The Application of Artificial Intelligence and Elastography to EBUS-TBNA Imaging Technology for the Prediction of Lymph Node Malignancy

McMaster University
Health Research Methodology
MSc Thesis

Nikkita Mistry

Title: The Application of Artificial Intelligence and Elastography to EBUS-TBNA Imaging Technology for the Prediction of Lymph Node Malignancy

Author: Nikkita Mistry, BSc McMaster University

Supervisor: Dr. Wael C. Hanna MDCM,MBA

Committee Members: Dr. Forough Farrokhyar, MPhil, PhD, Dr. Feng Xie, MSc, PhD,

Lay Abstract:

Non-Small Cell Lung Cancer (NSCLC) treatment decisions are made using vital information by performing biopsies to collect tissue from the lymph nodes near the lungs. The current method is called Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA), which involves a scope with a fine needle attached to it. This scope is led down the airway and guided by ultrasound to obtain the tissue needed to determine whether the lymph nodes have cancerous tissue. If the lymph nodes contain cancerous tissue, the patient may need chemotherapy; however, lung surgery may be the best treatment option if they do not. Many factors impact how successfully these tissue samples can be obtained, such as the skill and experience of the surgeon. These factors often lead to inconclusive results, making it difficult to make correct treatment decisions. Novel technologies such as Artificial Intelligence and Elastography are being used to diagnose lung cancer by interpreting images and providing information on tissue stiffness. We trained an Artificial Intelligence program to predict malignancy based on EBUS-TBNA images. Additionally, we trained the AI program to analyze Elastography images to aid us in understanding the relationship between the colour pattern of the elastography images and cancerous tissue. This thesis assesses how these novel technologies contribute to lung cancer diagnosis.

Abstract

Background: Before making any treatment decisions for patients with non-small cell lung cancer (NSCLC), it is crucial to determine whether the cancer has spread to the mediastinal lymph nodes (LNs). The preferred method for mediastinal staging is Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA). However, EBUS-TBNA has been reported to generate inconclusive results as much as 40% of the time. Since this jeopardizes good patient care, there is near-universal consensus on the need to develop and study a novel method for LN staging. Recent research has shown that AI and deep learning are used to accurately interpret images with comparisons to clinicians in radiology, pathology, and cardiology. Additionally, EBUS-Elastography is a novel modality which could be used as an adjunct to EBUS-TBNA for LN staging. This technology uses impedance ultrasonography to measure tissue stiffness.

Methods: There are three parts to this thesis. The first part involved the training, validating, and testing NeuralSeg, a deep neural network, to predict LN malignancy based on B-mode EBUS-TBNA images. The second part of this thesis involves EBUS-Elastography, defining the blue colour threshold and the optimal SAR cut-off value based on the blue threshold that most accurately distinguished benign and malignant LN. Finally, this thesis's third part involves validating part II's findings.

Results: Part I resulted in an overall accuracy of 80.63% (76.93% to 83.97%), a sensitivity of 43.23% (35.30% to 51.41%), a specificity of 96.91% (94.54% to 98.45%), a positive predictive value of 85.90% (76.81% to 91.80%), a negative predictive value of 79.68% (77.34% to

81.83%), and an AUC of 0.701 (0.646 to 0.755). Part II Level 60 was chosen as the blue threshold with an AUC of 0.89 (95% CI: 0.77-1.00), and the optimal SAR cut off was found to be 0.4959 with a sensitivity of 92.30% (95% CI: 62.10% to 99.60%) and a specificity of 76.50% (95% CI: 49.80% to 92.20%). Using the blue threshold and SAR cut-off, the results of part III resulted in an overall accuracy of 70.59% (95% (CI) 63.50% to 77.01%), the sensitivity of 43.04% (CI: 31.94% to 54.67%), and a specificity of 90.74% (CI: 83.63% to 95.47%).

Conclusion: It was observed that both AI and AI-powered EBUS-Elastography achieved high specificities on larger sample sizes, indicative that these methods may be helpful in identifying LN malignancy. However, due to the novelty of these technologies, more extensive multi-centre studies must be conducted before these processes can be standardized.

Acknowledgements

Firstly, I would like to thank my parents, my brother Pratik Mistry, and my best friend Anand Rai, for their constant support and encouragement throughout my academic career, without whom this would not be possible.

I would like to thank my supervisor, Dr. Wael Hanna for motivating me to achieve more and reach my fullest potential. Also for his guidance, support, and advice throughout my MSc experience

I would like to thank my thesis committee: Dr. Forough Farrokhyar, Dr. Feng Xie, and my external reviewer, Dr. Marko Simunovic, for their feedback and encouragememt.

I would like to thank the members of our research lab, Jacob Alaichi and Yogita Patel, your support and feedback was greatly appreciated.

I would like to thank the nurses and hospital staff for being so kind and friendly while I consented patients. Most importantly, I would like to thank the patients for participating in these research studies, without whom this research would not be possible.

Table of Contents

Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Abbreviations

AI	Aritifical Intelligence
AUC	Area Under the Curve
CHS	Central Hilar Structure
CI	Confidence Interval
CLNS	Canada Lymph Node Score
CNN	Convolutional Neural Network
CT	Computed Tomography
DNN	Deep Neural Network
EBUS-TBNA	Endobronchial Ultrasound Transbronchial Needle Aspiration
LN	Lymph Node
NPV	Negative Predictive Value
NSCLC	Non-Small Cell Lung Cancer
PET	Positron Emission Tomography
PPV	Positive Predictive Value
ROC	Recieiving Operating Curve
SAR	Stiffness Area Ratio

Diagnostic Test Definition and Equations

Sensitivity: quantifies a diagnostic test's ability to correctly identify subjects with the condition.[1]

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives}$$

Specificity: quantifies a diagnostic test's ability to correctly identify subjects without the condition.[1]

$$Specificity = \frac{True\ negatives}{False\ positives + True\ negatives}$$

Positive Predicative Value: the probability that the disease is present given a positive test result.[1]

$$PPV = \frac{Sensitivity * \pi}{Sensitivity * \pi + (1 - specificity) * (1 - \pi)}$$

Negative Predicative Value: the probability that the disease is absent given a negative test result.[1]

$$NPV = \frac{Specificity * (1 - \pi)}{Specificity * (1 - \pi) + (1 - sensitivity) * \pi}$$

1. Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. *Proceedings of Singapore healthcare*, *20*(4), 316-318.

Chapter 1: Background

Nikkita Mistry, BSc, Forough Farrokhyar, MPhil, PhD, Feng Xie, MSc, PhD, Waël C. Hanna, MDCM, MBA

The Severity of Lung Cancer

Despite many advances in health care and cancer research, cancer remains Canada's leading cause of death.[1] Lung cancer is the most diagnosed cancer among Canadians; of the estimated 225 800 new cancer cases, 29 800 are expected to be lung cancer. Lung cancer accounts for the highest number of cancer deaths, making up 25.5% of deaths surpassing breast cancer and prostate cancer.[1] Lung cancer impacts males more than females, mainly due to the difference in tobacco smoking in the past.[1] The amount of weight loss experienced by patients is another prognostic factor that may impact what treatment decisions should be made.[2] However, one of the most crucial prognostic factors when it comes to lung cancer survival is the stage at diagnosis. Lung cancer is most alarming because late-stage cancer has a significantly lower survival rate than early stages. The survival rate for stage 1A cancer is 49% compared to stage 4, which is only 1%.[3] Although the importance of staging is known, not all practitioners undergo this crucial step. In wealthy countries such as Australia, Canada, Denmark, Norway, Sweden and the UK, 5-26% of patients had no microscopic verification of the cancer stage.[4] This can lead to incorrect treatment decisions, which can be detrimental to patient health outcomes.

Lung Cancer Staging Techniques

Lung cancer lymph node (LN) staging can be characterized into two groups: non-invasive staging and invasive staging. Non-invasive staging includes medical imaging such as computerized tomography (CT) scans and positron emission tomography (PET) scans.[5] Although these scans provide vital information on LN size and metabolic activity; tissue confirmation is

absent. After undergoing non-invasive staging, if the LNs appear enlarged or fluorodeoxyglucose (FDG) avid, further investigation using invasive techniques is required, as shown in Figure 1.[5.] There are 14 different LN stations in the thoracic regions, as depicted in figure 2. Station 1 includes the low cervical, supraclavicular and sternal notch nodes.[6] The upper zone includes stations 2R, 2L, 3a, 3p, 4R and 4L, which are the upper paratracheal right and left, prevascular, retrotracheal, lower paratracheal right and left, respectively. The AP zone includes station5, the subaortic and station 6 para-aortic lymph nodes. Station 7 is the subcarinal nodes; station 8 is the paraesophageal, and station 9 is the pulmonary ligament.[6] The Hilar zone comprises stations 10 and 11, the hilar and interlobar zones, respectively. Lastly, the peripheral zone consists of station 12, lobar, station 13, segmental, and station 14, subsegmental.[6]

Lung Cancer Diagnostic Pathway

The diagnostic pathway starts with the initial computerized tomography (CT) scan; then, it is recommended to conduct a positron emission tomography (PET)-CT scan. First, the intrathoracic lymph nodes (LN) are observed and biopsied. Next, patients would undergo a mediastinoscopy which involves surgical staging so the physician can obtain tissue samples of the mediastinal lymph nodes (LN). This procedure is highly invasive, with a sensitivity of approximately 78%.[7] In recent times, more minimally invasive procedures have gained popularity. The current practices involve endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) or endoscopic ultrasound fine-needle aspiration (EUS-FNA) are conducted to stage the LNs.[8] EBUS-TBNA and EUS-FNA are endoscopic techniques that are far more minimally invasive than surgical staging techniques like mediastinoscopy, with a combined

sensitivity of 93%.[7] Additionally, important nodal stations are accessible by endoscopic techniques.[9] Due to these reasons, there has been an increased use of EBUS-TBNA and EUS-FNA compared to mediastinoscopy.

EBUS-TBNA Procedure

EBUS-TBNA involves a bronchoscope equipped with a convex type ultrasound probe combined with a biopsy needle. This technique allows for a real-time guided ultrasound when retrieving biopsy samples from intrathoracic LNs.[5] EBUS-TBNA is a widely accepted procedure for nodal staging. This procedure can be safely performed on patients even with COPD comorbidities.[5] It is minimally invasive and can achieve a diagnostic yield as high as the one of mediastinoscopy. In addition, EBUS-TBNA has a pooled sensitivity of 88-93% and specificity of as high as 100% based on published data from systematic reviews.[5] For these reasons, EBUS-TBNA is advantageous for nodal staging for patients with suspected or diagnosed NSCLC.

Staging through the sampling of thoracic LNs is a crucial step in the diagnostic pathway as treatment decisions are heavily dependent on the cancer stage. Patients with Non-Small Cell Lung Cancer (NSCLC) at stage I or II can be offered lung resection surgeries and may be offered adjuvant treatments based on the pathological staging.[8] However, treatment for stages III and IV are more complex involving chemoradiation and guided by histology and predictive biomarkers.[8] Given the significant difference in survival rate based on the stage of lung cancer, this provides a clear rationale that better screening for lung cancer can allow for early detection, potentially allowing patients to better respond to treatment and possibly avoid death or severe illness.[3]

Although nodal staging has advanced from mediastinoscopy to minimally invasive procedures such as EBUS-TBNA with high sensitivity, as high as 42% of nodal biopsies, remain inconclusive.[10] The implications of inconclusive biopsy results include patients coming in for repeat biopsies and treatment delays. In some cases, incorrect treatment decisions can be made, as the information needed to make optimal treatment choices is missing.[10] This low yield may discourage physicians from nodal staging and, in some cases, may skip this crucial step altogether. A study conducted by the American College of Surgeons Patient Care Evaluation report conducted over a decade after 2001, with just under 3000 patients, reported that only 22% of patients received nodal staging.[11] The low rates of nodal staging can be observed in figure 3; although there appears to be an upward trend in the usage of staging procedures, the percentage is still below 50%.[11] Non-invasively staged patients whose imaging results indicated they needed nodal staging significantly decreased survival compared to those who did not. The impact staging has on survival outcomes reveals a clear indication of the importance of nodal staging.[11] Based on these reports, there is a clear consensus that current nodal staging practices must be improved. Novel technologies could potentially aid in improving EBUS-TBNA yields or provide information to improve the pre-test probability of LN malignancy.

Artificial Intelligence and Medical Imaging

Artificial Intelligence (AI) is a novel technology infiltrating the field of medical imaging. AI and deep neural networks (DNN) mimic human cognition and have been applied to medical images to analyze and interpret results.[12] Not only are deep learning algorithms able to analyze

results, but the more data it is exposed to, the more it improves in the analysis. Over the years, there has been a substantial advance in radiology. Cross-sectional imaging with CT scans provides more information a radiologist will need to make a diagnosis; however, these advancements result in thousands of images for multiple body parts.[12] This can be time-consuming to go through, and the amount of data available for both human interpretation and software analysis continues to increase. Not all aspects of radiology can be analyzed by AI. However, more straightforward tasks that a radiologist can do with ease, such as identifying a lung nodule, can be delegated by an AI algorithm.[12]

DNN consists of three layers, as seen in figure 4.[13] There is an input layer, which is the initial data for the network to analyze. Then there are the hidden layers; there could be anywhere between 5 and 1000 hidden layers, each responding to different features of the image. Lastly, the output layer produces the results from the given inputs.[13]DNNs have been applied successfully to medical images, such as CT scans with the objective of detecting pulmonary nodules. The area under the curve (AUC) for the receiver operating curve (ROC), which is used to assess diagnostic criteria, ranged from 0.92 to 0.99, which shows excellent diagnostic capacity. [14]

Our Lab Groups Previous Work

Our research group has explored the application of AI and DNNs to EBUS-TBNA in predicting mediastinal LN malignancy. However, before applying AI and DDNs, our group investigated which nodal features are predictive of malignancy. Four nodal features are

particularly useful when it comes to diagnosing malignancy in mediastinal LNs. These features are: small axis diameter (size), central hilar structure, central necrosis, and margin status.[15] Based on these nodal features, a novel scoring system was created to reliably predict malignancy called the Canada Lymph Node Score (CLNS). This is a four-point binary system shown in figure 5.[15]A LN will be given a point each for the following conditions: small axis diameter greater than 10mm, central necrosis present, absent central hilum structure, and well-defined margins.[15] An overall score of 0-1 reflects a lower chance of malignancy, and 2-4 reflects a higher chance of malignancy. This study included 300 LNs, and the four-feature model showed good discriminating power c-statistic= 0.72 ± 0.04, 95%CI: 0.64-0.80.[15] The results of this study allowed our group to move forward with this work and apply AI algorithms to predict these nodal.

Our group took this research one step further by introducing AI to this work. A DNN called NeuralSeg was trained to segment the LN from the B-mode ultrasound images captured during EBUS-TBNA procedures. There was 300 LN for the derivation set, where NeuralSeg was trained to predict the binary outcome of each feature for the CLNS. There were 112 LNs used for the validation set, where the NeuralSeg algorithm was validated. The algorithm had a accuracy 73.8% (95% confidence interval [CI], 68.4%-78.7%) for the derivation set. For the validation set NeuralSeg had an accuracy of 72.9% (95% CI, 63.5%- 81.0%) and a specificity of 90.8% (95% CI, 81.9%-96.2%). These results indicate that NeuralSeg could be trained and validated to rule out malignancy in mediastinal LNs. The next step would be to train, test and validate NeuralSeg to directly predict LN malignancy and provide an overall prediction based on the B-mode images. Part I of this thesis uses a dataset of 3366 LNs to achieve this objective. EBUS-TBNA

and AI have shown some promising results; however, the input data feed to the NeuralSeg algorithm is limited to the nodal features extracted from the B-mode images.

Elastography

Although EBUS-TBNA technology has evolved, it is still challenging to predict malignancy using B-more imaging alone. There is a novel ultrasound-related technology called Elastography.[16] Elastography is a non-invasive technology that uses strain imaging techniques to produce visual colour maps based on tissue stiffness. The way elastography works is that the strain of the tissue is measured in response to mechanical stress, compression or vibration to estimate the tissue stiffness of the region of interest (ROI).[16] This information is coded in colour, where stiffer tissue is depicted in blue and softer tissue is depicted in red. Elastography may be a valuable tool in identifying malignancy because malignant tissue tends to be stiffer in nature compared to benign tissue. This is because malignant tissue has more cells per area compared to benign tissue.[17] This technology was first applied in breast cancer to evaluate breast masses. There was a significant difference observed in strain between malignant and benign masses.[18] Elastography was then applied to thyroid gland tumours. There were 52 thyroid glands analyzed using a strain index cut off of a value greater than 4, a 96% specificity and 82% sensitivity were achieved in diagnosing malignancy.[19] These results indicate the clinical utility elastography can provide in detecting malignancy.

Similar results were observed when applied to thoracic LNs in lung cancer patients. A recent systematic review conducted a meta-analysis on 17 studies with over 2000 mediastinal

LNs. Wu et al. discovered a pooled sensitivity of 0.90 (95% confidence interval (CI), 0.84-0.94) and a specificity 0.78 (95% CI, 0.74-0.81). Additionally, a pooled AUC of 0.86 was determined.[20] The results of this meta-analysis provide evidence that elastography can be promising in the field of lung cancer; however, how the studies varied in the way elastography images were analyzed.

The elastography analysis methods can be classified as either qualitative or quantitative. The qualitative methods include the 3-type classification and the 5-type classification. The quantitative methods include strain ratio, stiffness area ratio, and strain histogram. The stiffness area ratio can also be referred to as blue colour proportion. Below is a summary of each of the methods.

Qualitative Methods

3-type classification: The elastography patterns are assessed and then categorized into one of three classifications:

- Type 1, predominately non-blue, which can include green, yellow, and/or red
- Type 2, part blue, part non-blue
- Type 3, predominately blue

Type 3 would be the most indicative of malignant tissue in the LN. Figure 6 shows each type of classification.[21]

5-type classification: This method classifies the elastography patterns into one of five categories. Score 1 when the pattern shows a mixture of green-yellow-red, Score 2 is homogeneous predominately green, Score 3 is mixed blue-green-yellow-red, Score 4 is mixed blue-green, and Score 5 is homogenous predominately blue, as shown in Figure 7.[22]

Quantitative Methods

Strain Ratio: The largest possible area within the LN is selected, represented by A in figure 8, and a reference area is selected as B. The strain ratio is calculated as B/A.[23] With this method, a higher strain ratio is indicative of malignancy.

Stiffness Area Ratio (SAR): This method involves selecting the LN as the ROI, and the ratio is calculated based on the amount of stiff tissue, which is the blue colour pixels compared to the colour information in the ROI as seen in figure 9.[16] With this method a higher SAR is more indicative of malignancy.

Strain Histogram: This method uses the information provided on the bottom right of each panel shown in figure 10. A strain histogram is displayed, and the mean of the histogram is used as a quantitative measure to evaluate LN malignancy.[24]

When comparing qualitative and quantitative methods, quantitative methods are far more reproducible and less subjective. The most intuitive and straightforward technique is SAR of all the quantitive methods. This method's intuitive nature would encourage a potentially widespread

use of the technique. Additionally, this method has achieved high sensitivities and specificities.[16] Current literature has utilized this technique by investigating a SAR cut-off value that distinguishes benign and malignant LNs. For example, Uchimura and colleagues (2020) obtained a cut-off value of 0.41 based on the analysis of 149 LNs with 88.2% sensitivity, 80.2% specificity, and 83.9% diagnostic accuracy.[25] Nakajima and colleagues (2015) got a cut-off value of 0.31 based on 49 LNs with 81% sensitivity and 85% specificity.[16] Ma and colleagues (2018) had an AUC of 0.86 and a cut-off value of 0.37 based on 79 LNs and 92.30% sensitivity, 67.50% specificity, and 78.5% accuracy.[26] However, all these studies used manual methods to obtain the SAR using photoshop software, and Image J. EBUS-elastography can be used as a useful imaging modality remains debatable, as the technology is still relatively new and a method has yet to become standardized.[20] In chapters 3 and 4 of this thesis, we will explore the application of AI and DNN to obtain the SAR based on elastography colour maps to diagnose LN malignancy.

Thesis Objectives

The objective of this thesis is three-fold. Firstly, to train, validate and test a deep neural AI network algorithm (NeuralSeg; intervention) to predict the malignancy of a lymph node (outcome) based on ultrasonographic features examined by B-mode images collected from EBUS-TBNA. Secondly, to quantitively define the colour threshold of "blue" (Blue Threshold), which represents stiff LN tissue in the elastography colour maps and compute the optimal stiffness area ratio that can be used to differentiate benign LNs from malignant LNs based on the blue colour

threshold. Lastly, to validate the previously determined elastography feature, the stiffness area ratio is determined from this thesis and determines if it is predictive of malignancy compared to the surgical pathology.

Reference

1. Brenner, D. R. *et al.* Projected estimates of cancer in Canada in 2020. *Can. Med. Assoc. J.* **192**, E199 (2020).
2. Brundage, M. D., Davies, D. & Mackillop, W. J. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest* **122**, 1037–1057 (2002).
3. Care, C. T. F. on P. H. Recommendations on screening for lung cancer. *Cmaj* **188**, 425–432 (2016).
4. Walters, S. *et al.* Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. *Thorax* **68**, 551 (2013).
5. Nakajima, T., Yasufuku, K. & Yoshino, I. Current status and perspective of EBUS-TBNA. *Gen. Thorac. Cardiovasc. Surg.* **61**, 390–396 (2013).
6. Rusch, V. W. *et al.* The IASLC lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh edition of the TNM classification for lung cancer. *J. Thorac. Oncol.* **4**, 568–577 (2009).
7. Annema, J. T. *et al.* Mediastinoscopy vs endosonography for mediastinal nodal staging of lung cancer: a randomized trial. *Jama* **304**, 2245–2252 (2010).
8. Maconachie, R., Mercer, T., Navani, N. & McVeigh, G. Lung cancer: diagnosis and management: summary of updated NICE guidance. *BMJ Br. Med. J. Online* **364**, (2019).
9. De Leyn, P. *et al.* Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer†. *Eur. J. Cardiothorac. Surg.* **45**, 787–798 (2014).
10. Ortakoylu, M. G. *et al.* Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *J. Bras. Pneumol.* **41**, 410–414 (2015).

11. Osarogiagbon, R. U. *et al.* Invasive mediastinal staging for resected non–small cell lung cancer in a population-based cohort. *J. Thorac. Cardiovasc. Surg.* **158**, 1220-1229.e2 (2019).
12. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* **316**, 2353–2354 (2016).
13. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
14. Nam, J. G. *et al.* Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* **290**, 218–228 (2019).
15. Hylton, D. A. *et al.* The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. *J Thorac Cardiovasc Surg* **159**, 2499-2507.e3 (2020).
16. Nakajima, T. *et al.* Elastography for Predicting and Localizing Nodal Metastases during Endobronchial Ultrasound. *Respiration* **90**, 499–506 (2015).
17. Riegler, J. *et al.* Tumor Elastography and Its Association with Collagen and the Tumor Microenvironment. *Clin. Cancer Res.* **24**, 4455 (2018).
18. Moon, W. K., Chang, R.-F., Chen, C.-J., Chen, D.-R. & Chen, W.-L. Solid breast masses: classification with computer-aided analysis of continuous US images obtained with probe compression. *Radiology* **236**, 458–464 (2005).
19. Lyshchik, A. *et al.* Thyroid gland tumor diagnosis at US elastography. *Radiology* **237**, 202–211 (2005).

20. Wu, J. *et al.* Diagnostic value of endobronchial ultrasound elastography for differentiating benign and malignant hilar and mediastinal lymph nodes: a systematic review and meta-analysis. *Med. Ultrason.* (2021).
21. Izumo, T., Sasada, S., Chavez, C., Matsumoto, Y. & Tsuchida, T. Endobronchial Ultrasound Elastography in the Diagnosis of Mediastinal and Hilar Lymph Nodes. *Jpn. J. Clin. Oncol.* **44**, 956–962 (2014).
22. Sun, J. *et al.* Endobronchial Ultrasound Elastography for Evaluation of Intrathoracic Lymph Nodes: A Pilot Study. *Respiration* **93**, 327–338 (2017).
23. Korrungruang, P. & Boonsarngsuk, V. Diagnostic value of endobronchial ultrasound elastography for the differentiation of benign and malignant intrathoracic lymph nodes. *Respirology* **22**, 972–977 (2017).
24. Verhoeven, R. L. J., de Korte, C. L. & van der Heijden, E. H. F. M. Optimal Endobronchial Ultrasound Strain Elastography Assessment Strategy: An Explorative Study. *Respir. Int. Rev. Thorac. Dis.* **97**, 337–347 (2019).
25. Uchimura, K. *et al.* Quantitative analysis of endobronchial ultrasound elastography in computed tomography-negative mediastinal and hilar lymph nodes. *Thorac. Cancer* **11**, 2590–2599 (2020).
26. Ma, H. *et al.* Semi-quantitative Analysis of EBUS Elastography as a Feasible Approach in Diagnosing Mediastinal and Hilar Lymph Nodes of Lung Cancer Patients. *Sci. Rep.* **8**, 3571 (2018).
27. Kim, E. S. & Bosquée, L. The importance of accurate lymph node staging in early and locally advanced non-small cell lung cancer: an update on available techniques. *J. Thorac. Oncol.* **2**, S59–S67 (2007).

28. Churchill, I. F. *et al.* An Artificial Intelligence Algorithm to Predict Nodal Metastasis in Lung Cancer. *Submitt. Ann. Thorac. Surg.*

29. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data TKDD* **6**, 1–21 (2012).

30. Mustafa Basil *et al.* Deep Ensembles for Low-Data Transfer Learning. *2020* doi:https://doi.org/10.48550/arXiv.2010.06866.

31. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*. vol. 112 (Springer, 2013).

32. Statistics, I. S. IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. *Google Search* (2013).

33. Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv. CSUR* **27**, 326–327 (1995).

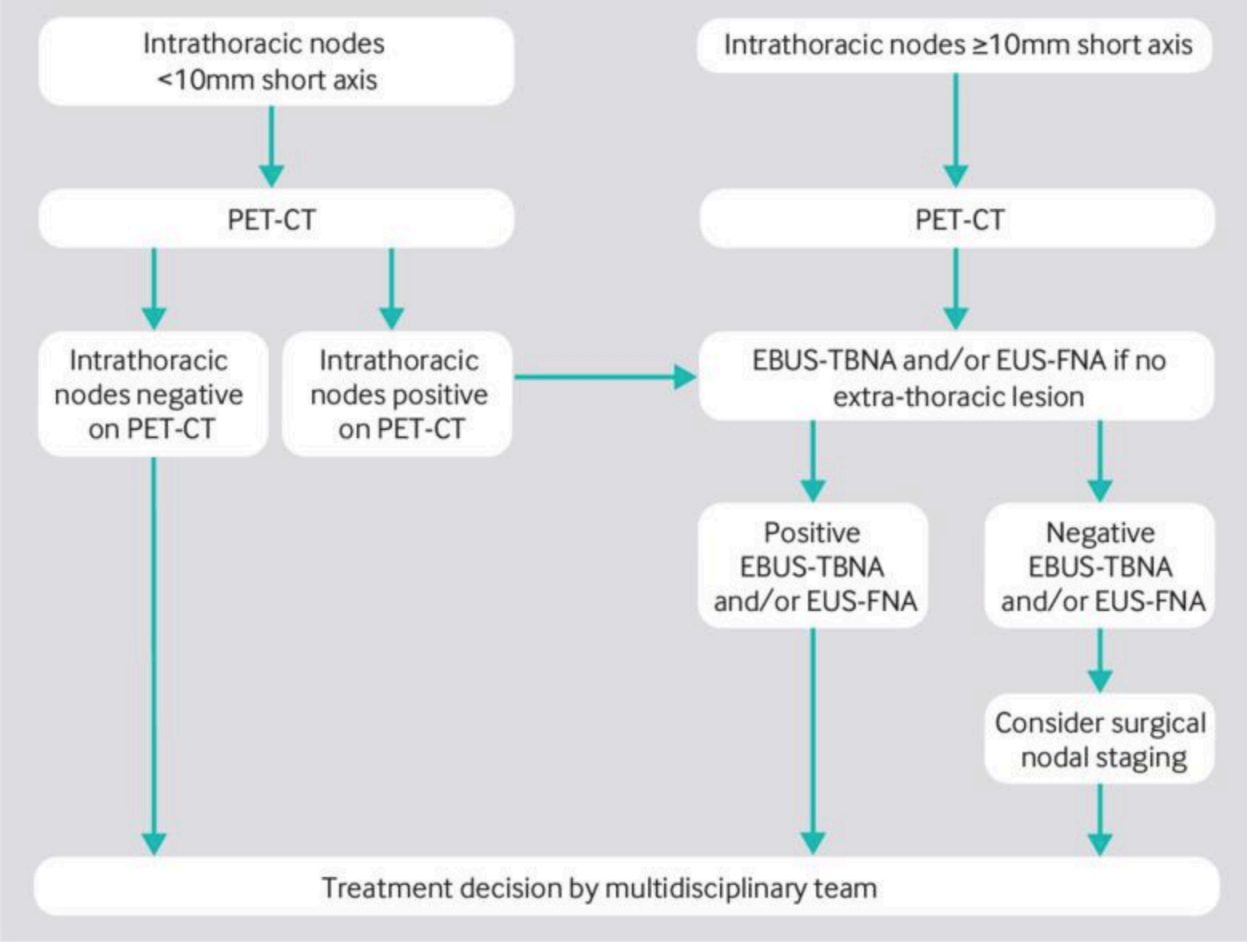
34. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **5**, 1315–6 (2010).

35. Yong, S. H. *et al.* Malignant thoracic lymph node classification with deep convolutional neural networks on real-time endobronchial ultrasound (EBUS) images. *2022* **11**, 14–23 (2022).

36. Verhoeven, R. L. J. *et al.* Predictive value of endobronchial ultrasound strain elastography in mediastinal lymph node staging: the E-predict multicenter study results. *Respiration* **99**, 484–492 (2020).

37. Divisi, D., Zaccagna, G., Barone, M., Gabriele, F. & Crisci, R. Endobronchial ultrasound-transbronchial needle aspiration (EBUS/TBNA): a diagnostic challenge for mediastinal lesions. *Ann. Transl. Med.* **6**, (2018).

38. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, 2015).
39. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform.* **48**, 193–204 (2014).
40. Olympus America Inc. Elastography Quick Reference. (2016).
41. Mower, W. R. Evaluating bias and variability in diagnostic test reports. *Ann. Emerg. Med.* **33**, 85–91 (1999).



PET-CT = positron emission tomography-computed tomography
EBUS-TBNA = endobronchial ultrasound-guided transbronchial needle aspiration
EUS-FNA = endoscopic ultrasound-guided fine needle aspiration

Figure 1. Diagnostic Pathway Flow Chart Including the Staging of Intrathoracic Lymph Nodes for Non-Small Cell Lung Cancer

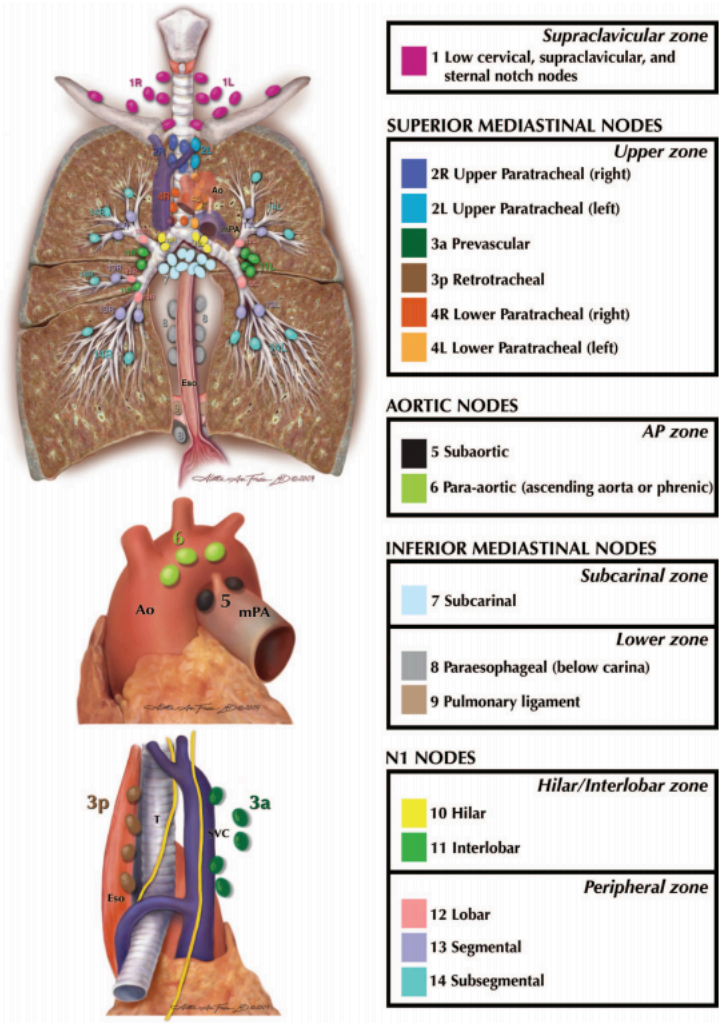


Figure 2. Thoracic Lymph Node Stations Organized into Zones

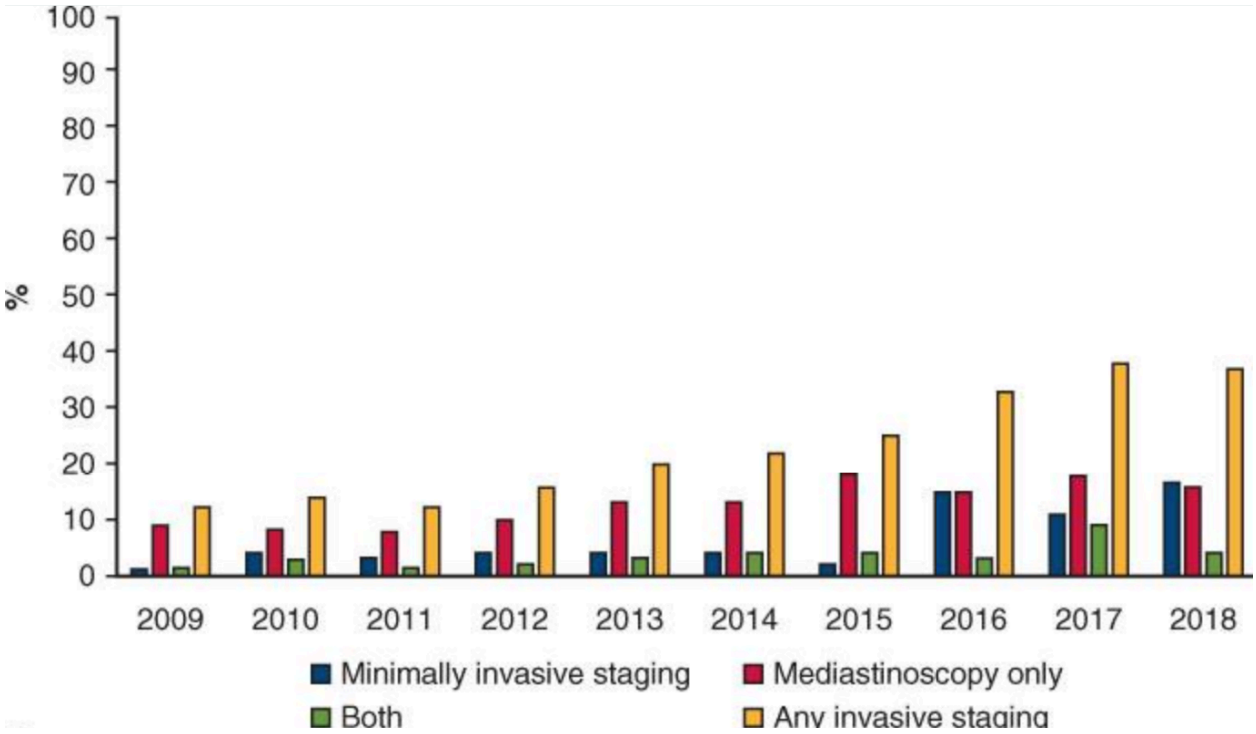


Figure 3. The Percentage of Patients that Received Nodal Staging Over the Span of a Decade (2009-2018)

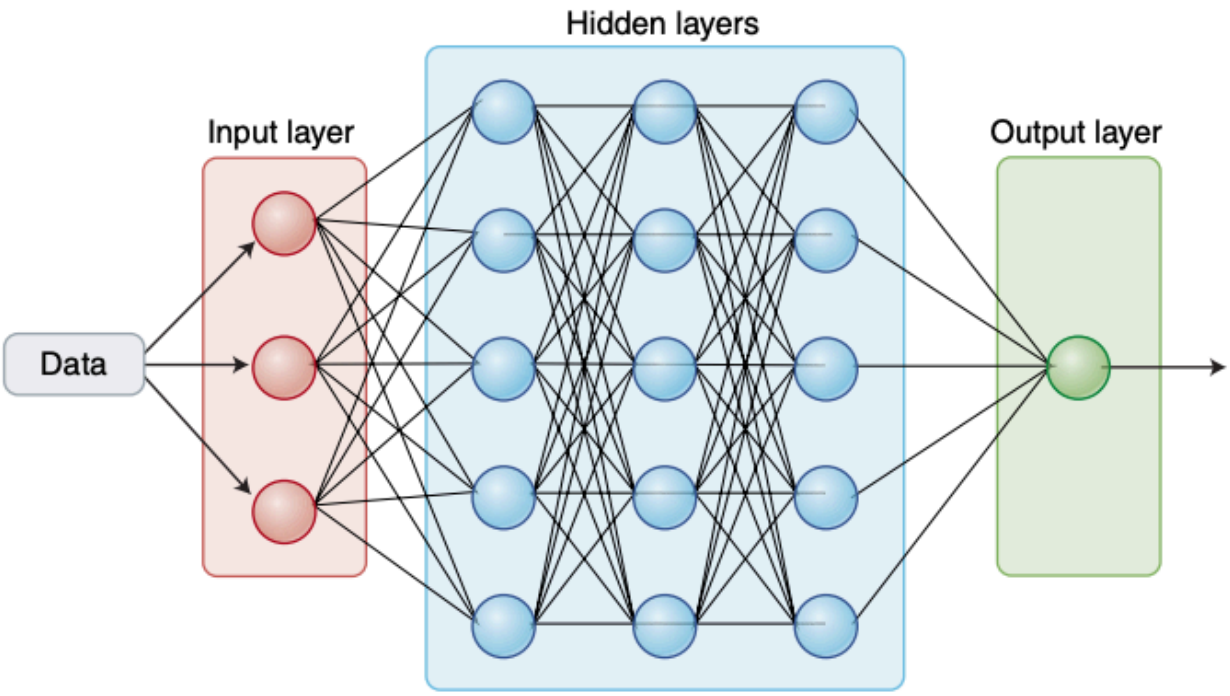


Figure 4. Simplified Structure of Deep Neural Network, Including the Input, Hidden, and Output Layers

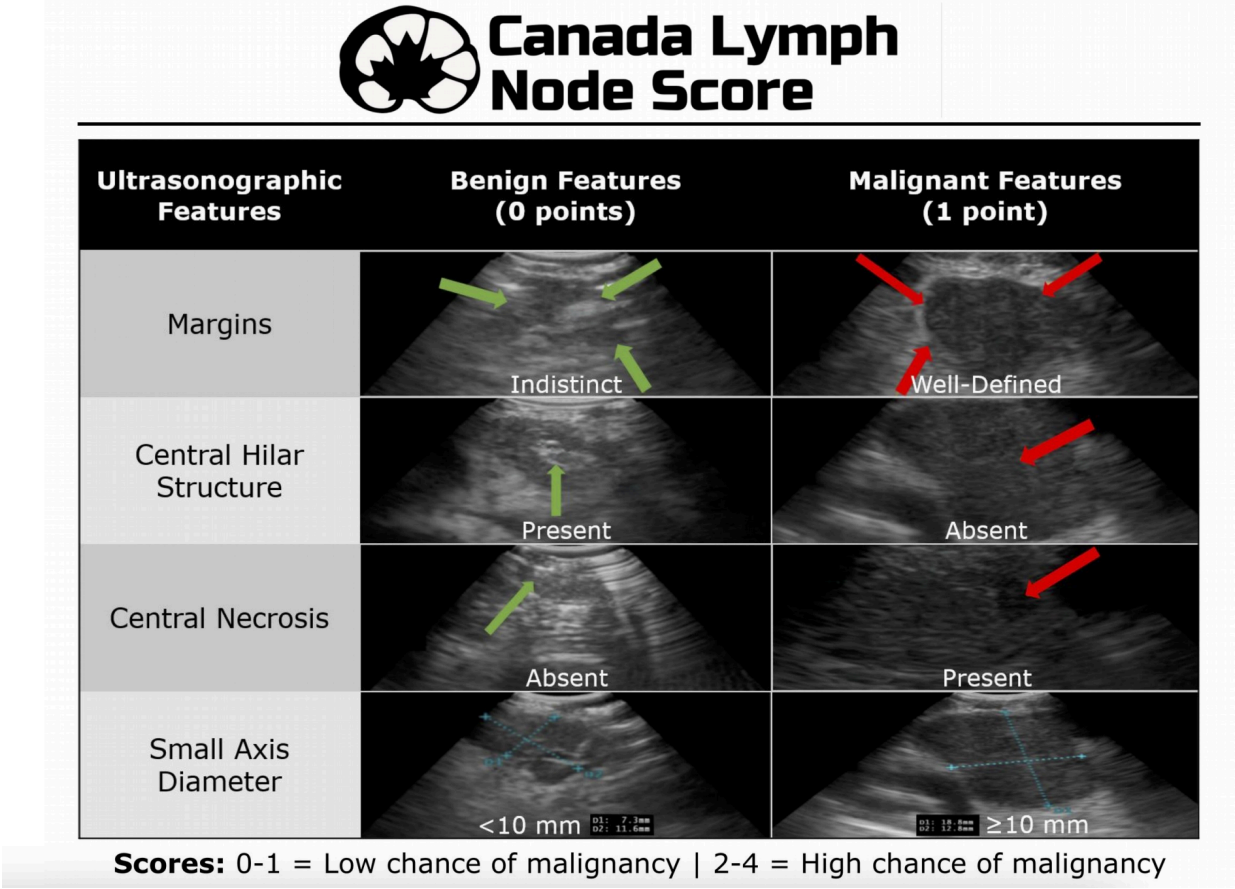


Figure 5. Visual Representation of the Canada Lymph Node Score

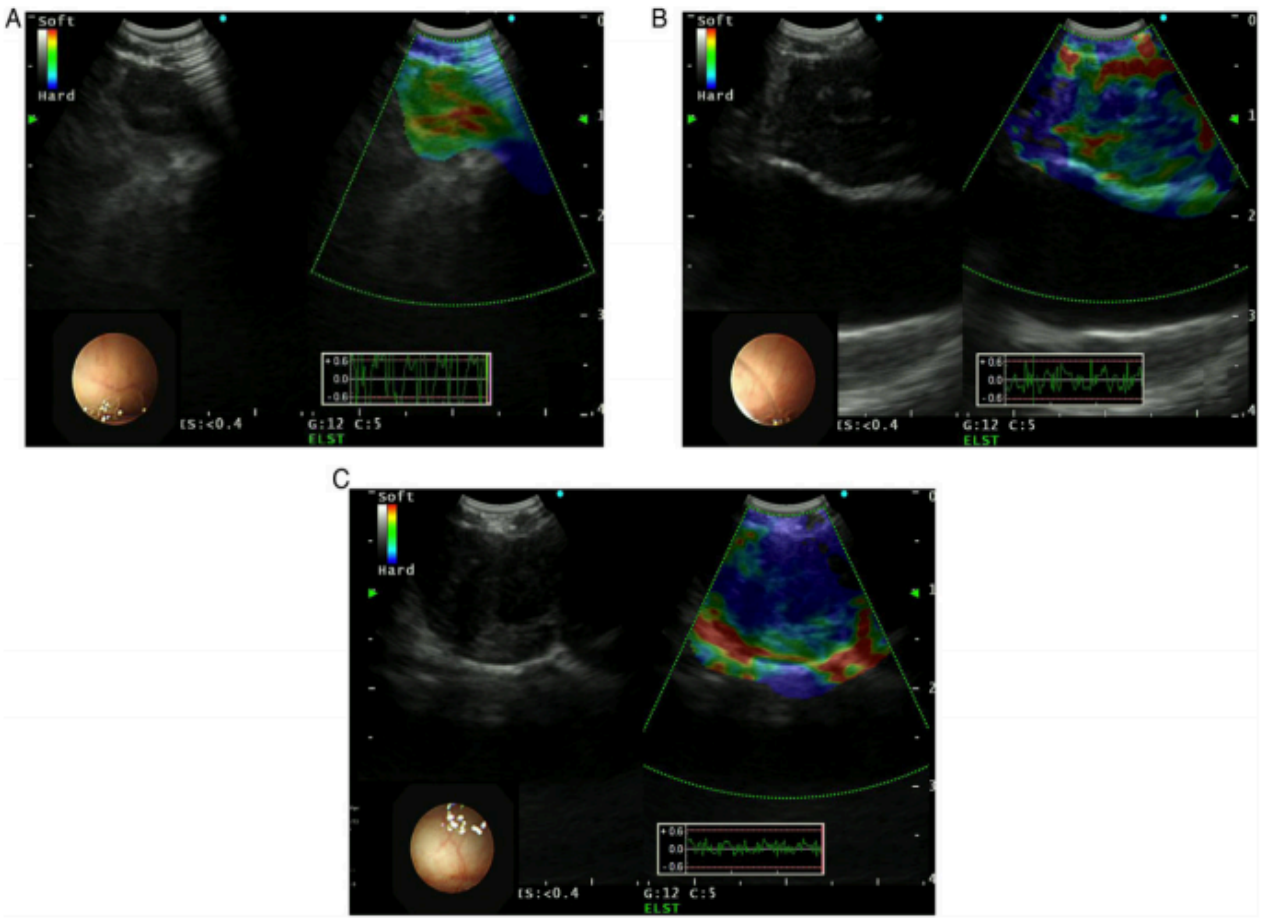


Figure 6. EBUS-Elastography analyzed using the 3-type classification: A) Type 1, predominately non-blue, which can include green, yellow, and/or red, B) Type 2, part blue, part non-blue, C) Type 3, predominately blue

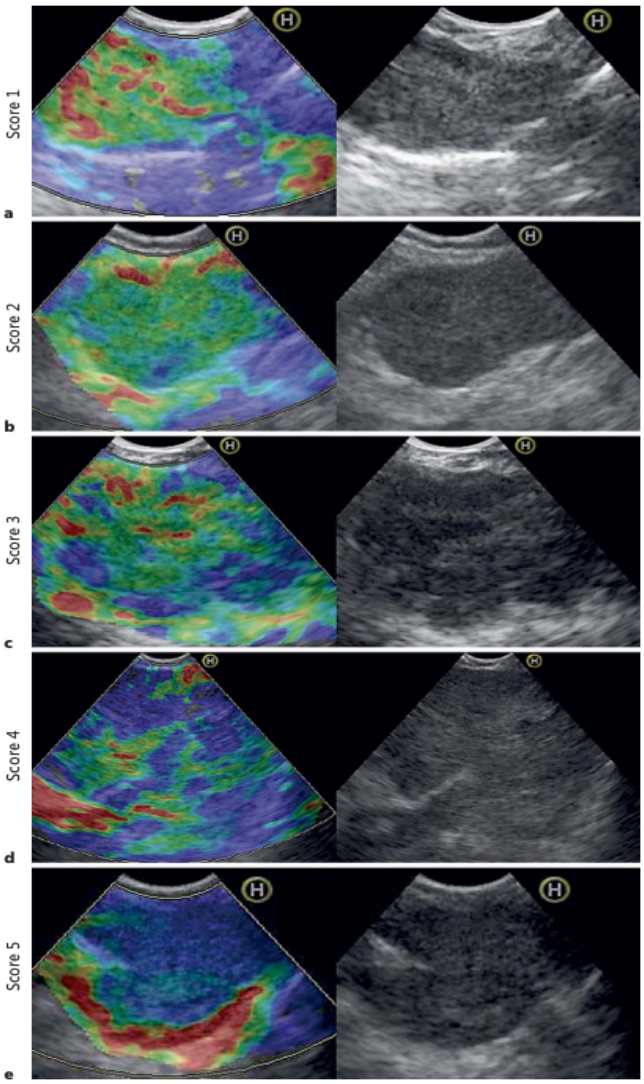


Figure 7. EBUS-Elastography analyzed using 5-type classification system Scores 1-5; a) Score 1 when the pattern shows a mixture of green-yellow-red, b) Score 2 is homogeneous predominately green, c) Score 3 is mixed blue-green-yellow-red, d) Score 4 is mixed blue-green, and e) Score 5 is homogenous predominately blue

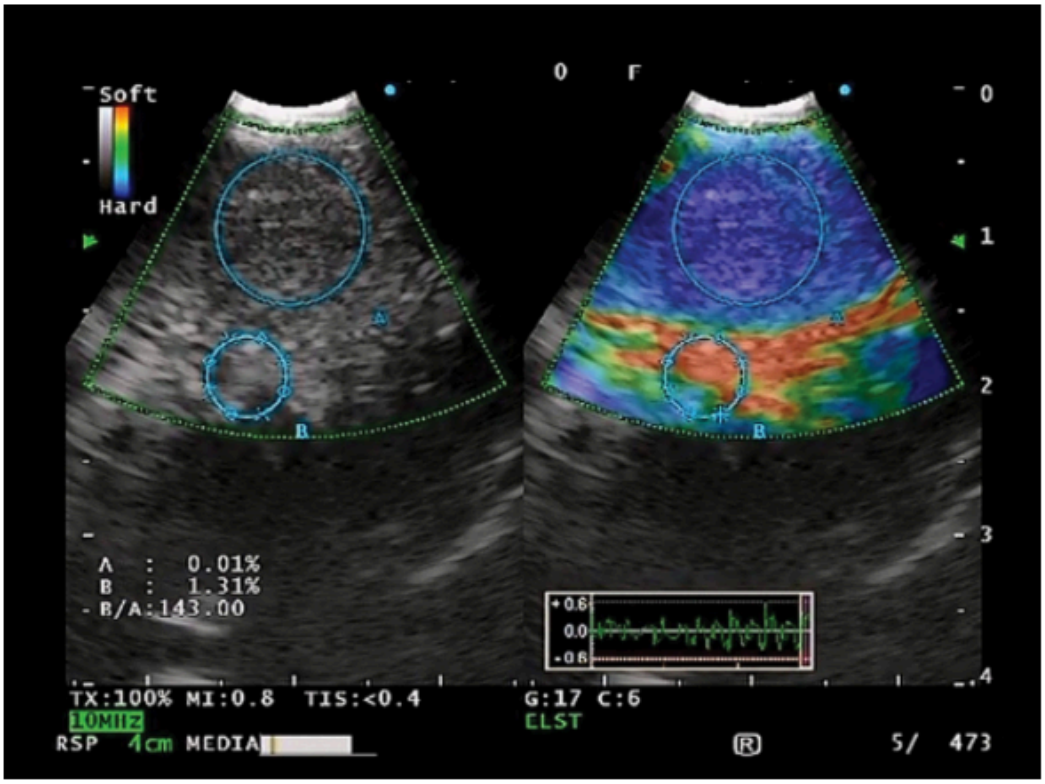


Figure 8. EBUS-Elastography analyzed using strain ratio, where circle A is the largest possible area within the LN is selected and a reference area is selected as circle B

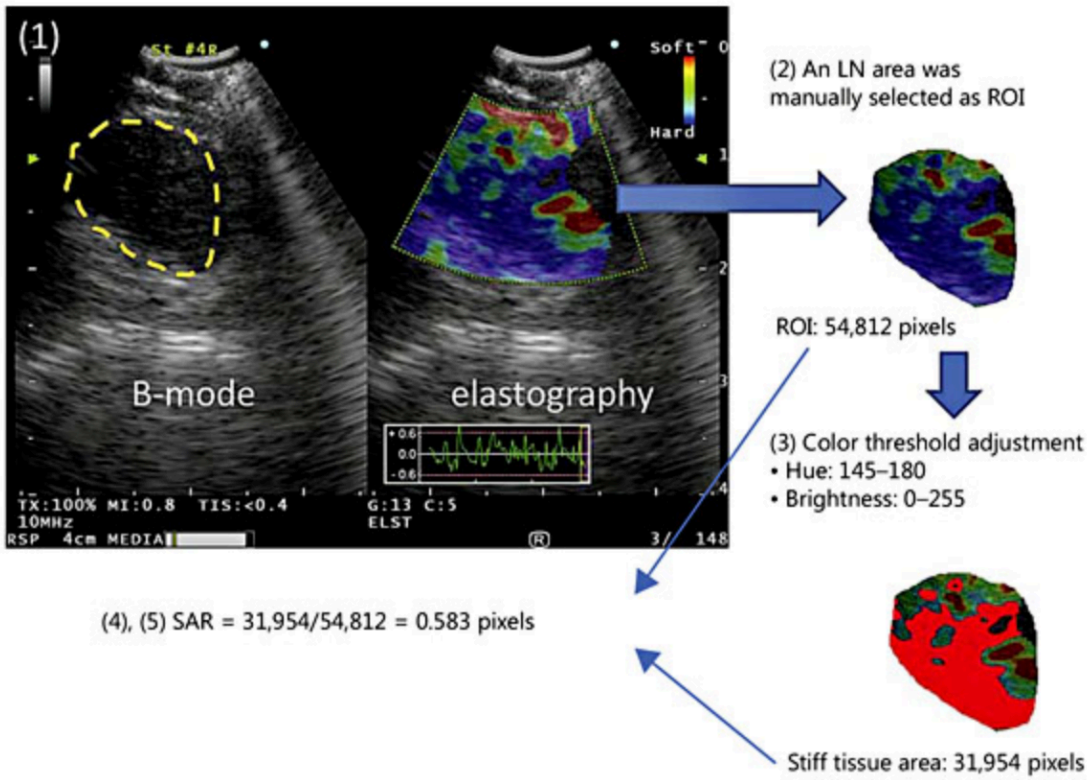


Figure 9. EBUS-Elastography analyzed using Stiffness Area Ratio, which is the number of blue pixels over the total amount of pixels

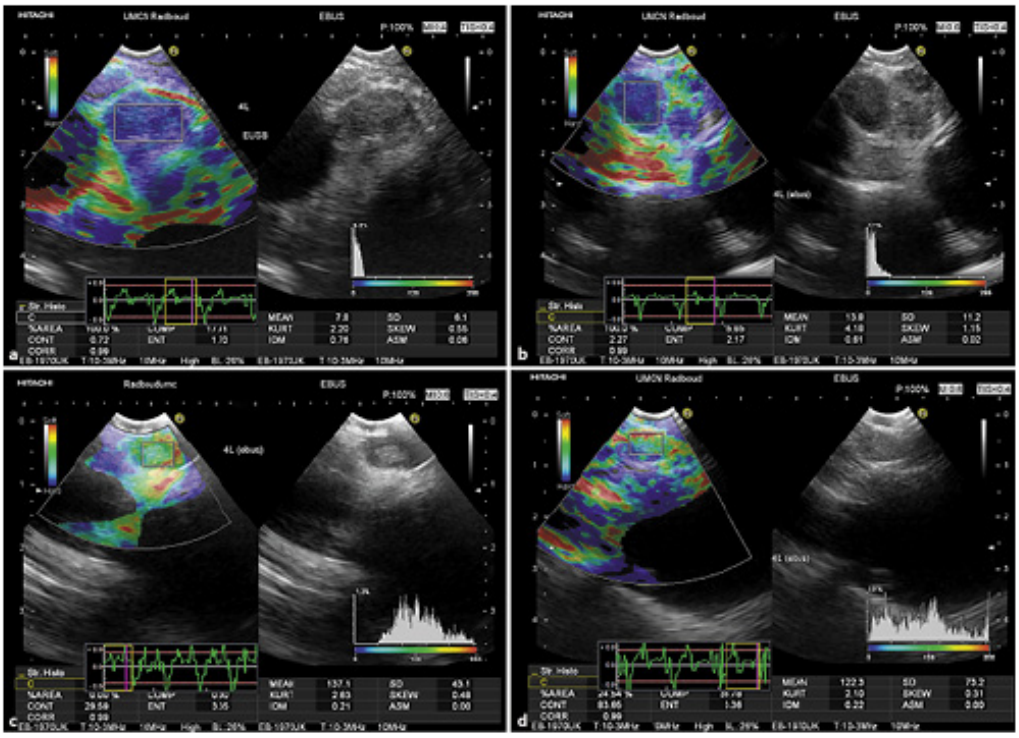


Figure 10. EBUS-Elastography analyzed using the Strain Histogram Mean, the strain histograms can be seen at the bottom right of the image

Chapter 2: The Application of an Artificial Intelligence Algorithm to predict Lymph Node Malignancy in Non-Small Cell Lung Cancer

Nikkita Mistry, BSc, Anthony A. Gatti, PhD,Yogita S. Patel, BSc, Forough Farrokhyar, MPhil, PhD, Feng Xie, MSc, PhD, Sam Cross, BSc, Waël C. Hanna, MDCM, MBA

Abstract

Background: Staging mediastinal lymph nodes (LNs) is an essential step in the non-small cell lung cancer (NSCLC) diagnostic pathway. Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA) is the current standard for mediastinal nodal staging. However, despite the advancements in staging procedures, as high as 40% of nodal biopsy results remain inconclusive. Artificial Intelligence (AI) and deep-learning computer neural networks are a technology gaining popularity in medicine, and recent research shows a high level of accuracy in interpreting images.

Objective: This study aims to train, validate and test a deep neural AI network algorithm, NeuralSeg, to predict the malignancy of a lymph node based on B-mode images collected from EBUS-TBNA compared to the final pathology results.

Methods: 3366 LN images which belonged to 943 patients obtained during EBUS-TBNA procedures were fed to a deep neural network, NeuralSeg. The endosonographer assigned a Canada Lymph Node Score (CLNS) for each LN. The algorithm segmented the LN, and LN small-axis diameter, central hilar structure, central necrosis, and margin status were extracted. The dataset set was split into the training/validation set (80% of the sample) and the testing set (20% of the sample). NeuralSeg is used on three different pools of pre-trained models: ResNet152V2, InceptionV3, and DenseNet201. The 5-fold cross-validation technique was used. The outputs from the three models were combined, and a final ensemble model was created.

Results: The final model had an overall accuracy of 80.63% (76.93% to 83.97%), a sensitivity of 43.23% (35.30% to 51.41%), a specificity of 96.91% (94.54% to 98.45%), a positive predictive value of 85.90% (76.81% to 91.80%), a negative predictive value of 79.68% (77.34% to 81.83%), and an AUC of 0.701 (0.646 to 0.755).

Conclusion: In conclusion, the deep neural network, NeuralSeg showed promising results with a high specificity and NPV to identify malignancy, a clinically valuable technique in staging mediastinal LNs. Although when the AI algorithm was compared to an expert endosonographer applying the CLNS, the AI algorithm was outperformed. This suggests that the AI model may benefit from further optimization.

Introduction

The staging of mediastinal lymph nodes (LNs) is a crucial step in the non-small cell lung cancer (NSCLC) diagnostic pathway.[9] Nodal staging aids in the determination of treatment decisions, whether a patient is a candidate for lung resection or if chemotherapy is the best route for treatment.[27] Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA) is the current standard for mediastinal nodal staging. However, despite the developments in staging procedures, as high as 40% of nodal biopsy results remain inconclusive.[10] The skill of the endoscopist and cytologists impact the accuracy of EBUS-TBNA.[28] This may be discouraging practitioners from conducting nodal staging altogether. For example, in a study conducted between 2009 and 2018 on 2,916 lung cancer patients, only 22% were compliant with LN staging guidelines.[11] Novel technologies may serve as a valuable adjunct to help improve diagnostic yields of EBUS-TBNA.

Artificial Intelligence (AI) and deep-learning computer neural networks are a technology gaining popularity in medicine. Computer simulation of human cognition by computer has become a valuable tool for interpreting medical images.[12] The amount of data in radiology imaging has increased. As medical imaging technology has advanced, the number of images produced is at an all-time high. Extracting crucial information from a large number of images can be time-consuming and difficult. A software can extract this information in order to provide physicians with valuable clinical data.[12] Recent research has shown that AI and deep learning are being used to accurately interpret images with comparisons to clinicians in radiology, pathology, and cardiology.[13]

Previous research has identified four ultrasonographic nodal features predictive of malignancy: LN small-axis diameter, central hilar structure, central necrosis, and margin status.[15] These features have been compiled into the Canada Lymph Node Score (CLNS), a four-feature score binary score used to enhance the prediction of malignancy during mediastinal EBUS-TBNA staging, where a higher score of 2 or greater is predictive of malignancy.[15] However, this score was tested across several sites. There was interrater variability between ultrasonographers when assigning the CLNS score.[15]This work was taken one step further by training a deep neural network, NeuralSeg, to segment the LN from the B-mode image and predict malignancy based on these ultrasonographic nodal features. NeuralSeg was able to achieve a diagnostic accuracy of 78.8% for predicting malignancy based on the CLNS. [28]

The combination of mediastinal LN imaging and AI has been proven feasible and can decrease operator-dependency, which often leads to inconclusive biopsy results. Building on previous research, this study will utilize the ensemble method, which involves the pre-training of the algorithm using different models, and then combining these models to create an ensemble model. In addition, this study aims to train, validate and test a deep neural AI network algorithm, NeuralSeg, to predict the malignancy of a lymph node based on B-mode images collected from EBUS-TBNA compared to the final pathology results.

Methods

After sorting the LN images and removing the images that did not have corresponding identifying information to obtain the pathology result, the final cohort consisted of 3366 LN images which belonged to 943 patients.

Source of Data

The training and testing dataset comes from a prospectively collected LN database. LN images were collected from St. Joseph's Healthcare Hamilton from January 19th, 2015, until June 7th, 2021. During EBUS-TBNA procedures, suitable static images were captured and stored on an external hard drive. In addition, information such as the patient's age, gender, LN station, date and time of the procedure was also recorded.

Patients

Patients were consecutively sampled at St. Joseph's Healthcare Hamilton and were provided informed consent. Patients diagnosed with suspected or confirmed NSCLC requiring mediastinal LN staging with EBUS-TBNA were eligible for this study. No exclusion criteria were applied. Enrollment took place between January 19th, 2015 and June 7th, 2021. Enrollment in the study did not intervene with the standard of care practices. Participants' enrollment ended at the same time as the EBUS-TBNA procedure.

Procedure

EBUS-TBNA was performed using an Olympus convex probe ultrasound endoscope (Olympus, Shinjuku-ku, Tokyo, Japan) and the EU-ME2 plus transducer (Olympus) under conscious sedation with midazolam and fentanyl. LN biopsy was performed by transbronchial needle aspiration with a 22-gauge needle. Specimen adequacy was confirmed by rapid onsite cytology. The endosonographer assigned a Canada Lymph Node Score (CLNS) for each LN based on LN small-axis diameter, central hilar structure, central necrosis, and margin status during the EBUS-TBNA procedure.[15]

Outcomes

The primary outcome is to test whether NeuralSeg can predict malignancy in a LN directly from the B-mode image. The gold standard for comparison will be the final pathology results

Predictors

Patient demographic information such as age, gender, and smoking status was collected for the training and testing dataset. Regarding LN characteristics, final pathology, LN station, short-axis measurements, and long-axis measurements were obtained. Ultrasonographic nodal features in the form of the Canada Lymph Node Score (CLNS) was collected in a binary manner based on the following criteria: LN small-axis diameter, central hilar structure, central necrosis, and margin status.[15] The endosonographer assigned a point for each of the following conditions: small axis length was greater than 10mm, absence of central hilar structure, presence of central necrosis, and margins were greater than 50%. The results for each feature were recorded and

compared between the training and testing dataset as well as the overall CLNS. NeuralSeg was blinded to the demographic information of the patients and the final pathology results.

Unit of Analysis

The unit of analysis for this study was the LNs rather than the patient, as the primary outcome is whether NeuralSeg could predict malignancy directly from the ultrasound B-mode image. The data will be organized by the patient rather than the individual image to avoid data leakage and prevent creating of an overly optimistic or an invalid model.[10]

Sample Size

This study was conducted using a sample size of about 3,368 lymph node B-mode images from EBUS-TBNA patients. This is a sufficient amount of data to train, validate and test the NeuralSeg algorithm to predict malignancy

Algorithm Design

LN segmentation was done using the previously trained NeuralSeg algorithm, a U-Net style convolutional neural network. NeuralSeg has been trained and validated for LN segmentation and LN feature extraction.[28] The LN region of interest (ROI) was automatically identified, and these ultrasonographic features: hilar structure of the LN, necrosis within the LN, and LN contour, were extracted.[15]

The dataset set was split into the training/validation set, which consisted of 80% of the sample size, and the remaining 20% was used as the testing set. The validation aims to ensure the model is not overfitting and find the optimal version of the model.

The ensemble model will be used to develop an algorithm for this study. A visual representation of the different ways to construct ensembles can be seen in Figure 1.[30] NeuralSeg used on different pre-trained model pools. This study used three common neural network models called ResNet152V2, InceptionV3, and DenseNet201. Each of these models was pre-trained on an open-source ImageNet dataset. ImageNet is a dataset that contains millions of images. Pre-training allows neural networks such as NeuralSeg to obtain a general understanding of the objects in the image. This method, in turn, improves the accuracy and robustness of the model.[30] These models are then re-trained and fine-tuned in a process referred to as transfer learning.[30] In this study, transfer learning occurred on 80% of our LN dataset.

The three models: ResNet152V2, InceptionV3, and DenseNet201, were trained individually and then combined to form an ensemble model.[30]The way this is done is the B-mode LN image was passed through each of the three models. Next, the outputs from the three models were combined and passed through several more layers of neural networks, and then a final prediction was made. Each individual model and the ensemble model were then trained five times in a process called 5-fold cross-validation.[31] The validation hold-out set approach was used, where one-fifth of the 80% sample was held out as the validation set and the remaining four-fifths were used as the training set.[31] Each image in the training set was used as a validation set only once.

Statistical Analysis

Descriptive statistics were provided for the demographic data of the study subjects. All analyses were done using Statistical Package for the Social Sciences (SPSS) 2020 version.[32] Baseline characteristics were compared between the training/validation and the testing group. Continuous parametric variables were compared using the student t-test. Continuous non-parametric variables were compared using Mann Whitney U test. Categorical variables were compared using the chi-square test. Diagnostic ability will be assessed using an index test and obtaining diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Receiving operating curve (ROC) analysis will be used to assess the model's accuracy by looking at the area under the curve (AUC). The McNemar test was used to observe if there are any statistically significant differences between the training and validation sets to assess overfitting

Results

The patients and their associated LN images were organized following Figure 2. 2704 LN images from 760 different patients made up 80% of the sample size, which was used for the training and validation set. The remaining 20% consisted of 662 LN images from 183 different patients and was used as the testing set. The baseline characteristics for both the training/validation set and the testing set can be found in Table 1. There was a significant difference between LN pathology between the two datasets (p=0.01). The training set consisted of 53.3% benign LN, 22.7% malignant, 9.9% inconclusive and 14.1% with no pathology. The

testing set consisted of 54.1% benign, 23.4% malignant, 12.7% inconclusive and 9.8 with no pathology. The only other significant difference was the presence of central necrosis, where the training set had 7.7%, and the testing set had 8.9%, p=0.02.

The three individual models: ResNet, Inception and DenseNet, were trained using a training and validation set which makes up 80% of the total data. Each model was trained five times where a different one-fifth of the data was held out for the validation set, and the remaining four-fifths were used to train the model. This method is referred to as the 5-fold cross-validation technique using the hold-out approach. Applying this method resulted in 5 different predictions for each of the three models; the results were averaged for each model and presented in Tables 2-4.

The sensitivities for the training sets for ResNet, Inception, and DenseNet are 29.68% (22.62% to 37.53%), 40.00% (32.22% to 48.17%), and 49.03% (40.93% to 57.18%), respectively. The specificities for training sets for ResNet, Inception, and DenseNet are 90.45% (86.91% to 93.30%), 93.54% (90.46% to 95.86%) and 93.54% (90.46% to 95.86%), respectively. There is a common trend of low sensitivities and high specificities for each of the models. The area under the curve (AUC) determined by the receiver operating curve (ROC) are 0.601 (0.545 to 0.657), 0.668 (0.612 to 0.723) and 0.713 (0.659 to 0.766), for ResNet, Inception, and DenseNet respectively. Of the three models, DenseNet had the highest AUC. The training set and validation set were compared for each of the models. No statistical difference was observed between the ResNet training and validation model, p=0.132. No statistical difference was observed between the Inception training and validation model, p=0.105. Lastly, no statistical difference was observed between the Inception training and validation model, p=0.999.

The three models were combined to create a final ensemble model and applied to the testing set on the final 20% of the total data. This ensemble model was analyzed the same way as the individual models, where the model was trained five times, a different one-fifth was left out as the validation set, and the remaining four-fifths were used for training. The results for the ensemble model can be seen in Table 5. Similar to the individual models, no statistical difference was observed between the training and validation set, p =0.335. The final model had an overall accuracy of 80.63% (76.93% to 83.97%), a sensitivity of 43.23% (35.30% to 51.41%), a specificity of 96.91% (94.54% to 98.45%), a positive predictive value of 85.90% (76.81% to 91.80%), a negative predictive value of 79.68% (77.34% to 81.83%), and an AUC of 0.701 (0.646 to 0.755). The ROC is shown in Figure 3.

The final AI ensemble model was compared to the Canada lymph node score assigned to each LN by an experienced endosonographer. The diagnostic results of the CLNS can be seen in Table 6. The CLNS had an overall accuracy of 79.38% (77.43 to 81.24%), a sensitivity of 81.56% (77.88% to 84.87%), a specificity of 78.54% (76.19% to 80.75%), a positive predictive value of 59.59% (56.85% to 62.27%), a negative predictive value of 91.65% (90.11% to 92.97%), and an AUC of 0.801 (0.777 to 0.824).

Discussion

This study demonstrates how a deep learning algorithm, NeuralSeg, can be trained, validated, and tested to predict LN malignancy in a non-small cell lung cancer population. Three individual models were trained and validated using 80% of the sample size; then, these models were combined to form an ensemble model, which was then tested on the remaining 20% of the sample. The final model had an overall accuracy of 80.63% (76.93% to 83.97%), a sensitivity of 43.23% (35.30% to 51.41%), a specificity of 96.91% (94.54% to 98.45%), a positive predictive value of 85.90% (76.81% to 91.80%), a negative predictive value of 79.68% (77.34% to 81.83%), and an AUC of 0.701 (0.646 to 0.755). These results demonstrate that NeuralSeg can effectively identify LN malignancy as it has a high specificity, which can be utilized in the mediastinal LN staging process.

A common issue with machine learning algorithms is the concept of overfitting. When an algorithm is trained on a particular training set, it may be trained to fit that specific dataset as the algorithm may pick up and memorize particular details relevant to that specific data set.[33] The issue that arises is then the algorithm that demonstrates overfitting will not be generalizable to other data, thus limiting its predictability. We tested for overfitting by comparing the training and validation results for each individual and ensemble model.[33] In all cases, there was no significant difference between the training and validation results, meaning the algorithm trained using the training set was generalizable to a new dataset which was the validation set.

These results indicate that NeuralSeg may be a helpful adjunct in the LN staging process. However, further optimization of the model is required before it can potentially outperform a human expert, reducing operator dependency.

Previously NeuralSeg was able to predict small axis length, margins, and central hilar structure, based on the binary CLNS. The overall accuracy was 72.87% (63.46% to 80.98%), sensitivity of 28.12% (13.75% to 46.75%), specificity of 90.79% (81.94% to 96.22%), PPV of 55.02% (33.27% to 75.00%), and a NPV of 75.92% (71.51% to 79.85%).[28] In this study, NeuralSeg was further trained using the ensemble method to provide an overall malignancy prediction on a larger sample set. The results of our study show an increase in diagnostic capabilities and smaller confidence intervals for NeuralSeg to predict LN malignancy, indicative of the improvements in the AI algorithm.

However, in this study, NeuralSeg was compared to the CLNS, which was conducted by an expert endosonographer; the CLNS outperformed the algorithm. Although NeuralSeg and CLNS had comparable overall accuracy, NeuralSeg had a higher specificity. In addition, CLNS had a higher sensitivity, PPV, NPV, and AUC. The AUC for NeuralSeg is 0.701 (0.646 to 0.755), which would be considered an acceptable diagnostic test, whereas CLNS is 0.801 (0.777 to 0.824), which is considered to be an excellent diagnostic test.[34] Although the results of this study show the improvements in NeuralSeg's predictive capabilities; the algorithm could be further optimized.

The application of artificial intelligence and deep neural networks is still novel in analyzing medical images. A study conducted by Yong and colleagues in January 2022 used traditional cross-entropy for classification loss, a technique used for the classification of images.[35] Based on these methods and a sample size of 2,394 LNs, a sensitivity, specificity and overall accuracy of 69.7%, 74.3%, and 72.0%, respectively, was obtained. Additionally, the AUC was 0.782. This research group took this work one step further. It optimized the model by modifying the algorithm using a new loss function, a technique used to improve an AI model to make more accurate predictions.[35] Using the new modified model, the sensitivity, specificity, and accuracy were improved to 72.7%, 79.0%, and 75.8%, respectively, and an AUC of 0.8.[35] The lack of human input makes using DNN appealing as the results of procedures such as EBUS-TBNA are heavily dependent on the operator. As this is still an emerging technology, new methods are being applied to improve DNN model accuracy.

One limitation of this study was the variation in LN baseline characteristics between the training and testing sets. These factors include LN pathology and central necrosis, both of which could have potentially impacted the algorithm's ability to make accurate malignancy predictions. Additionally, only four ultrasonographic features: LN small-axis diameter, central hilar structure, central necrosis, and margin status, were used to train the algorithm. Although these features are predictive of malignancy, other ultrasonographic features may contribute to the model's ability to predict malignancy. An interesting future direction would be to not only include more ultrasonographic features but include parameters beyond B-mode ultrasound. Another potential limitation is the variation in image quality. The static images are saved in real-time during the EBUS-TBNA procedures by the endosonographer.

Conclusion

In conclusion, the deep neural network, NeuralSeg showed promising results with a high specificity and NPV to identify malignancy, a clinically valuable technique in staging mediastinal LNs. Although when the AI algorithm was compared to an expert endosonographer applying the CLNS, the AI algorithm was outperformed. This suggests that further optimization of the AI model is required. A possible future direction is to take advantage deep neural network's ability to utilize many image features; therefore, incorporating LN parameters beyond EBUS-TBNA images may improve the model's predictability.

Reference

1. Brenner, D. R. *et al.* Projected estimates of cancer in Canada in 2020. *Can. Med. Assoc. J.* **192**, E199 (2020).
2. Brundage, M. D., Davies, D. & Mackillop, W. J. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest* **122**, 1037–1057 (2002).
3. Care, C. T. F. on P. H. Recommendations on screening for lung cancer. *Cmaj* **188**, 425–432 (2016).
4. Walters, S. *et al.* Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. *Thorax* **68**, 551 (2013).
5. Nakajima, T., Yasufuku, K. & Yoshino, I. Current status and perspective of EBUS-TBNA. *Gen. Thorac. Cardiovasc. Surg.* **61**, 390–396 (2013).
6. Rusch, V. W. *et al.* The IASLC lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh edition of the TNM classification for lung cancer. *J. Thorac. Oncol.* **4**, 568–577 (2009).
7. Annema, J. T. *et al.* Mediastinoscopy vs endosonography for mediastinal nodal staging of lung cancer: a randomized trial. *Jama* **304**, 2245–2252 (2010).
8. Maconachie, R., Mercer, T., Navani, N. & McVeigh, G. Lung cancer: diagnosis and management: summary of updated NICE guidance. *BMJ Br. Med. J. Online* **364**, (2019).
9. De Leyn, P. *et al.* Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer†. *Eur. J. Cardiothorac. Surg.* **45**, 787–798 (2014).
10. Ortakoylu, M. G. *et al.* Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *J. Bras. Pneumol.* **41**, 410–414 (2015).

11. Osarogiagbon, R. U. *et al.* Invasive mediastinal staging for resected non–small cell lung cancer in a population-based cohort. *J. Thorac. Cardiovasc. Surg.* **158**, 1220-1229.e2 (2019).
12. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* **316**, 2353–2354 (2016).
13. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
14. Nam, J. G. *et al.* Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* **290**, 218–228 (2019).
15. Hylton, D. A. *et al.* The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. *J Thorac Cardiovasc Surg* **159**, 2499-2507.e3 (2020).
16. Nakajima, T. *et al.* Elastography for Predicting and Localizing Nodal Metastases during Endobronchial Ultrasound. *Respiration* **90**, 499–506 (2015).
17. Riegler, J. *et al.* Tumor Elastography and Its Association with Collagen and the Tumor Microenvironment. *Clin. Cancer Res.* **24**, 4455 (2018).
18. Moon, W. K., Chang, R.-F., Chen, C.-J., Chen, D.-R. & Chen, W.-L. Solid breast masses: classification with computer-aided analysis of continuous US images obtained with probe compression. *Radiology* **236**, 458–464 (2005).
19. Lyshchik, A. *et al.* Thyroid gland tumor diagnosis at US elastography. *Radiology* **237**, 202–211 (2005).

20. Wu, J. *et al.* Diagnostic value of endobronchial ultrasound elastography for differentiating benign and malignant hilar and mediastinal lymph nodes: a systematic review and meta-analysis. *Med. Ultrason.* (2021).
21. Izumo, T., Sasada, S., Chavez, C., Matsumoto, Y. & Tsuchida, T. Endobronchial Ultrasound Elastography in the Diagnosis of Mediastinal and Hilar Lymph Nodes. *Jpn. J. Clin. Oncol.* **44**, 956–962 (2014).
22. Sun, J. *et al.* Endobronchial Ultrasound Elastography for Evaluation of Intrathoracic Lymph Nodes: A Pilot Study. *Respiration* **93**, 327–338 (2017).
23. Korrungruang, P. & Boonsarngsuk, V. Diagnostic value of endobronchial ultrasound elastography for the differentiation of benign and malignant intrathoracic lymph nodes. *Respirology* **22**, 972–977 (2017).
24. Verhoeven, R. L. J., de Korte, C. L. & van der Heijden, E. H. F. M. Optimal Endobronchial Ultrasound Strain Elastography Assessment Strategy: An Explorative Study. *Respir. Int. Rev. Thorac. Dis.* **97**, 337–347 (2019).
25. Uchimura, K. *et al.* Quantitative analysis of endobronchial ultrasound elastography in computed tomography-negative mediastinal and hilar lymph nodes. *Thorac. Cancer* **11**, 2590–2599 (2020).
26. Ma, H. *et al.* Semi-quantitative Analysis of EBUS Elastography as a Feasible Approach in Diagnosing Mediastinal and Hilar Lymph Nodes of Lung Cancer Patients. *Sci. Rep.* **8**, 3571 (2018).
27. Kim, E. S. & Bosquée, L. The importance of accurate lymph node staging in early and locally advanced non-small cell lung cancer: an update on available techniques. *J. Thorac. Oncol.* **2**, S59–S67 (2007).

28. Churchill, I. F. *et al.* An Artificial Intelligence Algorithm to Predict Nodal Metastasis in Lung Cancer. *Submitt. Ann. Thorac. Surg.*
29. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data TKDD* **6**, 1–21 (2012).
30. Mustafa Basil *et al.* Deep Ensembles for Low-Data Transfer Learning. *2020* doi:https://doi.org/10.48550/arXiv.2010.06866.
31. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*. vol. 112 (Springer, 2013).
32. Statistics, I. S. IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. *Google Search* (2013).
33. Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv. CSUR* **27**, 326–327 (1995).
34. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **5**, 1315–6 (2010).
35. Yong, S. H. *et al.* Malignant thoracic lymph node classification with deep convolutional neural networks on real-time endobronchial ultrasound (EBUS) images. *2022* **11**, 14–23 (2022).
36. Verhoeven, R. L. J. *et al.* Predictive value of endobronchial ultrasound strain elastography in mediastinal lymph node staging: the E-predict multicenter study results. *Respiration* **99**, 484–492 (2020).
37. Divisi, D., Zaccagna, G., Barone, M., Gabriele, F. & Crisci, R. Endobronchial ultrasound-transbronchial needle aspiration (EBUS/TBNA): a diagnostic challenge for mediastinal lesions. *Ann. Transl. Med.* **6**, (2018).

38. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, 2015).
39. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform.* **48**, 193–204 (2014).
40. Olympus America Inc. Elastography Quick Reference. (2016).
41. Mower, W. R. Evaluating bias and variability in diagnostic test reports. *Ann. Emerg. Med.* **33**, 85–91 (1999).

Figures

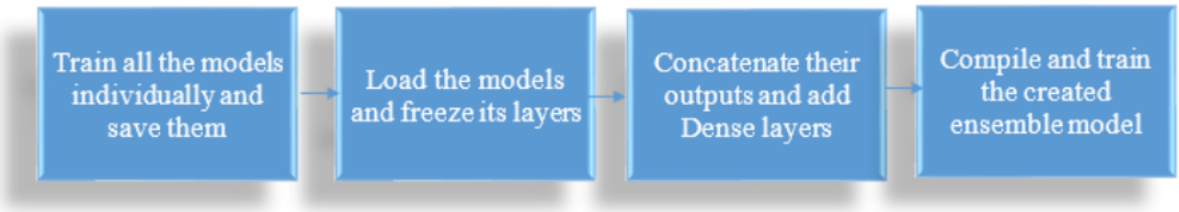


Figure 1. Visual flow chart of the different steps of creating an ensemble model

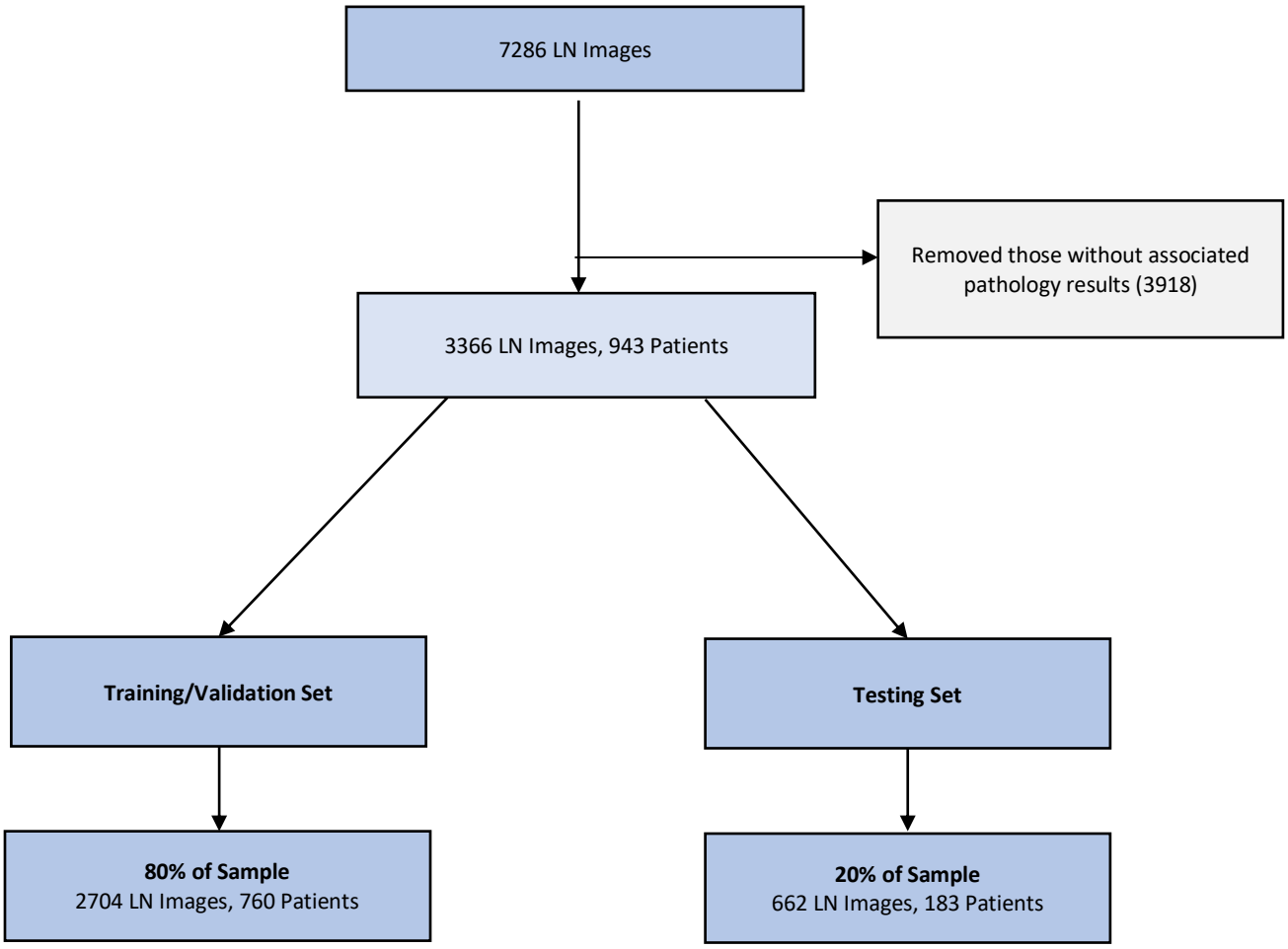


Figure 2. Flow diagram of the organization of the patients into the training/validation and the testing set

Table 1. Patient and LN demographics for the training/validation and the testing set

Variable	Training Set Patients (n=760) Lymph Nodes (n=2704)	Testing Set Patients (n=183) Lymph Nodes (n=662)	P-Value
Age (years) [mean ± SD]	67.40 ± 11.48	68.12 ± 11.13	0.467
Males, n (%)	472 (59.5)	109 (59.6)	0.982
Smoking Status, n (%)			0.869
Never Smoked	65 (8.6)	15 (8.2)	
Former Smoker	156 (20.5)	33 (18.0)	
Smoker	97 (12.8)	23 (12.6)	
Unknown	442 (58.2)	112 (61.2)	
LN Pathology, n (%)			0.01
Benign	1442 (53.3)	358 (54.1)	
Malignant	614 (22.7)	155 (23.4)	
Inconclusive	268 (9.9)	84 (12.7)	
No Pathology	380 (14.1)	65 (9.8)	
LN Station, n (%)			0.056
7	896 (33.1)	208 (31.4)	
4R	840 (31.1)	227 (34.3)	
4L	606 (22.4)	128 (19.3)	
10	78 (2.9)	14 (2.1)	
11	68 (2.5)	17 (2.6)	
Other (1,2,3,5,8,12)	216 (8.0)	68 (10.3)	
Short Axis Measurement (mm) [mean ± SD]	8.91 ± 5.13	9.20 ± 5.01	0.236
Long Axis Measurement (mm) [mean ± SD]	13.26 ± 5.11	13.53 ± 5.13	0.284
Ultrasound Malignancy Features, n (%)			
Short Axis (>10mm)	455 (20.5)	125 (18.9)	0.315
Margins (defined)	428 (19.3)	108 (16.3)	0.982
CHS (absent)	397 (17.9)	106 (16.0)	0.555
Central Necrosis (present)	171 (7.7)	59 (8.9)	0.026
Canada LN Score			0.959
0	1041 (46.9)	252 (38.1)	
1	312 (14.0)	81 (12.2)	
2	198 (8.9)	46 (6.9)	
3	235 (10.6)	62 (9.4)	
4	164 (7.4)	42 (8.7)	

Table 2. The diagnostic statistics for the training and validation set for the ResNet Model

ResNet	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Training	72.02% (67.91% to 75.87%)	29.68% (22.62% to 37.53%)	90.45% (86.91% to 93.30%)	57.50% (47.53% to 66.89%)	74.71% (72.62% to 76.69%)	0.601 (0.545 to 0.657)
Validation	72.74% (70.75% to 74.65%)	34.75% (30.88% to 38.66%)	88.90% (87.16% to 90.47%)	57.10% (52.60% to 61.49%)	76.20% (75.09% to 77.29%)	0.618 (0.590 to 0.646)

Table 3. The diagnostic statistics for the training and validation set for the Inception Model

Inception	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Training	77.30% (73.42% to 80.86%)	40.00% (32.22% to 48.17%)	93.54% (90.46% to 95.86%)	72.94% (63.46% to 80.71%)	78.17% (75.84% to 80.33%)	0.668 (0.612 to 0.723)
Validation	77.26% (75.39% to 79.06%)	42.90% (38.95% to 46.93%)	91.88% (90.35% to 93.24%)	69.21% (64.88% to 73.23%)	79.09% (77.91% to 80.23%)	0.674 (0.646 to 0.701)

Table 4. The diagnostic statistics for the training and validation set for the DenseNet Model

DenseNet	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Training	80.04% (76.31% to 83.42%)	49.03% (40.93% to 57.18%)	93.54% (90.46% to 95.86%)	76.77% (68.32% to 83.14%)	80.83% (78.28% to 83.14%)	0.713 (0.659 to 0.766)
Validation	78.97% (77.14% to 80.71%)	47.31% (43.30% to 51.35%)	92.44% (90.95% to 93.75%)	72.68% (68.56% to 76.45%)	80.48% (79.25% to 81.66%)	0.699 (0.672 to 0.726)

Table 5. The diagnostic statistics for the training and validation set for the Final Ensemble Model

Final Ensemble	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**
Testing Set	80.63% (76.93% to 83.97%)	43.23% (35.30% to 51.41%)	96.91% (94.54% to 98.45%)	85.90% (76.81% to 91.80%)	79.68% (77.34% to 81.83%)	0.701 (0.646 to 0.755)
Validation	78.48% (76.64% to 80.24%)	40.13% (36.22% to 44.13%)	94.80% (93.52% to 95.88%)	76.64% (72.06% to 80.67%)	78.82% (77.70% to 79.90%)	0.675 (0.647 to 0.702)

Table 6. The diagnostic statistics for the training and validation set for the CLNS

CLNS	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**
	79.38% (77.43 to 81.24%)	81.56% (77.88% to 84.87%)	78.54% (76.19% to 80.75%)	59.59% (56.85% to 62.27%)	91.65% (90.11% to 92.97%)	0.801 (0.777 to 0.824)

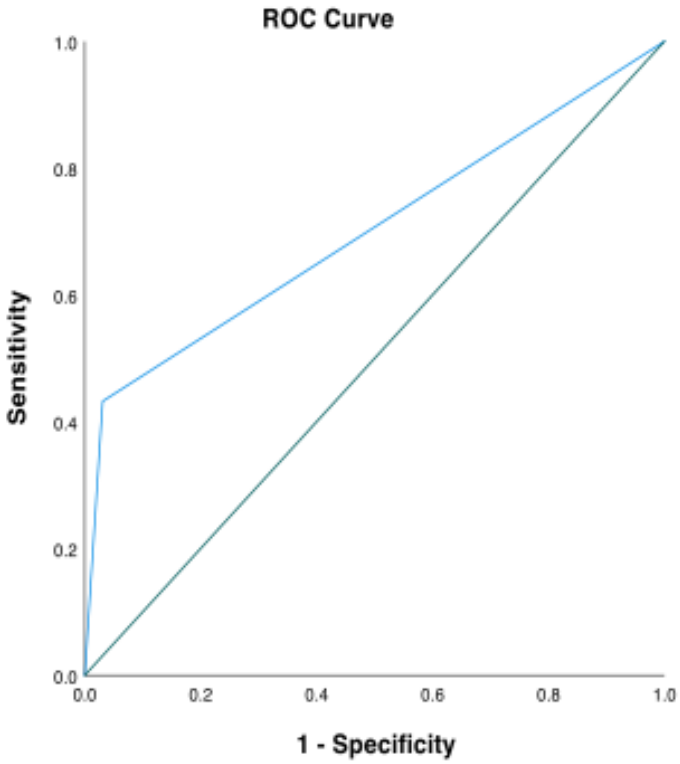


Figure 3. The ROC curve for the Testing Set using the Ensemble Model

Chapter 3: Lung Cancer Nodal Staging using EBUS-Elastography and AI: A Pilot Study

Nikkita Mistry, BSc, Anthony A. Gatti, PhD,Yogita S. Patel, BSc, Forough Farrokhyar, MPhil, PhD, Feng Xie, MSc, PhD, Isabella F. Churchill, MSc, Waël C. Hanna, MDCM, MBA

Abstract

Objectives: Elastography- the measurement of tissue stiffness- is a novel technology in the field of ultrasound. When used as an adjunct to endobronchial ultrasound (EBUS), elastography produces a colour map of the lymph node (LN), where the stiffer areas are depicted in blue and the softer areas in red. How this translates into information relating to cancer metastasis is still unknown. The primary aim of this study is to define the stiffness colour threshold (Blue Threshold) value that most accurately discriminates between malignant and benign LNs on elastography, as analyzed by a deep neural network. The secondary aim is to determine the stiffness area ratio (SAR) cut-off that can predict malignancy in any given LN.

Material and Methods: Thirty-one LNs were imaged using EBUS-Elastography, and then the images were fed to a trained deep neural network algorithm, which segmented the LNs and identified the percent of LN area above each of nine different stiffness colour thresholds. AUC was used to determine the optimal Blue Threshold. The choice of optimal SAR cut-off was also based on the ROC curve. Pathology results from nodal biopsies were considered the gold standard for determination of malignancy.

Results: The optimal Blue Threshold value was defined as level 60 on a color scale from 0-255, with an AUC of 0.891. The mean SAR was significantly different between benign and malignant

lymph nodes (0.37 vs. 0.67, p = 0.0002). The optimal SAR for predicting malignancy was 0.496, with a sensitivity of 92.3%, specificity of 76.5%, and overall accuracy of 83.3%.

Conclusions: A Blue Threshold of 60 allows for discrimination between benign and malignant LNs. A LN with SAR > 0.49 has a high probability of being malignant. These cut-offs will provide reference values for a future prospective study evaluating role of elastography in nodal staging.

Introduction

The current diagnostic pathway for non-small cell lung cancer (NSCLC) includes the staging of mediastinal lymph nodes (LNs). The nodal stage dictates the extent of the disease, the patient's prognosis, and treatment decisions.[1] The current standard for mediastinal LN staging is Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA). Despite multiple advances in EBUS-TBNA technology, as much as 40% of nodal biopsies remain inconclusive, and this often leads to repeat biopsies, delays in treatment, or in extreme cases, incorrect treatment decisions.[2] This low yield has prevented the widespread uptake of EBUS-TBNA, with many practitioners opting to omit mediastinal staging altogether. In a study conducted on 2,916 lung cancer patients between 2009 and 2018, only 22% received guideline-concordant LN staging.[3] As such, imaging adjuncts which can facilitate or improve the yield of EBUS-TBNA may have a positive impact on the diagnostic pathway of lung cancer patients.

Elastography, which has been explored in breast, thyroid, and lung cancer, is an emerging technology in the field of ultrasound oncology.[4] Elastography produces a colour map to visualize tissue elasticity, where the stiffer areas are depicted in blue, and the softer areas in red (Figure 1).[4] Malignant tissues, which contain more cells per area than normal tissues, are stiffer, and usually depicted in blue on elastography maps.[5] In the realm of LN staging, malignant LNs are expected to appear predominantly blue, whereas benign LNs will appear predominantly green or red.[6]

Elastography is promising in terms of its capacity to discriminate between benign and malignant LNs. Chen and colleagues (2018) completed a meta-analysis of seven studies that analyzed elastography patterns for the diagnosis of malignant intrathoracic LNs.[4] They reported a pooled sensitivity and specificity of 93% and 85%, respectively.[4] Of the different quantitative and qualitative methods used to analyze elastography images, stiffness area ratio (SAR) appeared to be the most intuitive. SAR is the proportion of blue relative to all the colors seen in the LN region of interest (ROI).[7] However, for the SAR to be determined adequately, one must first define what is the color threshold for "blue". In other words, of all the shades of blue, the one shade of blue that will be used as the numerator in the SAR, has not yet been quantitatively and reproducibly defined.[6,7,8]

In this work, we aimed to define the color threshold of "blue" that should be used to compute the SAR. We also defined the optimal SAR that can be used to differentiate benign LNs from malignant LNs.

Materials and Methods

Study Design

In this prospective pilot study, we collected and analyzed EBUS-Elastography LN images with their accompanying B-mode ultrasound images that were taken during EBUS-TBNA procedures. Images were stored as Red Green Blue (RGB) Joint Photographic Experts Group (JPEG) images

and analyzed by the NeuralSeg algorithm, a convolutional neural network (CNN) that uses machine learning for feature extraction and imaging analysis.[8]

Study Subjects

Patients who were diagnosed with suspected or confirmed NSCLC requiring mediastinal LN staging with EBUS-TBNA were eligible for this study. No exclusion criteria were applied. Enrollment took place between June and September 2019. Patients were consented at the time of the EBUS-TBNA procedure, and enrollment of the study did not intervene with standard of care practices. Participants' enrollment ended at the same time as the EBUS-TBNA procedure.

Source of Data

Retrospective analysis of prospectively collected EBUS-Elastography images. During EBUS-TBNA procedures Elastography was conducted prior to biopsy and suitable static images were captured and then stored on an external hard drive.

Unit of Analysis

The unit of analysis for this study was the LNs rather than the patient, as the primary outcome is whether NeuralSeg could calculate the SAR directly from the EBUS-Elastography image.

EBUS-TBNA Procedure

EBUS-TBNA was performed using an Olympus convex probe ultrasound endoscope (Olympus, Shinjuku-ku, Tokyo, Japan) and the EU-ME2 plus transducer (Olympus) under conscious sedation with midazolam and fentanyl. The Canada Lymph Node Score (CLNS) was assigned to each LN

prior to biopsy by transbronchial needle aspiration with a 22-gauge needle.[9] Specimen adequacy was confirmed by rapid onsite cytology.

EBUS-Elastography

EBUS-Elastography was also performed prior to biopsy. Strain graphs were used to confirm stable pressurization. The LN was identified as the ROI with a 1:1 ratio to surrounding mediastinal tissue. The B-mode and EBUS-Elastography images were displayed side-by-side, as seen in Figure 2, and these images were then captured and stored as JPEG images.

Outcomes

The primary outcome is to test whether NeuralSeg can calculate SAR in a LN directly from the EBUS-TBNA image, in order to determine the blue colour threshold and the optimal SAR cut-ff to identify LN maligancy. The gold standard for comparison will be the final pathology results

Image Analysis

First, to get isolated Elastography colour maps, we subtracted the B-Mode image (Figure 2 Left) from the Elastography image (Figure 2 Right). This processing step was performed to minimize the effects of the underlying B-mode image on Elastography analysis and SAR predictions. The Blue colour channel of this isolated Elastography colour map is presented in Figure 3 (Right).

NeuralSeg (Hamilton, ON, Canada), an Artificial Intelligence (AI) deep neural network, has been trained and validated for LN segmentation and LN feature extraction.[10] The B-mode image was used by NeuralSeg to segment the LN from the surrounding tissue (Figure 3).[10]

NeuralSeg automatically identified the LN (ROI) based on several ultrasonographic features: hilar structure of the LN, necrosis within the LN, and LN contour.[10] The segmented LN on the B-mode image was overlaid onto the EBUS-Elastography images to extract the LN stiffness measurements. After overlaying, we determined the SAR as the proportion of the LN area that was above a defined threshold on the blue channel of the RGB image (Blue Threshold). To determine which Blue Threshold value was best for calculating SAR and ultimately predicting LN malignancy, 9 different threshold values were analyzed on a scale of 0-255. The analysis started at 10, and increased in increments of 10, until 90. To improve robustness of these predictions, LN segmentations and the 9 coinciding SARs were predicted 5 times and the median SAR was used as the outcome for each threshold. The 5 predictions were derived from the 5-folds of a previous cross-validation study.[10]

These predictions were used to identify the Blue Threshold for predicting malignancy. The computer programmer analyzing the elastography images was blinded to personal identifiers, and the clinical and pathology data of the images.

Data Analysis

Descriptive statistics were provided for the demographic data of the study subjects. All analyses were done using Statistical Package for the Social Sciences (SPSS) 2020 version.[11]

I. ***Optimal Stiffness Colour Threshold Analysis (Blue Threshold)***

For each LN, SAR was calculated using the 9 different Blue Thresholds. For each of the 9 SAR predictions, using the pathology results, a Receiver Operating Characteristic (ROC) curve was conducted and Area Under the Curve (AUC) with 95% confidence interval are reported. AUC values and p-values were compared and pairwise comparison was conducted between the 9 stiffness colour thresholds to determine which threshold had the highest diagnostic accuracy (AUC). All LN outcomes were compared to the final pathology results from LN biopsies and/or surgical specimens, indicating whether the LN was malignant or benign.

II. ***Stiffness Area Ratio Analysis***

Using the SAR obtained using the optimal Blue Threshold, the optimal SAR cut-off was determined based on the value that had the highest sensitivity and specificity on the ROC curve and the Youden Index (sensitivity+specifcity-1), a summary statistic for the accuracy of the datapoints ROC, where 0 indicates no diagnostic value and 1 indicates no false positives or negatives.[12] An index table was created and negative predictive value, positive predictive value, false positives, false negatives, overall accuracy, positive likelihood ratio and negative likelihood ratios were calculated. Mann Whitney U test, a non-parametric test, was also conducted to compare the SAR between the benign and malignant LNs.

Results

Patients and LNs

Sixteen patients were enrolled, and 31 LNs were collected and analyzed in total. Patient characteristics and LN pathology are summarized in Table 1. Of the patients enrolled, 68.75% (11/16) were female and the mean age was 67.94 ± 10.43 years. Only 18.75% (3/16) of patients were non-smokers, 43.75% (7/16) were smokers, and 37.50% (6/16) were ex-smokers. The pathology of each LN is summarized in Table 2. The median number of LNs per patient was 2. The majority of LNs were benign 54.84% (17/31); 41.94% (13/31) were malignant, and the pathology results could not be obtained for 3.23% (1/31) of the LNs. Of the malignant LNs, the diagnosis was adenocarcinoma for 46.15% (6/13), squamous cell carcinoma for 15.38% (2/13), poorly differentiated carcinoma for 7.69% (1/13), and 23.08% (3/13) were grouped as other. LNs were obtained from varying stations (Table 1) with the majority from station 4R: 25.81% (8/31) and station 7: 35.48% (11/31).

Optimal Stiffness Colour Threshold (Blue Threshold)

The results of the AUC analysis for all nine predefined colour thresholds are summarized in Table 3. No LNs had colour pixels over 76.5 on the 0 to 255 colour scale. Of all the colour threshold levels, 40, 50, and 60 had the highest AUC and the ROC curves were significantly different from the other threshold values. Threshold level 60 was chosen as the optimal stiffness colour threshold (Blue Threshold) as it had the highest AUC of 0.89 (95% CI: 0.77-1.00) with a p-value < 0.001.

Stiffness Area Ratio

Using a Blue Threshold of 60 from the previous analysis, Table 2 shows the SAR for each of the LNs, accompanied by the LN station and pathology results. Table 4 shows the descriptive statistics

for stiffness area ratio for the benign and malignant LNs. The benign LNs had a mean SAR of 0.37 ±0.19, whereas the malignant LNs had a mean SAR of 0.67 ±0.14 (p-value < 0.001). Using a ROC, as shown in Figure 4, and the Youden Index, a cut-off value of 0.49 was determined, as it had the best Youden Index of 0.69 and a highest sensitivity of 92.30% (95% CI: 62.10% to 99.60%) and a specificity of 76.50% (95% CI: 49.80% to 92.20%) in predicting LN malignancy. Based on the 0.49 cut-off value, the overall accuracy in predicting malignancy was ROC a positive predictive value of 75.00%, negative predictive value of 92.90%, positive likelihood ratio of 3.93 and a negative likelihood ratio of 0.10 (Table 5). Using the 0.49 cut-off, 4 LNs were falsely diagnosed positive for malignancy on elastography when they were benign on pathology, and one LN was falsely diagnosed negative.

Discussion

We demonstrated that deep learning can automatically calculate SAR from EBUS-Elastography and determine LN malignancy. We systematically tested and identified the optimal methods for predicting LN malignancy using EBUS-Elastography. We demonstrated that the optimal Blue Threshold value of 60 for determining LN SAR produces an excellent AUC of 0.89.[13] Using this Blue Threshold, the optimal SAR cut-off (0.49 produced a sensitivity of 92.30% and specificity of 76.50%. Additionally, a positive predictive value and negative predictive value of 75.00% and 92.90%, respectively, were obtained. It is important to note that these values are influenced by prevalence, so as we continue this work on different populations, this is subject to change. This work identified that EBUS-Elastography is a meaningful tool for staging LN in NSCLC and can be automated using deep learning.

Elastography is still an emerging technology in the field of lung cancer staging, and multiple different methods of color map and tissue stiffness analysis continue to be explored. Chen and colleagues (2018) conducted a systematic review outlining the different methods of qualitative and quantitative analysis for elastography images.[4] Qualitative analysis includes the 3- and 5-type classification methods, which are based on the estimation of the preponderance of the color blue in the LN image. The 3-type method uses "non-blue," "part-blue," and "predominantly blue" as its determinants.[6] The 5-type method uses a similar approach and adds homogeneity and heterogeneity as two additional determinants.[14] The quantitative analysis methods include strain ratio and SAR. Strain ratio is calculated as the ratio of the colour pattern of the ROI to a reference area surrounding the LN.[15] SAR is calculated as the percentage of stiff areas (blue) in the ROI, based on a colour hue threshold.[7] In a systematic review by Chen et al, it was shown that quantitative approaches have higher diagnostic capabilities compared to qualitative.[4] Moreover, the quantification of the visual colour patterns produced by elastography can aid in the identification of malignant LNs with high sensitivity and specificity.[4]

In this work, we aimed to define the "Blue Threshold," which is the value used to determine the numerator in the equation to calculate SAR. Comparing the AUC of the nine thresholds from our study, blue colour threshold level 60 had the highest AUC of 0.89, classified as an excellent diagnostic test, as the AUC falls between 0.8-0.9.[13] Of the multiple studies reporting on the correlation of elastography and nodal metastasis, only one study by Nakajima et al. defined a color hue threshold, albeit without ROC analysis.[7,16,17] To our knowledge, no research has tested which

Blue Threshold produces the highest diagnostic capability in discriminating malignant LNs from benign LNs.

Using the optimal Blue Threshold value of 60, we also determined a SAR threshold for predicting nodal metastasis. A significant difference was observed in SAR between malignant and benign LNs, 0.67 and 0.37 (p-value < 0.001), respectively, with a cut-off of 0.49 being identified as the optimal discriminatory cut-off between benign and malignant lymph nodes. Our results are comparable to existing literature that quantified elastography images using SAR. Nakajima and colleagues (2015) analyzed 49 LNs and got a cut-off value of 0.31 with 81% sensitivity and 85% specificity.[7] Ma and colleagues (2018) analyzed 79 LNs, had an AUC of 0.86 and a cut-off value of 0.37 and 92.30% sensitivity, 67.50% specificity, and 78.5% accuracy.[16] Lastly, Uchimura and colleagues (2020) analyzed 149 LNs, obtained a cut-off value of 0.41 and 88.2% sensitivity, 80.2% specificity, and 83.9% diagnostic accuracy.[17] All three of these research studies used either Image J software or photoshop to select the ROI and analyze the colour patterns. Our study used a trained machine learning algorithm to automatically identify the ROI, which was the LN, and analyzed the colour patterns.

A recent study conducted by Verhoeven and colleagues in 2019 assessed the performance of qualitative and quantitative methods of analyzing elastography colour maps.[18] The qualitative methods included visual analogue scale (VAS) the 5-type classification and the second method was the modified version of the Tsukuba score, which has six different classifications. The quantitative methods included the mean strain elastography histogram and the second method is strain ratio. [18]This study showed the mean strain elastography histogram with a cut off of 78 and

obtained the best and most reproducible results. The sensitivity of 93%, specificity of 75%, and overall accuracy of 82%. Verhoeven and colleagues (2020) took this work one step further by conducting a multicentre study using the mean strain elastography histogram method with a cut off of 115, the AUC was 0.77, the sensitivity was 90% and the specificity of 43%.[19] Although SAR is more intuitive quantitative method other methods such as mean strain elastography histogram should continue to be explored.

The simulation of human cognition by AI and deep learning computer neural networks has become a valuable tool in medicine. The amount of data in radiology imaging has increased. This information can be extracted by software in order to provide physicians with valuable clinical data.[20] Recent research has shown that AI and deep learning can be used to accurately interpret images, in comparison to clinicians in radiology, pathology, and cardiology.[11] Korrungruang and colleagues (2017) raised the concern of the lack of standardization in the elastography image analysis.[20]2The lack of human input in selecting the LN and obtaining the SAR in our study allows for greater reproducibility of this method, contributing to the standardization of the technique.

The main limitation of this study is the small sample size, which is not large enough to generalize the results to the population of patients with suspected or diagnosed NSCLC. However, this was an exploratory proof of concept pilot study, and its results will be validated in a larger prospective trial (NCT04816981).

Conclusion

Deep learning was applied to EBUS-Elastography to automatically calculate SAR at nine predefined blue thresholds. A Blue Threshold of 60 and a SAR of 0.49 are optimal cut-offs to distinguish between benign and malignant LNs. Further prospective work with a larger sample size will be required to validate the findings of this pilot study.

References

1. De Leyn, P. *et al.* Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer†. *Eur. J. Cardiothorac. Surg.* **45**, 787–798 (2014).
2. Ortakoylu, M. G. *et al.* Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *J. Bras. Pneumol.* **41**, 410–414 (2015).
3. Osarogiagbon, R. U. *et al.* Invasive mediastinal staging for resected non–small cell lung cancer in a population-based cohort. *J. Thorac. Cardiovasc. Surg.* **158**, 1220-1229.e2 (2019).
4. Chen, Y.-F. *et al.* Endobronchial Ultrasound Elastography Differentiates Intrathoracic Lymph Nodes: A Meta-Analysis. *Ann. Thorac. Surg.* **106**, 1251–1257 (2018).
5. Riegler, J. *et al.* Tumor Elastography and Its Association with Collagen and the Tumor Microenvironment. *Clin. Cancer Res.* **24**, 4455 (2018).
6. Izumo, T., Sasada, S., Chavez, C., Matsumoto, Y. & Tsuchida, T. Endobronchial Ultrasound Elastography in the Diagnosis of Mediastinal and Hilar Lymph Nodes. *Jpn. J. Clin. Oncol.* **44**, 956–962 (2014).

7. Nakajima, T. *et al.* Elastography for Predicting and Localizing Nodal Metastases during Endobronchial Ultrasound. *Respiration* **90**, 499–506 (2015).
8. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, 2015).
9. Hylton, D. A. *et al.* The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. *J Thorac Cardiovasc Surg* **159**, 2499-2507.e3 (2020).
10. Churchill, I. F. *et al.* An Artificial Intelligence Algorithm to Predict Nodal Metastasis in Lung Cancer. *Submitt. Ann. Thorac. Surg.*
11. IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp.
12. Shan, G. Improved confidence intervals for the Youden index. *PloS One* **10**, e0127272 (2015).
13. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **5**, 1315–6 (2010).
14. Sun, J. *et al.* Endobronchial Ultrasound Elastography for Evaluation of Intrathoracic Lymph Nodes: A Pilot Study. *Respiration* **93**, 327–338 (2017).
15. Hernández Roca, M. *et al.* Diagnostic Value of Elastography and Endobronchial Ultrasound in the Study of Hilar and Mediastinal Lymph Nodes. *J. Bronchol. Interv. Pulmonol.* **26**, 184–192 (2019).

16. Ma, H. *et al.* Semi-quantitative Analysis of EBUS Elastography as a Feasible Approach in Diagnosing Mediastinal and Hilar Lymph Nodes of Lung Cancer Patients. *Sci. Rep.* **8**, 3571 (2018).

17. Uchimura, K. *et al.* Quantitative analysis of endobronchial ultrasound elastography in computed tomography-negative mediastinal and hilar lymph nodes. *Thorac. Cancer* **11**, 2590–2599 (2020).

18. Verhoeven, R. L. J., de Korte, C. L. & van der Heijden, E. H. F. M. Optimal Endobronchial Ultrasound Strain Elastography Assessment Strategy: An Explorative Study. *Respir. Int. Rev. Thorac. Dis.* **97**, 337–347 (2019).

19. Verhoeven, R. L. J. *et al.* Predictive value of endobronchial ultrasound strain elastography in mediastinal lymph node staging: the E-predict multicenter study results. *Respiration* **99**, 484–492 (2020).

20. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* **316**, 2353–2354 (2016).

21. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).

22. Korrungruang, P. & Boonsarngsuk, V. Diagnostic value of endobronchial ultrasound elastography for the differentiation of benign and malignant intrathoracic lymph nodes. *Respirology* **22**, 972–977 (2017).

Appendix

Table 1. Baseline Patient and Lymph Node Characteristics.

Patients	n = 16
Gender	
Male	5
Female	11 (68.75%)
Age, years	
Average	69.56 ± 10.43
Range	53-89
Smoking Status	
Non-Smoker	3 (18.75%)
Smoker	7 (43.75%)
Ex-Smoker	6 (37.50%)
Median LN per Patient	2 (1-3)
Pathology Diagnosis	
Benign	4
Malignant	
Adenocarcinoma	6
Squamous Cell Carcinoma	2
Poorly Differentiated Carcinoma	1
Other	3
LNs Station	
2R	2
4L	4
4R	9
7	11
8L	1
10R	1
11R	2
11L	2

Table 2. SAR and Pathology of Examined LNs by EBUS-Elastography.

LN Number	LN Station	Stiffness Area Ratio	Malignant (Y/N)
1	2R	0.32	-
2	4L	0.19	N
3	10R	0.62	Y
4	4R	0.5	Y
5	11R	0.85	Y
6	7	0.74	Y
7	4L	0.36	N
8	7	0.37	N
9	11L	0.67	Y
10	7	0.79	N
11	7	0.39	N
12	2R	0.81	Y
13	11R	0.68	N
14	7	0.41	Y
15	4R	0.20	N
16	4R	0.87	Y
17	4R	0.75	Y
18	4L	0.17	N
19	7	0.49	N
20	7	0.75	Y
21	4R	0.12	N
22	7	0.40	N
23	4R	0.17	N
24	4R	0.53	Y
25	8L	0.61	Y
26	11L	0.58	Y
27	4R	0.54	N
28	7	0.26	N
29	7	0.43	N
30	7	0.25	N
31	4L	0.51	N

Table 3. Results of AUC for 9 Predefined Blue Colour Thresholds.

Colour Threshold	AUC	Std. Error	Significance	95% CI
10	0.837	0.076	0.002	(0.689-0.985)
20	0.842	0.075	0.002	(0.695-0.989)
30	0.864	0.070	0.001	(0.727-1.000)
40	0.878	0.065	0.000	(0.750-1.000)
50	0.878	0.065	0.000	(0.750-1.000)
60	0.891	0.060	0.000	(0.774-1.000)
70	0.552	0.109	0.630	(0.339-0.765)
80	0.385	0.105	0.286	(0.178-0.591)
90	0.676	0.104	0.103	(0.473-0.880)

Table 4. Descriptive Statistics for Stiffness Area Ratio for Benign and Malignant LNs.

Pathology	N	Minimum	Maximum	Mean	Std. Deviation
Benign	17	0.12	0.79	0.37	0.19
Malignant	13	0.41	0.87	0.67	0.14

P= 0.0002, Mann Whitney U Test

Table 5. Diagnostic Statistics based on SAR cut off.

Sensitivity	Specificity	Overall Accuracy	Positive Predictive Value	Negative Predictive Value	Positive Likelihood Ratio	Negative Likelihood Ratio
92.30	76.50	83.30	75.0	92.90	3.93	0.10

Figures

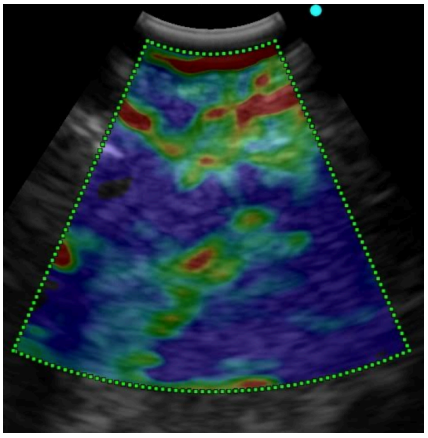


Figure 1. An example image, the B-mode image is in black and white, with an overlaid elastography colour map. Blue represented stiffer tissues and red represents softer tissues.

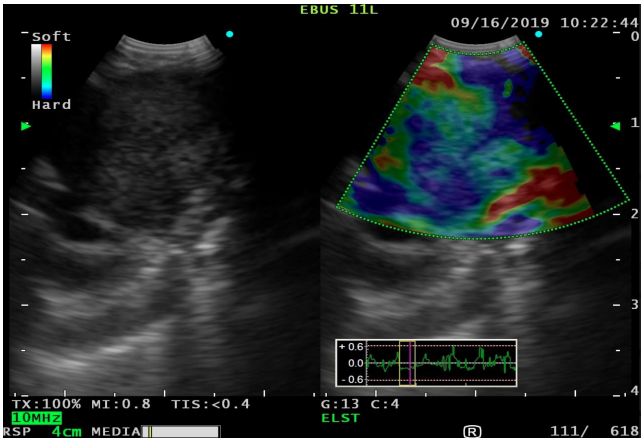


Figure 2. Image of B-mode Ultrasound (left) and EBUS-Elastography Colour Map (right). The B-mode image was used for segmentation, and the segmentation was then overlaid onto the colour map to extract the relevant stiffness information.

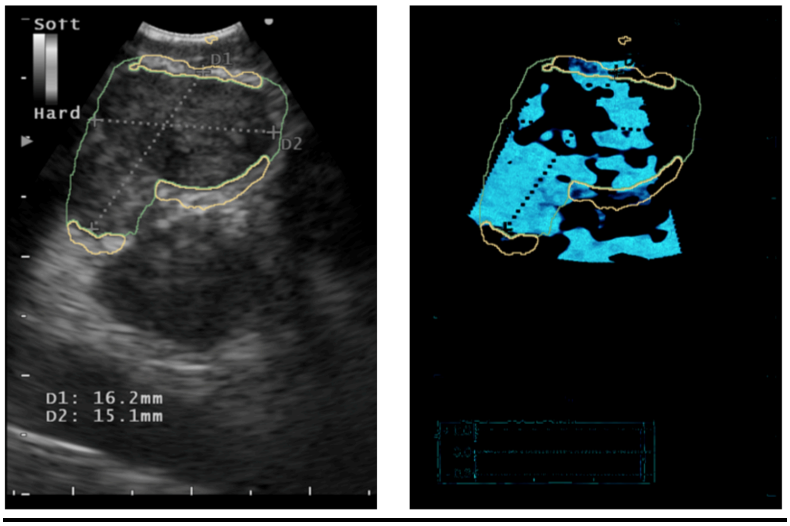


Figure 3. The left image shows the automated segmentation of the B-mode image using deep learning. The right image shows the blue channel of the isolated Elastography colourmap with the segmentation from the B-mode image overlaid. The isolated colourmap was produced by subtracting the B-mode image from the EBUS-Elastography colour map.

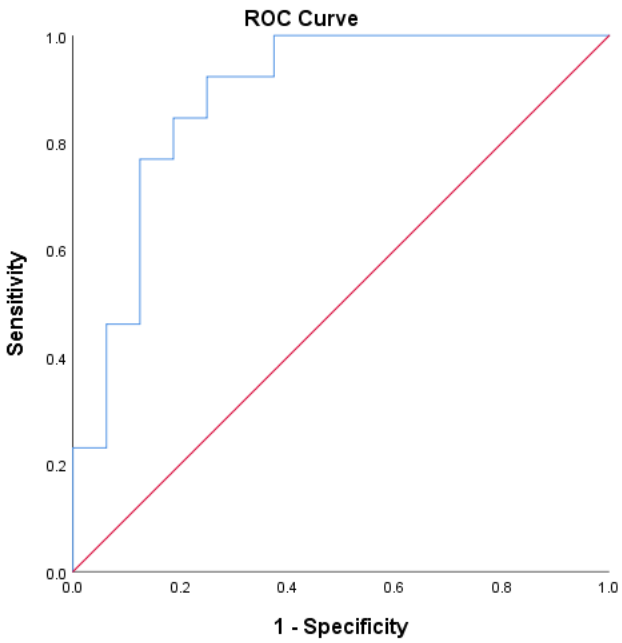


Figure 4. ROC for Optimal Stiffness Colour Threshold at Level 60

Chapter 4: Clinical Utility of Artificial Intelligence-Augmented Endobronchial Ultrasound Elastography in Lymph Node Staging for Lung Cancer

Nikkita Mistry, BSc, Anthony A. Gatti, PhD,Yogita S. Patel, BSc, Forough Farrokhyar, MPhil, PhD, Feng Xie, MSc, PhD, Waël C. Hanna, MDCM, MBA

Abstract

Background: The current mediastinal LN staging guideline for Non-Small Cell Lung Cancer (NSCLC) is Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA). Despite advances in staging technology, as high as 42% of biopsy results come back inconclusive. Elastography is non-invasive and uses strain technology in response to mechanical stress to create visual colour maps representing tissue stiffness. Red on the colour map represents soft tissue, and blue represents stiff tissue. A previous pilot study aimed to define the blue threshold required to calculate the SAR and determine the optimal SAR to distinguish benign and malignant LNs. Level 60 was chosen as the blue threshold based on the AUC of 0.891. Additionally, an optimal SAR of 0.496 was determined, and it had a sensitivity of 92.3%, specificity of 76.5%, and overall accuracy of 83.3%.

Objective: This study aims to validate the previously determined stiffness area ratio cut-off, 0.496, determined from the pilot study and to determine if it is predictive of malignancy compared to the surgical pathology.

Methods: This is a single-centre, prospective clinical trial where 210 EBUS-Elastography and B-mode Ultrasound LN images were collected during EBUS-TBNA procedures. These images were fed to a trained deep neural network algorithm, which segmented the LNs and identified the percent of LN area above threshold level 60. LN with a SAR above 0.496 were assigned malignant, and below 0.496 were benign. In addition, diagnostic statistics and ROC analysis were conducted. Pathology results from nodal biopsies were considered the gold standard for determining malignancy.

Results: There were 98 true negatives, 34 true positives, 10 false positives, and 45 false negatives. This resulted in an overall accuracy of 70.59% (95% Confidence Interval (CI) 63.50% to 77.01%), sensitivity of 43.04% (CI: 31.94% to 54.67%), a specificity of 90.74% (CI: 83.63% to 95.47%), a positive predictive value (PPV) of 77.27% (64.13% to 86.60%) and a negative predictive value (NPV) of 68.53% (64.05% to 72.70%). The AUC for level 60 was 0.820 (CI:0.758-0.883).

Conclusion: In this study, NeuralSeg, an artificial intelligence deep neural network, was successfully applied to Elastography images, and the SAR cut-off of 0.4959 at level 60 determined from a previous pilot study was validated. However, more extensive multi-centre studies must be conducted to standardize this process and optimize the algorithm.

Introduction

Lung cancer has the highest mortality among all cancers, making up 25.5% of all cancer deaths.[1] The survival rates for lung cancer differ significantly based on cancer stage, where stage 1A cancer has a 49% survival rate, and stage 4 has a 1% survival rate.[2] This shows how crucial the staging process is in the lung cancer diagnostic pathway. The current mediastinal LN staging guideline for Non-Small Cell Lung Cancer (NSCLC) is Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA). EBUS-TBNA is a safe procedure that causes minimal patient discomfort and has comparable diagnostic abilities compared to more invasive techniques such as mediastinoscopy.[3] Despite advances in staging technology, as high as 42% of biopsy results come back inconclusive. This leads to patients having to get repeat biopsies, treatment delays and possibly even incorrect treatment decisions.[4] This low yield seems to be discouraging physicians from conducting staging procedures. In a study conducted on over 2000 lung cancer patients, only 22% received mediastinal LN staging.[5]

Aside from the lack of utilization and low yield of EBUS-TBNA, relying on B-mode imaging alone remains challenging to predict malignancy. There is an emerging ultrasound technology called Elastography.[6] Elastography is a non-invasive and uses strain technology in response to mechanical stress to create visual colour maps that represent tissue stiffness. Red on the colour map represents soft tissue, and blue represents stiff tissue. [6] How Elastography can be applied to mediastinal staging is malignant tissue tends to be stiffer in nature, as there are more cells per area compared to benign.[7] This technology has been explored in the fields of breast cancer and thyroid cancer, achieving high sensitivities and specificities.[8,9]

A systematic review by Wu and colleagues evaluated EBUS-Elastography in differentiating benign and malignant LNs. This study included 2307 LNs and obtained a pooled sensitivity of 0.90 (95% CI, 0.84-0.94), pooled specificity of 0.78 (95% CI, 0.74-0.81) and AUC of 0.86 (95% CI, 0.82-0.88).[10] Different qualitative and quantitative methods were used to analyze the Elastography colour maps. For qualitative methods, images are observed and categorized using either a 3-type or 5-type classification method. [11,12] These methods are often subjective and dependent on the interpreter. Quantitative methods include strain ratio, stiffness area ratio (SAR), and strain histogram, which tend to be far more reproducible than qualitative methods.[13,6,14] Of these methods, the most intuitive is the SAR, as it is calculated based on the number of blue pixels compared to all the colour pixels in the region of interest (ROI), the LN.[6] However, this method has yet to become standardized as the numerator of the SAR, and the colour threshold of "blue" has not been defined.

This work builds on a previous pilot study which sought to define the blue threshold required to calculate the SAR and determine the optimal SAR used to distinguish benign and malignant LNs. Several blue colour thresholds were tested; level 60 from the 0-255 colour scale had the highest AUC of 0.891. Additionally, an optimal SAR of 0.496 was determined, and it had a sensitivity of 92.3%, specificity of 76.5%, and overall accuracy of 83.3%. The pilot study used a trained machine learning algorithm to automatically identify the ROI, which was the LN and analyzed the colour patterns. This study aims to validate the previously determined stiffness area ratio cut-off, 0.496, determined from the pilot study and to determine if it is predictive of malignancy compared to the surgical pathology.

Methods

Study Design

This is a single-centre, prospective clinical trial where EBUS-Elastography and B-mode Ultrasound LN images were collected at the time of the EBUS-TBNA procedures (NCT04816981). Images were stored on a hard drive as Red Green Blue (RGB) Joint Photographic Experts Group (JPEG) images. Data were prospectively collected, including patient demographics, LN station, and LN pathology. LN malignancy predictors include short-axis diameter, contour definition, central hilar structure absence, and central necrosis presence, also known as the Canada Lymph Node Score (CLNS). A convolutional neural network (CNN) that uses machine learning for feature extraction and imaging analysis, NeuralSeg, was used to analyze the images.[15]

Study Subjects

Patients undergoing EBUS-TBNA for confirmed or suspected NSCLC were eligible for this study. There were no exclusion criteria that were applied. Enrollment occurred from August 2021 to May 2022. Patients were consented prior to the EBUS-TBNA procedure, and enrollment in the study did not intervene with the standard of care practices. Patients were enrolled in a consecutive sample, and patient involvement concluded when the procedure ended. No follow-up was required after the study.

Source of Data

The dataset comes from a prospectively collected LN database. LN images were collected from St. Joseph's Healthcare Hamilton from August 2021 to May 2022. During EBUS-TBNA

procedures, suitable static images were captured and stored on an external hard drive. In addition, information such as the patient's age, gender, LN station, date and time of the procedure was also recorded.

Unit of Analysis

The unit of analysis for this study was the LNs rather than the patient, as the primary outcome is whether NeuralSeg could calculate the SAR directly from the EBUS-Elastography image.

Sample Size

This sample size was determined by the following calculation based on sensitivity: elastography for lymph node imaging is associated with a sensitivity of 92.3% Assuming this diagnostic value, and a prevalence of 18% malignancy in our population, an alpha of 0.05, and a marginal error of 0.085 with a 95% confidence interval, a sample size of 210 lymph nodes would be needed.[16]

EBUS-TBNA Procedure

An Olympus convex probe ultrasound endoscope (Olympus, Shinjuku-ku, Tokyo, Japan) and the EU-ME2 plus transducer (Olympus) was used to perform EBUS-TBNA under conscious sedation with midazolam and fentanyl. An expert endosonographer assigned the Canada Lymph Node Score (CLNS) before conducting the biopsy by transbronchial needle aspiration using a 22-gauge needle.[17]

EBUS-Elastography

Prior to the biopsy, EBUS-Elastography was performed. The region of interest (ROI) is the LN and should be a 1:1 ratio with the surrounding mediastinal tissue. The strain graph is used to confirm stable pressurization, as the wave should be between -0.06 and +0.06, as shown in Figure 1. The B-mode and Elastography images were displayed side-by-side, as seen in Figure 2. A suitable static image was captured and stored on an external hard drive.

Outcomes

The primary outcome is to test whether NeuralSeg can calculate SAR in a LN directly from the EBUS-TBNA image, in order to validate the findings from chapter 3. The gold standard for comparison will be the final pathology results

Image Analysis

The computer programmer training, NeuralSeg, an artificial intelligence (AI) deep neural network, was blinded to personal identifiers and the pathology and clinical data of the images. Images were captured and stored with the B-mode and Elastography image side by side. NeuralSeg was previously trained and validated to segment LNs and extract LN features.[18] The features include the hilar structure of the LN, necrosis within the LN, and LN contour.[17] NeuralSeg segmented the LN from the surrounding tissue on the B-mode side of the image. Then the segmentation was overlaid on the elastography side to extract the colour information from the LN.

The Elastography image was separated from the B-mode images to isolate the colour map to minimize the effects the B-mode image had on the SAR analysis. We focused on the blue channel of the RBG image to identify the optimal Blue threshold needed to calculate the SAR. We were able to get the proportion of LN area (SAR) above nine different thresholds on the 0-255 colour scale. The nine different thresholds ranged from 10 through 90, increasing in increments of 10.

Using a 5-fold cross-validation technique derived from a previous study, five SAR were obtained for each of the nine different thresholds.[18] The mean of these five predictions was used as the final SAR. This was done to improve the robustness of the results.

Results

Patients and Lymph Nodes

There were 124 patients enrolled, and 210 LN images were collected and analyzed. In Table 1, the patient demographics and LN characteristics are summarized. The mean age of the enrolled patients was 69.83± 9.95 years, and 51.6% (64/124) were male. In terms of smoking status, 8.1% (10/124) of patients never smoked, 43.5% (54/124) were former smokers, 35.5% (44/124), and 9.7% (12/124) the status was unknown. The mean body mass index (BMI) was 26.69± 6.14 kg/m^2. Common comorbidities among the patients include Chronic Inflammatory Lung Disease (COPD) for 24.2% (30/124), Atrial Fibrillation for 9.7% (12/124), Hypertension for 41.1% (51/124), and GERD for 8.1% (10/124). In addition, of the enrolled patients, 23.4% (29/124) had previous cancer, and 72.6% (90/124) did not.

The majority of the LNs analyzed were benign 53.1% (122/210), 37.4% (79/210), 8.1% (17/210) had inconclusive results, and the pathology could not be obtained for 0.9% (2/210). The majority of the LN were obtained from station 7: 38.4% (81/210), station 4R: 32.7% (69/210), and station 4L: 15.6% (33/210). The mean LN short axis measurement was 11.01 ± 5.75 mm, and the mean long axis measurements were 15.70 ± 6.00 mm. The Canada Lymph Node Score (CLNS) was assigned to each LN. The CLNS is a binary scoring system based on LN small-axis diameter, central hilar structure, central necrosis, and margin status.[17] A point was assigned to 46.4% (98/210) for the short axis diameter being larger than 10mm, 48.3% (102/210) for defined margins, 33.6% (71/210) for the absence of central hilar structure, and 20.9% (44/210) for the presence of central necrosis. The overall CLNS was 0 for 36.5% (77/210), 1 for 18.0% (38/210), 2 for 11.8% (25/210), 3 for 21.8% (46/210), and 4 for 10.4% (22/210).

Validation of the Stiffness Area Ratio Cut Off

Based on the SAR at level 60, those LN with a SAR greater than 0.4959 was assigned malignant, and those with a SAR below 0.4959 were assigned benign. These results were compared to pathology results, shown in Figure 3. There were 98 true negatives, 34 true positives, 10 false positives, and 45 false negatives. This resulted in an overall accuracy of 70.59% (95% Confidence Interval (CI) 63.50% to 77.01%), sensitivity of 43.04% (CI: 31.94% to 54.67%), a specificity of 90.74% (CI: 83.63% to 95.47%), a positive predictive value (PPV) of 77.27% (64.13% to 86.60%) and a negative predictive value (NPV) of 68.53% (64.05% to 72.70%), shown in Table 2. The ROC for level 60 can be seen in Figure 4. The mean SAR for the benign

LN was 0.25±0.17, and the mean SAR for malignant LNs was 0.45 ±0.15, and there was a significant difference between the two groups, $p<0.001$. The results can be seen in Table 3.

Secondary Analysis

Stiffness Colour Threshold (Blue Threshold)

The AUC analysis results can be seen in table __ for nine different threshold values between 10 and 90, increasing in increments of 10. There were no LNs that had a proportion higher than 76.5, therefore, these nine thresholds were used. Level 40, 50, and 60 had the highest AUC for the threshold values, 0.816 (CI: 0.754-0.878), 0.823 (CI: 0.762-0.884), 0.820 (CI:0.758-0.883), respectively. The AUC for each of the nine threshold levels can be seen in Table 4.

Logistic Regression exploring Prediction Accuracy

Logistic regression was conducted to assess how the following parameters impacted the accuracy of the LN diagnosis. The parameters include whether the strain wave was between -0.06 and +0.06, whether the LN was hypermetabolic, the positron emission tomography (PET) standardized uptake value (SUV), and LN size measured by short and long axis measurements. These results can be seen in table 5. Multivariate logistic regression was conducted for each parameter individually. The Wald statistic and the associated P-value were computed to assess the significance of individual beta coefficients in the model. The strain wave had a Wald X^2 of zero, and the rest of the parameters had a non-zero value.

Discussion

We demonstrated that a deep neural network algorithm could successfully calculate SAR from EBUS-Elastography images. In this study, we sought to validate the 0.4959 SAR cut-off at level 60 on a larger sample size. Any LN with a SAR greater than 0.4959 was assigned malignant, and if the proportion was below 0.4959, it was assigned benign. Using this blue threshold and SAR cut-off, we obtained an overall accuracy. 70.59% (95% Confidence Interval (CI) 63.50% to 77.01%), sensitivity of 43.04% (CI: 31.94% to 54.67%), and a specificity of 90.74% (CI: 83.63% to 95.47%). Additionally, a positive predictive value (PPV) of 77.27% (64.13% to 86.60%) and a negative predictive value (NPV) of 68.53% (64.05% to 72.70%) were obtained.

The diagnostic statistics differed from the pilot study results; however, the blue threshold AUC analysis shows a similar pattern. Threshold levels 40, 50 and 60 had the highest AUC compared to all the other threshold values, 0.816 (0.754-0.878), 0.823 (0.762-0.884), 0.820 (0.758-0.883), respectively. All three of the values constitute an excellent diagnostic test.[19] Using level 60, the mean SAR for benign LN was 0.25, and for malignant, it was 0.45, with a significant difference between the two groups. This gives us confidence that this threshold can accurately distinguish benign and malignant LNs. To our knowledge, no study aside from this study and the pilot study has explored defining the blue threshold. This is a promising step in standardizing the blue threshold to compute the SAR, thus improving the clinical utility of Elastography in diagnosing LNs.

Three other studies looked at SAR in mediastinal LNs; however, each used human input with the Image J software or Photoshop to select the region of interest and calculate SAR. The results of these studies are as follows, Uchimura and colleagues (2020) analyzed 149 LNs with a 0.41 SAR cut-off and 88.2% sensitivity, 80.2% specificity, and 83.9% diagnostic accuracy.[20] Nakajima and colleagues (2015) analyzed 49 LNs and got a cut-off value of 0.31 with 81% sensitivity and 85% specificity.[6] Finally, Ma and colleagues (2018) analyzed 79 LNs, with a cut-off value of 0.37 and 92.30% sensitivity, 67.50% specificity, and 78.5% accuracy. Compared to the existing literature, the sensitivity of our study is in the lower range. However, the specificity is high, indicating that this method can be helpful in ruling out malignancy. However, this study is valuable because it demonstrates the feasibility of NeuralSeg, a deep neural network's ability to segment the LN and calculate the SAR automatically.

SAR is a simple and intuitive way to understand Elastography colour maps that can achieve high sensitivities and specificities. However, other methods have also achieved good diagnostic results. In a study conducted by Verhoeven and colleagues (2019), they assessed four different techniques, two quantitative and two qualitative.[14] The method that performed the best was the mean strain elastography histogram. This technique achieved a 93% sensitivity and 75% specificity. This work was tested on a large sample of 525 LNs across five different centres.[21] Using mean strain Elastography histogram and a cut-off of 115, a sensitivity of 90%, specificity of 43%, and AUC of 0.77 was achieved. This study also combined the predicting probability of PET, EBUS LN size, and EBUS strain Elastography to detect LN malignancy.[21] This is an exciting avenue of research that can be applied to SAR and how it may impact the predictive

probability of LN malignancy when combined with clinical information such as PET and LN size.

Artificial Intelligence and deep neural networks and their application to medical imaging are becoming quite popular. AI and DNN mimic human cognition to analyze and interpret images.[22] The more data these algorithms are exposed to, the more they learn. Although certain more straightforward tasks such as identifying lung nodules can be accomplished by AI with relative ease, more challenging tasks can take more time.[22] In this study, NeuralSeg successfully segments and automatically calculates the SAR.

Furthermore, using AI input reduces the rater variability and increases the potential of this becoming a reproducible method in detecting LN malignancy. However, based on the study results, the DNN is still in the early stages of learning this task of detecting LN malignancy. Therefore, the algorithm can be further optimized with more data to improve its diagnostic capabilities.

An exploratory secondary analysis was conducted to explore how different parameters can impact the deep neural algorithm from making a correct prediction. Multivariate regression was conducted where the dependent variable was whether NeuralSeg made the correct prediction using 0.4959 at level 60 compared to pathology results. The independent variables are strain waves between -0.06 and +0.06, whether the LN is hypermetabolic, the positron emission tomography (PET) standardized uptake value (SUV), and LN size measured by short and long axis measurements. Although only the strain wave had a Wald statistic of zero, indicating it is

non-significant, the strain wave does provide clinical relevance. The strain wave allows the endosonographer to see if they are achieving stable pressurization visually.[23] Improving the EBUS-Elastography image quality by considering the strain wave and ROI is a future avenue for this research.

This study is not without limitations. Firstly, this is the experience from a single study. Therefore, it is essential to test whether this technique is truly reproducible by conducting a multi-centre study. Therefore, we will be one step closer to standardizing this process by conducting a multi-centre study. Additionally, only one method of Elastography analysis was used, the SAR. Although SAR is an intuitive quantitative method that has achieved by diagnostic results, Elastography is still a relatively new technology; therefore, all methods should be thoroughly analyzed. Lastly, the images are not consistent. Different factors like stable pressurization and the amount of the ROI in the elastography frame could potentially impact the results. Therefore, further studies should be conducted to obtain the best possible Elastography images.

Conclusion

In this study, NeuralSeg, an artificial intelligence deep neural network, was successfully applied to Elastography images to calculate the SAR automatically. As a result, the SAR cut-off of 0.4959 at level 60 determined from a previous pilot study was validated. This study is a meaningful step forward in the applicability of AI in detecting LN mediastinal malignancy. However, more extensive multi-centre studies must be conducted to standardize this process and optimize the algorithm.

References

1. Brenner, D. R. *et al.* Projected estimates of cancer in Canada in 2020. *Can. Med. Assoc. J.* **192**, E199 (2020).
2. Care, C. T. F. on P. H. Recommendations on screening for lung cancer. *Cmaj* **188**, 425–432 (2016).
3. Divisi, D., Zaccagna, G., Barone, M., Gabriele, F. & Crisci, R. Endobronchial ultrasound-transbronchial needle aspiration (EBUS/TBNA): a diagnostic challenge for mediastinal lesions. *Ann. Transl. Med.* **6**, (2018).
4. Ortakoylu, M. G. *et al.* Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *J. Bras. Pneumol.* **41**, 410–414 (2015).
5. Osarogiagbon, R. U. *et al.* Invasive mediastinal staging for resected non–small cell lung cancer in a population-based cohort. *J. Thorac. Cardiovasc. Surg.* **158**, 1220-1229.e2 (2019).
6. Nakajima, T. *et al.* Elastography for Predicting and Localizing Nodal Metastases during Endobronchial Ultrasound. *Respiration* **90**, 499–506 (2015).
7. Riegler, J. *et al.* Tumor Elastography and Its Association with Collagen and the Tumor Microenvironment. *Clin. Cancer Res.* **24**, 4455 (2018).
8. Moon, W. K., Chang, R.-F., Chen, C.-J., Chen, D.-R. & Chen, W.-L. Solid breast masses: classification with computer-aided analysis of continuous US images obtained with probe compression. *Radiology* **236**, 458–464 (2005).

9. Lyshchik, A. *et al.* Thyroid gland tumor diagnosis at US elastography. *Radiology* **237**, 202–211 (2005).
10. Wu, J. *et al.* Diagnostic value of endobronchial ultrasound elastography for differentiating benign and malignant hilar and mediastinal lymph nodes: a systematic review and meta-analysis. *Med. Ultrason.* (2021).
11. Izumo, T., Sasada, S., Chavez, C., Matsumoto, Y. & Tsuchida, T. Endobronchial Ultrasound Elastography in the Diagnosis of Mediastinal and Hilar Lymph Nodes. *Jpn. J. Clin. Oncol.* **44**, 956–962 (2014).
12. Sun, J. *et al.* Endobronchial Ultrasound Elastography for Evaluation of Intrathoracic Lymph Nodes: A Pilot Study. *Respiration* **93**, 327–338 (2017).
13. Korrungruang, P. & Boonsarngsuk, V. Diagnostic value of endobronchial ultrasound elastography for the differentiation of benign and malignant intrathoracic lymph nodes. *Respirology* **22**, 972–977 (2017).
14. Verhoeven, R. L. J., de Korte, C. L. & van der Heijden, E. H. F. M. Optimal Endobronchial Ultrasound Strain Elastography Assessment Strategy: An Explorative Study. *Respir. Int. Rev. Thorac. Dis.* **97**, 337–347 (2019).
15. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, 2015).
16. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform.* **48**, 193–204 (2014).

17. Hylton, D. A. *et al.* The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. *J Thorac Cardiovasc Surg* **159**, 2499-2507.e3 (2020).
18. Churchill, I. F. *et al.* An Artificial Intelligence Algorithm to Predict Nodal Metastasis in Lung Cancer. *Submitt. Ann. Thorac. Surg.*
19. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **5**, 1315–6 (2010).
20. Uchimura, K. *et al.* Quantitative analysis of endobronchial ultrasound elastography in computed tomography-negative mediastinal and hilar lymph nodes. *Thorac. Cancer* **11**, 2590–2599 (2020).
21. Verhoeven, R. L. J. *et al.* Predictive value of endobronchial ultrasound strain elastography in mediastinal lymph node staging: the E-predict multicenter study results. *Respiration* **99**, 484–492 (2020).
22. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* **316**, 2353–2354 (2016).
23. Olympus America Inc. Elastography Quick Reference. (2016).

Table 1. Patient and LN Demographics

Variable	Patients (n=124) Lymph Nodes (n=210)
Age (years) [mean ± SD]	69.83 ± 9.95
Males, n (%)	64 (51.6)
Smoking Status, n (%)	
Never Smoked	10 (8.1)
Former Smoker	54 (43.5)
Smoker	44 (35.5)
Unknown	12 (9.7)
Body Mass Index (BMI, kg/m^2)	26.69± 6.14
Comorbities, n (%)	
COPD	30 (24.2)
Atrial Fibrillation	12 (9.7)
Hypertension	51 (41.1)
GERD	10 (8.1)
Past Cancer	
Yes	29 (23.4)
No	90 (72.6)
LN Pathology, n (%)	
Benign	122 (53.1)
Malignant	79 (37.4)
Inconclusive	17 (8.1)
No Pathology	2 (0.9)
LN Station, n (%)	
7	81 (38.4)
4R	69 (32.7)
4L	33 (15.6)
10	2 (0.9)
11	3 (1.4)
Other (1,2,3,5,8,12)	22 (10.9)
Short Axis Measurement (mm) [mean ± SD]	11.01 ± 5.75
Long Axis Measurement (mm) [mean ± SD]	15.70 ± 6.00
Ultrasound Malignancy Features, n (%)	
Short Axis (>10mm)	98 (46.4)
Margins (defined)	102 (48.3)
CHS (absent)	71 (33.6)
Central Necrosis (present)	44 (20.9)

Canada LN Score	
0	77 (36.5)
1	38 (18.0)
2	25 (11.8)
3	46 (21.8)
4	22 (10.4)

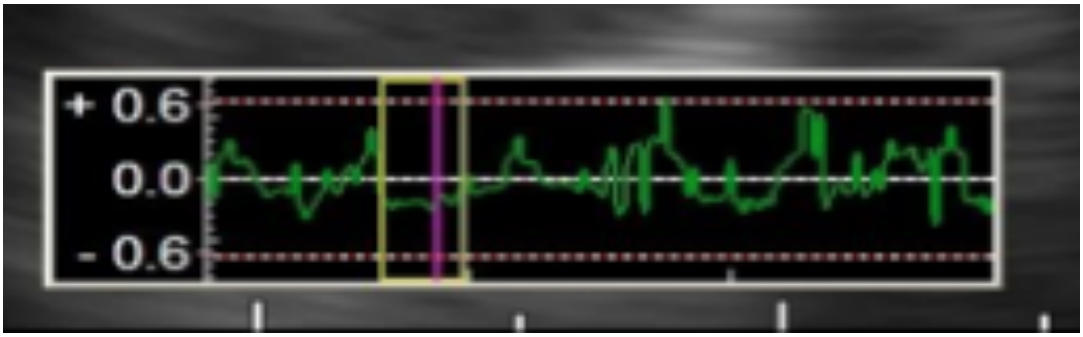


Figure 1. The strain graph used to achieve standard pressurization

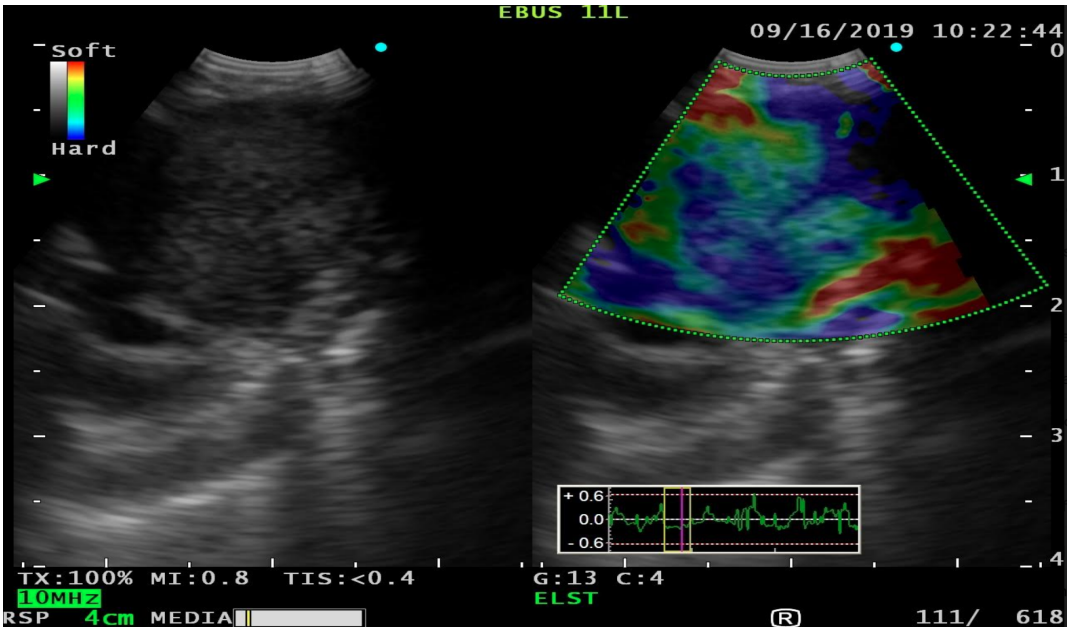


Figure 2. Image of B-mode and Elastography Image shown side-by-side

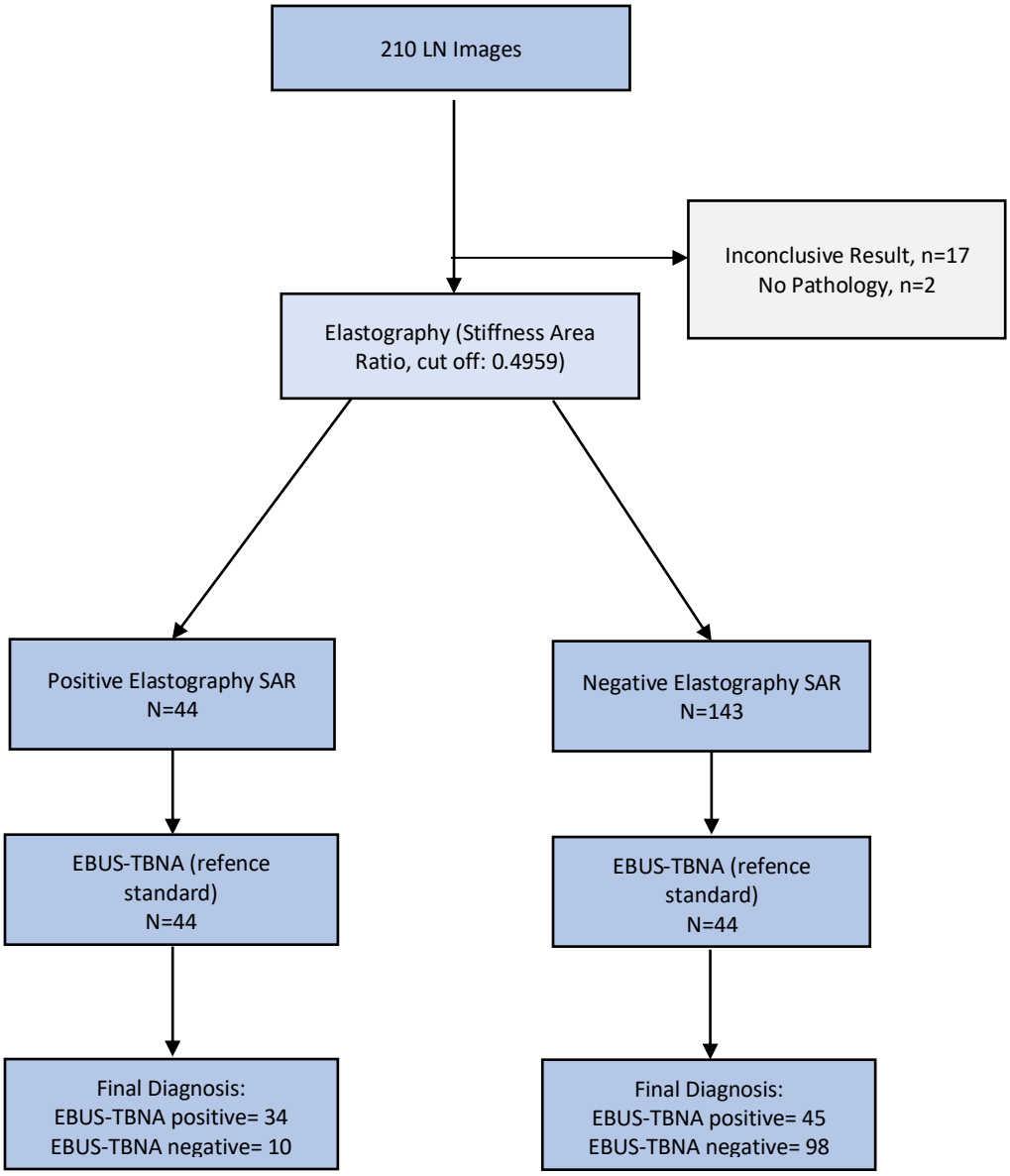


Figure 3. Flow chart of Elastography SAR diagnosis and EBUS-TBNA pathology diagnosis

Table 2. Diagnostic statistics based on 49.59% cut off at level 60

Accuracy	Sensitivity	Specificity	PPV	NPV
70.59% (63.50% to 77.01%)	43.04% (31.94% to 54.67%)	90.74% (83.63% to 95.47%)	77.27% (64.13% to 86.60%)	68.53% (64.05% to 72.70%)

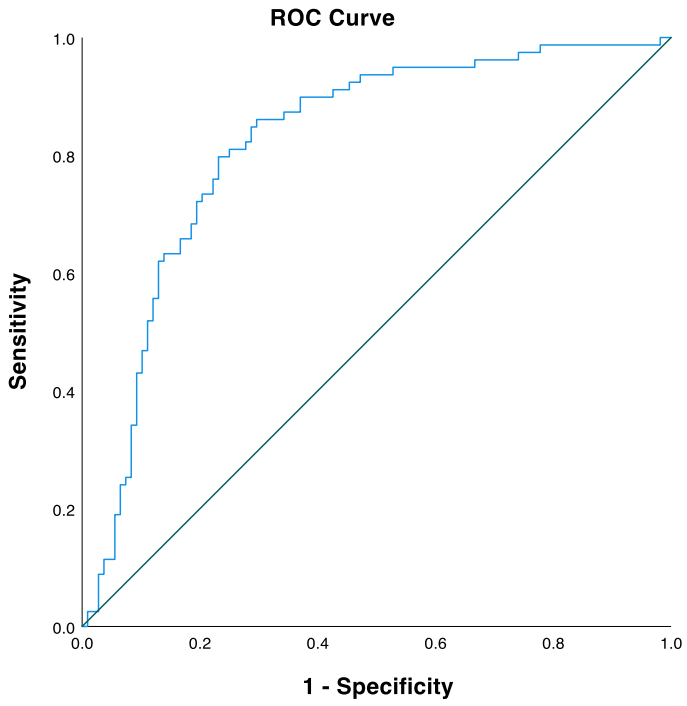


Figure 4. The ROC for Optimal Blue Threshold Level 60

Table 3. Descriptive Statistics for Stiffness Area Ratio for Benign and Malignant LNs.

Pathology	N	Minimum	Maximum	Mean	Std. Deviation
Benign	108	0.016	0.81	0.25	0.17
Malignant	79	0.037	0.78	0.45	0.15

P<0.001, Mann Whitney U Test

Table 4. Results of AUC for 9 Predefined Blue Colour Thresholds.

Colour Threshold	AUC	Std. Error	Significance	95% CI
10	0.799	0.033	0.000	(0.735-0.863)
20	0.805	0.032	0.000	(0.742-0.868)
30	0.809	0.032	0.000	(0.746-0.872)
40	0.816	0.032	0.000	(0.754-0.878)
50	0.823	0.031	0.000	(0.762-0.884)
60	0.820	0.032	0.000	(0.758-0.883)
70	0.678	0.041	0.000	(0.597-0.759)
80	0.405	0.041	0.026	(0.324-0.486)
90	0.413	0.042	0.042	(0.331-0.494)

Table 5. Multivariate Logistic Regression exploring the impact on prediction accuracy

Variable	B	S.E	Wald	df	Significance	Exp (B)
Strain Wave	0.008	0.547	0.000	1	0.988	1.008
Hypermetabolic LN	0.192	0.832	0.053	1	0.817	1.212
PET SUV	0.033	0.080	0.171	1	0.679	1.034
Short Axis	0.075	0.083	0.822	1	0.365	1.078
Long Axis	-0.122	0.068	3.219	1	0.073	0.885
Constant	1.389	1.882	0.544	1	0.461	4.009

Chapter 5: Summary of Findings, Methodological Challenges, and Future Directions

Nikkita Mistry, BSc, Forough Farrokhyar, MPhil, PhD, Feng Xie, MSc, PhD, Waël C. Hanna, MDCM, MBA

Lung cancer has the highest mortality among all the type of cancer, making up 25.5% of deaths.[1] Late-stage lung cancer has a significantly lower survival rate compared to early-stage cancer. Stage 4 has only 1 % survival rate whereas stage 1A has a 49% survival rate.[2] Staging is a critical step in the NSCLC diagnostic pathway, with the current standard being EBUS-TBNA.[3] Despite advancements in staging technology, as high as 40% of nodal biopsies are inconclusive and as low as 22% of patients receiving mediastinal staging.[4,5] This provides a clear rationale that mediastinal staging procedures need to be improved. Technologies like Artificial Intelligence and Elastography have shown promise in serving as useful adjuncts to improve staging, which can in turn improvement treatment decisions making and reduce the number of repeat biopsies.[6,7]

Summary of Findings

Part 1: The Application of an Artificial Intelligence Algorithm to predict Lymph Node Malignancy in Non-Small Cell Lung Cancer

The first part of this thesis aimed to train, validate, and test a deep neural AI network to predict LN malignancy based on B-mode EBUS-TBNA images. Using a sample size of 3366 LN, separated in the training/validation (80%) and the testing set (20%). The ensemble method was used, where three individual models were trained using 5-fold cross-validation. The outputs of these models were combined to create the final ensemble model which is applied to testing set.

The final model had an overall accuracy of 80.63% (76.93% to 83.97%), a sensitivity of 43.23% (35.30% to 51.41%), a specificity of 96.91% (94.54% to 98.45%), a positive predictive value of 85.90% (76.81% to 91.80%), a negative predictive value of 79.68% (77.34% to 81.83%), and an AUC of 0.701 (0.646 to 0.755). The results of the final model was compared to the CLNS assigned by an expert endosonographer. The AI ensemble model was outperformed by the CLNS. This shows that although the AI model has potential there is still improvements to be made.

Part II: Lung Cancer Nodal Staging using EBUS-Elastography and AI: A Pilot Study

The aim of second part of this thesis was two-fold. Firstly, we aimed to define the blue colour threshold. Secondly, we aimed to find the optimal SAR cut-off value based on the blue threshold that most accurately distinguished benign and malignant LNs. A sample size of 31 LNs were fed to a trained deep neural network, NeuralSeg. Nine different thresholds from 10 to 90 were analyzed, increasing in increments of 10. Level 60 was chosen as the blue threshold has it had the highest AUC of 0.89 (95% CI: 0.77-1.00). Using this threshold, the optimal SAR cut off was found to be 0.4959 with a sensitivity of 92.30% (95% CI: 62.10% to 99.60%) and a specificity of 76.50% (95% CI: 49.80% to 92.20%). NeuralSeg was successfully applied to EBUS-Elastography images to automatically calculate SAR the predefined threshold values. This paper was the first to define the blue threshold.

Part III: Clinical Utility of Artificial Intelligence-Augmented Endobronchial Ultrasound Elastography in Lymph Node Staging for Lung Cancer

The third part of this thesis aimed to validate the SAR cut off value of 0.4959 determined from part II. The sample size consisted of 201 EBUS-Elastography LN images. The images were fed to NeuralSeg, which segmented the LN and identified the percentage of LN above level 60. LN that had a SAR above 0.4959 were assigned malignant, and those that were lower were assigned benign. Results were compared to pathology results. This resulted in an overall accuracy of 70.59% (95% Confidence Interval (CI) 63.50% to 77.01%), sensitivity of 43.04% (CI: 31.94% to 54.67%), a specificity of 90.74% (CI: 83.63% to 95.47%), a positive predictive value (PPV) of 77.27% (64.13% to 86.60%) and a negative predictive value (NPV) of 68.53% (64.05% to 72.70%). The AUC for level 60 was 0.820 (CI:0.758-0.883). Although AI-powered EBUS-Elastography shows promising result, more extensive multi-centre studies should be conducted before this method can be standardized.

Challenges

Overfitting

A major problem with machine learning tasks for an AI algorithm is overfitting. When an algorithm is in its learning stages it is trained on a training data set and then applied to a testing data set to make predictions. However, overfitting occurs when the algorithm memorizes specific details of the training set rather than making general predictions.[8] This can result in good results

for a training set but poor results when the algorithm is exposed to new data, diminishing the predicative capabilities and generalizability.[8] There are different methods to combat overfitting, one of which is cross-validation.[9] 5-fold cross validation was applied to each part of this thesis, where a different one-fifth of the training data was held out for the validation set and the remaining four-fifths were used for the training set. The average of these predictions were used.[9]

Variability

It is important to acknowledge there are sources of variability which can impact the performance of the diagnostic methods assessed in this thesis. Results are subject to differ based on different patient populations.[10] A strength of the studies in this thesis was there was no preferential selection and consecutive sampling was used for all three parts of this study. Although variability does exist in the way these methods were conducted. For example, when assigning CLNS there may be inter-rater variability. Additionally, when conducting EBUS-Elastography there is room for variability in achieving standard pressurization and obtain a good image. All three parts of this thesis were the experiences of a single centre, conducted by a single endosonographer. In order for these methods to be widely used larger multi-centre studies must be conducted to ensure the true reproducibly of these techniques.

Future Directions

Further optimization using different techniques beyond 5-fold cross validation can be applied to NeuralSeg to achieve greater diagnostic results. Yong and collegues (2022) have had success

with the new loss function technique and observed improvements in sensitivity, specificity, overall accuracy and AUC.[11]

Additionally, taking advantage of DNN ability to extract many different image features combining EBUS-TBNA ultrasonographic features and Elastography on large studies including multicentre studies is an exciting avenue of research that may increase the pre-test probability of LN malignancy.

In conclusion novel technologies like AI and AI-powered EBUS-Elastography has shown promising and feasible results in terms of LN malignancy detection. It was observed that both AI and AI-powered EBUS-Elastography achieved high specificities, indicative that these methods may be useful in ruling out LN malignancy. However due to the novelty of these technologies further studies must be conducted before these processes can be standardized and applied to bedside.

References

1. Brenner, D. R. *et al.* Projected estimates of cancer in Canada in 2020. *Can. Med. Assoc. J.* **192**, E199 (2020).
2. Care, C. T. F. on P. H. Recommendations on screening for lung cancer. *Cmaj* **188**, 425–432 (2016).
3. Nakajima, T., Yasufuku, K. & Yoshino, I. Current status and perspective of EBUS-TBNA. *Gen. Thorac. Cardiovasc. Surg.* **61**, 390–396 (2013).
4. Ortakoylu, M. G. *et al.* Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *J. Bras. Pneumol.* **41**, 410–414 (2015).
5. Osarogiagbon, R. U. *et al.* Invasive mediastinal staging for resected non–small cell lung cancer in a population-based cohort. *J. Thorac. Cardiovasc. Surg.* **158**, 1220-1229.e2 (2019).
6. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* **316**, 2353–2354 (2016).
7. Nakajima, T. *et al.* Elastography for Predicting and Localizing Nodal Metastases during Endobronchial Ultrasound. *Respiration* **90**, 499–506 (2015).
8. Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv. CSUR* **27**, 326–327 (1995).
9. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*. vol. 112 (Springer, 2013).
10. Mower, W. R. Evaluating bias and variability in diagnostic test reports. *Ann. Emerg. Med.* **33**, 85–91 (1999).

11. Yong, S. H. *et al.* Malignant thoracic lymph node classification with deep convolutional neural networks on real-time endobronchial ultrasound (EBUS) images. *2022* **11**, 14–23 (2022).