48)

	Sens.	Spec.	PPV	NPV	Acc. (95% CI)	TN	TP	FN	FP
Overall dataset (n = 490, disease prevalence 0.48)									
Size ≥10 mm	0.75	0.64	0.66	0.73	0.70 (0.65–0.74)	163	178	59	90
Size ≥8 mm	0.89	0.46	0.61	0.81	0.67 (0.62–0.71)	116	210	27	137
Echogenicity	0.62	0.81	0.75	0.69	0.72 (0.67–0.76)	204	147	90	49
Shape	0.68	0.60	0.61	0.67	0.64 (0.59–0.68)	151	162	75	102
Margin	0.83	0.35	0.54	0.69	0.58 (0.54–0.63)	88	197	40	165
CHS	0.91	0.29	0.54	0.77	0.59 (0.54–0.63)	73	215	22	180
CNS	0.15	0.98	0.85	0.55	0.58 (0.53–0.62)	247	35	202	6
US Canada score >3	0.64	0.82	0.77	0.71	0.73 (0.69–0.77)	208	152	85	45
US Canada score >4	0.12	0.99	0.94	0.55	0.57 (0.62–0.52)	251	29	208	2
Mean strain <115	0.90	0.44	0.60	0.82	0.66 (0.62–0.70)	111	213	24	142

PPV, positive predictive value; NPV, negative predictive value; Acc., accuracy; CI, confidence interval; TN, true negative; TP, true positive; FN, false negative; FP, false positive; CHS, central hilar structure; CNS, central necrosis sign.

Table 2. Performance of prospectively scored ultrasound features for predicting lymph node malignancy in the subsets of data with lymph node short axis size <8 and ≥8 mm

	Sens.	Spec.	NPV	Acc. (95% CI)	TN	TP	FN	FP
Subset <8 mm (n = 143, disease prevalence 0.19)								
Echogenicity	0.37	0.84	0.85	0.75 (0.67–0.82)	97	10	17	19
Shape	0.44	0.78	0.86	0.71 (0.63–0.79)	90	12	15	26
Margin	0.67	0.34	0.82	0.41 (0.32–0.49)	40	18	9	76
CHS	0.81	0.28	0.86	0.38 (0.30–0.46)	32	22	5	84
CNS	0.07	1.00	0.82	0.83 (0.75–0.88)	116	2	25	0
US Canada score >3	0.07	1.00	0.82	0.83 (0.75–0.88)	116	2	25	0
Mean strain <115	0.78	0.47	0.90	0.53 (0.45–0.62)	55	21	6	61
Subset ≥8 mm (n = 347, disease prevalence 0.61)								
Echogenicity	0.65	0.78	0.59	0.70 (0.65–0.75)	107	137	73	30
Shape	0.71	0.45	0.50	0.61 (0.55–0.66)	61	150	60	76
Margin	0.85	0.35	0.61	0.65 (0.60–0.70)	48	179	31	89
CHS	0.92	0.30	0.71	0.67 (0.62–0.72)	41	193	17	96
CNS	0.16	0.96	0.43	0.47 (0.42–0.53)	131	33	177	6
US Canada score >3	0.71	0.67	0.61	0.70 (0.65–0.75)	92	150	60	45
US Canada score >4	0.14	0.99	0.43	0.47 (0.53–0.61)	135	29	181	2
Mean strain <115	0.91	0.41	0.76	0.71 (0.66–0.76)	56	192	18	81

PPV, positive predictive value; NPV, negative predictive value; Acc., accuracy; CI, confidence interval; TN, true negative; TP, true positive; FN, false negative; FP, false positive; CHS, central hilar structure; CNS, central necrosis sign.

obtaining the recently introduced Canada lymph node score. Individual lymph node cytopathology, subsequent surgical results, and/or 6-month clinical follow-up were used as reference standard.

2. For the retrospective assessment of variability in compounded endoscopic B-mode feature scoring for predicting lymph nodal malignancy, repeated scoring of a subset of the prospectively collected lymph nodes was performed. The subset included 200 lymph node images randomly selected from all prospectively included lymph nodes. The prevalence of disease was unknown to the observers. Only one single B-mode still image was presented, and participants were asked to classify it into "malignant" or "benign." Scoring was performed with in-house developed Mevislab software (MeVis Medical Solutions AG version

2.8.2, Bremen, Germany; Fig. 2). A size measurement tool was available. Metadata (i.e., node location and patient and image characteristics) were removed. The observers did not receive any feedback on their lymph node rating. Intraobserver scorings were performed with at least 1 day in between scorings and a new random image order.

Analysis

Analysis was performed with R and Rstudio [20]. To assess the influence of size on B-mode feature scoring, analysis was performed on the overall dataset and the subsets in which lymph node sizes were <8 mm and ≥8 mm. To assess the performance of B-mode features descriptive characteristics, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy along with counts are reported. For the continuous variable short axis size, the receiver operator characteristic-area under the curve (ROC-AUC) is furthermore presented.

For retrospective analysis on observer variability, a division into differently sized lymph nodes was also made (8-mm cutoff). Knowing there is considerable variation in individual observer learning curves in EBUS [21–23], we chose to further classify observer expertise into novice (<50 endoscopic US staging procedures), intermediate (50–400 procedures), and expert (>400 procedures). Raw agreement (%) and Gwet's agreement coefficient (AC1) – correcting for by chance agreement – were calculated to assess the agreement of measurements [24, 25]. An AC1 value of 1 would indicate perfect agreement, 0 would indicate at chance agreement, and <0 would indicate an agreement worse than as expected by chance. With no universal cutoff for determining good agreement but knowing its potential high impact, an AC1 minimum of 0.70 was considered desirable for interobserver research and an AC1 minimum of 0.85 for intraobserver agreement.

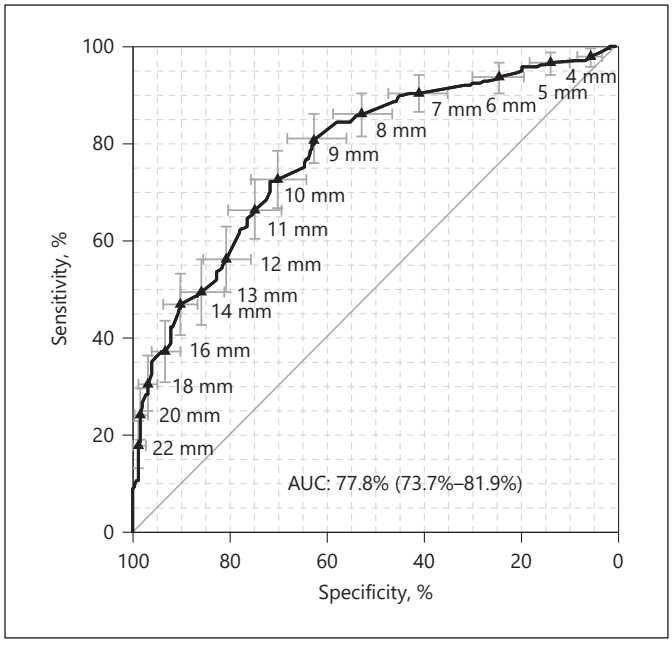

Fig. 1. ROC of ultrasound-measured lymph node short axis size; 95% confidence intervals of sensitivity and specificity at distinct short axis sizes are given in gray. ROC, receiver operator characteristic; AUC, area under the curve (95% confidence interval between brackets).

Results

The E-predict multicenter trial enrolled 525 lymph nodes from 327 study subjects between May 2016 and July 2018. For patient and lymph node characteristics, see [18]. For 490 lymph nodes (with 347 and 143 lymph nodes ≥8 mm and <8 mm, respectively), a full B-mode feature measurement was available for analysis. Overall prevalence of malignancy was 48%. Prevalence of malignancy in the ≥8-mm and <8-mm subsets was 61 and 19%, respectively. See Tables 1 and 2 for a summary on multicenter B-mode feature scoring performance and online suppl. Table E.1 (for all online suppl. material, see www.karger.com/doi/10.1159/000516505) for individual center performance.

B-Mode Size

Currently, the best made objective traditional B-mode feature is short axis size. In line with routine clinical use, it is of value in the first identification and estimation of lymph nodes at risk of malignancy. As depicted in Figure 1, ROC-AUC of short axis size was 0.778 (95% confidence interval: 0.737–0.819). The clinically often used 8-mm cutoff identified 210 out of 237 malignant lymph nodes and had 137 false positives (sensitivity 89%, specificity 46%, PPV 61%, and NPV 81%). A slightly larger 10-mm cutoff had higher overall accuracy but also identified only 178 out of 237 malignant lymph nodes. As a first measure for identification and subsequent deciding on aspiration, an 8-mm criterion is an objective feature with relatively high sensitivity and NPV (Tables 1–2; Fig. 1). A <8-mm size cutoff for including every identified lymph node would however increase the sensitivity even further, although at a significant cost in specificity (Fig. 1).

Descriptive B-Mode Features

In our multicenter scoring, several B-mode features initially show predictive value in the overall dataset (Table 1). When however further looking into individual center outcomes, large differences in predictive value of several B-mode features were seen between centers (online suppl. Table E.1). Maximal sensitivity differences between centers were >20% for features echogenicity, shape, and margin (overall sensitivity being 62, 68, and 83%, re-

Table 3. Observer variability in retrospective scoring of a random subset of lymph nodes

	All lymph nodes			Malignant lymph nodes			Benign lymph nodes		
	all	<8 mm	≥8 mm	all	<8 mm	≥8 mm	all	<8 mm	≥8 mm
Lymph nodes, *n*	200	69	131	99	15	84	101	54	47
Interobserver									
Expert (*n* = 5)	74/0.48	71/0.54	76/0.59	79/0.66	77/0.59	80/0.70	69/0.40	69/0.53	68/0.39
Interm. (*n* = 3)	71/0.42	73/0.57	69/0.47	76/0.58	78/0.57	75/0.61	66/0.35	72/0.58	59/0.21
Novice (*n* = 3)	58/0.17	63/0.28	56/0.12	58/0.17	64/0.32	57/0.15	58/0.17	63/0.28	52/0.08
Intraobserver									
Expert 1	86/0.76	93/0.92	82/0.65	82/0.64	80/0.76	82/0.65	90/0.87	96/0.96	83/0.72
Expert 2	86/0.72	84/0.78	87/0.79	92/0.87	93/0.88	92/0.88	80/0.65	81/0.77	79/0.58
Interm. 1	80/0.60	78/0.69	80/0.69	84/0.72	87/0.76	83/0.74	75/0.52	76/0.68	74/0.57

Observer variability when only looking at malignant/benign lymph nodes and sizes <8 and ≥8 mm are presented. The raw percent agreement is first presented, followed by the Gwet's AC1 coefficient.

spectively). A maximal specificity difference of 24.7 and 66.2% was furthermore seen in the features shape and margin, respectively (overall specificities 60 and 35%; online suppl. Table E.1). The features central hilar structure and a central necrosis sign had more similarity in outcome across centers. The central necrosis sign feature is not often found present in the different centers, but if so, a high risk of malignancy is warranted (specificities in all but one center: 91.7–100%, overall: 98%). The central hilar structure showed most conformity over all centers of classically used B-mode features other than size. With overall sensitivity 91%, specificity 29%, PPV 54%, and NPV 77%, only 22 out of 237 malignant lymph nodes would have gone unnoticed in this dataset. Oppositely, with specificity 29% and PPV 54%, it has a high number of false positives (180/253). The retrospectively calculated Canada lymph node score, recombining multiple B-mode features, showed relatively high accuracy (73%). Its clinical value in deciding on aspiration however seemed limited in our dataset, missing 85 out of 237 malignant nodes (sensitivity 64%).

Clinical Decision-Making Workup

Considering the scenario where all lymph nodes ≥8 mm should be aspirated regardless of other imaging features – as it is the firstly available and the most objective feature enabling a risk of prediction – further analysis of B-mode features for deciding on aspiration as specified by size was deemed necessary (Table 2). In the subset of lymph nodes <8 mm, prevalence of malignancy was 19% (*n* = 143). With the exception of "margin" and "central hilar structure," all sensitivities are below 50%, indicating more than half of malignancies go unnoticed. The central hilar structure had the most chance of identifying lymph node malignancy in <8-mm nodes with sensitivity 81% and NPV 86% (identifying 22 out of 27 lymph nodes). It however also resulted in a high number of false positives (84 out of 116 benign lymph nodes; Table 1). The ultrasound Canada score again had high overall accuracy, but considering disease prevalence did not prove useful in further workup on malignancy (sensitivity 7%).

In the subset of lymph nodes ≥8 mm, the prevalence of malignancy was 61% (*n* = 347). With this high overall risk of malignancy in this subset, the need of aspiration regardless of features seems warranted. Yet, if a further stratification needs to be made, the central hilar structure proved to be more useful with high sensitivity (92%) and moderate NPV (72%). Overall accuracy of lymph nodal risk estimation was however the highest by echogenicity (72%) and a Canada score >3 (69%). Yet, both echogenicity and the Canada score do not seem accurate enough for reliable clinical decision-making, as they are neither sensitive nor specific enough in excluding or including malignant and benign findings with values <78% (Table 2; online suppl. Table E1).

Compounded B-Mode Scoring – Observer Variability

For retrospective observer scoring, 5 experts, 3 intermediate observers, and 3 novices were included. Two experts and 1 intermediate observer performed scoring twice to enable intraobserver variability assessment (Table 3). All observers scored the dataset of 200 lymph nodes within 20 min.

Table 4. Explorative observer prediction accuracy when asked to classify retrospective lymph node B-mode ultrasound images as either malignant or benign based on a compounding of visible B-mode features

	Accuracy			Sensitivity			Specificity			PPV			NPV		
	all	<8 mm	≥8 mm	all	<8 mm	≥8 mm	all	<8 mm	≥8 mm	all	<8 mm	≥8 mm	all	<8 mm	≥8 mm
Experts	0.67	0.69	0.65	0.73	0.35	0.79	0.61	0.79	0.40	0.66	0.39	0.71	0.70	0.81	0.51
Expert 1	0.71	0.81	0.65	0.55	0.20	0.61	0.86	0.98	0.72	0.79	0.75	0.80	0.66	0.82	0.51
Expert 2	0.71	0.71	0.70	0.77	0.33	0.85	0.64	0.81	0.45	0.68	0.33	0.73	0.74	0.81	0.62
Expert 3	0.63	0.61	0.63	0.78	0.47	0.83	0.48	0.65	0.28	0.59	0.27	0.67	0.69	0.81	0.48
Expert 4	0.66	0.67	0.65	0.79	0.40	0.86	0.52	0.74	0.28	0.62	0.30	0.68	0.72	0.82	0.52
Expert 5	0.64	0.67	0.62	0.75	0.33	0.82	0.52	0.76	0.26	0.61	0.28	0.66	0.68	0.80	0.44
Intermediates	0.66	0.72	0.62	0.70	0.42	0.75	0.61	0.80	0.40	0.65	0.37	0.70	0.68	0.83	0.44
Interm. 1	0.65	0.71	0.62	0.78	0.40	0.85	0.52	0.80	0.21	0.62	0.35	0.66	0.71	0.83	0.43
Interm. 2	0.60	0.70	0.55	0.67	0.40	0.71	0.53	0.78	0.26	0.58	0.33	0.63	0.62	0.82	0.33
Interm. 3	0.72	0.75	0.70	0.66	0.47	0.69	0.78	0.83	0.72	0.75	0.44	0.82	0.70	0.85	0.57
Novices	0.51	0.54	0.49	0.51	0.40	0.53	0.50	0.57	0.43	0.50	0.21	0.62	0.52	0.77	0.34
Novice 1	0.58	0.57	0.59	0.61	0.40	0.64	0.55	0.61	0.49	0.57	0.22	0.69	0.59	0.79	0.43
Novice 2	0.47	0.52	0.44	0.45	0.40	0.46	0.48	0.56	0.38	0.46	0.20	0.57	0.47	0.77	0.29
Novice 3	0.48	0.52	0.46	0.47	0.40	0.49	0.49	0.56	0.40	0.47	0.20	0.59	0.49	0.77	0.31

PPV, positive predictive value; NPV, negative predictive value.

Interobserver Variability

In the overall dataset, the experts ($n = 5$) had the highest agreement, with Gwet's AC1 0.48 (raw agreement 74%; Table 3). Intermediate ($n = 3$) AC1 was 0.42 (raw agreement 71%). The novices ($n = 3$) showed an AC1 of only 0.17 (raw agreement 58%). Division of the overall dataset into subsets with <8- and ≥8-mm size did not change raw agreement >5% as compared to the overall dataset.

Analysis of observer variability against lymph node pathology shows experts and intermediates have more disagreement in scoring benign lymph nodes than malignant nodes (Table 3). Intermediate and expert observers had an AC1 of 0.58–0.66 (76–79% raw agreement) in malignant lymph nodes and an AC1 of 0.35–0.40 (66–69% raw agreement) in benign lymph nodes. These findings remained after differentiating by size.

Intraobserver Variability

The experts showed moderate intraobserver agreement in the overall dataset, with an AC1 of 0.72–0.76 (raw agreements 86%) while the intermediate had an AC1 of 0.60 (raw agreement 80%). The Gwet's AC1 in intraobserver scoring of <8-mm lymph nodes varied from 0.69 to 0.92 (raw agreement 78–93%), whilst the >8-mm lymph nodes had an AC1 of 0.65–0.79 (80–87% raw agreement).

Observer Accuracy

The observer variability scoring on still images was used to explore B-mode compounding accuracy. Accuracy of predicting lymph node malignancy based on still B-mode images varied from 0.47–0.58 in novice to 0.60–0.72 and 0.63–0.71 in intermediate and expert, respectively. A division of the dataset by an 8-mm size cutoff did not show unequivocal changes in overall accuracy (Table 4). Further evaluation of compound scoring shows low overall sensitivity and PPV in <8-mm lymph nodes (both <0.42), indicating difficulty in identifying malignant nodes in this cohort. Oppositely, a low specificity and NPV in ≥8-mm lymph nodes in the majority of observers indicates that benign nodes are often not identified in the cohort of enlarged lymph nodes.

Discussion

This study evaluates whether individual B-mode features and a compounding thereof can be used to accurately and reproducibly predict lymph node malignancy in a dataset of 490 aspirated lymph nodes for which follow-up was available. While we find that the classically first available ultrasound short axis size is the most objective feature with reasonable predictive value (AUC-ROC 0.78), we also show that not one additional nor a combination of B-mode features is a do-it-all solution for predicting lymph node ma-

lignancy. Based on our findings and incorporating the purpose of endoscopic ultrasound-guided needle aspiration as a systematic staging procedure in suspected lung cancer, we conclude that preferably all assessed lymph node regions should be aspirated regardless of size or features. At least, all lymph nodes ≥8 mm should be subject to sampling regardless of other features due to a high prevalence of malignancy in this subgroup of nodes, but preferably smaller. In <8-mm lymph nodes, a central hilar structure (sensitivity 81% and NPV 86%) could also be used to stratify risk of malignancy, but the low specificity (28%) and considerable observer interpretation differences in feature scoring suggest it is of suboptimal outcome.

Our retrospective analysis of observer variability using still images as a representation of compounded B-mode features shows that malignancy risk estimation cannot reliably be performed on still B-mode images. We find a relevant inter- and intraobserver agreement difference. Furthermore, while exploratory, we find compounded feature scoring had an accuracy of <72%. We therefore recommend not to decide on risk of malignancy or further workup by compounding of only B-mode features.

Varying results have been reported on endoscopic B-mode lymph node imaging scoring agreement across different observers and within observers. Schmid-Bindert et al. [17] retrospectively studied B-mode feature scoring through blinded procedural videos and reported interobserver agreements of 88.6% and higher ($n = 2$). Hylton et al. [8] also had a database of videos and in multiple observers ($n = 12$) reported a raw agreement different for the individual features; ranging from 62.6% for echogenicity to 81.7% for the central necrosis sign (Gwet's AC1 0.25–0.77). Our retrospective study determined the reproducibility of endoscopist B-mode feature compounding for predicting malignancy rather than individual B-mode feature scoring. In that regard, our study design was dissimilar to the aforementioned. In studying compounding, we find there is considerable interobserver variability, leading to 20% and higher expert raw disagreement over all lymph node subsets. Intraobserver variability is a further concern. Intraobserver agreement of the experts and intermediate was ≤86% in the overall dataset when presented with the exact same image (AC1 0.72–0.76). Observer compounding of B-mode features is subject to significant variability. However, knowing assessment of stationary images is far from ideal and irrevocable, a 0.60–0.72 accuracy on predicting nodal malignancy by experts and intermediates also implicates prediction accuracy was better than at chance agreement. Analysis however shows a low sensitivity and PPV for prediction in the subset of lymph nodes <8-mm size. Unfortunately, this subset is the most important for further workup risk stratification and preventing understaging.

The currently available B-mode features do not provide sufficient information for a highly accurate and reproducible prediction of lymph node malignancy. As we recently reported and show here, strain elastography may enable a semiquantitative way of predicting malignancy and increasing sensitivity and NPV better than B-mode characteristics scoring [18]. Future research should be conducted, focusing on implementation of computer-aided diagnosis systems which could enable more consistent outcomes and possibly more accurate predictions. Integration of (other) (semi)quantitative ultrasound features not easily objectified by the endoscopist should furthermore be investigated [26], along with the possibility of combining it with multiple modalities (such as strain elastography and/or PET-CT/CT findings [18, 27]). Such systems may further allow tailoring to individual anatomical lymph node regions, that is, weighing factors such as size or a presence of hilar structure differently.

Limitations

The prospectively collected B-mode images and the scoring thereof by the observers were done in routine clinical practice during a multicenter trial on strain elastography. As systematic study inclusion and scoring of all assessed nodes was not mandatory, inclusion of lymph nodes could have been biased. A second limitation is the retrospective nature of our analysis on reproducibility and having only still B-mode ultrasound images available for scoring. Last, retrospective scoring was performed without taking notice of metadata (e.g., lymph node location and PET/CT information). This could have affected exploratively assessed scoring accuracy but is not expected to negatively affect scoring variability.

Conclusion

Aside from ultrasound B-mode short axis size, the currently used B-mode features do not provide sufficient information for good prediction of lymph nodal malignancy, and reproducibility in their scoring shows an issue. In systematic staging, preferably all lymph nodes should be sampled. If any decision-making on which nodes need to be aspirated is made aside from FDG-PET findings, the objective short axis size seems a relevant predictor. A ≥8-mm size has shown to be a clinically feasible cutoff and could be used as a starting point. In systematic sampling, even a smaller lymph

node should however be considered. Future research should focus on implementation of computer-aided diagnosis systems and further integration of multimodality information for possibly improving nodal malignancy prediction.

Acknowledgment

We gratefully acknowledge our E-predict study team Piero Candoli, Michela Bezzi, Alessandro Messi, Mark Krasnik, and Jouke Annema for contributing to the collection of the dataset used for this study.

Statement of Ethics

The research was conducted ethically in accordance with the latest World Medical Association Declaration of Helsinki. The E-predict study from which the here presented data were obtained is registered and can be found on ClinicalTrials.gov (identifier: NCT02488928). This study was approved by the medical ethics committees in both the Netherlands and Italy, where subject inclusion took place after having obtained informed consent.

Conflict of Interest Statement

Roel L.J. Verhoeven reports grants from Pentax Medical Europe, during the conduct of the study, and grants from AstraZeneca Oncology, grants from Philips Medical, personal fees and nonfinancial support from Medtronic, and grants from Ankie Hak Fund, outside the submitted work. Rocco Trisolini, Fausto Leoncini, and Chris L. de Korte report nothing to disclose. Erik H.F.M. van der Heijden reports grants and personal fees from Pentax Medical Europe, during the conduct of the study, and grants from AstraZeneca Oncology, grants and nonfinancial support from Philips Medical, personal fees and nonfinancial support from Medtronic, and personal fees from Cook Medical, outside the submitted work.

Funding Sources

This work was kindly supported by unrestricted research grants from the Ankie Hak fund, Astra Zeneca Oncology Netherlands, and Pentax Medical Europe.

Author Contributions

R.V., J.S., and E.V.D.H. had full access to all the data in the study, drafted the first version of the manuscript, and take responsibility for the integrity of the data, collection of data, and the accuracy of the data analysis, including and especially any adverse effects. F.L. and R.T. substantially contributed to data collection, analysis, interpretation, and revision of the manuscript. C.K. contributed substantially to study design, interpretation, and revision of the manuscript.

References

1 Vilmann P, Clementsen PF, Colella S, Siemsen M, De Leyn P, Dumonceau JM, et al. Combined endobronchial and oesophageal endosonography for the diagnosis and staging of lung cancer. European Society of Gastrointestinal Endoscopy (ESGE) Guideline, in cooperation with the European Respiratory Society (ERS) and the European Society of Thoracic Surgeons (ESTS). Eur Respir J. 2015 Jul;46(1):40–60.

2 Herth FJ, Eberhardt R, Krasnik M, Ernst A. Endobronchial ultrasound-guided transbronchial needle aspiration of lymph nodes in the radiologically and positron emission tomography-normal mediastinum in patients with lung cancer. Chest. 2008;133(4):887–91.

3 Van Der Heijden EHFM, Casal F, Trisolini R, Steinfort P. Guideline for the acquisition and preparation of conventional and endobronchial ultrasound-guided transbronchial needle aspiration specimens for the diagnosis and molecular testing of patients with known or suspected lung cancer. Respiration. 2014; 88(6):500–17.

4 De Leyn P, Dooms C, Kuzdzal J, Lardinois D, Passlick B, Rami-Porta R, et al. Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer. Eur J Cardiothorac Surg. 2014 May; 45:787.

5 Fujiwara T, Yasufuku K, Nakajima T, Chiyo M, Yoshida S, Suzuki M, et al. The utility of sonographic features during endobronchial ultrasound-guided transbronchial needle aspiration for lymph node staging in patients with lung cancer: a standard endobronchial ultrasound image classification system. Chest. 2010;138(3):641–7.

6 Hylton DA, Turner J, Shargall Y, Finley C, Agzarian J, Yasufuku K, et al. Ultrasonographic characteristics of lymph nodes as predictors of malignancy during endobronchial ultrasound (EBUS): a systematic review. Lung Cancer. 2018;126:97–105.

7 Evison M, Morris J, Martin J, Shah R, Barber PV, Booton R, et al. Nodal staging in lung cancer: a risk stratification model for lymph nodes classified as negative by EBUS-TBNA. J Thorac Oncol. 2015;10(1):126–33.

8 Hylton DA, Turner S, Kidane B, Spicer J, Xie F, Farrokhyar F, et al. The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. J Thorac Cardiovasc Surg. 2020;159:2499.

9 Wang Memoli JS, El-Bayoumi E, Pastis NJ, Tanner NT, Gomez M, Huggins JT, et al. Using endobronchial ultrasound features to predict lymph node metastasis in patients with lung cancer. Chest. 2011;140(6):1550–6.

10 Satterwhite LG, Berkowitz DM, Parks CS, Bechara RI. Central intranodal vessels to predict cytology during endobronchial ultrasound transbronchial needle aspiration. J Bronchol Interv Pulmonol. 2011;18(4):322–8.

11 Alici IO, Yılmaz Demirci N, Yılmaz A, Karakaya J, Özaydın E. The sonographic features of malignant mediastinal lymph nodes and a proposal for an algorithmic approach for sampling during endobronchial ultrasound. Clin Respir J. 2016;10(5):606–13.

12 Shafiek H, Fiorentino F, Peralta AD, Serra E, Esteban B, Martinez R, et al. Real-time prediction of mediastinal lymph node malignancy by endobronchial ultrasound. Arch Bronconeumol. 2014;50(6):228–34.
13 Schmid-Bindert G, Jiang H, Kähler G, Saur J, Henzler T, Wang H, et al. Predicting malignancy in mediastinal lymph nodes by endobronchial ultrasound: a new ultrasound scoring system. Respirology. 2012;17(8):1190–8.
14 Wahidi MM, Herth F, Yasufuku K, Shepherd RW, Yarmus L, Chawla M, et al. Technical aspects of endobronchial ultrasound-guided transbronchial needle aspiration: CHEST guideline and expert panel report. Chest. 2016;149(3):816–35.
15 Bousema JE, Dijkgraaf M, Papen-Botterhuis EN, Schreurs H, Maessen J, Van Der Heijden EHFM, et al. MEDIASTinal staging of non-small cell lung cancer by endobronchial and endoscopic ultrasonography with or without additional surgical mediastinoscopy (MEDIASTrial): study protocol of a multicenter randomised controlled trial. BMC Surg. 2018;18(1):27.
16 Crombag LMM, Dooms C, Stigt JA, Tournoy KG, Schuurbiers OCJ, Ninaber MK, et al. Systematic and combined endosonographic staging of lung cancer (SCORE study). Eur Respir J. 2019;53(2):1800800.
17 Schmid-Bindert G, Jiang H, Kähler G, Saur J, Henzler T, Wang H, et al. Predicting malignancy in mediastinal lymph nodes by endobronchial ultrasound: a new ultrasound scoring system. Respirology. 2012;17(8):1190–8.
18 Verhoeven RLJ, Trisolini R, Leoncini F, Candoli P, Bezzi M, Messi A, et al. Predictive value of endobronchial ultrasound strain elastography in mediastinal lymph node staging: the E-predict multicenter study results. Respiration. 2020;99(6):484–92.
19 Fujiwara T, Yasufuku K, Nakajima T, Chiyo M, Yoshida S, Suzuki M, et al. The utility of sonographic features during endobronchial ultrasound-guided transbronchial needle aspiration for lymph node staging in patients with lung cancer: a standard endobronchial ultrasound image classification system. Chest. 2010;138(3):641–7.
20 R Core Team. R: A language and environment for statistical computing; 2019. Available from: https://www.r-project.org.
21 Kemp SV, El Batrawy SH, Harrison RN, Skwarski K, Munavvar M, Rosell A, et al. Learning curves for endobronchial ultrasound using cusum analysis. Thorax. 2010;65(6):534–8.
22 Wahidi MM, Hulett C, Pastis N, Shepherd RW, Shofer SL, Mahmood K, et al. Learning experience of linear endobronchial ultrasound among pulmonary trainees. Chest. 2014;145(3):574–8.
23 Sehgal IS, Dhooria S, Aggarwal AN, Agarwal R. Training and proficiency in endobronchial ultrasound-guided transbronchial needle aspiration: a systematic review. Respirology. 2017;22(8):1547–57.
24 Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(1):29–48.
25 Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. the problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9.
26 Nguyen P, Bashirzadeh F, Hundloe J, Salvado O, Dowson N, Ware R, et al. Optical differentiation between malignant and benign lymphadenopathy by grey scale texture analysis of endobronchial ultrasound convex probe images. Chest. 2012;141(3):709–15.
27 Verhoeven RLJ, De Korte CL, Van Der Heijden EHFM. Optimal endobronchial ultrasound strain elastography assessment strategy: an explorative study. Respiration. 2019;97(4):337–47.