

**USE OF ARTIFICIAL INTELLIGENCE TO AUGMENT CLINICIANS IN  
MEDIASTINAL STAGING FOR LUNG CANCER**

**THE USE OF ARTIFICIAL INTELLIGENCE FOR THE DEVELOPMENT AND  
VALIDATION OF A COMPUTER-AIDED ALGORITHM FOR THE  
SEGMENTATION OF LYMPH NODE FEATURES FROM THORACIC  
IMAGING**

By: ISABELLA FRANCESCA CHURCHILL, BSc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the  
Requirements for the Degree Masters of Science

McMaster University © Copyright by Isabella Francesca Churchill, July 15, 2020

MSc. Thesis- I.F. Churchill; McMaster University- Health Research Methodology

McMaster University MASTER OF SCIENCE (2020) Hamilton, Ontario (Health Research Methodology)

**TITLE:** The Use of Artificial Intelligence for the Development and Validation of a Computer-Aided Algorithm for the Segmentation of Ultrasonographic Lymph Node Features from Thoracic Imaging

**AUTHOR:** Isabella Francesca Churchill, BSc. (Brock University)

**SUPERVISOR:** Dr. Wael C. Hanna, MDCM, MBA

**NUMBER OF PAGES:** XV, 160

**LAY ABSTRACT:**

Before deciding on treatment for patients with lung cancer, a critical step in the investigation is finding out whether the lymph nodes in the chest contain cancer cells. This is accomplished through medical imaging of the lymph nodes or taking a biopsy of the lymph node tissue using a needle attached to a scope that is entered through the airway wall. The purpose of these tests is to ensure that lung cancer patients receive the optimal treatment option. However, imaging of the lymph nodes is heavily reliant on human interpretation, which can be error prone. We aimed to critically analyze and investigate the use of Artificial Intelligence to enhance clinician performance for image interpretation. We performed a search of the medical literature for the use of Artificial Intelligence to diagnosis lung cancer from medical imaging. We also taught a computer program, known as NeuralSeg, to learn and identify cancerous lymph nodes from ultrasound imaging. This thesis provides a significant contribution to the Artificial Intelligence literature and provides recommendations for future research.

**ABSTRACT:**

**Background-** Mediastinal staging is the rate-limiting step prior to initiation of lung cancer treatment and is essential in identifying the most appropriate treatment for the patient. However, this process is often complex and involves multiple imaging modalities including invasive and non-invasive methods for the assessment of lymph nodes in the mediastinum which are error prone. The use of Artificial Intelligence may be able to provide more accurate and precise measurements and eliminate error associated with medical imaging.

**Methods-** This thesis was conducted in three parts. In Part 1, we synthesized and critically appraised the methodological quality of existing studies that use Artificial Intelligence to diagnosis and stage lung cancer from thoracic imaging based on lymph node features. In Part 2, we determined the inter-rater reliability of segmentation of the ultrasonographic lymph node features performed by an experienced endoscopist (manually) compared to NeuralSeg (automatically). In Part 3, we developed and validated a deep neural network through a clinical prediction model to determine if NeuralSeg could learn and identify ultrasonographic lymph node features from endobronchial ultrasound images in patients undergoing lung cancer staging.

**Results-** In Part 1, there were few studies in the Artificial Intelligence literature that provided a complete and detailed description of the design, Artificial Intelligence architecture, validation strategies and performance measures. In Part 2, NeuralSeg and the experienced endosonographer possessed excellent inter-rater correlation (Intraclass Correlation Coefficient = 0.76, 95% CI= 0.70 – 0.80,  $p < 0.0001$ ). In Part 3, NeuralSeg's

algorithm had an accuracy of 73.78% (95% CI: 68.40% to 78.68%), a sensitivity of 18.37% (95% CI: 8.76% to 32.02%) and specificity of 84.34% (95% CI: 79.22% to 88.62%).

**Conclusions-** Analysis of staging modalities for lung cancer using Artificial Intelligence may be useful for when results are inconclusive or uninterpretable by a human reader. NeuralSeg's high specificity may inform decision-making regarding biopsy if results are benign. Prospective external validation of algorithms and direct comparisons through cut-off thresholds are required to determine their true predictive capability. Future work with a larger dataset will be required to improve and refine the algorithm prior to trials in clinical practice.

**ACKNOWLEDGEMENTS:**

First, I would like to thank my parents for giving me the opportunity to pursue educational opportunities. Without their support and encouragement, I would not be where I am today.

I would also like to thank my thesis supervisor, Dr. Wael Hanna, who guided me throughout my MSc studies and provided me with constant feedback, advice and valuable direction. Thank you for challenging me and showing me all that I can achieve.

I would also like to thank my thesis committee members, Dr. Forough Farrokhyar and Dr. Grigorios Leontiadis for their support, patience and advice.

To the research lab members, Yogita Patel, Danielle Hylton and Kerrie Sullivan. Thank you for your feedback and insightful perspectives. My MSc experience would not have been the same without you.

To my friends and colleagues in the Health Research Methodology program and at McMaster University. Thank you for your constant support and encouragement. It was a privilege to learn in such a collaborative and enriching environment.

Last, but certainly not least, thank you to all the patients that enrolled in my thesis. Without you, this work would not have been possible.

**TABLE OF CONTENTS**

**OVERVIEW OF THESIS:** ..... 1

**CHAPTER 1: BACKGROUND**..... 4

**1.1 Lung Cancer: The Canadian Setting** ..... 4

**1.2 Current Lung Cancer Detection and Investigation Guidelines** ..... 5

**1.3 Ultrasonographic Features and Predictive Tools for Lymph Node Malignancy**  
    ..... 9

**1.4 Research Group’s Preliminary Work**..... 11

**1.5 Rationale for Artificial Intelligence in Lung Cancer Staging**..... 12

**1.6 Machine and Deep Learning Framework**..... 13

**1.7 Objectives**..... 14

**CHAPTER 1 TABLES AND FIGURES:** ..... 16

**CHAPTER 2: APPLICATION OF RADIOMICS TO PREDICT LYMPH NODE  
MALIGNANCY FOR THE STAGING OF NON-SMALL CELL LUNG CANCER IN  
THORACIC MEDICAL IMAGING: A SYSTEMATIC REVIEW**..... 20

**ABSTRACT:** ..... 20

**2.1 BACKGROUND** ..... 22

**2.1.1 Target Condition** ..... 23

**2.1.2 Index Tests**..... 24

**2.2. Clinical Pathways**..... 24

**2.2.1 Role of Index Tests**..... 25

**2.2.2 Alternative Tests** ..... 25

**2.2.3 Rationale** ..... 26

**2.2.4 Objectives**..... 27

**2.3 METHODS** ..... 28

**2.3.1 Criteria for Considering Studies for this Review**..... 28

**2.3.2 Search Methods for Identification of Studies** ..... 31

**2.3.3 Data Collection and Analysis** ..... 32

**2.4 RESULTS** ..... 33

**2.4.1 Results of Search** ..... 33

**2.4.2 Included Studies** ..... 34

**2.4.3 Methodological Quality of Included Studies** ..... 35

**2.4.4 Findings**..... 37

**2.5 DISCUSSION** ..... 39

**2.5.1 Summary of Main Results**..... 39

**2.5.2 Strengths and Weaknesses of Review** ..... 40



<b>2.5.3 Applicability of Findings to Review Question</b> .....	<b>41</b>
<b>2.6 AUTHOR’S CONCLUSION</b>	
<b>2.6.1 Implication for Practice</b> .....	<b>42</b>
<b>2.6.2 Implication for Research</b> .....	<b>43</b>
<b>CHAPTER 2 TABLES AND FIGURES:</b> .....	<b>44</b>
<b>CHAPTER 3: COMPARISON BETWEEN MANUAL AND AUTOMATIC SEGMENTATIONS OF ULTRASONOGRAPHIC LYMPH NODE FEATURES OBSERVED DURING ENDOBRONCHIAL ULTRASOUND: ASSESSMENT OF INTER-RATER RELIABILITY</b> .....	<b>49</b>
<b>ABSTRACT:</b> .....	<b>49</b>
<b>3.1 INTRODUCTION</b> .....	<b>51</b>
<b>3.2 METHODS</b> .....	<b>53</b>
<b>3.2.1 Study Design</b> .....	<b>53</b>
<b>3.2.2 Participants</b> .....	<b>53</b>
<b>3.2.2 Testing Methods</b> .....	<b>54</b>
<b>3.2.3 Statistical Analysis</b> .....	<b>58</b>
<b>3.3 RESULTS</b> .....	<b>59</b>
<b>3.4 DISCUSSION</b> .....	<b>61</b>
<b>CHAPTER 3 TABLES AND FIGURES:</b> .....	<b>65</b>
<b>CHAPTER 4: DEVELOPMENT AND VALIDATION OF DEEP NEURAL NETWORK FOR PREDICTING LYMPH NODE MALIGNANCY USING THE CANADA LYMPH NODE SCORE IN LUNG CANCER PATIENTS UNDERGOING MEDIASTINAL STAGING</b> .....	<b>76</b>
<b>ABSTRACT:</b> .....	<b>76</b>
<b>4.1 INTRODUCTION</b> .....	<b>78</b>
<b>4.2 METHODS</b> .....	<b>80</b>
<b>4.2.1 Source of Data</b> .....	<b>80</b>
<b>4.2.2 Participants</b> .....	<b>81</b>
<b>4.2.3 Outcomes</b> .....	<b>82</b>
<b>4.2.4 Predictors</b> .....	<b>83</b>
<b>4.2.5 Sample Size</b> .....	<b>84</b>
<b>4.2.6 Missing Data</b> .....	<b>85</b>
<b>4.2.7 Statistical Analysis</b> .....	<b>85</b>
<b>4.2.8 Risk Groups</b> .....	<b>86</b>
<b>4.2.9 Development and Validation</b> .....	<b>86</b>
<b>4.3 RESULTS</b> .....	<b>89</b>
<b>4.3.1 Participants</b> .....	<b>89</b>
<b>4.3.2 Model Development and Specification</b> .....	<b>90</b>
<b>4.3.3 Model Discrimination and Calibration</b> .....	<b>91</b>

<b>4.3.4 Model Performance</b> .....	91
<b>4.4 DISCUSSION</b> .....	92
<b>CHAPTER 4 TABLES AND FIGURES:</b> .....	96
<b>CHAPTER 5: SUMMARY OF FINDINGS AND METHODOLOGICAL CHALLENGES</b> .....	109
<b>5.1 THESIS FINDINGS AND LESSONS LEARNED</b> .....	109
<b>5.2 METHODOLOGICAL CHALLENGES AND MITIGATION STRATEGIES</b> .....	110

## **TABLES, FIGURES AND APPENDICES**

### **CHAPTER 1**

<b>Figure 1.</b> Age-standardized incidence rates (ASIRs) for selected cancers, in Canada (excluding Quebec), 1984-2020 by sex.....	<b>16</b>
<b>Figure 2.</b> Age-standardized mortality rates (ASMRs) for selected cancers in Canada, 1984-2020, by sex.....	<b>16</b>
<b>Figure 3.</b> Clinical care pathway map for diagnostic imaging of lung cancer.....	<b>17</b>
<b>Figure 4.</b> International Association for the Study of Lung Cancer (IASCL) lymph node map.....	<b>18</b>
<b>Figure 5.</b> Canada Lymph Node Score Criteria.....	<b>19</b>
<b>Figure 6:</b> Schematic graphical representation of Artificial Intelligence, machine learning and deep learning. ....	<b>19</b>

### **CHAPTER 2**

<b>Table 1.</b> Eligibility criteria of studies for inclusion in systematic review.....	<b>44</b>
<b>Table 2.</b> Performance and validation characteristics of included studies.....	<b>45</b>
<b>Figure 1.</b> PRISMA flow diagram of screening, eligibility and inclusion process.....	<b>47</b>
<b>Figure 2.</b> Risk of bias assessment.....	<b>48</b>

### **CHAPTER 3**

<b>Figure 1:</b> Study design to assess inter-rater reliability and accuracy.....	<b>65</b>
<b>Figure 2.</b> Dice Similarity Coefficient (DSC) diagram representing spatial overlap.....	<b>66</b>
<b>Figure 3:</b> Example of lymph node station 4R assessed with Canada Lymph Node Score criteria during manual segmentation.....	<b>67</b>

<b>Figure 4:</b> NeuralSeg with its convolutional neural network components and feedforward framework.....	<b>68</b>
<b>Figure 5.</b> Flow diagram of lymph node images through study.....	<b>69</b>
<b>Figure 6:</b> Inter-rater comparisons as measured by Dice Similarity Coefficient Scores...	<b>70</b>
<b>Figure 7:</b> Correlation matrix for comparison of spatial overlap of segmentations.....	<b>71</b>
<b>Table 1:</b> Patient characteristics and pathological results for lymph nodes.....	<b>72</b>
<b>Table 2:</b> Dice Similarity Coefficients of Canada Lymph Node Score for manual and automatic segmentations of lymph node and entire contents.....	<b>73</b>
<b>Table 3.</b> Diagnostic statistics of ultrasonographic lymph node features and $\geq 2$ CLNS determined by NeuralSeg (gold standard = endosonographer).....	<b>74</b>
<b>Table 4:</b> Assessment of performance measurements for the Canada Lymph Node Scores on a binary scale ( $\geq 2$ considered malignant) according to each rater.....	<b>75</b>

#### **CHAPTER 4**

<b>Table 1:</b> Clinical characteristics and pathology results for derivation and validation sets.....	<b>96</b>
<b>Table 2.</b> Ultrasonographic feature presence in malignant and benign lymph nodes in derivation set identified by NeuralSeg (reference standard = histopathology).....	<b>97</b>
<b>Table 3.</b> Univariate analysis of binary ultrasonographic features with logistic regression for derivation set.....	<b>98</b>
<b>Table 4.</b> Univariate analysis of continuous ultrasonographic features with logistic regression for derivation set.....	<b>98</b>
<b>Table 5.</b> Multivariate analysis of ultrasonographic features with logistic regression for derivation set.....	<b>98</b>
<b>Table 6.</b> Comparison of NeuralSeg’s predictive algorithm with previously published studies with lymph node segmentation algorithms.....	<b>99</b>

**Table 7.** Diagnostic statistics of ultrasonographic lymph node features and  $\geq 2$  CLNS determined by NeuralSeg in derivation set (gold standard = endosonographer).....100

**Table 8.** Diagnostic statistics and discrimination of  $\geq 2$  CLNS assigned by NeuralSeg (reference standard = histopathology)..... 101

**Figure 1.** Study design.....102

**Figure 2:** A representation of NeuralSeg’s network with a U-Net style architecture....103

**Figure 3.** Flow of participants through study.....104

**Figure 4.** Boxplots displaying the predicted probabilities for lymph nodes that were malignant versus those that were benign based on a multivariate model for the derivation set.....105

**Figure 5.** NeuralSeg multivariate model calibration plot for derivation set.....106

**Figure 6.** Receiver operator characteristic curve for derivation set.....107

**Figure 7.** Receiver operator characteristic curve for validation set.....108

**APPENDICES**

**Appendix 1.** Search Strategy.....125

**Appendix 2.** QUADAS-2 Guideline for Risk of Bias Assessment.....131

**Appendix 3.** Characteristics of Included Studies and Risk of Bias Assessment [*ordered by study ID*].....134

**Appendix 4.** Ongoing and Awaiting Classification Studies.....160

**LIST OF ABBREVIATIONS**

<b>AI</b>	Artificial Intelligence
<b>ANN</b>	Artificial Neural Network
<b>AUC</b>	Area Under the Curve
<b>CAD</b>	Computer Aided Diagnosis
<b>CAT</b>	Computed Axial Tomography
<b>CHS</b>	Central Hilar Structure
<b>CI</b>	Confidence Interval
<b>CLNS</b>	Canada Lymph Node Score
<b>CNN</b>	Convolutional Neural Network
<b>CT</b>	Computed Tomography
<b>CTFPC</b>	Canadian Task Force on Preventative Health Care
<b>DSC</b>	Dice Similarity Coefficient
<b>EBUS-TBNA</b>	Endobronchial Ultrasound Transbronchial Needle Aspiration
<b>FDG</b>	Fludeoxyglucose (18F)
<b>GRAAS</b>	Guidelines for Reporting Reliability and Agreement Studies
<b>ICC</b>	Intraclass Correlation Coefficient
<b>MRI</b>	Magnetic Resonance Imaging
<b>NPV</b>	Negative Predictive Value
<b>NSCLC</b>	Non-Small Cell Lung Cancer
<b>OR</b>	Odds Ratio
<b>PET</b>	Positron Emission Tomography

<b>PPV</b>	Positive Predictive Value
<b>PRISMA</b>	Preferred Reporting in Systematic Reviews and Meta- Analysis
<b>ROC</b>	Receiver Operator Characteristic
<b>TRIPOD</b>	Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis

## **DECLARATION OF ACADEMIC ACHIEVEMENT**

Isabella F. Churchill was involved in the conception of the design of the study and was responsible for all aspects of this thesis, including data collection, analysis and is the primary author of the subsequent manuscripts published based on this thesis work. All submitted work is original.

Dr. Wael C. Hanna was the primary supervisor of the thesis. He was involved in the conception and design of the study, review of statistical analysis and review of manuscripts.

Drs. Forough Farrokhyar and Grigorios Leontiadis advised on statistical analysis and study design. They also provided valuable feedback for this thesis.

Alexander Simone assisted with screening; Kerrie Sullivan assisted with screening, data abstraction and risk of bias assessment; and Yogita Patel assisted with resolving discrepancies for Chapter 2.

Anthony Gatti assisted with the pre-processing of images as well as the training and running of NeuralSeg for Chapter 3 and 4.



**OVERVIEW OF THESIS:**

This master's thesis is composed of five chapters and contains three separate papers that will be submitted to peer reviewed medical journals. The chapters are briefly outlined below.

**Chapter 1:** The first chapter presents an overview of the Canadian lung cancer literature. The state of lung cancer diagnosis and staging is summarized and various diagnostic modalities for thoracic oncology are highlighted. This chapter further discusses the recent work of various ultrasonographic features used for lymph node (LN) malignancy prediction that laid the foundation for this master's thesis. An introduction to the use of Artificial Intelligence in diagnostic imaging is also explained.

**Chapter 2:** This systematic review aims to critically appraise the current use of radiomics in medical imaging for the staging and diagnosis of nodal involvement in lung cancer. In this chapter, we aim to determine the diagnostic accuracy of radiomics for mediastinal staging in patients undergoing computed tomography, positron emission tomography or endobronchial ultrasound imaging procedures for lung cancer. Overall, a synthesis of data from 19 studies is provided and diagnostic statistics on various segmentation algorithms to predict lymph node malignancy are reported. However, many of these studies had unclear reporting and only one study externally validated their model.

**Chapter 3:** This chapter aimed to determine if a deep neural network could learn to identify and segment four ultrasonographic lymph node features (short axis, margins, central hilar structure, central necrosis) that were part of the Canada Lymph Node Score in order to reduce the operator dependency associated with the tool. We compared manual and automatic segmentation of ultrasonographic lymph node features and found that the deep neural network was able to segment the features with greater accuracy than the manual segmentations produced by the endosonographer. However, we identified that the proposed prediction model was needed and that the model would require further validation.

**Chapter 4:** We developed and validated a deep neural network known as NeuralSeg to determine if it was capable of predicting LN metastasis through the segmentation of ultrasonographic LN features observed during EBUS imaging. This study was conducted in two phases: a derivation phase followed by a validation phase. In the derivation phase, LN images were segmented twice by a blinded experienced endosonographer using 3D Slicer and a 5-fold cross-validation was used for training NeuralSeg. In the validation phase, LN images were prospectively collected to test the algorithm. NeuralSeg showed excellent performance in identifying malignant LNs from EBUS images. However, future research with a larger dataset will be required to improve and refine the algorithm prior to trials in clinical practice

**Chapter 5:** The concluding chapter summarizes the findings and methodological challenges that were encountered over the course of this study. We found that the deep

neural network was able to identify and segment ultrasonographic lymph node features with high accuracy. In terms of methodological frameworks, emphasis was placed on bias associated with diagnostic studies and the consideration of sample size, population of interest, internal and external validation strategies, overfitting and sources of measurement for machine learning.

## **CHAPTER 1: BACKGROUND**

### **1.1 Lung Cancer: The Canadian Setting**

Lung cancer is the most prevalent cancer in Canada, is more fatal than colon, breast, and prostate cancers combined, and results in 18.4% of all cancer deaths (Bray et al., 2018). Lung cancer accounts for 14% of all cancer cases in Canada; with 1 in 11 males and 1 in 15 females expected to be diagnosed with lung cancer in their lifetime (Brenner et al., 2020). In 2016, the Canadian Task Force on Preventative Health Care (CTFPC) released a guideline that recommended annual screening for lung cancer in high risk adults aged 55-74 years using low dose computed tomography (CT) (Care, 2016). There has been an increased incidence of lung cancer in Canada due to the increased lung cancer screening (Akhtar-Danesh & Finley, 2015; Evans et al., 2016).

In 2019, it was estimated that 29,300 individuals were diagnosed with lung cancer in Canada (Brenner et al., 2020). Incidence rates for lung cancer by age group followed an upward trend, implying that the likelihood of developing lung cancer increases with age (Smith et al., 2019). However, this trend may have been misleading as it was mostly driven by an overall increase in lung cancer rates in women aged 65 years and older. In 2020, lung cancer is projected to be the most commonly diagnosed cancer with an estimated 29,800 cases (Brenner et al., 2020). Lung cancer is anticipated to be the leading cause of death for both males and females, accounting for 25% and 26% of cancer deaths, respectively (Bray et al., 2018; Smith et al., 2019). Interestingly, the age-standardized incidence rate is

expected to decrease for both males and females as the result of smoking cessation and screening programs. However, the age-standardized mortality rate is expected to be 26% higher (219.7 per 100,000 males and 164.2 per 100,000 females) (**Figure 1; Figure 2**). It is predicted that lung cancer will be the second most frequently diagnosed cancer (12%) by 2028-2032, with 80-85% of all cases consisting of non-small cell lung cancer cases (NSCLC) (Brenner et al., 2020). Additional efforts to improve uptake of existing programs, as well as to advance research, prevention, screening and treatment, are needed to manage the disease burden.

## **1.2 Current Lung Cancer Detection and Investigation Guidelines**

Various biopsy and imaging techniques are used to diagnose and stage lung cancer. In cases where lung cancer is suspected, the preliminary diagnostic modality is chest radiography or CT and a clinical assessment. Key aspects of this assessment include ascertaining the patient's medical history, physical examination, standard blood work and pulmonary function tests. From these investigations, further diagnostic tests may be performed according to the lung cancer diagnosis clinical pathway outlined by National Cancer Care Network and Cancer Care Ontario (**Figure 3**).

Patients may also present to their primary care provider with symptoms indicative of lung cancer's clinical vignette, such as hemoptysis, cough, unintentional weight loss, loss of appetite, chest, rib or shoulder pain, and/or dyspnea. If symptoms persist for more than three weeks or if patients have identified risk factors (Cancer Care Ontario, n.d.) (i.e.

family history of cancer, smoker etc.) they are commonly referred for chest imaging via CT, as dictated by cancer imaging guidelines (Sampsonas, Kakoullis, Lykouras, Karkoulias, & Spiropoulos, 2018; Silvestri et al., 2013). If the diagnostic results are suspicious lung cancer, patients undergo an organized diagnostic assessment which may include positron emission tomography (PET), abdominal CT, magnetic resonance imaging (MRI) brain scans or bone scans. The location of the primary tumour, patient preferences, and the fitness of the patient may require a change to the diagnostic and staging pathway, which augments the complexity of the NSCLC clinical process.

Accurately determining the stage of lung cancer is critical to ensure patients are offered the best and most appropriate treatment options. Mediastinal staging is typically completed via CT and PET scans. Depending on the imaging findings, patients may then undergo endobronchial ultrasound-transbronchial needle aspiration (EBUS-TBNA) to assist with treatment decisions. If the disease has not spread to either the parabranchial, interlobar or hilar LNs (N1), ipsilateral LNs (N2) nodes or contralateral mediastinal, contralateral hilar or supraclavicular LNs (N3) (Figure 3), and the patient is otherwise considered fit for surgery, resection is often the treatment of choice (Barnes, See, Barnett, & Manser, 2017; Rena, 2016). Patients with N1-N3 nodal involvement usually undergo chemoradiation therapy (Burdett et al., 2015). Consequently, accurate mediastinal staging can ensure that the right treatment is given to the right patient and justifies the reason why it is the rate-limiting step prior to initiation of lung cancer treatment.

The staging system most commonly used for NSCLC is the American Joint Committee on Cancer TNM System. This system is used to describe the amount and spread of cancer throughout the patient's body. T describes the tumour size, N describes the spread of cancer to nearby lymph nodes, and M describes the spread of cancer to distant sites of the body, known as metastasis. Currently, the 8th version of this staging system has resulted in important modifications to the stage classification, including the creation of several new stage groups (Carter et al., 2018; Goldstraw et al., 2016).

The gold standard for diagnostically assessing LNs for lung cancer is an invasive approach through surgical staging known as cervical mediastinoscopy (Silvestri et al., 2013). This procedure is performed in an operating room under general anesthesia and provides access to upper and lower paratracheal (2R, 2L, 4R and 4L) and subcarinal (7) LNs (**Figure 4**). A systematic review used to update staging guidelines found that the median sensitivity of standard cervical mediastinoscopy was 78% and the median negative predictive value (NPV) was 91%, where approximately 42-57% of the false negative cases were due to nodes that were not accessible by a traditional mediastinoscopy (Silvestri et al., 2013).

EBUS-TBNA is a minimally invasive approach that is completed in an out-patient setting without general anesthesia. EBUS-TBNA allows access to the upper and lower paratracheal (2R, 2L, 4R, 4L), subcarinal (7), hilar (10R, 10L) and interlobar (11R, 11L) LNs- allowing for the investigation of more LNs compared to mediastinoscopy (Wahidi et

al., 2016). A meta-analysis has shown the median sensitivity for EBUS-TBNA to be 89% and the NPV to be 91% (Silvestri et al., 2013). Further, a multi-centre randomized controlled trial with 241 patients compared surgical staging alone combined with EBUS-TBNA followed by surgical staging if the needle approach was negative (Annema et al., 2010). The sensitivities of surgery, endosonography, and endosonography followed by surgery if the surgery was negative) were 79%, 85% and 94%, respectively. As such, EBUS-TBNA has been shown to outperform traditional cervical mediastinoscopy in regards to diagnostic statistics (Navani et al., 2012).

The use of EBUS-TBNA as the initial diagnostic and staging procedure in patients has garnered substantial support in thoracic surgery (Wahidi et al., 2016) and now is recommended by various guidelines (Sampsonas et al., 2018; Silvestri et al., 2013). However, this method is associated with wait times up to four weeks and increased costs resulting from needles, biopsies, specimens, cytotechnology time, pathologist time, endoscopy time and access to the required equipment. Additionally, the sensitivity of the EBUS for mediastinal LN staging depends on various factors which include skill of the operator, the skill of the cytotechnologist, the skill of the pathologist, the size of the LNs, the gauge of the needle, and the pretest probability of cancer. Consequently, inconclusive or non-diagnostic results are obtained in as many as 42.14% of EBUS-TBNA cases (Ortakoylu et al., 2015a). As treatment decisions cannot be made without accurate staging, this brings the lung cancer treatment cycle to a stand-still. Inconclusive and non-diagnostic results also mandate the need for a repeated EBUS-TBNA, which exacerbates the vicious



cycle of treatment delay, increased healthcare costs, and a potential increase in patient morbidity.

### **1.3 Ultrasonographic Features and Predictive Tools for Lymph Node Malignancy**

Several ultrasonographic features can be observed during EBUS-TBNA for LN malignancy prediction: 1) length of short axis; 2) shape; 3) margins; 4) echogenicity; 5) central hilar structure; and 6) central necrosis. A systematic review conducted by our research group at McMaster University identified the optimal ultrasonographic features for predicting malignancy in mediastinal LNs for clinical utility. A total of 13 studies with 1061 LNs (487) LNs were examined and all of the features were assessed by the endosonographer performing the procedure.

A. **Short Axis Length:** length equal to or greater than 10 mm is thought to be associated with malignancy. Gogia and colleagues confirmed that a short axis length less than 10 mm was considered an independent predictor of benign LNs in a multivariate regression model (Risk Ratio- 1.31, 95% Confidence Interval (CI): 1.107-1.549,  $p=0.002$ ) (Gogia et al., 2016).

B. **Shape:** determined in several studies by calculating the ratio of the long axis to the short axis of the LN. A ratio of less than 1.5 is considered round in shape and a ratio equal to or greater than 1.5 is considered oval in shape. Jhun et al. determined that

round shape was a significant predictor of malignancy in a univariate regression analysis (Jhun et al., 2014).

- C. **Margin Status:** may be categorized as well defined (>50% of border hyperechoic) or poorly defined. Well defined margin status is a predictor for malignancy while poorly defined margin status is a predictor of benign LNs. An analysis performed by Gogia and colleagues showed that the absence of well-defined margins has a 96% specificity for predicting benign LNs (Gogia et al., 2016).
- D. **Echogenicity:** can either be considered homogenous or heterogenous based on the grayscale texture of the LN from EBUS. Heterogeneous echogenicity was found to be a predictor of malignancy while homogenous echogenicity was found to be a predictor of a benign LN. In two multivariate analyses performed by Jhun et al. and Evison et al., heterogeneous echogenicity was found to be a significant predictor for LN malignancy (OR = 48, 95% CI: 8-282, P<0.001; OR= 3.1, 95% CI 1.4-6.7, p=0.005) (Evison et al., 2015; Jhun et al., 2014)
- E. **Central Hilar Structure (CHS):** is an ultrasonographic feature, with its presence being predictive of a benign LN and its absence being predictive of malignancy. Studies reported that CHS was an independent predictor of benign LNs (p=0.03) and the absence of CHS resulted in a sensitivity ranging from 89-99% sensitivity

and 90-92% NPV for the prediction of malignant disease (Fujiwara et al., 2010; Shafiek et al., 2014) .

F. ***Central Necrosis:*** is defined as the presence of centrally located hypoechoic structure within a LN. The presence of central necrosis is a predictor of malignancy while its absence is a predictor for benign LNs. Fujiwara and colleagues found central necrosis to be a significant predictor for pathologically confirmed LNs ( $p < 0.001$ ) and associated with a hazard ratio of 5.64 (95% CI: 3.40-9.38) by logistic regression (Fujiwara et al., 2010).

Overall, the results of the systematic review demonstrated that ultrasonographic features may assist during EBUS and diagnostic processes relating to LN biopsy. Additionally, the use of a predictive score may prevent the need for repeat EBUS procedures when initial biopsy results are inconclusive.

#### **1.4 Research Group's Preliminary Work**

Based on the systematic review of 13 studies published by our research group in 2018, it was determined that a composite of ultrasonographic features should be used when attempting to determine mediastinal disease (Hylton et al., 2018). A multicenter prospective validation clinical trial, coordinated by our research team, identified four ultrasonographic LN features that were clinically relevant predictors of malignancy

(Hylton et al., 2019). In order to improve upon the accuracy of predicting LN malignancy, our research group developed a highly specific predictive tool, investigating four ultrasonographic features predictive of malignancy (small axis length, margin, central hilar structure and central necrosis) (Hylton et al., 2019). Together, these four features formed the Canada Lymph Node Score (CLNS; **Figure 5**). However, when used in multiple centres across Canada, experienced endosonographers agreed on the CLNS diagnosis only 22.54% of the time (Hylton et al., 2019). This lack of consensus between endosonographers demonstrated that there was a high operator dependency associated with the tool.

### **1.5 Rationale for Artificial Intelligence in Lung Cancer Staging**

Despite the fact that nodal biopsies are considered the “gold standard” of LN staging by clinical guidelines, recent population level data from the United States has shown that as many as 50% of patients with lung cancer have been sent to treatment without an attempt at LN biopsies (Boffa et al., 2017; Little et al., 2005a). This finding suggests that the thoracic surgery community has largely abandoned nodal biopsies. However, most lung cancer surgeons would agree that abandoning biopsies is not beneficial for patients. Further, when biopsy results are inconclusive or insufficient, for cytological interpretation, EBUS-TBNA procedures need to be repeated or the patient must undergo a mediastinoscopy (Jalil, Yasufuku, & Khan, 2015). EBUS-TBNA samples are deemed inconclusive for pathological diagnosis in as high as 42.14% of cases and 29.85% of patients with inconclusive results are referred to mediastinoscopy or undergo repeat EBUS-

TBNA biopsy (NICE, 2019). Accordingly, there is a near unanimous agreement on the need to develop and study other methods of nodal staging.

Historically, trained physicians visually assessed medical imaging for detection, characterization, and monitoring of disease. However, Artificial Intelligence (AI) methods have been found to excel at automatically recognizing complex patterns and features, providing quantitative assessments of imaging characteristics. Human error associated with diagnostic tools has also spurred research towards the development of computer-aided algorithms with hopes to eliminate operator dependency as it is believed that AI can produce more precise measurements compared to humans. For this reason, in the early 1980s computer-aided diagnosis (CAD) systems were brought to assist doctors to improve the efficiency of medical image interpretation (Doi, 2007). The use of AI may be able to provide more accurate and precise measurements and eliminate error associated with CLNS. The effective implementation of AI for nodal staging could be used to develop an algorithm suitable for use in clinical settings.

## **1.6 Machine and Deep Learning Framework**

A special focus has been placed on novel methods to develop more accurate identification of lung cancer characteristics using deep machine learning and radiomics, which are both a form of AI (Henzler, 2017) (**Figure 6**). Deep learning allows computational models that are composed of multiple processing layers to learn

representations of data with multiple levels of abstraction (Lecun, Bengio, & Hinton, 2015). Information obtained from training allows deep learning algorithms to recognize patterns and perform accurate segmentations. Similarly, radiomics are textural mathematical constructs that capture the spatial appearance of the tissue of interest (shape and texture) on different types of images using texture (Parekh & Jacobs, 2019). Traditionally, radiomic features provide information about the grey-scale patterns, interpixel relationships, shape, and spectral properties within regions of interest on medical images. Feature extraction is the key step to adopt machine learning and various methods of feature extraction for different types of cancer have been investigated (Munir, Elahi, Ayub, Frezza, & Rizzi, 2019). However, these methods based on feature extraction have weaknesses and AI has faced critical appraisal. Concerns raised in this field include whether the study designs are biased in favour of the new technology, whether the findings are generalizable, and whether the study was performed in silico or in a clinical environment. Therefore, the degree the study results are applicable to the real-world setting have been questioned. As such, it is important to use rigorous methodology when designing and conducting machine learning studies.

## **1.7 Objectives**

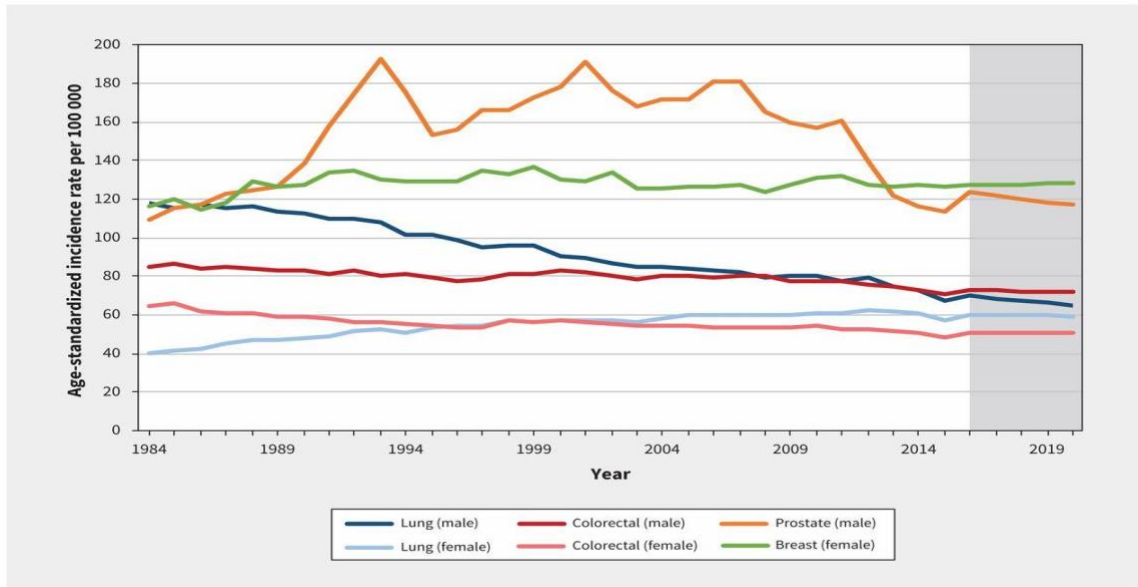
The use of AI and deep-learning computer neural networks have been shown to enhance clinician performance, predominantly through rapid and accurate image interpretation (Jha & Topol, 2016; Topol, 2019). Recent studies have investigated the use

of AI for the prediction of LN and tumour malignancy in positron emission tomography and computed tomography for NSCLC staging and diagnosis (H. et al., 2019; H. Wang et al., n.d.; Wnuk et al., 2014). These studies have demonstrated that computer-aided algorithms may result in higher performance diagnostics through the convergence of humans and AI. Therefore, as a potential solution to overcome the user dependency of the CLNS, we propose an innovative and novel approach: the use of NeuralSeg, a deep neural network, to segment ultrasonographic LN features to predict malignancy.

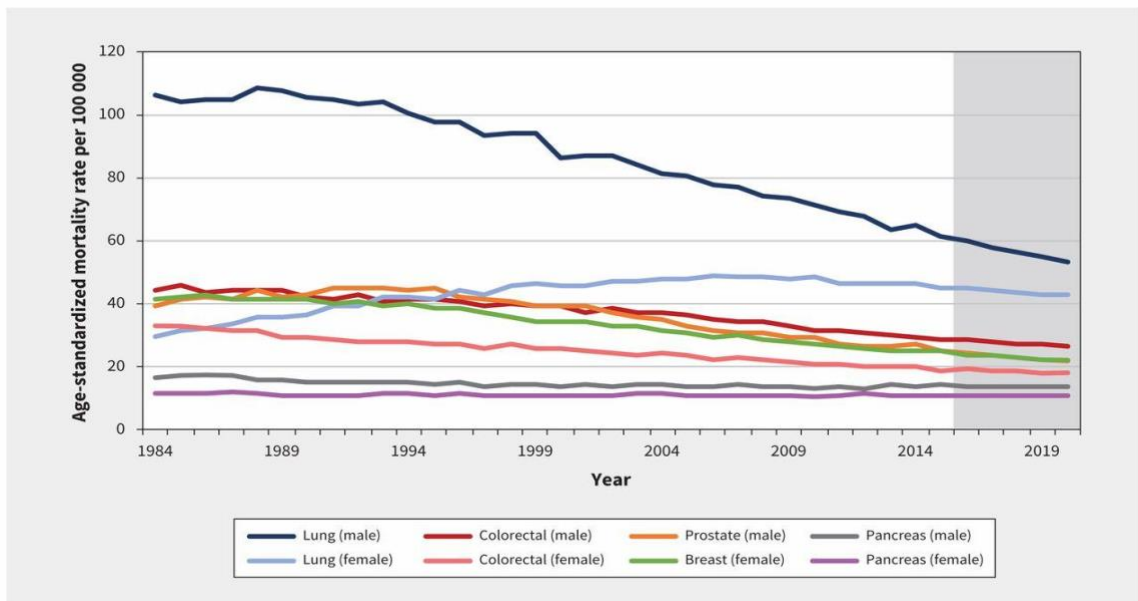
This thesis will focus on the use of machine learning, specifically a deep neural network, for the diagnosis and staging of lung cancer through medical imaging. We will aim to appraise the current literature on the use of radiomics, a form of deep learning, in the lung cancer population as well as develop and validate an algorithm capable of identifying ultrasonographic LN features observed during EBUS to predict malignancy.

The primary objective of this thesis was to therefore determine whether a deep neural AI network (NeuralSeg) can 1) segment ultrasonographic LN features from an existing derivation set of LN images examined during EBUS-TBNA and; 2) correctly apply the CLNS to a new validation set of LNs it has never seen before. The secondary objectives of this thesis are to 1) compare of the accuracy and reliability of the segmentation performed by NeuralSeg to the segmentation performed by the experienced endosonographer; 3) critically appraise the AI literature for the use of deep learning for diagnostic capabilities to diagnose and stage lung cancer.

**CHAPTER 1 TABLES AND FIGURES:**



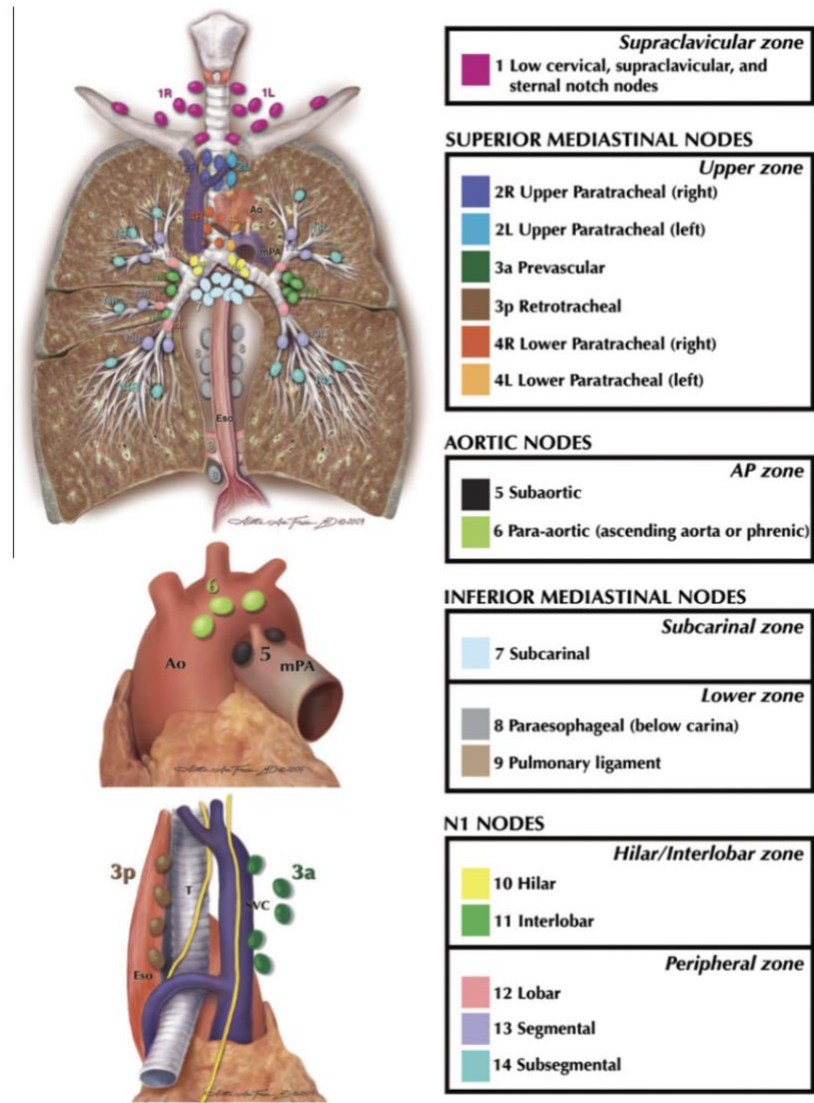
**Figure 1.** Age-standardized incidence rates (ASIRs) for selected cancers, in Canada (excluding Quebec), 1984-2020 by sex. Retrieved from Brennar et al. (2020)




**Figure 2.** Age-standardized mortality rates (ASMRs) for selected cancers in Canada, 1984-2020, by sex. Shading indicates projected data. Retrieved from Brennar et al (2020).


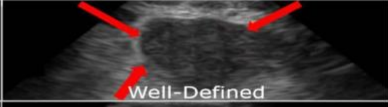



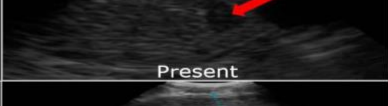
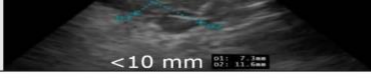
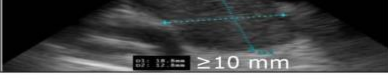






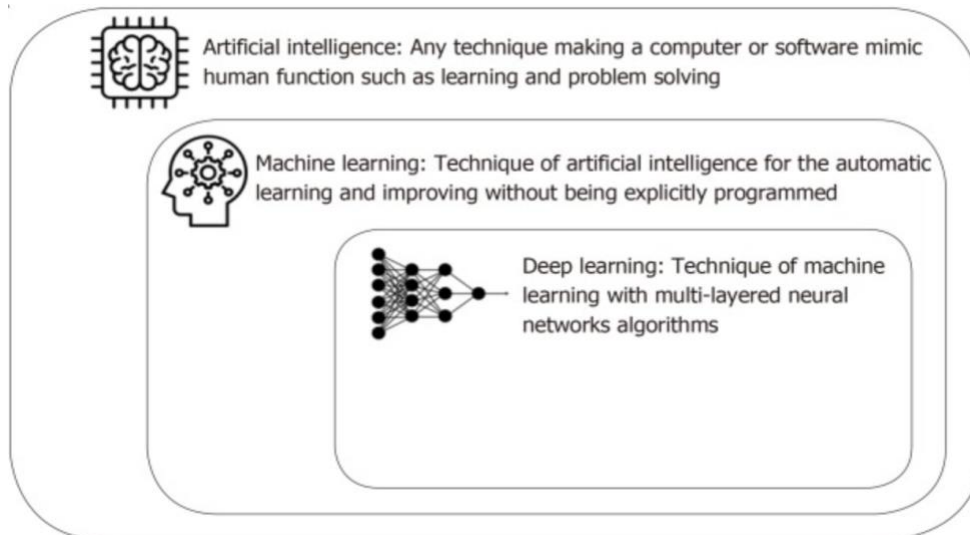
**Figure 4.** International Association for the Study of Lung Cancer (IASCL) lymph node map. Retrieved from Rusch et al. (2009)

 **Canada Lymph Node Score**

Ultrasonographic Features	Benign Features (0 points)	Malignant Features (1 point)
Margins	 Indistinct	 Well-Defined
Central Hilar Structure	 Present	 Absent
Central Necrosis	 Absent	 Present
Small Axis Diameter	 <10 mm	 ≥10 mm

**SCORE:** 0-1 = Low chance of malignancy | 2-4 = High chance of malignancy

**Figure 5. Canada Lymph Node Score Criteria.** Benign and malignant criteria of all four ultrasonographic lymph node features with binary scoring. A total score of four can be achieved for each lymph node examined.



**Figure 6:** Schematic graphical representation of Artificial Intelligence, machine learning and deep learning. Retrieved from Yang 2019

**CHAPTER 2: APPLICATION OF RADIOMICS TO PREDICT LYMPH NODE MALIGNANCY FOR THE STAGING OF NON-SMALL CELL LUNG CANCER IN THORACIC MEDICAL IMAGING: A SYSTEMATIC REVIEW**

Churchill, I.F., Sullivan, K., Simone, A., Patel, Y.S., Leontiadis, G., Farrokhyar, F., Gatti, A.A., Hanna, W.C.

**ABSTRACT:**

**Background:** Medical imaging is one of the most valuable sources of diagnostic information, but it is heavily reliant on human interpretation, which can be error prone. Radiomics demonstrate the potential for objectivity in highlighting suspicious regions in images; detecting indeterminate nodules and tissues; and addressing the high positive rates that may lead to overdiagnosis.

**Research Question:** We aimed to answer two research questions: 1) what is the accuracy of radiomics with CT, PET and EBUS for mediastinal staging of NSCLC? 2) How does the accuracy of radiomics with CT, PET and EBUS compare to these imaging modalities without radiomics?

**Study Design and Methods:** The literature was systematically searched using Cochrane Central Register of Controlled Trials, MEDLINE, EMBASE and Web of Science for observational studies between the databases' inception and January 2020.

**Results:** The literature search identified 4,954 potentially relevant studies (after 1,073 duplicates were removed). After screening abstracts (n=4,954) and full texts (n=72), 19 studies were included, 17 which provided full reports, while two recorded provided data from conference findings. Overall, 3265 patients were enrolled with a total of 3472 LNs (70% malignancy). The most common radiomic approach to assess images was a Support Vector Machine (5/19, 26%) followed by an Artificial Neural Network (4/19, 21%).

Sensitivity and specificity were the second most commonly reported diagnostic statistics with 11 studies reporting both measures and area under the curve (AUC) was the most reported measure, with 13 studies using AUC to discern discrimination potential. Sensitivities for algorithms ranged from 52-99%, while specificity for the algorithms ranged from 62-94%. Clinicians reported similar sensitivities and specificities ranging from 72-95% and 52-92%, respectively. AUC c-statistics were only reported for algorithms and not the clinician. C-statistics ranged from 0.64-0.94 suggesting that the algorithms possessed good discrimination potentials.

**Interpretation:** As data could not be pooled, only a summary of the literature could be provided. The estimation and comparison of the reported statistics of an index test for each imaging modality should be interpreted with caution as they may not have been evaluated at a common threshold. Analysis of staging modalities for lung cancer using radiomics may be useful for when results are inconclusive or uninterpretable by a human reader. However, prospective external validation of these algorithms and direct comparisons through cut off thresholds is required to determine their true predictive capability.

**PROSPERO ID:** CRD42020162952

## 2.1 BACKGROUND

Accurately determining the stage of NSCLC is important in order to ensure that patients are offered the best treatment options (Department of Health, 2011). Consequently, mediastinal staging is the rate-limiting step prior to initiation of NSCLC treatment and is essential in identifying the most appropriate treatment for the patient. However, this process is often complex and involves multiple imaging modalities including invasive and non-invasive imaging methods for the assessment of lymph nodes in the mediastinum. Mediastinal staging is usually undertaken by computed tomography (CT) scans, positron emission tomography (PET) scans and endobronchial ultrasound-transbronchial needle aspiration (EBUS-TNA) (Cancer Care Ontario, n.d.). Other factors that may influence both the diagnostic and treatment pathway include the location of the tumour, the extent of cancer spread to the mediastinal lymph nodes (LNs), and the pulmonary fitness of the patient.

Many research studies have focused on using non-invasive  $^{18}\text{F}$ -FDG PET/CT images for the diagnosis of mediastinal LN metastasis, where judgments are mostly based on thresholding image features and metabolic features (i.e. standard uptake values). However, in the past 10 years, the median sensitivity for mediastinal LN NSCLC diagnosis using  $^{18}\text{F}$ -FDG PET/CT was 62%, due to the low spatial resolution this modality possesses, resulting in large false-negative rates (Silvestri et al., 2013). Furthermore, invasive methods, such as EBUS-TBNA, are not routinely performed in every patient, especially those with occult lymph nodes. However, in circumstances where EBUS-TBNA is

indicated, as much as 40% of biopsy results are inconclusive (Ortakoylu et al., 2015a). In order to improve these diagnostic tests a more sophisticated classification strategy is needed.

### **2.1.1 Target Condition**

Worldwide, NSCLC is more fatal than colon, breast and prostate cancers combined. Currently, lung cancer accounts for 14% of all cancers in Canada. Incidence rates for NSCLC by age group follow an upward trend, implying that the likelihood of developing NSCLC increases with age (Smith et al., 2018). In 2016, the Canadian Task Force on Preventive Health Care (CTFPHC) released a guideline recommending annual screening for lung cancer in high risk adults aged 55-74 years using low dose CT (Care, 2016). Accordingly, the expansion of lung cancer screening programs is resulting in an increased incidence of NSCLC across Canada (Akhtar-Danesh & Finley, 2015; Evans et al., 2016).

The diagnosis of NSCLC is made by a variety of different biopsies and imaging techniques, some of which yield information about both diagnosis and staging (NICE, 2019). The need to consider the location of the primary tumour, patient preferences, and the fitness of the patient may require a change to the diagnostic and staging pathway, thus augmenting the complexity of the process. If the disease has not spread to either the ipsilateral mediastinal nodes, subcarinal (N2) nodes, or both, and the patient is otherwise considered fit for surgery, resection is often the treatment of choice (Barnes et al., 2017). Those patients who are found to have unresectable NSCLC will usually have undergone a number of tests to identify affected LNs and confirm the pathological stage of their cancer.

Therefore, the reference standard for this review will consist of a number of tests that can yield pathological information to provide cytohistological confirmation of the tumour extent. These reference standards will include tumour review boards, expert consensus, nodal and lung pathology.

### **2.1.2 Index Tests**

In the era of precision medicine, radiomics in medical imaging is an emerging field offering vast potential. Radiomics is a complex multi-step process that aims to find associations between qualitative and quantitative information extracted from both medical and clinical imaging (Gillies, Kinahan, & Hricak, 2016). This process may help in clinical decision-making and outcome prediction through the segmentation of imaging features, by serving as a decision support tool. Research has shown the capacity of radiomic analyses to help distinguish cancerous from benign tissues and can help to determine cancer staging (Bi et al., 2019). As such, radiomics may be a non-invasive modality in combination with NSCLC imaging to identify nodal disease and may be useful for discriminating malignant characteristics to facilitate decision making.

## **2.2. Clinical Pathways**

Patients may present with signs suspicious for lung cancer such as hemoptysis; cough; weight loss or loss of appetite; chest, rib or shoulder pain; and/or dyspnea (Cancer Care Ontario, n.d.). If symptoms are present for greater than three weeks or if patients have identified risk factors, they are usually referred first for chest imaging via CT, as dictated by cancer imaging guidelines (NICE, 2019). If results remain suspicious of NSCLC,



patients undergo an organized diagnostic assessment where they may undergo PET, abdominal CT, magnetic resonance imaging (MRI) brain scans or bone scans. Depending on the imaging findings, patients may then undergo EBUS-TBNA to assist with treatment decisions. Exclusion of N1-N3 nodal involvement dictates patients would most benefit from surgical resection, while those with N1-N3 nodal involvement usually undergo chemoradiation therapy (Burdett et al., 2015).

### **2.2.1 Role of Index Tests**

Radiomics can rapidly extract innumerable quantitative features from digital medical images. Accordingly, radiomics can support decision-making in both the staging and diagnosis of NSCLC. If radiomics is demonstrated to be a clinical adjunct to PET-CT and EBUS clinical pathway, then it is envisioned that the diagnostic performance of radiomic-augmented imaging will be superior to the comparator tests alone. This may also eliminate the need for current biopsies and result in a substantial decrease in healthcare costs.

### **2.2.2 Alternative Tests**

Other imaging modalities can provide similar information to PET-CT, and these include contrast-enhanced MRI and single photon emission-computed tomography (SPECT) (Schmidt-Hansen et al., 2014). However, neither of these tests are commonly used in the lung cancer pathway.

### **2.2.3 Rationale**

Medical imaging is one of the most valuable sources of diagnostic information, but it is heavily reliant on human interpretation, which can be error prone. Radiomics demonstrate the potential for objectivity in highlighting suspicious regions in images; detecting indeterminate nodules and tissues; and addressing the high positive rates that may lead to overdiagnosis. However, for the purpose of this review, the focus of radiomics will be to obtain accurate information from the process of lymph node staging. CT and PET help determine which lymph nodes should be biopsied based on lymphadenopathy ( $\geq 1$  cm) and hypermetabolism, respectively. Both CT and PET are associated with specificities above 85% and a false-negative rate ranging between 20-25% (Herth, 2013; Navani et al., 2015). However, approximately 40% of enlarged mediastinal lymph nodes on CT are benign and 25% of hypermetabolic lymph nodes are false positives (Navani et al., 2015). Therefore, relying on CT and PET for mediastinal staging alone can both under stage and over stage patients. Clinicians must have a clear idea of the likelihood of false positive and negative PET-CT results, in order to best manage patients and advise them on whether a biopsy is necessary. A false negative rate that is consistently above 20% would cause clinicians to question the utility of the test. Additionally, EBUS-TBNA has been reported to generate inconclusive results in as much as 40% of the time. Therefore, there is a need to eliminate the operator dependency of these diagnostic imaging tests in order to improve diagnostic accuracy.

#### **2.2.4 Objectives**

To the best of our knowledge, radiomics for LN staging and diagnostic imaging applications in NSCLC have yet to be systematically summarized and reviewed in the clinical literature. In this review, we aimed to critically appraise the current use of radiomics in medical imaging for the staging and diagnosis of nodal involvement in NSCLC.

**Primary Objective:** The primary objective of this review was to determine the diagnostic accuracy of radiomics for mediastinal staging in patients undergoing CT, PET or EBUS imaging procedures for NSCLC. We aimed to address the following research questions:

- 1) What is the accuracy of radiomics with CT, PET and EBUS for mediastinal staging of NSCLC?
- 2) How does the accuracy of radiomics with CT, PET and EBUS compare to these imaging modalities without radiomics?

**Secondary Objective:** The secondary objective of this review was to investigate heterogeneity originating from imaging modality and study design in patients with suspected or confirmed NSCLC undergoing CT, PET or EBUS procedures.

## **2.3 METHODS**

This systematic review was written according to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Campbell et al., 2015) and followed the PRISMA guidelines (Moher et al., 2016).

### **2.3.1 Criteria for Considering Studies for this Review**

Eligibility criteria for studies are summarized in **Table 1**.

#### **2.3.1.1 Types of Studies**

Studies eligible for this systematic review included both prospective and retrospective observational studies. Quasi-randomized trials were considered for eligibility. However, none were identified. No protocols or editorials were included. Abstracts were included when data was able to be extracted.

#### **2.3.1.2 Participants**

Trial participants were 18 years of age or older who were undergoing nodal staging and diagnosis of NSCLC at a secondary or tertiary care facility. This systematic review was inclusive towards studies that compared NSCLC patients to those with other conditions or healthy controls, however none of these studies were identified. No exclusion criteria was applied towards NSCLC patients that presented with various comorbidities.

### **2.3.1.3 Index Tests**

The index tests were defined as the complementary addition of radiomics to the various diagnostic medical imaging procedures (*refer to Section 2.3.1.4*). Radiomics is best described as the process of converting medical images into quantifiable mineable data such that it can assist in clinical decision making (Gillies et al., 2016). Radiomics encompasses a wide array of algorithmic features including segmentation, deep machine learning, and convolutional neural networks. This systematic review focused on radiomics, specific in regard to diagnosis and staging (i.e. computer-assisted diagnosis), rather than those with a prognostic or treatment-response purpose (i.e. survival analyses, radiotherapy, chemotherapy response) (Hatt et al., 2011). Radiogenomics were excluded from this review as the algorithmic assistance in identifying molecular biomarkers is conducive of prognostic and treatment-response based studies.

### **2.3.1.4 Comparator Tests**

Traditional diagnostic imaging procedures for NSCLC, without the inclusion of radiomics, were considered the comparator tests with their results interpreted by a radiologist or endosonographer. These standard imaging procedures encompassed:

- *Positron Emission Tomography (PET)* – whole body imaging or exclusive towards the thorax;

- *Computerized Tomography (CT) or Computerized Axial Tomography (CAT)* – whole body imaging or exclusive towards the thorax;
- *Endobronchial Ultrasound (EBUS)*.

#### **2.3.1.5 Target Condition**

NSCLC was the target condition with no exclusion towards any of its clinical stages or histological types. This systematic review was interested in studies involved in the diagnosis of NSCLC compared to benignity or other health conditions. Small Cell Lung Cancer (SCLC) was excluded given its faster onset of metastatic spread and different staging characterization (Kalemkerian & Schneider, 2017).

#### **2.3.1.6 Reference Standards**

- Various gold reference standards were accepted as establishing the presence or absence of NSCLC. They included:
- *Tumour Review Boards*: in which a consensus is reached amongst experts in the lung cancer field.
- *Lung Pathology*: histopathology of the lung obtained during lung resection (surgical pathology), EBUS-TBNA biopsies, bronchial brush or bronchoalveolar lavage.
- *Nodal Pathology*: histopathology of sampled LNs during lung resection (surgical pathology), EBUS-TBNA biopsies and mediastinoscopy.

## **2.3.2 Search Methods for Identification of Studies**

### **2.3.2.1 Electronic Searches**

We searched the Cochrane Central Register of Controlled Trials (CENTRAL; inception to Cochrane Central Register of Controlled Trials 2020, Issue 1), Ovid MEDLINE® and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily (1946 to January 14, 2020), EMBASE (1974 to 2020 Week 2), and Web of Science (All Databases; 1926 to January 2020). Databases were not limited from any publication date nor will there be any restrictions placed on language or publication status. To account for ongoing trials, ClinicalTrials.gov will be searched up until January 14, 2020 with the inclusion of any type of trial design (*Appendix 1*).

### **2.3.2.2 Searching Other Resources**

Relevant systematic reviews and literature reviews identified during title and abstract screening were hand searched to identify any potential eligible trials and ensure literature saturation. Moreover, for studies and conference abstracts that demonstrated uncertainty in their eligibility or provide insufficient details, the authors were contacted by email to obtain further details and/or additional data.

### **2.3.3 Data Collection and Analysis**

#### **2.3.3.1 Selection of Studies**

Three review authors (IC, KS, and AS) independently screened the title and abstract of the articles derived from the search strategy, in which one review author screened all titles (IC) for a paired comparison. In cases of disagreement, a fourth reviewer (YP), determined if the study was eligible for full text screening. The three review authors (IC, KS and AS) completed full text screening where they independently assessed the potentially eligible studies for inclusion for a paired comparison. Disagreement during full text screening was resolved through discussion and when agreements could not be achieved, a fourth reviewer (YP) made the final decision. The entire screening process was conducted through the online systematic review screening software, Covidence ©.

#### **2.3.3.2 Data Extraction and Management**

For each study, two review authors (IC and KS) independently extracted data to obtain the following information:

1. *Study characteristics* (e.g. setting, study author, study design, type of trial, funding, country, year of publication, participants)
2. *Performance and validation* (e.g. algorithm information, reference test, index test, comparator test, validation method)
3. *Accuracy* (Area under the curve (AUC), false positive, false negative, true positive, true negative, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, odd ratios)



Data extraction was performed using a data collection form created on Microsoft® Office Excel. Extractions discrepancies were compared and discussed. When agreement was not reached, consultation with a third reviewer (YP) resolved discrepancies.

### **2.3.3.3 Assessment of Methodological Quality**

Two review authors (IC and KS) independently assessed the methodological quality of each included study, using QUADAS-2 (Whitting Group). The assessment consisted of four domains: patient selection; index test(s); reference standard; and flow and timing (*Appendix 2*).

### **2.3.3.4 Statistical Analysis and Data Synthesis**

Diagnostic statistical data were not able to be pooled for studies due to reporting inconsistencies. However, we avoided meta-analysis and qualitatively synthesized and reported the information from included studies.

## **2.4 RESULTS**

### **2.4.1 Results of Search**

There were no publication date restrictions on the search strategy. As a result, the search included relevant records between inception and January 2020. The literature search identified 4,954 potentially relevant studies (after 1,073 duplicates were removed) from the following databases: EMBASE (n=1,972), MEDLINE Ovid and Epub Ahead of Print, In-Process and Other Non-Indexed Citations and Daily (n=1,083), CENTRAL (n=815) and

Web of Science (n=2,047) and clinicaltrials.gov (n=96). Hand-searching identified 14 other potentially relevant studies. After screening abstracts (n=4,954) and full texts (n=72), 19 studies were included, 17 which provided full reports (Ferreira-Junior et al., 2020; Flechsig et al., 2017; Gao et al., 2015; He et al., 2019; Inoue et al., 2011; X. Liu et al., 2019; Na et al., 2018; Pham, 2018; Pham, Watanabe, Higuchi, & Suzuki, 2017; Song, Cai, Eberl, Fulham, & Feng, 2011; Song, Cai, Kim, & Feng, 2012; Tagaya, Kurimoto, Osada, & Kobayashi, 2008; Teoh et al., 2016; Toney & Vesselle, 2014; Vesselle, Turcotte, Wiens, & Haynor, 2003a; H. Wang et al., 2017; Zhong et al., 2018), while two provided data from conference findings (Bella, Dancewicz, Szczęsny, & Kowalewski, 2013; Wang et al., 2018). In addition, a total of five studies (Genseke, Wielenberg, Schreiber, & Walles, 2019; He et al., 2017; Wnuk et al., 2014; Zhao & Shi, 2018; Zhu, Xu, Xiao, & Zhou, 2019) are awaiting classification and three studies are currently ongoing (NCT03849040, NCT03648151, NCT04000620). The details of the screening and selection process are illustrated in **Figure 1**.

#### **2.4.2 Included Studies**

The main characteristics of the eligible studies, which were published from 2003 through 2019, are reported in *Appendix 3 (Characteristics of Included Studies and Risk of Bias)*. Overall, 3265 patients were enrolled with a total of 3472 LNs (70% malignancy). Studies were conducted in nine different countries, with a mean age of 64 years (SD= 5 years) and 59% (SD=12%) of the patients were male. There was a mean of 175 patients enrolled per study (range: 14 to 717), in which two patient cohorts were each used twice.

Most of these studies (11/19, 58%) were published after 2015. The available evidence came primarily from retrospective studies (12/19, 63%), which enrolled patients from Asian countries in 10 studies (53%). The target condition was lung cancer in 18 out of 19 studies (95%), whereas in the remaining study the authors focused on lung diseases such as non-small cell lung cancer, small cell lung cancer and sarcoidosis (Tagaya et al., 2008). In regard to imaging modalities, seven studies (37%) assessed CT images, one study (5%) assessed EBUS images, two studies (11%) assessed PET images and nine studies (53%) assessed both PET/CT images. Regions of interest for segmentation varied throughout the studies. Eight out of the 19 studies (50%) selected LNs of their region of interest, four studies (21%) segmented tumours as their region of interest and six studies (32%) segmented both tumours and LNs as their region of interest. One study did not report the algorithm's region of interest (Bella, 2013).

#### **2.4.3 Methodological Quality of Included Studies**

Overall, the methodological quality of studies was considered to be low as displayed in the risk of bias QUADAS-2 results summary (**Figure 2**). There was concern for at least one study in each domain based on the analysis of the available data. Four studies were considered high risk of bias for patient selection (X. Liu et al., 2019; Toney & Vesselle, 2014; Vesselle et al., 2003a; Zhong et al., 2018). One did not describe how patients were enrolled and only included patients with clinical N stage N0 (X. Liu et al., 2019). Two studies were deemed high risk for the same reasons: although a prospective design was employed, consecutive patient enrollment was not utilized and there was

exclusion for stage IV disease (Toney & Vesselle, 2014; Vesselle et al., 2003a). Similarly, another study was also at high risk due to patient exclusion criteria as there was strict inclusion (i.e. exclusion of >N0) (Zhong et al., 2018).

Three studies were deemed high risk for patient selection, no study was deemed high risk for the index test domain, four studies were deemed high risk for the reference standard domain and five studies were deemed high risk for flow and timing. Innoue (2011), Liu (2018) and Zhong (2018) were considered to be high risk as patients were not consecutively sampled, nor were their exclusion criteria justifiable. Veselle (2003) was considered high risk as the feedback of surgical nodal staging results was provided to the index test assessor. Innoue (2011), Song (2011), Song (2012) and Veselle (2003) were at high risk of bias as it was unclear if the reference standard was able to correctly classify the target condition and there was no mention of blinding. Finally, five studies possessed a high risk of bias for the flow and timing domain as it was unclear on how clinical observations played a role in timeline and there was a lack of justification for patients included in analysis. Overall, uninterpretable index test or reference standards results were rarely reported. In most studies, the interval between the index test and reference test were unreported. However, we believe that the length of time between the index test and reference standard would not undermine the reliability or accuracy of the results and treatment plan, since the diagnosis of malignant LNs is usually an indication for chemoradiation, and thus pathological evaluation would have occurred within a very short

time since EBUS staging (some days/ a few weeks). As a result, this would unlikely be sufficient time for the stage of the disease to change.

#### **2.4.4 Findings**

Performance and validation characteristics are presented in **Table 2**.

##### **2.4.4.1 Diagnostic Statistics**

Only one study provided information regarding true positive, false positive and false negative results. However, as the number of LNs in the sample and the true negative results were not reported, we were unable to construct contingency tables nor were we able to calculate diagnostic statistics. However, the authors provided a malignancy prediction accuracy of 91% for the algorithm and 78% for the experienced clinician. Ten of the 19 studies reported the accuracy of the algorithm. Accuracies ranged from 56-99% depending on the software and region of interest, while accuracies for clinicians' segmentations possessed a smaller range from 78-92%. Two studies reported odds ratios for their prediction models with increased odds of 4.546 (95% confidence interval= 2.347-8.806,  $p < 0.001$ ) (He, 2019) and decreased odds of 0.35 of malignancy (95% CI= 0.21-0.59,  $p$ -value=NR) compared to standardized diagnostic imaging. (Liu, 2018) Negative predictive values (NPV) and positive predictive values (PPV) were reported for one study (Pham 2017a), for their various algorithms that were assessed. Overall, their algorithm's NPV and PPV ranged from 62%-86% and 62-93%, respectively. Sensitivity and specificity were the second most commonly reported diagnostic statistics with 11 studies reporting both

measures and area under the curve (AUC) was the most reported measure, with 13 studies using AUC to discern discrimination potential. Sensitivities for algorithms ranged from 52-99%, while specificity for the algorithms ranged from 62-94%. Clinicians reported similar sensitivities and specificities ranging from 72-95% and 52-92%, respectively. AUC c-statistics were only reported for algorithms and not the clinician. C-statistics ranged from 0.64-0.94 suggesting that the algorithms possessed good discrimination potentials.

#### **2.4.4.2 Study Methods and Validation Reporting**

Sixteen out of 19 studies (84%) used a form of validation (i.e. cross-validation [15/16] or bootstrapping [1/16]). However, of the 19 studies, only one study externally validated their radiomic software (Teoh et al., 2016). No studies reported a pre-specified sample size calculation. 10 out of the 19 studies (53%) specified that image pre-processing occurred to exclude low-quality images and prepare images for segmentation. Three studies (16%) also tested the scenario where the clinicians' segmentation of the regions of interest were compared to the deep learning algorithm. Years of experience of the clinician completing the segmentations ranged from 5-15 years, suggesting learning curve bias and that some clinicians may have been more experienced than others. In some studies, rater's segmentations with differing years of experience were compared to determine if the learning curve affected the results produced by both the clinicians and the algorithm.

### **2.4.4.3 Index and Reference Test Reporting**

The most common radiomic approach to assess images was a Support Vector Machine (5/19, 26%) followed by an Artificial Neural Network (4/19, 21%). Seven studies adopted multiple algorithms to determine segmentation capabilities. Reference standards used a range of tests in line with the lung cancer diagnosis pathway. Eight studies (42%) used surgical pathology, three studies (16%) used EBUS-TBNA, two studies used the ground truth from expert radiologists, five studies (26%) used histopathological confirmation (unspecified) as their reference standard test.

## **2.5 DISCUSSION**

### **2.5.1 Summary of Main Results**

Although early attempts at computerized analysis of medical images were made in the 1960s, serious and systematic investigation on computer aided diagnosis began in the 1980s with a fundamental change in the concept for utilization of the computer output, from automated computer diagnosis to computer-aided diagnosis (Doi, 2007). However, the first reported study (Aquino et al., 2003) for the prediction of LN malignancy in lung cancer was published in 2003. Registration of CT and FDG–PET of 45 datasets with 130 LNs significantly improved the specificity of detecting metastatic disease. In addition, registration improved the radiologic staging of lung cancer patients when compared with CT or FDG–PET alone (Aquino et al., 2003). Following this study, the use of radiomics for identifying specific features present on thoracic imaging were explored.

Overall, our results showed that the radiomic algorithms performed as well if not better than clinicians at predicting LN malignancy. Three other systematic reviews examining the diagnostic accuracy of malignancy prediction from medical imaging were identified (Jethanandani et al., 2018; X. Liu et al., 2019; Traverso, Wee, Dekker, & Gillies, 2018). However, none of these systematic reviews investigated the use of radiomics for LN malignancy prediction to stage lung cancer alone. Differences in assessment measures, search strategies, consideration of meta-analyses and interpretation of results were clearly demonstrated between the systematic reviews and our own. For instance, our study was the only one to include a risk of bias assessment using QUADAS-2, a verified risk of bias tool, which allowed us to derive more conclusions regarding the current best evidence. Of the three reviews, Jethanandani (2018) was the most similar to ours as it was the only one to assess the use of radiomics in medical imaging for the prediction of one type of cancer nor did they appear to impose language restrictions upon the search strategy. Despite the improved comprehensiveness of their search strategy, the ability to pool data and the improved quality of evidence for studies, our systematic review arrived at the overall same conclusion as the other systematic reviews. That is the evidence suggests that radiomic algorithms have high accuracy in predicting malignancy from medical imaging.

### **2.5.2 Strengths and Weaknesses of Review**

The strength of this systematic review is that it is fully bias controlled. The complete process was conducted independently and in duplicate. Two review authors independently carried out title and abstract screening, full text screening, data abstraction



and risk of bias assessment. Publication bias may exist as results were not able to be pooled, and as a result could not be assessed using the funnel plot method. Additionally, unpublished studies were identified through clinical trial registries, mitigating the chance of publication bias. However, data from unpublished studies was not obtained and it is possible that studies with low diagnostic accuracy were not published. As such, publication bias cannot be ruled out. Nevertheless, we did our utmost to reduce such bias in the review process incorporating a comprehensive search strategy that had no publication date restriction nor any language restrictions. Another limitation was failing to quantitatively synthesize data from included studies due to inconsistencies and lack of sufficient information. We however did a comprehensive qualitative review. Finally, the included studies have low reporting quality. This is evident from number of “unclear risk” in our quality assessment. Although our review might have been limited by “unclear risk” for many domains in the risk of bias assessment, we tried to limit this option to circumstances in which no information was provided to make an informed judgement.

### **2.5.3 Applicability of Findings to Review Question**

We are confident that our comprehensive search strategy identified all relevant observational studies investigating the diagnostic accuracy of radiomics for lung cancer staging. In addition to searching various databases, a large effort was put into identifying grey literature and hand searching the citations of relevant systematic reviews. Our systematic review attempted to formulate a research question that was applicable to the general lung cancer population and measuring outcomes objectively when possible (e.g.

sensitivity, specificity, AUC, OR, NPV, PPV and accuracy). In terms of generalizability, our review did not exclude patients based on the stage of their disease nor the confirmation of lung cancer in order to capture the full range of the disease. The results presented from our study qualitatively answered the review questions. However, we were unable to make direct comparisons due to the inability to pool data. Additionally, some studies limited their inclusion criteria and as a result may have impacted the generalizability of results (Y. Liu et al., 2018; Toney & Vesselle, 2014; Vesselle, Turcotte, Wiens, & Haynor, 2003b; Zhong et al., 2018)

## **2.6 AUTHOR'S CONCLUSION**

### **2.6.1 Implication for Practice**

As data could not be pooled, only a narrative summary of the literature could be provided. The estimation and comparison of the reported statistics of an index test for each imaging modality should be interpreted with caution as they may not have been evaluated at a common threshold. Analysis of staging modalities for lung cancer using radiomics may be useful in clinical practice for when results are inconclusive or uninterpretable by a clinician. However, prospective external validation of these algorithms and direct comparisons through cut off thresholds is required to determine their true predictive capability.

### **2.6.2 Implication for Research**

Nearly all of the studies evaluated the performance of radiomic algorithms for diagnostic analysis of medical images were designed as feasibility studies and did not have the design features that are recommended for robust validation of the real-world clinical performance of radiomic algorithms. In order to make accurate comparison of deep learning methods, it is important to develop standards for study protocols and reporting that recognize specific challenges of deep learning to ensure quality and interpretability of future studies.

**CHAPTER 2 TABLES AND FIGURES:**

**Table 1.** Eligibility criteria of studies for inclusion in systematic review

<b>Inclusion</b>	<b>Exclusion</b>
18 years of age	Small Cell Lung Cancer
Observational (prospective & retrospective), randomized & quasi-randomized controlled studies	Case reports
At least one study arm contains NSCLC patients	Radiogenomics (molecular biomarkers)
Radiomics are used (e.g. algorithms, segmentation, computer-assisted diagnosis [CAD], convolutional neural networks, deep learning, machine learning)	Studies where main objective involves survival-prediction, prognostic or treatment-response (i.e. radiotherapy, SBRT, chemotherapy)
CT, PET or EBUS	MRI and X-Rays (i.e. chest radiographs)
Must assess nodal disease (not just lung nodules)	

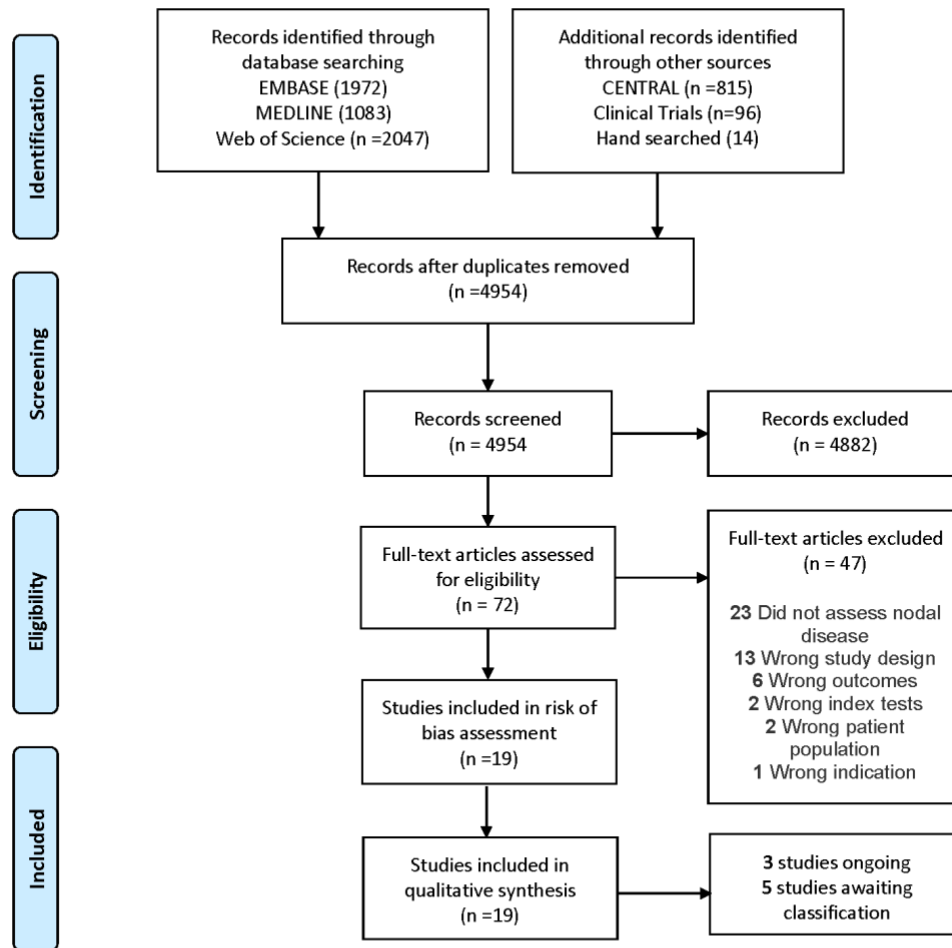
**Table 2.** Performance and validation characteristics of included studies.

Author	Year	Target Condition	Reference Test	Index Test	Comparator Test	Manual Segmentation Training?	Segmentation Software (if manual segmentation)	Years of Training of Human Rater (if specified)	Region of Interest (ROI) (if specified)	Features	Image Pre-Processing?	Radiomic Software Name	Internal Validation (Y/N)	External Validation (Y/N)
Bella	2013	Lung Cancer	EBUS-TBNA or Mediastinoscopy or Lymphadenectomy via Thoracotomy	Artificial Neural Networks (ANNs)	CT/PET	NR	NR	NR	Mediastinal LNs	LN weight, SUV, length and volume	NR	NR	Yes	NR
Ferreira-Junior	2009	Lung Cancer	Biopsy or Surgical Resection	Volumetric Segmentation	CT	Yes	3D Slicer	12 years	Lung lesions/ neoplasms	2465 shape, first-order, second order, and higher-order attributes, gray-level intensity, histogram, co-occurrence matrix, neighborhood intensity matrix, run-length matrix, Tamura texture, Laplacian of Gaussian filters, Gabor filters, Fourier transform, Haar wavelet, fractal dimension and shape	Yes	GrowCut	NR	NR
Flechsig	2017	Lung Cancer	Histologically confirmed	Volumetric CT Histogram Analysis with Semi-automated Segmentation	FDG-PET/CT	Yes	NR	5 years	LN's	Density, short axis and volume, SUVmax	Yes	NR	Yes-Post-procedural validation	NR
Gao	2015	Lung Cancer (NSCLC)	Surgical Resection (Pathological Results)	Support Vector Machine (SVM)	FDG-PET/CT	Yes	NR	NR	LN's	512 histogram vectors from CT images and 534 vectors from PET and CT. Texture features with gray-level co-occurrence matrix.	Yes	NR	Yes	No
He	2019	Lung Cancer (NSCLC) with lymphadenectomy	Surgical Resection (Pathological Results)	Radiomic-based Predictive Risk Score	CT	Yes	NR	15 years and 12 years (Segmentation: 10+ years)	Lung tumor	591 quantitative features (92 features showed independence)	Yes	Inhouse radiomics analysis software with algorithms implanted in Matlab.	Yes-Bootstrapping	No
Inoue	2011	Lung Cancer	Histologically confirmed or clinical observation over a year	3D- ordered subset expectation maximization (OSEM)	PET/ CT + 2D-OSEM, PET/CT + FORE + OSEM	No- human raters used a scoring system	NA	"Experienced" (Not specified)	LN's	Max SUV of tumor and LN metastasis. Contrast ratio, image noise, signal to noise ratios.	NR	3D-OSEM Algorithm, (VUE point, GE Healthcare)	Yes- Phantom Study	No
Liu	2018	Lung Cancer (Adenocarcinoma)	Surgical Resection (Pathological Results)	Cognition Network Technology	CT	Yes	Definiens Developer XD	3-6 years	Lung tumours	219 tumor features	Yes	NR	Yes- 5-fold-cross validation	No
Na	2018	Lung Cancer (NSCLC)	Pathological Reports	Convolutional Neural Network (CNN) w/ XGBoost classifier	PET/CT	NR	NR	NR	Lung tumours	Tumour size; SUV max	Yes	NR	Yes-5-fold-cross validation	No
Pham*a	2017	Lung Cancer	Surgical Resection (Histological Results)	Gray-Level Co-Occurrence Matrix (GLCM), Unsupervised neural network (deep learning) +/- Semi-variogram features	CT	Yes	NR	NR	LN's	GLCM features (contrast, correlation, energy, homogeneity) + 20 texture features. SV vector features, texture features.	NR	NR	Yes- 10-fold cross validation	No
Pham*b	2017	Lung Cancer	Biopsy Proven (Histological Analysis)	Gray-Level Co-Occurrence Matrix (GLCM, Support	CT	Yes	NR	8 years	LN's	GLCM features (contrast, correlation,	NR	Publicly available Matlab	Yes- 10-fold cross validation	No

				vector Machine (SVM)						energy, homogeneity) + 20 texture features		Program for GLCM features		
Song	2011	Lung Cancer	Ground truth = expert radiologist identifying ROI	Support Vector Machine (SVM)	PET/CT	No- human raters used a scoring system	NA	NR	Lung tumor, LNs, mediastinum, lung lobe	(1) texture features: the mean, standard deviation and kurtosis of the Gabor filtered T and N areas for both CT and PET; (2) shape features: the volume, eccentricity, extent and solidity of T and N; and (3) spatial features: the distance to the chest wall and mediastinum for tumor, and distance to two lung fields for lymph nodes, normalized by the size of the tumor or lymph node itself	Yes	NR	Yes	No
Song	2012	Lung Cancer	Ground truth = expert radiologist identifying ROI	Support Vector Machine (SVM)	PET/CT	Yes	NR	Senior expert (has read over 8000 PET-CT images)	Lung tumor and LNs	Intensity, spatial and contextual features	Yes	Matlab	Yes- Bootstrapping	No
Tagaya	2008	Lung Cancer and Sarcoidosis	Histologically confirmed	Supervised Layered Artificial Neural Networks (ANNs)	EBUS-TBNA	No	NA	Total = 5 surgeons (3 without any experience, 1 with two years and 1 with 5 year's experience)	LNs	5, 10, or 15 ROIs were randomly selected from each image. As a result, a total of 30, 60, or 90 ROIs for metastasis, and a total of 15, 30, or 45 ROIs for sarcoidosis were extracted	Yes. B-mode images were prepared into 640 x 480-pixel still images, which were then converted to bitmap files	ANN software programmed by one of the authors	Yes - used teaching images	No
Teoh	2016	Lung Cancer (NSCLC)	Histopathological confirmation (Surgical or EBUS-TBNA)	PET/CT with Bayesian Penalised Likelihood (BPL) Reconstruction	PET/CT with Ordered Subset Expectation Maximum (OSEM) Reconstruction	No - rater scored LN status based on the degree of FDG-uptake compared in the LN to the background FGD uptake	NA	Senior radiologist with 4 years of radiology (and 1 year of PET/CT) experience	LNs	Signal-to-background (SBR), signal-to-noise (SNR), SUVmax	Yes. Reconstruction based on the algorithm	Q. Clear, GE Healthcare	No	No
Toney	2014	Lung Cancer (NSCLC)	Surgical staging via bronchoscopy and mediastinoscopy	Supervised Artificial Neural Networks (ANNs)	PET/CT	No - rater extracted SUV and size measurements for each PET/CT scan	NA	10 years	Lung tumour, LNs	SUV of primary tumour and most metabolically active LN for each stations. Background SUV was used to correct ROI SUVs. Lymph node size based on short-axis diameter.	NR	R Core Team	Yes - 100 fold cross validation	No
Vesselle	2003	Lung Cancer (NSCLC)	Surgically proven N status (Pathology Reports)	Supervised Artificial Neural Networks (ANNs)	PET	Yes	NR	NR	Primary tumour, hypermetabolic LNs	Primary tumour size and SUVmax, normal lung and mediastinal SUV, nodal SUVmax	Yes - reconstruction using PET Advance system	DOS platform v4R1 of NevProp, GNU Public License	Yes - 2 fold cross validation	No
Wang	2017	Lung Cancer (NSCLC)	Pathological diagnosis	5 tests: 1- Random forest 2- support vector machine (SVM) 3- adaptive boosting 4 - back-propagation artificial neural network 5 -	PET/CT	Yes	NR	Two of the four readers had over 10 years' experience	LNs and its vicinity	Short axis, area, volume, CT mean, CT contrast, SUV mean, SUV max, SUVstd, 1st order texture features, 2nd order texture features, high order texture features	Yes - PET reconstruction by iterative algorithm using CT image attenuation correction	Classical machine learning - MATLAB R2014b, CNN - AlexNet from the Keras	Yes - 10-fold cross validation	No

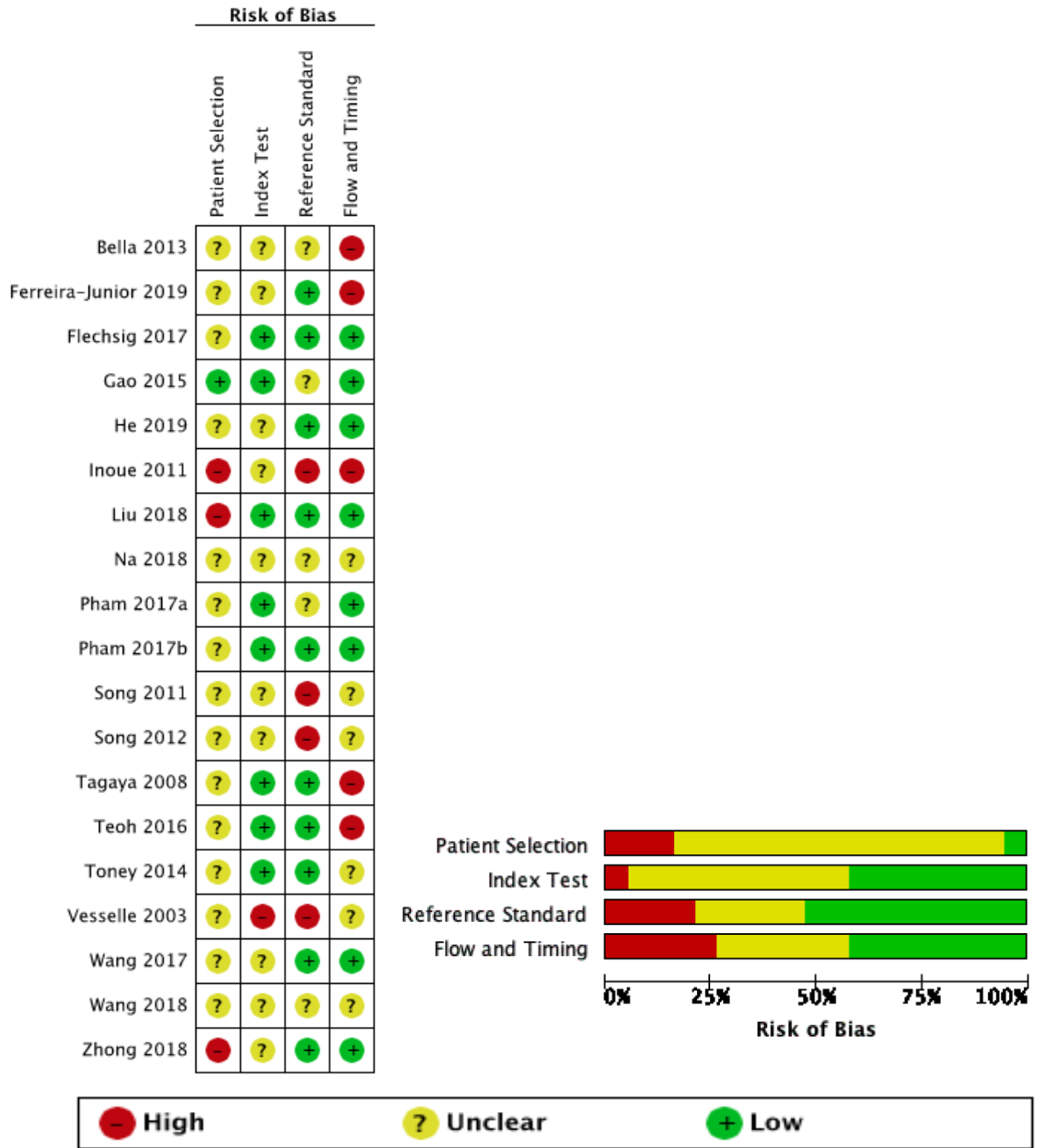
				convolutional neural networks (CNN)							library for Python			
<b>Wang</b>	<b>2018</b>	Lung Cancer (Squamous Cell)	Pathological diagnosis from mediastinal lymphadenectomy	Support vector machine (SVM)	CT	No	NR	NR	Tumours, LNs	Laplacian of Gaussian, gray-level co-occurrence matrix, texture and heterogeneity features	NR	NR	Yes - leave one out cross validation	No
<b>Zhong</b>	<b>2018</b>	Lung Cancer (Adenocarcinoma)	Histopathological confirmation	Support vector machine (SVM)	CT	Yes	Analysis Kit, v30.0, GE Healthcare	NR	Tumours	First order features, gray-level co-occurrence and gray-level run length matrices, wavelet features.	NR	NR	Yes - 10-fold cross validation	No

**NSLC**= Non-small Cell Lung Cancer; **EBUS-TBNA**=endobronchial ultrasound transbronchial aspiration; **NR**= not reported; **LNs** = lymph nodes; **ROI** = region of interest; **CT**= computed tomography; **PET**= positron emission tomography; **SUV** = standard uptake values; **SVM**= support vector machine; **ANN**= artificial neural network; **OSEM**= ordered subset expectation maximization; **GLCM**= Gray-Level Co-Occurrence Matrix; **CNN**= convolutional neural network



**Figure 1.** PRISMA flow diagram of screening, eligibility and inclusion process.





**Figure 2.** Risk of bias assessment **Left:** Risk of bias assessment for each included study for patient selection, index test, reference standard and flow and timing domains. **Right:** Overall risk of bias for each domain.

**CHAPTER 3: COMPARISON BETWEEN MANUAL AND AUTOMATIC SEGMENTATIONS OF ULTRASONOGRAPHIC LYMPH NODE FEATURES OBSERVED DURING ENDOBRONCHIAL ULTRASOUND: ASSESSMENT OF INTER-RATER RELIABILITY**

Churchill, I.F., Gatti, A.A., Hylton, D.A., Sullivan, K., Patel, Y.S. Farrokhyar, F., Leontiadis, G., Hanna, W.C.

**ABSTRACT:**

**Background-** The endosonographic Canada Lymph Node Score (CLNS) has a 96% accuracy for predicting malignancy in mediastinal lymph nodes (LNs). However, its applicability is limited because ultrasound is operator dependent and only achieves inter-rater reliability in 22% of cases. We hypothesized that operator dependency can be eliminated by a deep learning neural network that can learn the CLNS and correctly identify ultrasonographic LN features.

**Methods-** Endobronchial ultrasound images from patients undergoing lung cancer staging were retrospectively explored. The CLNS was applied in real-time to LNs by the endosonographer and static images were captured. LN images were segmented twice by the blinded experienced endosonographer using 3D Slicer and a 5-fold cross-validation was used for training and testing NeuralSeg. Dice Similarity Coefficients (DSC) were used to measure accuracy, Intraclass Correlation Coefficient (ICC) for agreement between NeuralSeg's and the endosonographer's accuracy, and diagnostic statistics to evaluate the

performance of the algorithm. Pathological specimens were used as the gold standard for diagnostic performance.

**Results-** In total, 298 LNs (18% malignant) from 140 patients were available for analysis. The expert endosonographer achieved a mean DSC of 0.77 (SD=0.21), and NeuralSeg a mean DSC of 0.68 (SD=0.21), with excellent inter-rater correlation (Intraclass Correlation Coefficient = 0.76, 95% CI= 0.70 – 0.80,  $p < 0.0001$ ). The percent sensitivity, specificity and accuracy were 18.37% (95% CI: 8.76-32.02%), and 84.34% (95% CI: 79.22-88.62%) and 73.78% (95% CI: 78.68%), respectively.

**Interpretation-** We demonstrated that segmentations performed between the endoscopist and NeuralSeg were found to be similar. NeuralSeg also able to rule out malignancy in benign LNs with a high specificity. However, the development of a machine learning risk prediction model and external validation of this algorithm is required to determine its true predictive capability.

### 3.1 INTRODUCTION

Lung cancer is the leading cause of cancer mortality worldwide, and effective treatment highly depends on the accuracy of information obtained from the process of mediastinal staging (Silvestri et al., 2013). Patients whose cancer spreads to the mediastinal lymph nodes (LNs) are best treated with chemotherapy and radiation, whereas patients with benign mediastinal LNs are best treated with surgery (Barnes et al., 2017). Consequently, accurate mediastinal staging can ensure that the most appropriate course of treatment is undertaken. Thus, mediastinal staging is the rate-limiting step prior to initiation of lung cancer treatment.

Mediastinal staging is usually undertaken via endobronchial ultrasound transbronchial needle aspiration (EBUS-TBNA), as is recommended by various guidelines (Cancer Care Ontario, n.d.; NICE, 2019; Sampsonas et al., 2018; Silvestri et al., 2013). During EBUS-TBNA, the operator can report on certain ultrasonographic features that are predictive of malignancy. Unfortunately, the sensitivity of EBUS-TBNA for mediastinal staging is highly dependent on the skill of the operator and on various other factors (i.e. cytopathologist skill or adequacy of sample). As such, inconclusive or non-diagnostic results are obtained in as many as 42.14% of EBUS-TBNA cases, which brings the lung cancer treatment cycle to a stand-still (Ortakoylu et al., 2015b).

The identification of ultrasonographic features and their capability of predicting LN malignancy has led to the development of predictive diagnostic tools (Hylton et al., 2018).

One tool known as the Canada Lymph Node Score (CLNS)- a four point score- has 96% accuracy for predicting malignancy in mediastinal LNs examined during EBUS (Hylton et al., 2019). However, its applicability is limited as it only achieves an inter-rater reliability in 22% of cases. This lack of consensus between endosonographers shows that there is a high operator dependency associated with the tool.

Human error associated with diagnostic tools has spurred research towards the development of computer-aided algorithms to help eliminate operator dependency as it is believed that Artificial Intelligence (AI) is able to produce more precise measurements compared to humans (Topol, 2019). As there exists a significant need to eliminate operator dependency associated with the CLNS, the use of AI may provide more accurate and precise measurements. The effective implementation of AI for nodal staging could be used to develop a deep neural network capable of being used in a clinical setting.

We hypothesized that operator dependency can be eliminated by a deep learning neural network, known as NeuralSeg, that can learn from LN images and correctly identify ultrasonographic LN features. Accordingly, the aim of this study was to determine if a novel deep learning neural network could segment ultrasonographic LN features according to the CLNS as accurately as an experienced endosonographer.

## **3.2 METHODS**

This study was written according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Kottner et al., 2011).

### *3.2.1 Study Design*

The study design is presented in **Figure 1**. A prospective library of 300 EBUS videos were retrospectively explored. These videos were recorded as part of a prospective clinical trial to develop the CLNS (Hylton et al., 2019). Static images of the most appropriate LN slice were created from the videos. At baseline, clinical features on primary Non-Small Cell Lung Cancer (NSCLC) (age, gender, LN stations, other staging modalities) and the acquisition date of the EBUS imaging were recorded.

### *3.2.2 Participants*

The cohort of LNs images were retrieved from 140 patients undergoing staging for suspected or confirmed lung or esophageal cancer between August 2016 to September 2017 at a designated thoracic cancer surgery tertiary site. No exclusion criteria were applied, except for neoadjuvant chemotherapy, in order to avoid nodal down-staging as a confounding variable. Consecutive patients were evaluated and screened for eligibility prior to study entry. Informed consent was obtained from each patient prior to their procedure.

### *3.2.2 Testing Methods*

#### *EBUS-TBNA Procedure*

EBUS-TBNA, ultrasonographic feature identification, and video recording were completed by the same endosonographer (WCH). After the administration of midazolam and fentanyl, an Olympus endoscope (Olympus, Shinjuku-Ku, Tokyo, Japan) with a convex-type probe EBUS and EU-ME1 transducer was inserted through the mouth into the trachea. LNs were identified using anatomical landmarks in the airway and mediastinum. The axes of the LNs were measured with calipers on the frozen images. The other ultrasonographic features were identified visually and scored in real-time. Transbronchial needle aspiration with a 22- gauge needle was then performed to obtain a biopsy of the LN under ultrasound guidance. The specimen was spread onto glass slides, fixed, and air-dried. The dried slides were evaluated via rapid-on-site examination by a cytopathologist to determine if the specimens were adequate for pathological analysis. Pathological reports for each LN biopsy were obtained.

#### *Assessment of Ultrasonographic LN Features*

Four ultrasonographic criteria were considered malignant based on the following definitions:

1. *Small axis length*:  $\geq 10$  mm predictive of malignancy
2. *Central hilar structure*: Absence of central hilar structure (missing, flat, central, echogenic structure in the LN)

3. *Central necrosis*: Presence of central necrosis (presence of central hypoechoic structure in the LN)
4. *Margins*: >50% margin (distinguished by majority echogenic line delimiting the LN)

A LN with a score of  $\geq 2$  was considered to be highly suspicious for malignancy based on the prediction model developed by Hylton and colleagues (Hylton et al., 2020).

#### *Manual Segmentations by Endosonographer*

Static images obtained from EBUS videos were converted to DICOM format to perform manual segmentations. LN images were segmented by an experienced endosonographer (WCH), with 7 years of experience, to produce the gold standard for assessment by NeuralSeg. Manual segmentations of the 4 CLNS features predictive of malignancy as well as the entire node were performed using 3D Slicer (3D Slicer V4.10.2, Boston, MA) (**Figure 3**). Each of the blinded LN images were segmented twice in order to determine expert level reliability. The endosonographer assessing the images was blinded to the personal identifiers, imaging, and pathology results associated with each LN. Additionally, images were shuffled and assigned random identification numbers to ensure that repeated images were not segmented in a defined order and to prevent diagnostic review bias (Schmidt & Factor, 2013a). The Dice Similarity Coefficient (DSC), a statistical validation metric, was used to evaluate the performance of the reproducibility of manual segmentations. The value of a DSC ranges from 0, indicating no spatial overlap between



two sets of binary segmentation results, to 1, indicating complete overlap (**Figure 2**). The DSC is calculated as follows:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}$$

### *Deep Neural Network Architecture*

The proposed imaging process algorithm, NeuralSeg, is a convolutional neural network (CNN), that is a feedforward network in which the signal is processed directly without any loops or cycles (**Figure 4**). Segmentations of all tissues were computed using a combination of two algorithms. One algorithm was trained to segment the node and the contour and included the necrosis and central hilar structure within the node segmentation to create an aggregate node. The second algorithm was utilized to determine the presence of the central hilar structure and necrosis and segmented all tissues of interest (central hilar structure, necrosis, node, margin).

Both segmentation algorithms utilized a U-Net style convolutional neural network (CNN). The network input was an image shaped 512x512 pixels which was down sampled, to fit in graphics processing unit memory, from the original image shape of 1300x975. The network output was a three-dimensional probability map, where dimensions 1 and 2 were 512x512, the same as the input image, and the third dimension included n levels that coincided with the probability of each pixel belonging to the n tissue of interest. The n probabilities for each pixel always summed to 1.0, each pixel was classified according to

the tissue it had the highest probability of belonging to. After classifying each pixel to the appropriate tissue, segmentations were resampled to be the same shape as the original image (1300x975) using nearest neighbour interpolation.

### *Supervised Learning*

The network was trained using a batch sizes of 8, the Adam optimizer with a learning rate of  $10^{-3}$ , and a custom loss function which summed the negative dice similarity coefficients (DSC) of each of the tissues. Image augmentation including random rotation of up to 6 degrees and translation of up to 20% was employed. To enable robust estimation of the accuracy of predictions, a 5-fold cross-validation scheme was used. During training, both segmentations produced by the expert segmenter for each node were used. After training was complete, the holdout (testing) LNs were predicted only one time. Due to high intra-segmenter variability, while training the aggregate node and margin algorithm, only LNs that had an intra-segmenter DSC for the aggregate node which were greater than 0.8 were used. This strategy was employed to reduce noise in learning the optimal segmentation and due to the importance of the aggregate node in calculating 2/4 of the CLNS features.

The CLNS features were calculated for each segmentation produced by the expert segmenter, as well as for each segmentation produced autonomously by the trained network. Presence of the central hilar structure was determined by segmentation of  $>50$  pixels ( $<0.004\%$  of the pixels in the full image) to belong to the central hilar class. Presence

of necrosis was determined by segmentation of >5 pixels (<0.0004% of the pixels in the full image) to belong to the necrosis class. Different thresholds were used for the central hilar structure and central necrosis due to their differences in imbalance both within and between images.

### *3.2.3 Statistical Analysis*

LNs were used as the unit of analysis as machine learning methods were employed to detect the reliability and accuracy of the specific ultrasonographic features of each LN (Jiang, Yang, Wang, Li, & Sun, 2020). Sample size was estimated for ICC with precision for hypothesis testing (Walter, Eliasziw, & Donner, 1998; Zou, 2012). The minimum acceptable reliability ICC was set to 0.60 and the expected reliability to 0.7, with an alpha error of 0.05, power of 0.8 and an expected missing data error of 12% due to artifacts on images. As such, for two raters per subject, a sample size of 296 was calculated. There were no indeterminate results nor were there any data missing from the reference standard.

The data are presented as mean  $\pm$  standard deviation, median (range), or number (percentage). DSC scores were used to determine the accuracy of automated segmentations through spatial overlap. DSC scores were compared using an inter-rater correlation matrix. To provide a representative estimate of the automated segmentation accuracies, the automated segmentation accuracies were compared to the endosonographer accuracies using Intraclass Correlation Coefficients (ICC). Sensitivity, specificity and percent

correctly classified for real-time scoring, manual segmentations and automatic segmentations were calculated. Statistical analysis was performed using R software (R Foundation for Statistical Computing, 2013, Vienna, Austria).

### 3.3 RESULTS

Patient baseline characteristics and pathological data of biopsied and scored LNs are presented in **Table 1**. In total, 298 LNs from 140 patients were segmented (**Figure 5**). The average age of participants was 69.92 (standard deviation [SD]=10.64), with 54.29% (n=76) being male. Standard of care mandates that patients undergo diagnostic imaging prior to mediastinal staging. As such, 99.29% of patients underwent imaging via chest computed tomography (CT) or positron emission tomography (PET) scan prior to EBUS. Of the 298 LNs sampled, the median number of LNs biopsied during EBUS was 3 (range=1-4), and the most commonly LN stations biopsied were 7 (n=125, 41.94%) and 4R (n=84, 28.19%).

After pathological assessment, lung masses were considered malignant in 109 (77.9%) cases and benign in 31 (22.1%) cases. Lung cancer and esophageal cancer were confirmed in 77 (70.6%) and 32 (22.9%) of the malignant cases, respectively. With respect to the 298 sampled LNs, malignancy was present in 49 (16.4%) of the LNs and benign in 249 (83.6%) of LNs. According to standard of care treatment guidelines, 56 (40.0%) of

patients received surgery and 84 (60.0%) received other treatments such as chemoradiation or immunotherapy, based on the stage of their cancer.

Analysis of segmentation spatial overlap demonstrated that the expert endosonographer achieved a mean DSC of 0.77 (SD=0.21), and NeuralSeg a mean DSC of 0.68 (SD=0.21) (**Table 2**). The segmentations between the expert endosonographer and NeuralSeg were compared and were found to possess excellent inter-rater correlation (ICC = 0.76, 95% CI= 0.70 – 0.80,  $p<0.0001$ ) (**Figure 6**) (Koo & Li, 2016). The inter-rater correlation matrix for spatial overlap of segmentations revealed that there was a strong positive correlation between segmentations produced by NeuralSeg and the endosonographer's segmentations both times ( $r=0.71$ ,  $p<0.001$ ) (**Figure 7**).

Diagnostic statistics were evaluated for real-time scoring by the endosonographer, the endosonographer's segmentations and NeuralSeg's segmentations. Standard diagnostic performance measures are presented in **Table 3 and 4**. A CLNS of equal to or greater than 2 produced an sensitivity, specificity and accuracy of 77.55% (95% Confidence Interval [CI]:63.38-88.23%), 53.82% (95% CI: 47.41-60.13%), and 57.61% (95% CI: 51.78-63.29%) for the real-time assessment by the endosonographer in the endoscopy suite; 46.94% (95% CI: 32.53-61.73%), 84.74% (95% CI: 79.66-88.97%) and 78.69% (95% CI: 73.60-83.20%) for the manual segmentations performed by the endosonographer; and 18.37% (95% CI: 8.76-32.02%), and 84.34% (95% CI: 79.22-88.62%) and 73.78% (95% CI: 78.68%) for NeuralSeg's segmentations, respectively.

### **3.4 DISCUSSION**

In this study, a deep CNN was trained to apply the CLNS in order classify benign and malignant LNs. By means of the designed CNN, LN detection as well as benign vs. malignant distinction was performed with good accuracy and high specificity. Results showed there was no difference between NeuralSeg and the endosonographer (gold standard) as measured by ICC. Moreover, NeuralSeg performed with higher accuracy than the CLNS completed in real-time, suggesting that it may have eliminated the human error associated with the predictive tool.

True delineation of regions of interest is important to precision medicine. As in previous studies, clinicians have been able to identify regions of interest through the manual segmentation of lung masses and/or lymph nodes in thoracic imaging for lung cancer (Zhang, Ferriera-Junior, Fleshig, Wang 2017, Vesselle, Pham 2017a, Liu). However, these studies did not report a spatial overlap measurement comparing manual segmentations completed by the clinician to the automatic segmentations produced by the deep neural network, or repeated segmentations by the same rater. Thus, inter-rater and intra-rater reliability and accuracy could not be determined through spatial overlap. One study investigated lung cancer through tumour segmentation using radiomic features (Owens et al., 2018). The authors reported on the two semi-automatic tools' intra-observer reliability with a mean DSCs of 0.88 (SD= 0.06) and 0.88 (SD=0.08) for each model respectively, while they compared inter-observer reliability using an ICC. Similarly, Zhu and colleagues segmented multiple organs at risk in CT images for lung cancer and found

that their deep neural networks were able to segment lung LNs with the best delineation (DSC>0.90) (Li et al., 2016). The DSC of our proposed tool was slightly lower than these models with a DSC of 0.68 (SD=0.21) and manual segmentations with a DSC of 0.77 (SD=0.21). However, the literature has reported on a range of DSC scores [0.60-0.90] (Owens, Zhang, Xu), depending on the size, imaging modality, radiomic features and region(s) of interest.

Accuracy in machine learning may be defined as the number of false positives, speed or automation level (Huidrom, Jina Chanu, & Manglem Singh, 2018). We defined accuracy as the overall probability that a LN is correctly classified based on the ability of NeuralSeg to identify ultrasonographic LNs compared to expert endosonographer (gold standard) or pathology reports (reference standard). According to this definition, we found that NeuralSeg was able to identify ultrasonographic LN features more accurately compared to the clinician rater in real-time. This may have been due to the elimination of human error associated with feature identification and uncertainties (Owens et al., 2018). Segmentation accuracies have ranged from 56-91% in the literature for the deep neural network. Our results were in the middle of this range, suggesting that our model possessed good accuracy.

This study possessed many strengths. Our study was pragmatic as exclusion criteria was not limited. As such, our results may be generalizable and as we investigated all stages of esophageal and lung cancer. Additionally, we used an experienced endoscopist that was

familiar with the CLNS to mitigate the possibility of a learning curve associated with the tool. Finally, diagnostic review bias was minimized through blinding of both the personal identifiers and pathology associated with each LN during segmentation. Images were randomized and assigned random identification numbers to ensure that repeated images were not segmented in a defined order allowing for a non-biased assessment of the outcome being assessed (Schmidt & Factor, 2013b).

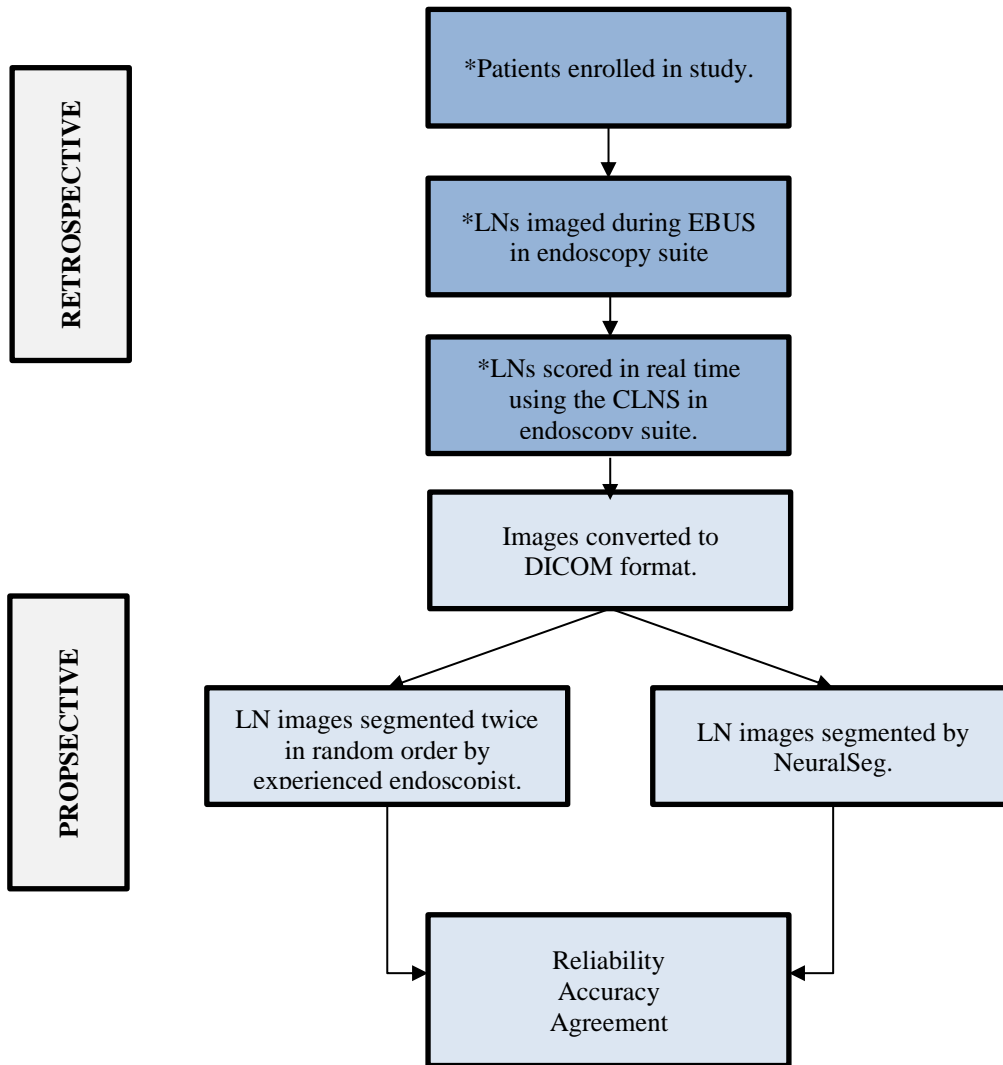
This study is not without limitations. First, since this study was retrospective in nature, the sample size was small and lacked external validation. Sample size is an important factor to consider when using inferential statistics such as ICC. Small sample sizes may lack precision and may generate large confidence intervals (Jones, Carley, & Harrison, 2003). Second, a relatively small number of malignant LNs was present in the sample. This low proportion of malignant LNs may have affected the accuracy when predicting the CLNS for each LN. Further, as segmentation was performed manually to train the algorithm, the segmentations may be susceptible to subject factors (systemic bias in the placement of the boundary or learning curve associated with manual segmentation software) (Warfield, Zou, & Wells, 2008). Finally, as the dataset was obtained from a single institution in one country, results may not be generalizable to other settings. As ICC depends on the heterogeneity of LNs in the patient population of the study, populations that are more heterogeneous will yield higher ICC values than more homogenous populations (Bartlett & Frost, 2008). Due to this limitation, both patient population



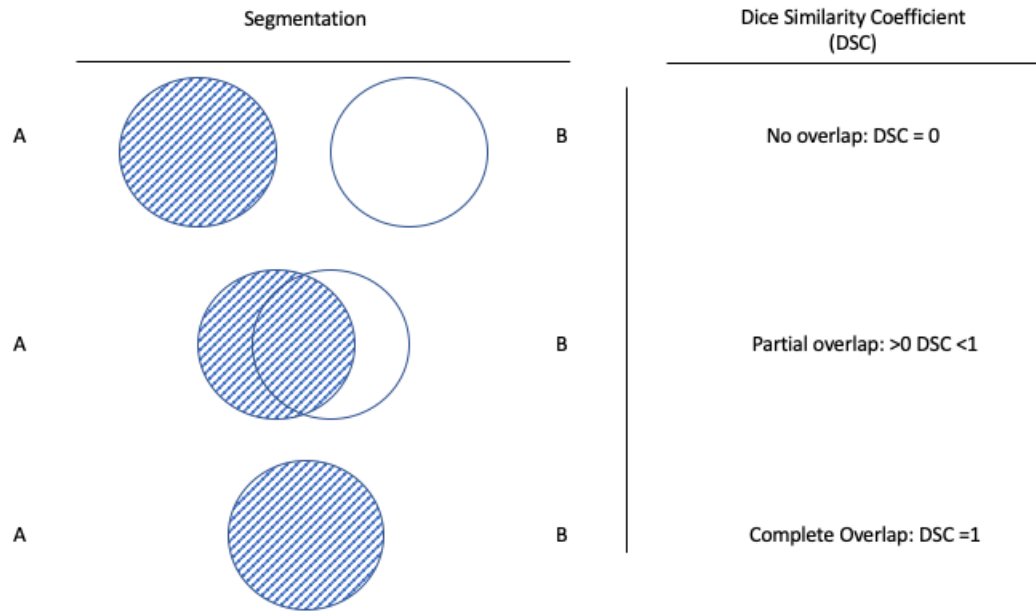
demographics and LN characteristics were reported to give an idea of the between- LN heterogeneity.

To our knowledge, this is the first study to compare manual to automatic segmentation of LN features observed during EBUS. Our findings showed segmentations performed between the endoscopist and NeuralSeg were found to be similar. Further, the algorithm is also able to rule out malignancy in benign LNs with a high specificity when a cut off of a CLNS  $\geq 2$  is used. However, the development of a machine learning risk prediction model and external validation of this algorithm is required to determine its true predictive capability. An important future study would be to evaluate the effect that contouring can play in building outcome models to improve feature reliability.

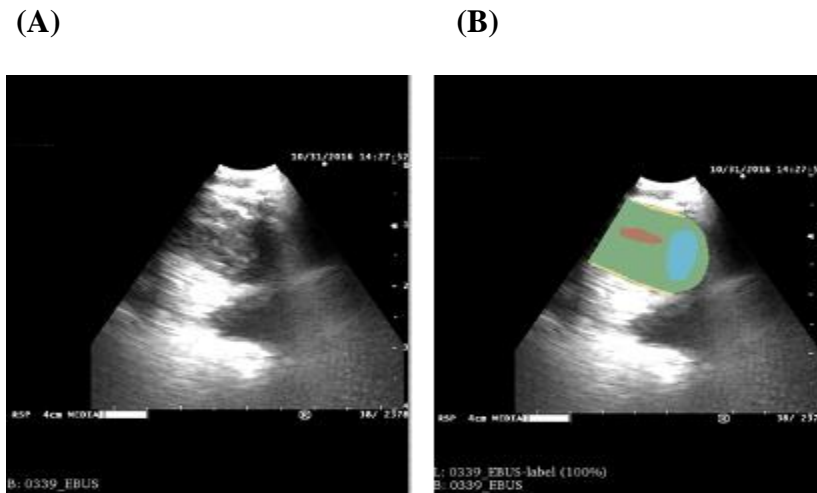
**CHAPTER 3 TABLES AND FIGURES:**



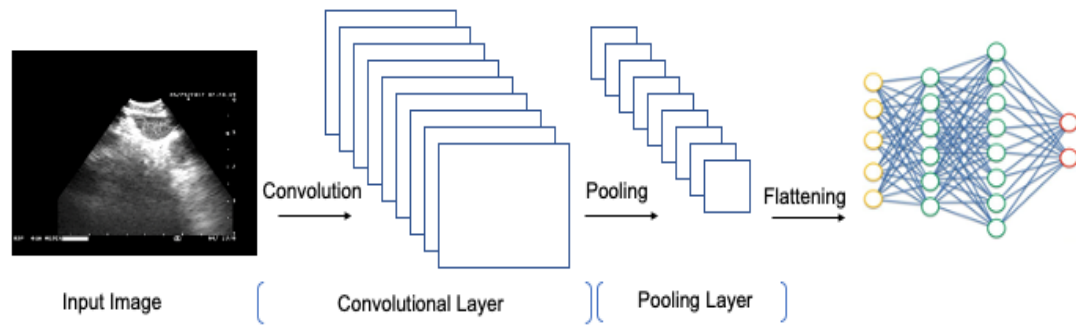
**Figure 1: Study design to assess inter-rater reliability and accuracy.** A library of EBUS images were explored retrospectively. Segmentation occurred after the collection of images. \*Denotes steps that occurred retrospectively.



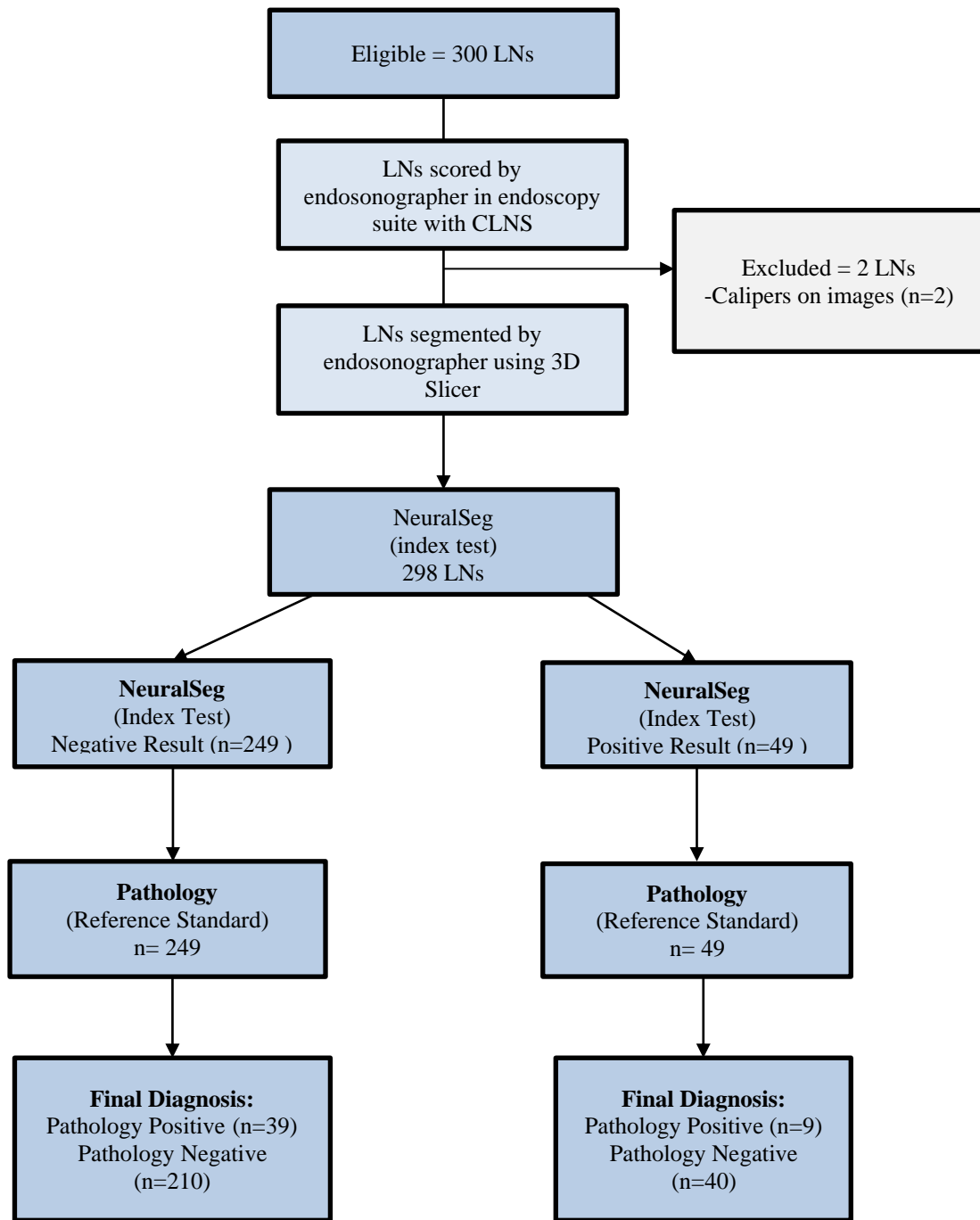
**Figure 2.** Dice Similarity Coefficient (DSC) diagram representing spatial overlap. **Left:** visual representation of segmentation overlap. DSC score is calculated by the DSC is a measure of overlap between the two segmentations being compared and is defined as:  $DSC = \frac{2|A \cap B|}{(|A \cup B|)}$ . **Right:** DSC calculated based on spatial overlap. A DSC of 0 means that no pixels were the same between the two segmentations, and a DSC of 1.0 indicates perfect overlap of every pixel.



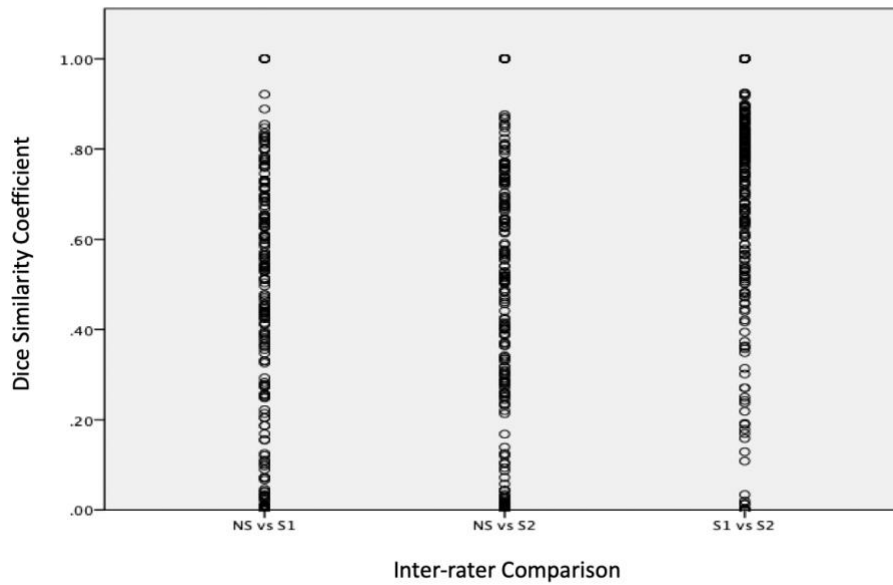
**Figure 3: Example of lymph node station 4R assessed with Canada Lymph Node Score criteria during manual segmentation. (A) Still image of lymph node imaged on EBUS during mediastinal staging (B) Segmentation of lymph node performed by an experienced endoscopist using 3D Slicer for size (green), margin (yellow), central hilar structure (brown) and central necrosis (blue).**



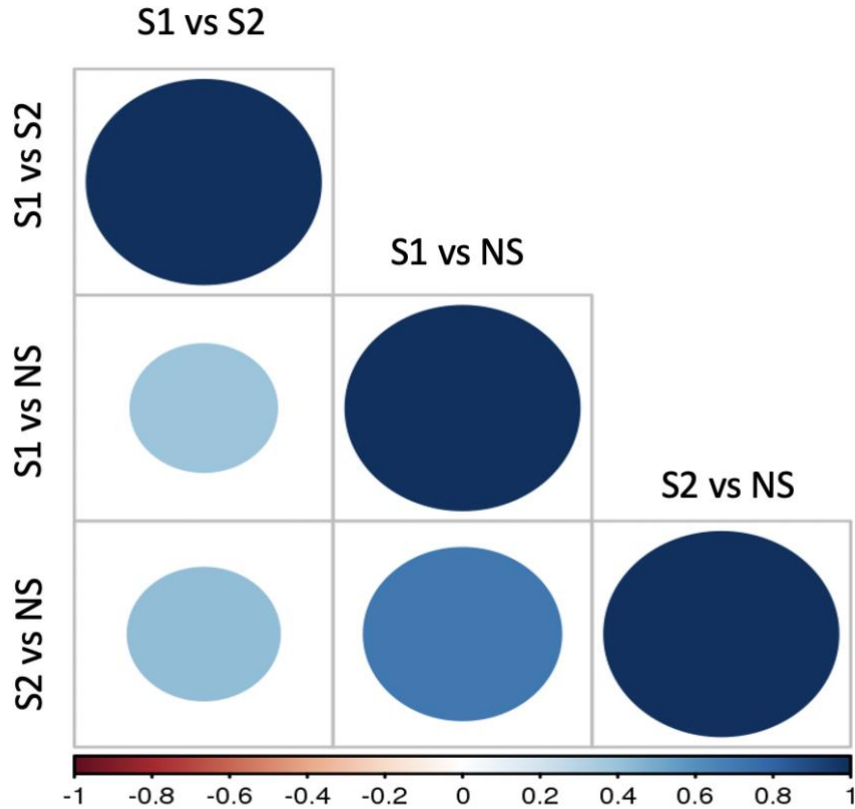
**Figure 4: NeuralSeg with its convolutional neural network components and feedforward framework.** Endobronchial ultrasound images were used as an input to learn lymph node features. The input and output layers as well as hidden layers are displayed in a series of convolutional layers.



**Figure 5. Flow diagram of lymph node images through study.** 300 potential lymph node images were retrospectively explored, and 298 lymph node images were available for segmentation analysis. All results were compared to pathological analysis (gold standard reference test).



**Figure 6: Inter-rater comparisons as measured by Dice Similarity Coefficient Scores.**  
\*NS= NeuralSeg; S1= 1st manual segmentation by endosonographer; S2= 2nd manual segmentation by endosonographer



**Figure 7: Correlation matrix for comparison of spatial overlap of segmentations.** Correlation coefficients were (left to right/ top to bottom:  $r=1.00$ ;  $r=0.39$ ;  $r=0.42$ ; B2)  $r=1.00$ ;  $r=0.71$ ;  $r=1.00$ ). \*S1= 1st manual segmentation by endosonographer; S2= 2nd manual segmentation by endosonographer



**Table 1: Patient characteristics and pathological results for lymph nodes.**

<b>Variable</b>	<b>Patients (n=140) LN's sampled (n=298)</b>
<b>Age (years) [mean ± SD]</b>	69.92 ±10.64
<b>Males, n (%)</b>	76 (54.3)
<b>Pre-planned imaging modalities completed</b>	
Chest CT or PET, n (%)	139 (99.3%)
Head CT, n (%)	10 (7.1%)
MRI, n (%)	27 (19.3%)
<b>Median (range) of LN's scored/biopsied per patient</b>	3 (1-4)
<b>Scored/ Biopsied LN Stations (n=298)</b>	
7, n (%)	125 (41.9%)
4R, n (%)	84 (28.2%)
4L, n (%)	57 (19.1%)
10, n (%)	14 (4.7%)
11, n (%)	7 (2.3%)
Other (1,2L,2R,12), n (%)	5 (1.7%)
<b>Pathology Diagnosis: Lung Mass</b>	
Primary Lung Cancer	77 (55.0%)
Primary Esophageal Cancer	32 (22.9%)
Benign Cases	31 (22.1%)
<b>Pathology Diagnosis: LNS</b>	
Malignant, n (%)	49 (16.4%)
Benign, n (%)	249 (83.5%)
<b>Malignant Ultrasonographic Features based on CLNS Unblinded (in Endoscopy Suite) (n=298)</b>	
Small axis, n, (%)	65 (21.8%)
Margins, n (%)	150 (50.3%)
Central Hilar Structure, n (%)	150 (50.3)
Central Necrosis, n (%)	137 (46.0%)
<b>Treatment</b>	
Surgery, n, (%)	56 (40.0%)
Other treatment, n (%)	84 (60.0%)

SD= standard deviation; LN= lymph node

**Table 2: Dice Similarity Coefficients of Canada Lymph Node Score for manual and automatic segmentations of lymph node and entire contents.**

<b>Rater Comparison</b>	<b>Dice Similarity Coefficient</b>
<b>S1 vs S2</b>	$0.7677 \pm 0.2131$
<b>NeuralSeg vs S1</b>	$0.6847 \pm 0.2101$
<b>NeuralSeg vs S2</b>	$0.6818 \pm 0.2175$

\***S1**= 1st manual segmentation; **S2**= 2nd manual segmentation

**Table 3. Diagnostic statistics of ultrasonographic lymph node features and  $\geq 2$  CLNS determined by NeuralSeg (gold standard = endosonographer)**

Feature/Score	Total	TP	TN	FP	FN	Sensitivity (95% CI)	Specificity (95% CI)	**Accuracy (95% CI)
<b>Small Axis Length</b> >10 mm (1) < 10 mm (0)	298	37	99	134	28	56.92% (44.04% to 69.15%)	42.49% (36.06% to 49.11%)	44.80% (39.06% to 50.64%)
<b>Margins</b> >50% (1) <50% (0)	298	33	123	25	117	22.00% (15.65% to 29.49%)	83.11% (76.08% to 88.76%)	73.33% (67.93% to 78.27%)
<b>CHS</b> Absent (1) Present (0)	298	39	129	20	111	26.00% (19.19% to 33.79%)	86.58% (80.03% to 91.60%)	76.88% (71.69% to 81.54%)
<b>Central Necrosis</b> Present (1) Absent (0)	0	-	-	-	-	-	-	-
<b>CLNS</b> <2 (Benign) $\geq 2$ (Malignant)	298	33	130	15	120	21.57% (15.34% to 28.94%)	89.66% (83.51% to 94.09%)	78.76% (73.67% to 83.27%)

**CHS** = central hilar structure; **CI** = confidence interval

\*\* accuracy determined based off of prevalence of malignancy

**Table 4: Assessment of performance measurements for the Canada Lymph Node Scores on a binary scale ( $\geq 2$  considered malignant) according to each rater. Pathology report used as gold standard comparison.**

Rater	Performance Measurement							
	Total	TP	FP	TN	FN	Sensitivity (95% CI)	Specificity (95% CI)	**Accuracy (95% CI)
<b>Real-time Assessment (Endoscopy Suite)</b>	298	38	115	134	11	77.55% (63.38% to 88.23%)	53.82% (47.41% to 60.13%)	57.61% (51.78% to 63.29%)
<b>Blinded Assessor (Manual Segmentations)</b>	298	23	38	211	26	46.94% (32.53% to 61.73%)	84.74% (79.66% to 88.97%)	78.69% (73.60% to 83.20%)
<b>NeuralSeg Assessment (Automatic Segmentations)</b>	298	9	40	210	39	18.37% (8.76% to 32.02%)	84.34% (79.22% to 88.62%)	73.78% (68.40% to 78.68%)

\*CI = confidence interval; TP= true positive; FP= false positive; TN= true negative; FN= false negative

\*\*Accuracy calculated based on prevalence

**CHAPTER 4: DEVELOPMENT AND VALIDATION OF DEEP NEURAL NETWORK FOR PREDICTING LYMPH NODE MALIGNANCY USING THE CANADA LYMPH NODE SCORE IN LUNG CANCER PATIENTS UNDERGOING MEDIASTINAL STAGING**

Churchill, I.F., Gatti, A.A. Hylton, D.A., Sullivan, K., Patel, Y.S. Farrokhvar, F., Leontiadis, G., Hanna, W.C.

**ABSTRACT:**

**Background-** NeuralSeg, a deep learning neural network, has a specificity of 83.34% for ruling in benign lymph nodes (LNs) observed during endobronchial ultrasound procedures. However, its applicability is limited as it requires validation to determine its true predictive capability. Our study sought to develop and externally validate NeuralSeg, a deep neural network, capable of predicting LN metastasis through the segmentation of ultrasonographic LN features observed during EBUS imaging.

**Methods-** We conducted this study in two phases, a derivation phase followed by a validation phase. In the derivation phase, LN images were retrospectively explored. The images were segmented twice by a blinded experienced endosonographer using 3D Slicer and a 5-fold cross-validation was used for training NeuralSeg. In the validation phase, LN images were prospectively collected to test the algorithm. Logistic regression, c-statistic and receiver operator characteristic curve were used to test the performance, and discrimination, respectively. Pathologic specimens from EBUS biopsies/surgical resections were used as the ground truth.

**Results-** In total, 298 LNs (16.4% malignant) from 140 patients were used for derivation and 108 LNs (29.8% malignant) from 47 patients for validation. Overall, NeuralSeg had an accuracy of 73.78% (95% CI: 68.40% to 78.68%), a sensitivity of 18.37% (95% CI: 8.76% to 32.02%) and specificity of 84.34% (95% CI: 79.22% to 88.62%). Further, external validation showed that NeuralSeg had higher diagnostic discrimination in the validation sample compared to the derivation sample (c-statistic= 0.60 [0.47-0.27] vs c-statistic=0.51 [0.42-0.63]).

**Interpretation-** NeuralSeg had excellent performance in identifying malignant LNs from EBUS images. We demonstrated that an AI algorithm is able to rule out malignancy in benign LNs with a specificity of 84.34% and an accuracy of 73.78%. Its high specificity may inform decision-making regarding biopsy if results are benign. Future work with a larger dataset will be required to improve and refine the algorithm prior to trials in clinical practice.

## 4.1 INTRODUCTION

Lung cancer is the most commonly diagnosed cancer, accounting for 11.6% of cancer cases and the leading cause of cancer mortality worldwide, thus warranting the need for accurate mediastinal staging in order to determine treatment accordingly. Effective treatment of lung cancer is almost entirely dependent upon the accuracy of information obtained from the process of lymph node (LN) staging (Ortakoylu et al., 2015a), which is usually undertaken by the procedure of Endobronchial Ultrasound Transbronchial Needle Aspiration (EBUS-TBNA), as is recommended by various guidelines (Sampsonas et al., 2018; Silvestri et al., 2013). However, EBUS-TBNA has been reported to generate inconclusive results as much as 40% of the time (Silvestri et al., 2013). This significant percentage of inconclusive results occurs as the sensitivity of the procedure is dependent on multiple factors including the skill of the operator, the skill of the cytotechnologist, the skill of the pathologist, the size of the LNs, the gauge of the needle, and the pretest probability of cancer. Despite a copious amount of research conducted over the past decade, the diagnostic yield of EBUS-TBNA has not improved. As a result, the medical community is beginning to abandon LN biopsies, and as many as 50% of lung cancer patients are being sent to treatment without nodal staging (Boffa et al., 2017; Little et al., 2005b) The exclusion of nodal staging jeopardizes good patient care. There is near-universal consensus on the need to develop and study prediction models for LN staging.

Deep learning, a form of Artificial Intelligence (AI), offers considerable promise for improving the accuracy and speed of diagnosis through medical imaging via the

extraction of quantitative image descriptors non-invasively (Doi, 2007). Recent advances in deep learning have provided insight into tumour detection and subtype classification (Jha & Topol, 2016; X. Liu et al., 2019; Topol, 2019). Most notably, the recent studies have indicated that radiomics predictive models have been accepted as reliable tools to quantify risk by incorporating and illustrating important factors for and prediction (Huang et al., 2016; H. Wang et al., 2017). Computed tomography (CT) and/or positron emission tomography (PET) radiomics features assessments have been applied and demonstrated to be useful for nodal involvement prediction in patients with Non-Small Cell Lung Cancer (NSCLC) (Ferreira-Junior et al., 2020; Flechsig et al., 2017; He et al., 2019; Pham et al., 2017), and one study has assessed the use of deep learning for LN disease for lung cancer and sarcoidosis (Tagaya et al., 2008). However, to our knowledge, no published study has determined whether the individual prediction of LN metastasis from lung cancer patients' mediastinal EBUS images could be achieved by a radiomics through the use of clinical scoring system.

In search for a method to optimize the prediction of LN malignancy, our research group developed, validated, and published the Canada Lymph Node Score (CLNS)- a very accurate 4-point scale used to predict malignancy based on ultrasonographic LN characteristics that are observed during the EBUS procedure. The CLNS was prospectively validated across 7 centres in Canada and was found to be highly specific (99.59%) for ruling in cancer in the LNs (Hylton et al., 2019). We also demonstrated that a deep neural network is able to identify and score mediastinal LNs as accurately (Chapter 3). However,



the development of a deep neural network risk prediction model and external validation of this algorithm is required to determine its true predictive capability.

As such, the aim of the study was twofold: 1) to develop a deep neural network known as NeuralSeg, in order to identify and segment ultrasonographic LN features based on a validated four-point score and 2) to validate NeuralSeg to see if it able to predict LN malignancy in patients undergoing EBUS to accurately stage lung cancer.

## **4.2 METHODS**

This study was written according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD) reporting guidelines (Collins, Reitsma, Altman, & Moons, 2015; Moons et al., 2015). This study was composed of two phases. Phase A utilized a derivation set to develop the algorithm and Phase B used a validation set to prospectively validate the algorithm. The study design is presented in **Figure 1**.

### ***4.2.1 Source of Data***

In Phase A, a derivation set of retrospectively explored LN images was used to develop the deep neural network and a set of prospectively collected LN images were used to validate a deep neural network, NeuralSeg, to predict LN malignancy (NCT03849040). Retrospective images were collected between August 2016 to September 2017. In Phase

B, prospective images were collected between April and September 2019. All images in both phases collected ultrasonographic LN features and images from consecutive patients undergoing EBUS for mediastinal investigation of confirmed or suspected lung cancer were collected. EBUS procedures was recorded and static images of the most appropriate LN slice were captured and saved onto an external hard drive. At baseline, clinical features of lung cancer (age, gender, LN stations, imaging modalities) and the acquisition date of the EBUS imaging were documented. This study received ethics clearance from prior to conducting research (HiREB #5636).

#### ***4.2.2 Participants***

Consecutive patients were evaluated and screened for eligibility prior to study entry. Informed consent was obtained from each patient prior to their procedure. Based on staging results, patients with cancer present in the hilar or mediastinal LNs were referred to chemoradiation therapy and patients with cancer absent from the hilar or mediastinal lymph nodes were referred for surgical resection.

*Phase A:* The cohort of LNs images were retrieved from adult patients undergoing staging for suspected or confirmed lung or esophageal cancer at a designated thoracic cancer surgery tertiary site. No exclusion criteria were applied, except for neoadjuvant chemotherapy, in order to avoid nodal down-staging as a confounding variable.

*Phase B:* Adult patients ( $\geq 18$  years) with suspected or confirmed lung cancer (based on CT and/ or PET investigation) undergoing mediastinal staging at a tertiary thoracic cancer centre, were enrolled in the study. There were no exclusion criteria in order to not limit the stage and range of disease.

#### ***4.2.3 Outcomes***

For both Phase A and B, LN malignancy (outcome) was determined based on EBUS biopsy/ surgical specimen pathological results (gold standard). During EBUS, transbronchial needle aspiration with a 22- gauge needle was performed to obtain a biopsy of the LN under ultrasound guidance. The specimen was spread onto glass slides, fixed, and air-dried. The dried slides were evaluated via rapid-on-site examination by a cytopathologist to determine if the specimens were adequate for pathological analysis. If patients underwent surgical resection for their lung cancer, evaluation of nodal status was undertaken through the excision or biopsy of all LNs surrounding the tumour and sent for histopathological analysis. Pathology reports from EBUS biopsies and surgical specimens were obtained within 3 weeks from the date of both EBUS and surgery. The presence or absence of cancer in LNs was determined without knowledge of the ultrasonographic LN features (predictors used for the study).

#### **4.2.4 Predictors**

In both Phase and B, the following baseline data were extracted for each patient: age, gender, imaging modalities, LN stations, LN features, date of EBUS procedure and planned treatment. Clinical features were obtained from patient charts. Ultrasonographic LN features were identified by both the endosonographer and NeuralSeg. The endosonographer used the Canada Lymph Node Score- a highly accurate four-point scale used to predict LN malignancy (Hylton et al., 2019) to identify ultrasonographic LN features in the endoscopy suite. NeuralSeg then used its previously developed algorithm to apply the CLNS through the segmentation of ultrasonographic LN features to the EBUS images. NeuralSeg was blinded to both the CLNS measured by the endosonographer as well as demographic and clinical characteristics of the patients. A LN with a score of  $\geq 3$  was considered to be highly suspicious for malignancy. Ultrasonographic criteria were considered malignant based on the following definitions:

- *Small axis length*:  $\geq 10$  mm predictive of malignancy
- *Central hilar structure*: Absence of central hilar structure (missing, flat, central, echogenic structure in the LN)
- *Central necrosis*: Presence of central necrosis (presence of central hypoechoic structure in the LN)
- *Margins*:  $>50\%$  margin (distinguished by majority echogenic line delimiting the LN)

#### *4.2.5 Sample Size*

We calculated the study sample sizes needed according to the requirement of malignancy as the outcome of interest and LNs as the unit of analysis. LNs were chosen to determine sample size rather than number of patients as machine learning methods were employed. As such, although we were interested in determining the correct staging of each patient, our primary outcome was interested in determining the discrimination and diagnostic potential of the algorithm to determine malignancy in each LN.

##### *Phase A:*

The derivation sample size was extrapolated from the computer-aided diagnostic literature and predicted based on the number of events per variable needed for a prediction model (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012). In accordance with our previous findings (Hylton et al.), we calculated a LN sample based on a prevalence of 0.18, four independent predictors (based on the CLNS scoring criteria) and at least 10 events per variable. The total needed sample size was calculated to be 223 LNs. However, due to the use of machine learning, it was decided to use the entire dataset of LN images available for the training and testing of the algorithm.

##### *Phase B:*

The validation sample size was calculated based on accuracy whereby assuming a 95.00% confidence interval z-score (1.96), 0.05 accuracy level, a prevalence of 0.18 and specificity of 0.96 (Jones et al., 2003). It was determined that a sample size of 72 LNs would enable diagnostic statistics to be calculated accurately.

#### ***4.2.6 Missing Data***

If an outcome was missing, or an image was of unsatisfactory quality, the patient and LN data were excluded from analysis for both the derivation and validation set.

#### ***4.2.7 Statistical Analysis***

The data are presented as mean  $\pm$  standard deviation, median (range), or number (percentage). Baseline continuous variables were compared between the derivation and validation cohort using the Student's t-test (parametric) or Mann-Whitney U test (non-parametric), and categorical variables compared using Chi-Square. All statistical tests used two-sided hypotheses with p-values less than 0.05 considered statistically significant. Statistical analysis was performed using R software (R Studio, 2013, Vienna, Austria).

##### *Phase A:*

Pearson's chi-square test was used to test the likelihood of the presence or absence of certain ultrasonographic features being independently associated with malignant or benign LNs. Logistic regression was used to develop the regression model, the Hosmer-Lemeshow test was used to evaluate the model's calibration and the receiver operator characteristic (ROC) curves with corresponding c-statistics to evaluate the model's discrimination. Multivariable logistic regression with backward stepwise selection with a *p*-value greater than 0.05 for removal of variables was used, but predictors that we considered to have great clinical relevance based on the CLNS were forced back into the model.

### *Phase B*

Sensitivity, specificity and accuracy, negative predictive value (NPV), positive predictive value (PPV) and area under the curve (AUC) for NeuralSeg's rating of the CLNS were calculated. Predictions of malignancy produced using automated segmentations were compared to the ground truth findings of malignancy (pathology reports) and benchmarked against predictions produced using the CLNS.

#### ***4.2.8 Risk Groups***

A diagnostic rule was derived to estimate the probability of LN malignancy. Thresholds for ruling in and ruling out LN malignancy were introduced based on a score of 2 or more on the CLNS (Hylton et al., 2019).

#### ***4.2.9 Development and Validation***

##### *Deep Neural Network Architecture*

Segmentations of all images were computed using a combination of two algorithms. One algorithm was trained to segment the node and the contour and included the necrosis and central hilar structure within the node segmentation, known as the aggregate node. The second algorithm was utilized to determine the presence of the central hilar structure and necrosis and segmented all tissues of interest (central hilar structure, necrosis, node, contour). Both segmentation algorithms utilized a U-Net style convolutional neural network (CNN). The network architecture is described in **Figure 1**.

The network input was an image shaped 512x512 pixels which was downsampled, to fit in graphics processing unit (GPU) memory, from the original image shape of 1300x975. The network output was a three-dimensional probability map, where dimensions 1 and 2 were 512x512, the same as the input image, and the third dimension included  $n$  levels that coincided with the probability of each pixel belonging to the  $n$  tissue of interest. The  $n$  probabilities for each pixel always summed to 1.0, each pixel was classified according to the tissue it had the highest probability of belonging to. After classifying each pixel to the appropriate tissue, segmentations were resampled to be the same shape as the original image (1300x975) using nearest neighbour interpolation.

For both algorithms, the network was trained using a batch size of 8, the Adam optimizer with a learning rate of  $10^{-3}$ , and a custom loss function which summed the negative dice similarity coefficients (DSC) of each of the tissues. Image augmentation including random rotation of up to 6 degrees and translation of up to 20% was employed. To enable robust estimation of the accuracy of predictions, a 5-fold cross-validation scheme was used.

The definition of variables and setting of the study were identical between the development and validation of the algorithm. However, inclusion criteria slightly differed as patients with esophageal cancer were also included and patients undergoing adjuvant chemotherapy were excluded from the derivation set.



*Phase A- Retrospective Algorithm Development: Derivation Set*

During training, segmentations produced by the expert endosonographer for each LN were used. After training was complete, the holdout (testing) LNs were predicted only one time. Using cross validation for this assessment provides testing errors/accuracies for the entire dataset, allowing for a more robust estimate of accuracy. Due to high intra-segmenter variability, while training the aggregate node and margin algorithm, only LNs that had an intra-segmenter Dice Similarity Coefficient for the aggregate node which were greater than 0.8 were used. This strategy was employed to reduce noise in learning the optimal segmentation and due to the importance of the aggregate node in calculating 2/4 of the CLNS features.

*Phase B- Prospective Prediction Validation Set*

Prediction of the CLNS score for the prospective dataset was completed using an ensemble of the 5 networks trained for the initial cross-validation. First, each LN was segmented using both of the trained algorithms (1. aggregate node and contour, 2. node, central hilar structure, necrosis, margin) for each of the 5 cross-validations. Next, the appropriate segmentations were used to calculate a set of CLNS features for each of the 5 cross-validations. The final CLNS feature score was determined for each node using the median of the 5 predictions.

## 4.3 RESULTS

### 4.3.1 Participants

The flow of patients for both Phase A and B is presented in **Figure 3**. Patient baseline characteristics and pathological data of biopsied and scored LNs for the derivation and validation set are presented in **Table 1**. The following sets of LNs were used:

- *Phase A- Derivation Set:* 298 LNs from 140 patients were used for training and testing of the deep neural network algorithm.
- *Phase B- Validation Set:* 108 LNs from 47 patients were used for the validation of the deep neural network algorithm.

Overall, there were no differences in the derivation and validation samples except for the proportion of patients that underwent MRI imaging (19.3% vs 59.6%,  $p=0.0002$ ), the proportion of malignant cases and LNs (55.0% vs 89.9% [lung cancer]; 16.4% vs 28.6% [malignant LNs]  $p<0.0001$ ), and the proportion of malignant features (small axis features [21.8% vs 44.6%,  $p<0.00001$ ], margin features [50.3% vs 29.5%,  $p=0.00016$ ], central hilar structure features [50.3% vs 32.1%,  $p=0.00096$ ] and central necrosis features [46.0% vs 21.4% , $p<0.0001$ ]).

The presence and absence of ultrasonographic features in malignant and benign and LNs in the derivation set were examined. Results were not statistically significant for small

axis length (Chi-Square=0.354,  $p=0.636$ ), margins (Chi-Square=0.033,  $p=0.845$ ) and CHS (Chi-Square=0.259,  $p=0.695$ ), suggesting that the features were independent of pathological outcome (**Table 2**). Predictive probabilities for LNs that were benign versus those that were malignant were presented in **Figure 4**. Central necrosis could not be evaluated as NeuralSeg did not identify this ultrasonographic feature.

#### ***4.3.2 Model Development and Specification***

In order to develop a predictive model using the derivation set in Phase A, univariate analysis was conducted to identify which features were independent predictors of malignancy (**Table 3 and 4**). None of the binary ultrasonographic features were found to be statistically significant (small axis, OR=1.209 [95% CI=0.646-2.264],  $p=0.552$ ; margins OR=1.209 [95% CI=0.646-2.264],  $p=0.552$ ; and CHS OR=1.209 [95% CI=0.646-2.264],  $p=0.552$ ). However, when continuous variables produced by NeuralSeg were examined, it was found that small axis length was significant ( $p=0.010$ ). As such, small axis length measured on a continuous scale was included in the model and the other features were investigated in a backwards entry process. None of the variables were found to be significant. Therefore, variables were forced back into the model based on the clinical significance determined by the CLNS. The final model is presented in **Table 5** and the calibration of the model is presented in **Figure 5**. The Hosmer and Lemeshow Test showed that the data fit the model well (Chi-square= 5.84,  $p\text{-value}=0.666$ )

### ***4.3.3 Model Discrimination and Calibration***

#### *Phase A*

The derivation set showed that a LN with a small axis measured on a continuous scale a margin >10mm measured on a continuous scale, and the absence of a central hilar structure had an increased odds of 2.392 (95% CI: 1.321-4.322), 3.972 (95% CI: 0.417-37.851) and 1.415 (95% CI: 0.657-3.045) for being malignant, respectively. However, only the small axis length was shown to be significant (p=0.004).

#### *Phase A and B*

Model discrimination for the derivation set (0.631 [95% CI: 0.543-0.719]) and validation set (0.748 [95% CI: 0.648-0.847]) are presented in **Figure 6 and 7**, respectively. Performance of our proposed algorithm compared to other algorithms in the literature are presented in **Table 6**.

### ***4.3.4 Model Performance***

#### *Phase A*

Diagnostic statistics were calculated for the accuracy between NeuralSeg and the endosonographer (**Table 7**). NeuralSeg was found to have an accuracy of 44.80% (95% CI: 39.06% to 50.64%) for small axis length, 73.33% (95% CI: 67.93% to 78.27%) for margin and 76.88% (95% CI: 71.69% to 81.54%) for CHS. When a CLNS of  $\geq 2$  was taken into account, NeuralSeg was found to have an accuracy of 78.76% (95% CI: 73.67% to

83.27%), specificity of 84.34% (95% CI: 79.22% to 88.62%) and Negative Predictive Value of 84.34% (95% CI: 84.26% to 86.22%).

#### *Phase B*

NeuralSeg was found to have an accuracy of 72.87% (95% CI: 63.46% to 80.98%), 73.33% (95% CI: 67.93% to 78.27%) for margin and 76.88% (95% CI: 71.69% to 81.54%), specificity of 90.79% (95% CI: 81.94% to 96.22%) and Negative Predictive Value of 75.92% (95% CI: 71.51% to 79.85%).

Diagnostic statistics and discrimination of  $\geq 2$  CLNS assigned by NeuralSeg for the derivation and validation set are presented in **Table 8**.

#### **4.4 DISCUSSION**

NeuralSeg was able to accurately identify and segment ultrasonographic features in LNs examined by EBUS. As a result, diagnostic statistics produced by NeuralSeg were similar to those found in the literature (Flechsig et al., 2017; Gao et al., 2015; Pham et al., 2017; Tagaya, Kurimoto, Osada, & Kobayashi, n.d.; H. Wang et al., 2017) Overall, NeuralSeg had an accuracy of 73.78% (95% CI: 68.40% to 78.68%), a sensitivity of 18.37% (95% CI: 8.76% to 32.02%) and specificity of 84.34% (95% CI: 79.22% to 88.62%). Our algorithm showed that NeuralSeg was in the middle of the range of accuracy (56-91%) and the higher end for specificity (23-93%) compared to other algorithms (**Table 6**). Although we were able to generate a prediction model based on NeuralSeg's

segmentation capabilities, central necrosis was not included in the model. As a result, risk groups (i.e. cut-off value of CLNS 2) could not be applied to the regression model. As such, the production of a modified regression model would be required in order to determine a cut-off that would be clinically significant.

Several studies have stressed the importance of test-retest evaluations for diagnostic prediction models. Major clinical journals such as *Journal of the American Medical Association* and *New England Journal of Medicine* have appreciated the reporting of model discrimination and calibration in independent samples and others recommend that a full independent external validation with data not available at the time of prediction model development is important (Steyerberg & Harrell, 2016). A systematic review found that when comparing performance validation on internal versus external validation, internal validation was shown to overestimate diagnostic accuracy for both the healthcare professional and deep learning algorithm. This highlights the need for an out-of-sample external validation for predictive models (X. Liu et al., 2019). Based on our study design, we were able to validate NeuralSeg on a new sample of LNs it has never seen before. Further, our external validation showed that NeuralSeg had higher diagnostic discrimination in the validation sample compared to the derivation sample (c-statistic= 0.60 [0.47-0.27] vs c-statistic=0.51 [0.42-0.63]), suggesting that prediction model performed better in the validation set in regard to sensitivity and specificity.

Four measures were taken in order to ensure the protection against sources of bias. First, consecutive sampling was employed to obtain the prospective validation set. Consecutive sampling is typically better than convenience sampling in controlling sampling bias as patients are enrolled in a systematic manner. Second, statistical analysis and sample size was determined a priori, thereby reducing the risk of an underpowered result. Third, inclusion criteria was not limited as to increase generalizability and mitigate diagnostic spectrum bias (Schmidt & Factor, 2013a). Finally, all patients received both the index test (NeuralSeg) and the reference standard (histopathology) in order to prevent verification bias.

However, this study is not without its limitations. First, we only tested four ultrasonographic features instead of an exhaustive list of radiomics features. One ultrasonographic feature that is worth mentioning is the margin feature (Echegaray, Bakr, Rubin, & Napel, 2018; Echegaray et al., 2016). On the basis of its construction, we expect that the margin may be correlated with the short axis feature. For example, the short axis length would be influenced by the smoothness of the LNs's boundary, with smoother boundaries making it more difficult to delineate where the LN margins are located. Because both axis length and margin features are dependent on the LN boundary, we believe that boundary features may exhibit similar feature variability. Second, as necrosis was not identified by NeuralSeg, there were too few events per the variable in order to include this feature in the prediction model. As such, we were only able to create a prediction model using three ultrasonographic features. Third, as there were varying baseline differences for

important features between the derivation and validation cohort. This may have affected the ability to make accurate predictions in the validation cohort. Finally, this study possessed a small sample size for machine learning. As a result, the study may have been underpowered to generate precise diagnostics statistics, as it was evident from wide confidence intervals.

We demonstrated that an AI algorithm that may be able to rule out malignancy in benign LNs with a specificity of 84.34% and an accuracy of 73.78%. NeuralSeg has the potential to decrease and eliminate retesting and ensure more rapid access to cancer treatment by shortening staging time and improving patient outcomes. Additionally, this prediction model may also obviate the need for many tissue samples that are currently obtained for the staging of lung cancer. This in turn will substantially decrease healthcare costs associated with tissue biopsies (needles, sampling equipment, storage, pathological analysis, databases etc.). As biopsies are estimated to cost \$1,120 per procedure, it is estimated that \$3,496,000 to \$5,360,000 per year is spent on EBUS procedures in Canada (Canadian Agency for Drugs and Technologies in Health (CADTH), 2010). Therefore, this new tool could save approximately between \$1,398,400 and \$2,144,000 on repeat biopsies due to inconclusive results. Future work with a larger dataset will be required to improve and refine the algorithm prior to trials in clinical practice.



**CHAPTER 4 TABLES AND FIGURES:****Table 1: Clinical characteristics and pathology results for derivation and validation sets.**

<b>Variable</b>	<b>DERIVATION</b> Patient Population (n=140) LNs (n=298)	<b>VALIDATION</b> Patient Population (n=47) LNs (n=108)	<b>P-Value</b>
<b>Age (years) [mean ± SD]</b>	69.92 ±10.64	70.64 ± 11.02	0.6912
<b>Males, n (%)</b>	76 (54.3)	22 (46.8%)	0.3735
<b>Pre-planned imaging modalities completed</b>			
Chest CT or PET, n (%)	139 (99.3%)	46 (91.5%)	0.4122
Head CT, n (%)	10 (7.1%)	3 (6.4%)	0.8572
MRI, n (%)	27 (19.3%)	28 (59.6%)	<b>0.0002</b>
<b>Median (range) of LNs scored per patient</b>	3 (1-4)	3 (1-5)	--
<b>Scored LN Stations</b>			
7, n (%)	125 (41.9%)	36 (31.9%)	0.0703
4R, n (%)	84 (28.2%)	39 (34.5%)	0.1902
4L, n (%)	57 (19.1%)	25 (22.1%)	0.4715
10, n (%)	14 (4.7%)	4 (3.5%)	0.6171
11, n (%)	7 (2.3%)	3 (2.7%)	0.8493
Other (1,2L,2R,12), n (%)	5 (1.7%)	5 (4.5%)	0.1031
<b>Pathology Diagnosis: Mass</b>			
Lung Cancer	77 (55.0%)	42 (89.9%)	<b>&lt;0.0001</b>
Esophageal Cancer	32 (22.9%)	NA	--
Benign Cases	31 (22.1%)	5 (10.6%)	<b>&lt;0.0001</b>
<b>Pathology Diagnosis: LNs</b>			
Malignant, n (%)	49 (16.4%)	32 (28.6%)	<b>&lt;0.0001</b>
Benign, n (%)	249 (83.6%)	76 (67.9%)	<b>&lt;0.0001</b>
Insufficient, n (%)	NR	4 (3.6%)	--
<b>Malignant Features**</b>			
Small axis, n, (%)	65 (21.8%)	50 (44.6%)	<b>&lt;0.00001</b>
Margins, n (%)	150 (50.3%)	33 (29.5%)	<b>0.00016</b>
*CHS, n (%)	150 (50.3)	36 (32.1%)	<b>0.00096</b>
Central Necrosis, n (%)	137 (46.0%)	24 (21.4%)	<b>&lt;0.00001</b>
<b>Treatment</b>			
Surgery, n, (%)	56 (40.0%)	20 (42.6%)	0.7566
Other treatment, n (%)	84 (60.0%)	27 (57.4%)	

\*CHS= central hilar structure; SD= standard deviation; CT= computed tomography; MRI= magnetic resonance imaging; PET= positron emission tomography

\*\* Based on unblinded scoring by endosonographer using CLNS in endoscopy suite

**Table 2. Ultrasonographic feature presence in malignant and benign lymph nodes in derivation set identified by NeuralSeg (reference standard = histopathology)**

<b>Feature</b>	<b>Malignant*</b> n=49 LNs	<b>Benign*</b> n=249 LNs	<b>Pearson's Chi Square Statistic</b>	<b>P-Value</b>
<b>Small Axis Length</b> >10 mm < 10 mm	30 (61.2%) 19 (38.8%)	141 (56.6%) 108 (43.4%)	0.354	0.636
<b>Margins</b> >50% <50%	10 (20.4%) 39 (79.6%)	48 (19.3%) 201(80.7%)	0.033	0.845
<b>CHS</b> Absent Present	38 (77.6%) 11 (22.4%)	201(80.7%) 48 (19.3%)	0.259	0.695
<b>Central Necrosis</b> Present Absent	NR NR	NR NR	-	-

\*Histopathologically confirmed

LNs= lymph nodes; CHS = central hilar structure

**Table 3. Univariate analysis of binary ultrasonographic features with logistic regression for derivation set.**

<b>Ultrasonographic Feature (Binary)</b>	<b>Odds Ratio</b>	<b>95% Confidence Interval</b>	<b>Standard Error</b>	<b>P-Value</b>
Small Axis	1.209	0.646-2.264	0.320	0.552
Margins	1.074	0.501-2.302	0.389	0.855
Central Hilar Structure	1.212	0.578-2.544	0.378	0.611

**Table 4. Univariate analysis of continuous ultrasonographic features with logistic regression for derivation set.**

<b>Ultrasonographic Feature (Continuous)</b>	<b>Odds Ratio</b>	<b>95% Confidence Interval</b>	<b>Standard Error</b>	<b>P-Value</b>
Small Axis	2.023	1.186-3.453	0.273	0.010
Margins	1.293	0.192-8.717	0.974	0.792

**Table 5. Multivariate analysis of ultrasonographic features with logistic regression for derivation set.**

<b>Ultrasonographic Feature</b>	<b>Odds Ratio</b>	<b>95% Confidence Interval</b>	<b>Standard Error</b>	<b>P-Value</b>
Small Axis (continuous)	2.392	1.321-4.322	0.303	0.004
Margins (continuous)	3.972	0.417-37.851	1.150	0.230
Central Hilar Structure	1.415	0.657-3.045	0.391	0.375
Constant	0.037	0.009-0.087	0.718	<0.001

**Table 6. Comparison of NeuralSeg’s predictive algorithm with previously published studies with lymph node segmentation algorithms**

<b>Study</b>	<b>Proportion of Malignant LNs</b>	<b>Imaging Modality</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Accuracy</b>	<b>Negative Predictive Value</b>	<b>Positive Predictive Value</b>
<b>Fleschig (2017)</b>	52%	PET/CT	92%	77%	NR	NR	NR
<b>Gao (2015)</b>	NR	PET/CT	52-95% (depending on model)	60-75% (depending on model)	56-86% (depending on model)	NR	NR
<b>Pham (2017a)</b>	NR	CT	76-89% (depending on model)	60-93% (depending on model)	NR	62-89% (depending on model)	62-93% (depending on model)
<b>Pham (2017b)</b>	49%	CT	75%	90%	NR	NR	NR
<b>Tagaya (2008)</b>	73%	EBUS	21-100% (depending on model)	23-88% (depending on model)	91%	NR	NR
<b>Wang (2017)</b>	91%	PET/CT	72-86% (depending on model)	84-87% (depending on model)	80-85%	NR	NR
<b>Proposed Model</b>	16%	EBUS	18%	84%	74%	84%	18%

**CT**= computed tomography; **PET**=positron emission tomography; **EBUS**= endobronchial ultrasound; **NR**= not reported

**Table 7. Diagnostic statistics of ultrasonographic lymph node features and  $\geq 2$  CLNS determined by NeuralSeg in derivation set (gold standard = endosonographer)**

Feature/Score	Total	TP	TN	FP	FN	Sensitivity (95% CI)	Specificity (95% CI)	**Accuracy (95% CI)
<b>Small Axis Length</b> >10 mm (1) < 10 mm (0)	298	37	99	134	28	56.92% (44.04% to 69.15%)	42.49% (36.06% to 49.11%)	44.80% (39.06% to 50.64%)
<b>Margins</b> >50% (1) <50% (0)	298	33	123	25	117	22.00% (15.65% to 29.49%)	83.11% (76.08% to 88.76%)	73.33% (67.93% to 78.27%)
<b>CHS</b> Absent (1) Present (0)	298	39	129	20	111	26.00% (19.19% to 33.79%)	86.58% (80.03% to 91.60%)	76.88% (71.69% to 81.54%)
<b>Central Necrosis</b> Present (1) Absent (0)	0	-	-	-	-	-	-	-
<b>CLNS</b> <2 (Benign) $\geq 2$ (Malignant)	298	33	130	15	120	21.57% (15.34% to 28.94%)	89.66% (83.51% to 94.09%)	78.76% (73.67% to 83.27%)

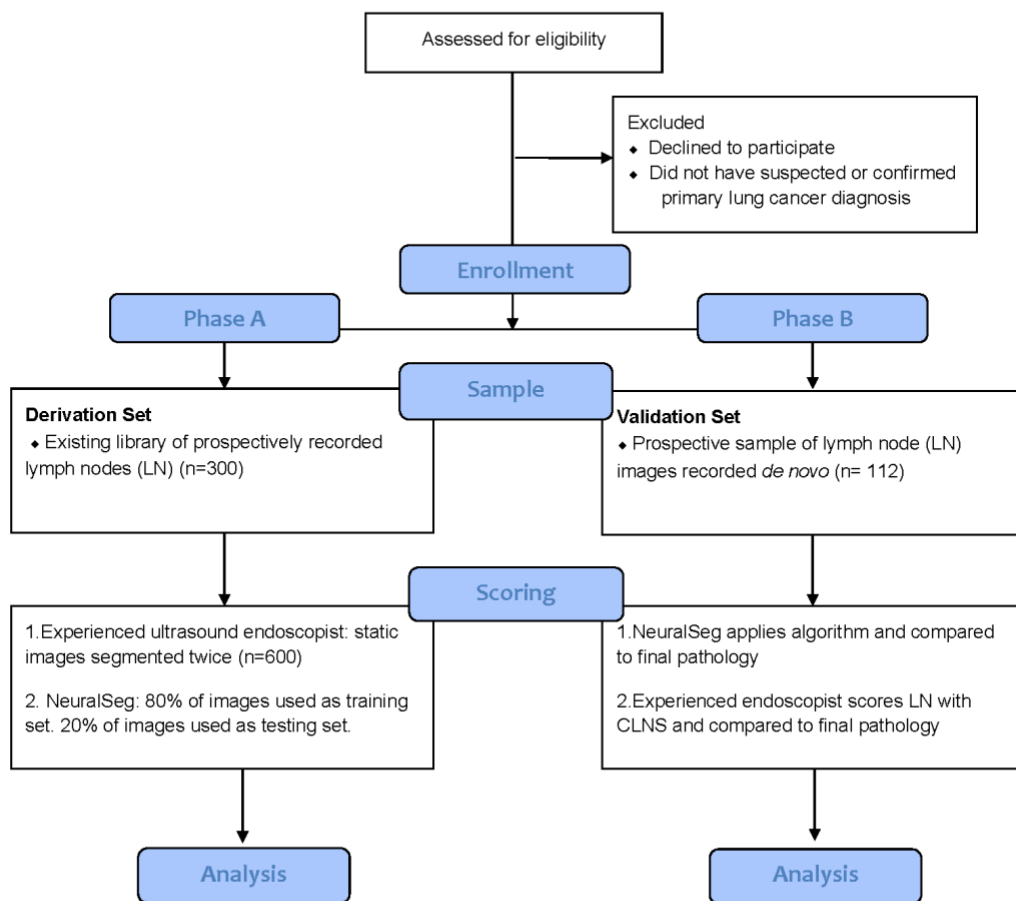
**CHS** = central hilar structure; **CI** = confidence interval

\*\* accuracy determined based off of prevalence of malignancy

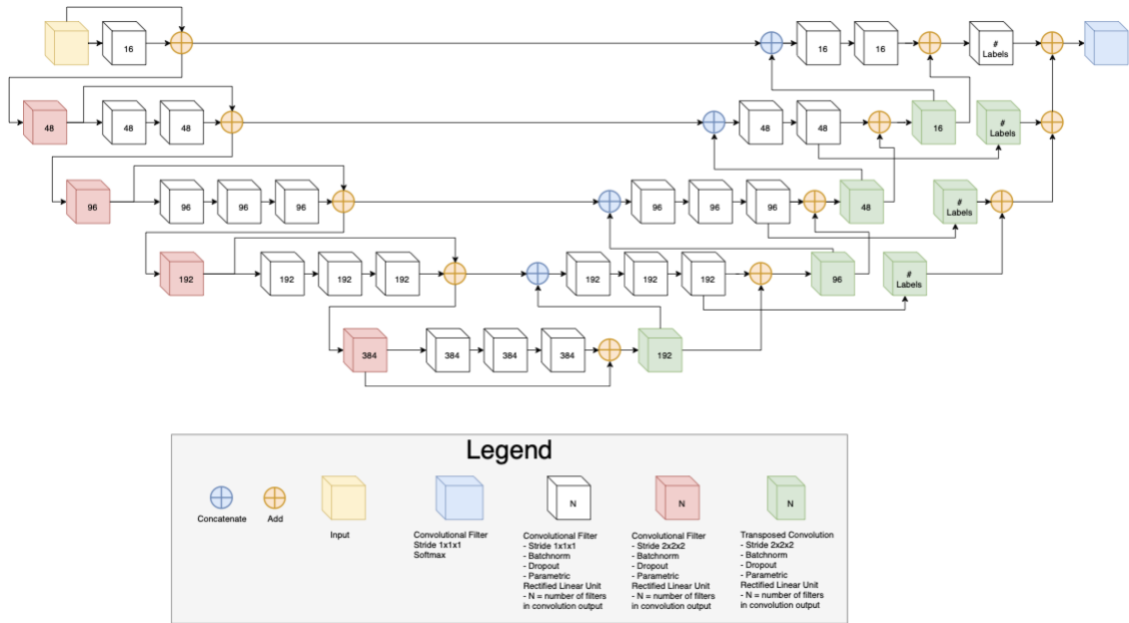
**Table 8. Diagnostic statistics and discrimination of  $\geq 2$  CLNS assigned by NeuralSeg (reference standard = histopathology)**

<b>Measure</b>	<b>Derivation Set (n=298)</b>	<b>Validation Set (n=108)</b>
<b>Area Under the Curve (95% CI)</b>	0.63 (0.54 to 0.72)	0.75 (0.65 to 0.85)
<b>Sensitivity (95% CI)</b>	18.37% (8.76% to 32.02%)	28.12% (13.75% to 46.75%)
<b>Specificity (95% CI)</b>	84.34% (79.22% to 88.62%)	90.79% (81.94% to 96.22%)
<b>Positive Predictive Value (95% CI)</b>	18.26% (10.38% to 30.11%)	55.02% (33.27% to 75.00%)
<b>Negative Predictive Value (95% CI)</b>	84.43% (82.46% to 86.22%)	75.92% (71.51% to 79.85%)
<b>Positive Likelihood Ratio (95% CI)</b>	1.17 (0.61 to 2.26)	3.05 (1.24 to 7.49)
<b>Negative Likelihood Ratio (95% CI)</b>	0.97 (0.84 to 1.12)	0.79 (0.63 to 0.99)
<b>Accuracy (95% CI)</b>	73.78% (68.40% to 78.68%)	72.87% (63.46% to 80.98%)

\*CI= confidence interval



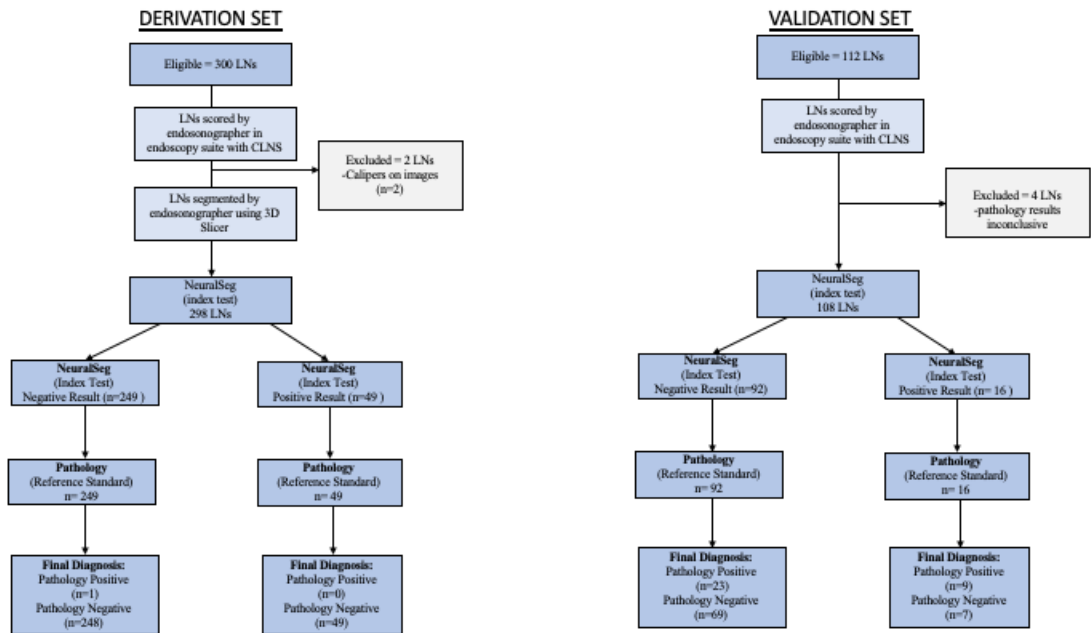
**Figure 1. Study design.** \*Phase A commences before Phase B.



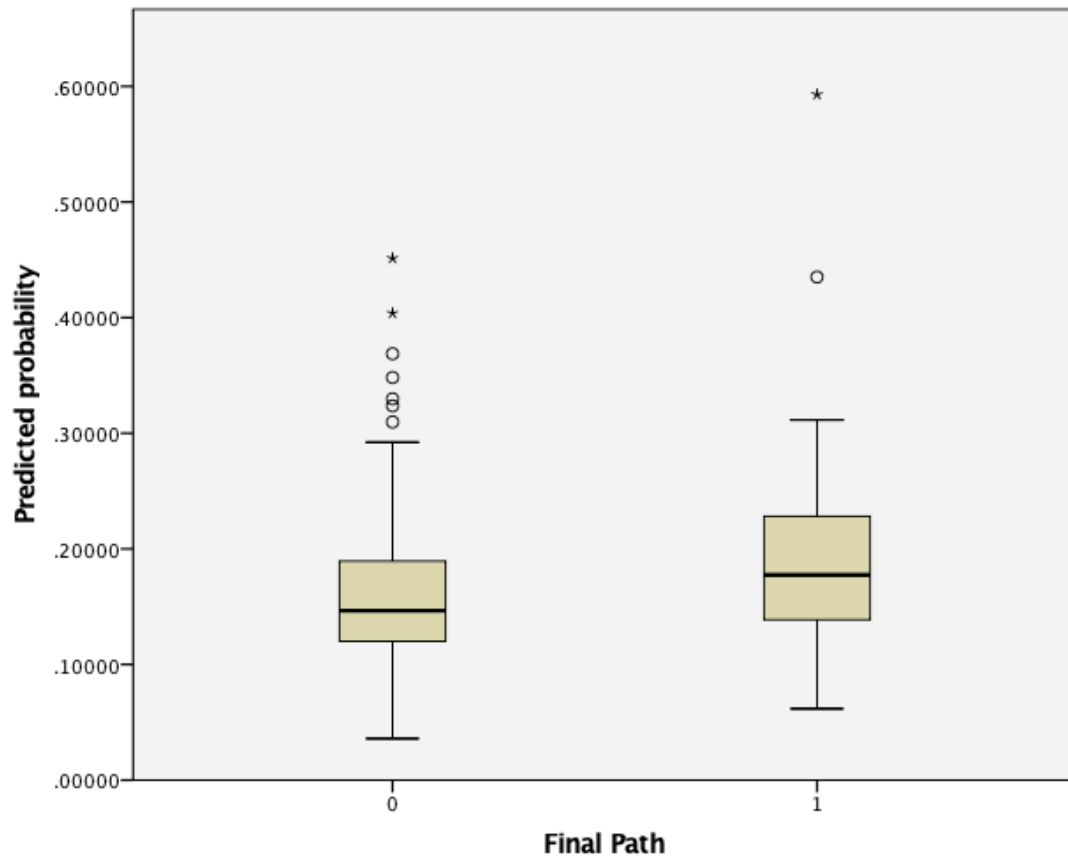
**Figure 2: A representation of NeuralSeg’s network with a U-Net style architecture.**

The figure legend describes all symbols. The network followed a U-Net style architecture and included long residual connections from the compression to decompression branches. The network used short residual connections at each stage in the compression and decompression branches. All short residual connections used a summation methodology, and all long residual connections used concatenation. In each stage of the compression branch each image dimension was compressed to be  $\frac{1}{2}$  its previous size by using a stride of 2 with a regular convolution, and in each stage of the decompression branch each image dimension was doubled to be 2 times its previous size by using a stride of 2 with a transpose convolution. The number of outputted filters from each convolution operation are printed on the associated symbol; in general, the number of filters tripled after the first down convolution and then doubled after each successive down convolution until the image dimensions reached their smallest size (after 4 down convolutions). Then, each stage of the decompression branch included the same number of filters as the equivalent level on the compression branch. All convolutions are described by their respective symbol, generally they all used a convolution filter of 3x3, and were followed by batch normalization, dropout (probability of being dropped = 0.6), and a parametric rectified linear unit (PReLU). In addition to traditional U-Net compression and decompression branches, the network included a form of deep-supervision. The deep supervision more directly passed data from the deep layers directly to the final output, using the same number of filters (equivalent to the number of labels in the image) at each stage. The main difference between the current architecture and that described previously is that we used a PReLU activation instead of a logistic or softmax function. The final convolution filter used a softmax activation to give probabilities of each pixel belonging to each of the potential tissue classes.

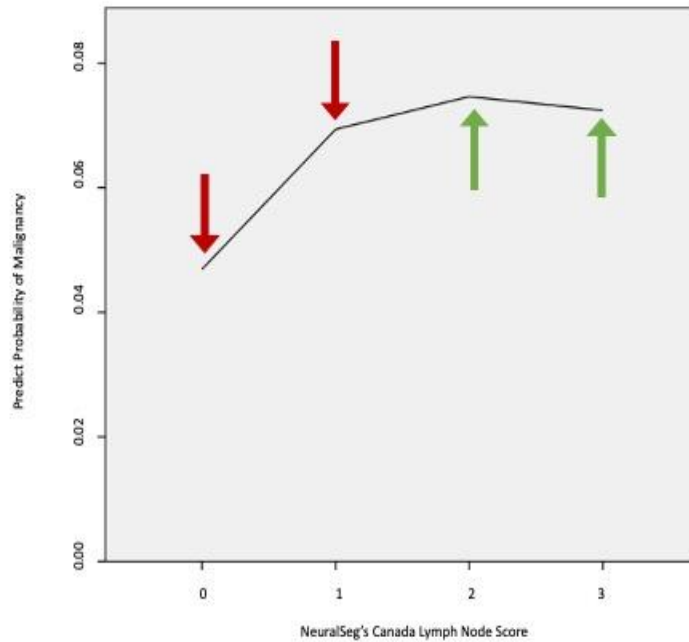




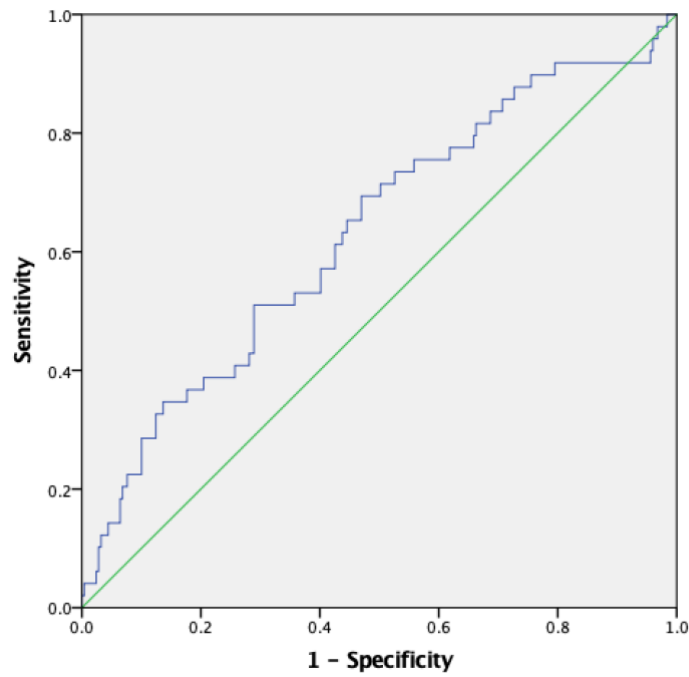
**Figure 3. Flow of participants through study. Phase A:** development of NeuralSeg using a derivation set of 298 lymph node images. **Phase B:** validation of NeuralSeg using a new sample of 108 prospectively collected lymph node images. Phase A occurred before Phase B.



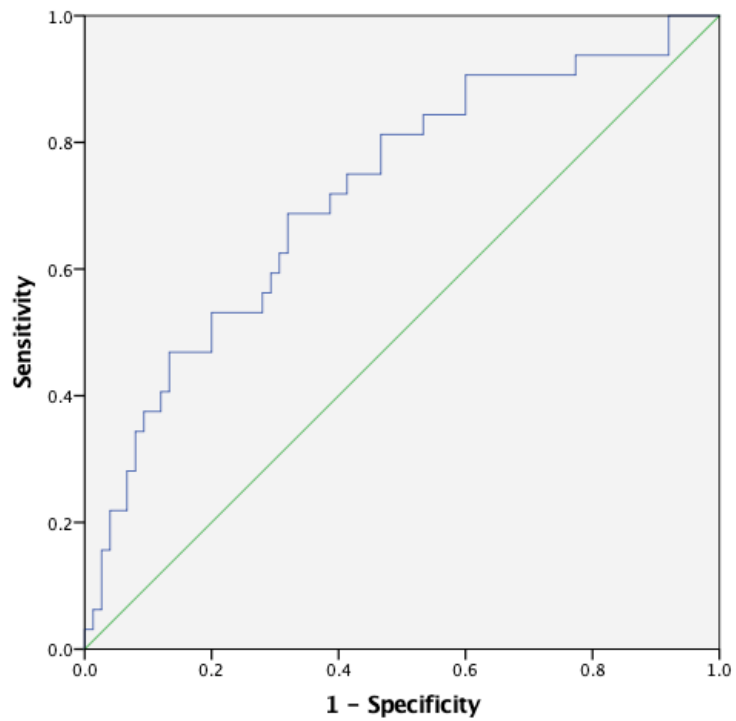
**Figure 4.** Boxplots displaying the predicted probabilities for lymph nodes that were malignant versus those that were benign based on a multivariate model for the derivation set. Malignancy (category = 1) versus benign (category = 0). Medians of predicted probabilities differed between benign and malignant lymph nodes ( $p=0.29$ ) Outliers denoted with a circle are more than 1.5 the interquartile range.



**Figure 5. NeuralSeg multivariate model calibration plot for derivation set.** NeuralSeg CLNS Score of equal to or greater than 2 is considered highly suspicious for malignancy. NS score of 2 has a 0.70 probability of malignancy NeuralSeg CLNS of 3 has a 0.67 probability of malignancy. Hosmer and Lemeshow Test goodness of fit (Chi-square= 5.84, p-value=0.666)



**Figure 6. NeuralSeg multivariate model receiver operator characteristic curve for derivation set (c-index = 0.631, st. error=0.045, 95% confidence interval=0.543-0.719).**



**Figure 7. NeuralSeg multivariate model receiver operator characteristic curve for validation set (c-index = 0.748, st. error= 0.051, 95% confidence interval=0.648-0.847).**

## **CHAPTER 5: SUMMARY OF FINDINGS AND METHODOLOGICAL CHALLENGES**

### **5.1 THESIS FINDINGS AND LESSONS LEARNED**

One of the major hurdles for successful translation of deep learning algorithms from research to practice in precision medicine is their interpretability to physicians. NeuralSeg aimed to learn a binary four-point score capable of predicting LN malignancy. However, due to NeuralSeg's limited ability to identify all four ultrasonographic features, it was determined that the predictive model that was developed and validated only contained three features (small axis length, margins and CHS) two of which were continuous measurements. Initially, the CLNS was developed as a binary scoring system in order to assist clinicians in predicting LN malignancy. This binary scale allowed for easy interpretability and ease of administration during practice. In contrast, NeuralSeg's predictive model included two features with continuous measurements. Consequently, this modified predictive model may be difficult to implement in a clinical setting due to the potential need to have the algorithm score the LNs in real time. This may pose as a challenge in a clinical setting.

## **5.2 METHODOLOGICAL CHALLENGES AND MITIGATION STRATEGIES**

Several methodological challenges associated with diagnostic studies and machine learning were encountered throughout the process of this thesis. The explanation of the challenges and their mitigation strategies and/or solutions are presented below.

### ***“Lucky Good Fit” and Overfitting***

The use of a smaller dataset in machine learning can lead to a “lucky good fit” when determining the goodness of fit for the prediction model (James, Witten, Hastie, & Tibshirani, 2000; Mutasa, Sun, & Ha, 2020). In order to mitigate this methodological challenge, we sought to have more data in the training phase of the prediction model and employed K-fold cross validation (Rodriguez-Roisin, 2000). Additionally, the K-fold cross validation was also used to prevent overfitting. Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test mean square error (James et al., 2000). This occurs as a result of the model learning details and noise specific to the training set (Yamashita, Nishio, Do, & Togashi, 2018). Further, we also performed a routine check on the training data to monitor the loss and accuracy on the training set and used testing to ensure proper performance evaluation of the algorithm.

### ***Appropriate Internal and External Validation***

When conducting prediction model studies using machine learning, it is important to use rigorous methodologies. In our study we used a five-fold-cross validation. Using

cross validation for this assessment provides testing errors/accuracies for the entire dataset, providing a more robust estimate of accuracy. Furthermore, this method included training the algorithm from the project's outset using 5 different sets of data and therefore represents its generalizability to learning from different inputs (James et al., 2000). However, before considering whether to use a clinical prediction model, it is essential that its predictive performance be empirically evaluated in datasets that were not used to develop the model(Steyerberg & Harrell, 2016). This process is often referred to as external validation. Predictive performance is typically characterised by evaluating a model's calibration and discrimination(Steyerberg & Harrell, 2016). Calibration is the agreement between predicted and observed risks, whilst discrimination is the ability of the model to differentiate between patients with different outcomes(Collins et al., 2015). In our study, we assessed the discrimination and calibration of both the development and validation samples in order to objectively evaluate the models and determine if the prediction model may be applied in a clinical setting.

### ***Generalizability***

The term generalizability refers to the extension of research findings and conclusions from a study conducted on a sample population to the population at large (Debray et al., 2015). In order for this to occur, it is important to use a diverse dataset for study samples. In our study, inclusion criteria were not limited so that the range of disease and all stages of lung cancer would be captured. Additionally, we used consecutive sampling so as to avoid sampling bias and further ensure the study sample was



representative of the patient population. Further, considering the role of generalizability is important in retest studies as some studies have specific case mix groupings or low prevalence of disease (Tugwell & Knottnerus, 2015).

### ***Diagnostic Review Bias***

These types of bias occur when the interpretation of the reference test is not independent of the index test, which weakens the results of retrospective studies (Schmidt & Factor, 2013a). A rigorous study would require either reporting that the results are blinded, or that the cases were reviewed again to obtain a blinded diagnosis. In our study, the endosonographer assessing the images was blinded to the personal identifiers, imaging, and pathology results associated with each LN. Additionally, images were shuffled and assigned random identification numbers to ensure that repeated images were not segmented in a defined order and to prevent diagnostic review bias.

### ***Spectrum Bias***

A spectrum effect occurs when there is a variation in the performance of tests for the prediction for the diagnosis of disease among population subgroups (Usher-Smith, Sharp, & Griffin, 2016). A patient's probability of a disease is in part determined by the test's result (Guyatt et al., 2000). As such, it is essential to have a reliable estimate of the test's performance in order to make good decisions and ensure appropriate patient management. Unfortunately, studies tend to report weighted average estimates across

broad patient populations and result in inaccurate predictions at the individual and population level due to case mixes.

To avert such a problem and guarantee clinicians can use a specific tool in practice, investigators should test all relevant subgroups and be explicit in the case mix in the study sample. In this study, we were able to stratify patients based on their risk of malignancy (CLNS greater than or equal to 2) and we also reported the prevalence of LN malignancy in each sample. The aforementioned steps taken are critical as a test developed in a population with a higher prevalence of disease will typically have a lower sensitivity and higher specificity when applied to a new population with lower disease prevalence. However, it is critical to note that the opposite occurred in our study, in which the derivation sample had a lower LN malignancy prevalence (16.4%) compared to the validation sample (26.8%).

### ***Verification Bias***

Standard methods for assessing the accuracy of diagnostic tests require the determination of true disease status for each study patient. As such, if the decision to verify a patient is influenced by the knowledge of a test result, or if only certain cases are verified, this may introduce biased results, known as verification bias (Cronin & Vickers, 2008). In order to mitigate verification bias, only patients that had both the index test and reference standard were included in the study. Verification bias can affect the accuracy of an index test as partial verification bias will underestimate the number of false negative patients and

thus carries a risk of overestimating the sensitivity (O'Sullivan, Banerjee, Heneghan, & Pluddemann, 2018).

## REFERENCES

- Akhtar-Danesh, N., & Finley, C. (2015). Temporal trends in the incidence and relative survival of non-small cell lung cancer in Canada: A population-based study. *Lung Cancer*, *90*(1), 8–14. <https://doi.org/10.1016/j.lungcan.2015.07.004>
- Annema, J. T., Van Meerbeeck, J. P., Rintoul, R. C., Doooms, C., Deschepper, E., Dekkers, O. M., ... Tournoy, K. G. (2010). Mediastinoscopy vs endosonography for mediastinal nodal staging of lung cancer: A randomized trial. *JAMA - Journal of the American Medical Association*, *304*(20), 2245–2252. <https://doi.org/10.1001/jama.2010.1705>
- Aquino, S. L., Asmuth, J. C., Alpert, N. M., Halpern, E. F., Fischman, A. J., Aquino, S. L., ... Halpern, E. F. (2003). *Improved Radiologic Staging of Lung Cancer with 2-[18 F]-Fluoro-2-Deoxy-D-Glucose-Positron Emission Tomography and Computed Tomography Registration*. *Journal of Computer Assisted Tomography* (Vol. 27).
- Barnes, H., See, K., Barnett, S., & Manser, R. (2017, April 21). Surgery for limited-stage small-cell lung cancer. *Cochrane Database of Systematic Reviews*. John Wiley and Sons Ltd. <https://doi.org/10.1002/14651858.CD011917.pub2>
- Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics and Gynecology*, *31*(4), 466–475. <https://doi.org/10.1002/uog.5256>
- Bella, P. W. M., Danciewicz, M., Szczęśny, T., & Kowalewski, J. (2013). P-151 EVALUATION OF METASTASES IN MEDIASTINAL LYMPH NODES BASED ON POSITRON EMISSION TOMOGRAPHY-COMPUTED TOMOGRAPHY SCANNING IN NON-SMALL CELL LUNG CANCER. *Interactive CardioVascular and Thoracic Surgery*, *17*(suppl\_1), S40–S40. <https://doi.org/10.1093/ICVTS/IVT288.151>
- Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., ... Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*. <https://doi.org/10.3322/caac.21552>
- Boffa, D., Fernandez, F. G., Kim, S., Kosinski, A., Onaitis, M. W., Cowper, P., ... Furnary, A. P. (2017). Surgically Managed Clinical Stage IIIA–Clinical N2 Lung Cancer in The Society of Thoracic Surgeons Database. In *Annals of Thoracic Surgery* (Vol. 104, pp. 395–403). Elsevier USA. <https://doi.org/10.1016/j.athoracsur.2017.02.031>
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. <https://doi.org/10.3322/caac.21492>
- Brenner, D. R., Weir, H. K., Demers, A. A., Ellison, L. F., Louzado, C., Shaw, A., ... Smith, L. M. (2020). Projected estimates of cancer in Canada in 2020. *CMAJ*, *192*(9), E199–E205. <https://doi.org/10.1503/cmaj.191292>
- Burdett, S., Pignon, J. P., Tierney, J., Tribodet, H., Stewart, L., Le Pechoux, C., ... Liang, Y. (2015, March 2). Adjuvant chemotherapy for resected early-stage non-small cell

- lung cancer. *Cochrane Database of Systematic Reviews*. John Wiley and Sons Ltd. <https://doi.org/10.1002/14651858.CD011430>
- Campbell, J. M., Klugar, M., Ding, S., Carmody, D. P., Hakonsen, S. J., Jadotte, Y. T., ... Munn, Z. (2015). Diagnostic test accuracy. *International Journal of Evidence-Based Healthcare*, 13(3), 154–162. <https://doi.org/10.1097/XEB.0000000000000061>
- Canadian Agency for Drugs and Technologies in Health (CADTH). (2010). Endobronchial ultrasound for lung cancer diagnosis and staging: a review of the clinical and cost-effectiveness. *CADTH Technology Overviews*, 1(2), e0115. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22977405>
- Cancer Care Ontario. (n.d.). *Lung Cancer Diagnosis Pathway Map*.
- Care, C. T. F. on P. H. (2016). Recommendations on screening for lung cancer. *CMAJ*, 188(6), 425–432. <https://doi.org/10.1503/cmaj.151421>
- Carter, B. W., Lichtenberger, J. P., Benveniste, M. K., de Groot, P. M., Wu, C. C., Erasmus, J. J., & Truong, M. T. (2018). Revisions to the TNM Staging of Lung Cancer: Rationale, Significance, and Clinical Application. *RadioGraphics*, 38(2), 374–391. <https://doi.org/10.1148/rg.2018170081>
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, 162(1), 55–63. <https://doi.org/10.7326/M14-0697>
- Cronin, A. M., & Vickers, A. J. (2008). Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: A simulation study. *BMC Medical Research Methodology*, 8(1), 75. <https://doi.org/10.1186/1471-2288-8-75>
- Debray, T. P. A., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W., & Moons, K. G. M. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3), 279–289. <https://doi.org/10.1016/j.jclinepi.2014.06.018>
- Department of Health. (2011). *Improving Outcomes: A Strategy for Cancer*.
- Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4–5), 198–211. <https://doi.org/10.1016/j.compmedimag.2007.02.002>
- Echegaray, S., Bakr, S., Rubin, D. L., & Napel, S. (2018). Quantitative Image Feature Engine (QIFE): an Open-Source, Modular Engine for 3D Quantitative Feature Extraction from Volumetric Medical Images. *Journal of Digital Imaging*, 31(4), 403–414. <https://doi.org/10.1007/s10278-017-0019-x>
- Echegaray, S., Nair, V., Kadoch, M., Leung, A., Rubin, D., Gevaert, O., & Napel, S. (2016). A Rapid Segmentation-Insensitive “Digital Biopsy” Method for Radiomic Feature Extraction: Method and Pilot Study Using CT Images of Non-Small Cell Lung Cancer. *Tomography (Ann Arbor, Mich.)*, 2(4), 283–294. <https://doi.org/10.18383/j.tom.2016.00163>
- Evans, W. K., Flanagan, W. M., Miller, A. B., Goffin, J. R., Memon, S., Fitzgerald, N., & Wolfson, M. C. (2016). Implementing low-dose computed tomography screening for lung cancer in Canada: Implications of alternative at-risk populations, screening

- frequency, and duration. *Current Oncology*, 23(3), e179–e187.  
<https://doi.org/10.3747/co.23.2988>
- Evison, M., Morris, J., Martin, J., Shah, R., Barber, P. V., Booton, R., & Crosbie, P. A. J. (2015). Nodal staging in lung cancer: A risk stratification model for lymph nodes classified as negative by EBUS-TBNA. *Journal of Thoracic Oncology*, 10(1), 126–133. <https://doi.org/10.1097/JTO.0000000000000348>
- Ferreira-Junior, J. R., Koenigkam-Santos, M., Magalhães Tenório, A. P., Faleiros, M. C., Garcia Cipriano, F. E., Fabro, A. T., ... de Azevedo-Marques, P. M. (2020). CT-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms. *International Journal of Computer Assisted Radiology and Surgery*, 15(1), 163–172. <https://doi.org/10.1007/s11548-019-02093-y>
- Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1). <https://doi.org/10.1186/1472-6947-12-8>
- Flechsig, P., Frank, P., Kratochwil, C., Antoch, G., Rath, D., Moltz, J., ... Giesel, F. L. (2017). Radiomic Analysis using Density Threshold for FDG-PET/CT-Based N-Staging in Lung Cancer Patients. *Molecular Imaging and Biology*, 19(2), 315–322. <https://doi.org/10.1007/s11307-016-0996-z>
- Fujiwara, T., Yasufuku, K., Nakajima, T., Chiyo, M., Yoshida, S., Suzuki, M., ... Yoshino, I. (2010). The Utility of Sonographic Features During Endobronchial Ultrasound-Guided Transbronchial Needle Aspiration for Lymph Node Staging in Patients With Lung Cancer. *Chest*, 138(3), 641–647. <https://doi.org/10.1378/chest.09-2006>
- Gao, X., Chu, C., Li, Y., Lu, P., Wang, W., Liu, W., & Yu, L. (2015). The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from 18F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer. *European Journal of Radiology*, 84(2), 312–317. <https://doi.org/10.1016/j.ejrad.2014.11.006>
- Genseke, P., Wielenberg, C., Schreiber, J., Waller, M. K. (2019). Quantitative F-18-FDG-PET/CT in preoperative staging of lung cancer as a potential target for machine learning - a prospective study. Retrieved June 8, 2020, from [http://jnm.snmjournals.org/content/60/supplement\\_1/1341.short](http://jnm.snmjournals.org/content/60/supplement_1/1341.short)
- Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2), 563–577. <https://doi.org/10.1148/radiol.2015151169>
- Gogia, P., Insaf, T. Z., McNulty, W., Boutou, A., Nicholson, A. G., Zoumot, Z., & Shah, P. L. (2016). Endobronchial ultrasound: Morphological predictors of benign disease. *ERJ Open Research*, 2(1). <https://doi.org/10.1183/23120541.00053-2015>
- Goldstraw, P., Chansky, K., Crowley, J., Rami-Porta, R., Asamura, H., Eberhardt, W. E. E., ... Yokoi, K. (2016). The IASLC lung cancer staging project: Proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM Classification for lung cancer. *Journal of Thoracic Oncology*, 11(1), 39–51. <https://doi.org/10.1016/j.jtho.2015.09.009>

- Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., ... Richardson, W. S. (2000, September 13). Users' guides to the medical literature: XXV. Evidence-based medicine: Principles for applying the users' guides to patient care. *Journal of the American Medical Association*. American Medical Association. <https://doi.org/10.1001/jama.284.10.1290>
- H., S., S., M., J., M., P., C., L., S., F., A., ... Ahmed, F. (2019). Prediction of lymph node maximum standardized uptake value in patients with cancer using a 3D convolutional neural network: A proof-of-concept study. *American Journal of Roentgenology*, 212(2), 238–244. <https://doi.org/10.2214/AJR.18.20094> LK - [http://elinks.library.upenn.edu/sfx\\_local?sid=EMBASE&issn=15463141&id=doi:10.2214%2FAJR.18.20094&atitle=Prediction+of+lymph+node+maximum+standardized+uptake+value+in+patients+with+cancer+using+a+3D+convolutional+neural+network%3A+A+proof-of-concept+study&stitle=Am.+J.+Roentgenol.&title=American+Journal+of+Roentgenology&volume=212&issue=2&spage=238&epage=244&aulast=Shaish&aufirst=Hiram&aunit=H.&aufull=Shaish+H.&coden=AJROA&isbn=&pages=238-244&date=2019&aunit1=H&aunit](http://elinks.library.upenn.edu/sfx_local?sid=EMBASE&issn=15463141&id=doi:10.2214%2FAJR.18.20094&atitle=Prediction+of+lymph+node+maximum+standardized+uptake+value+in+patients+with+cancer+using+a+3D+convolutional+neural+network%3A+A+proof-of-concept+study&stitle=Am.+J.+Roentgenol.&title=American+Journal+of+Roentgenology&volume=212&issue=2&spage=238&epage=244&aulast=Shaish&aufirst=Hiram&aunit=H.&aufull=Shaish+H.&coden=AJROA&isbn=&pages=238-244&date=2019&aunit1=H&aunit)
- Hatt, M., Cheze-Le Rest, C., Van Baardwijk, A., Lambin, P., Pradier, O., & Visvikis, D. (2011). Impact of tumor size and tracer uptake heterogeneity in 18F-FDG PET and CT non-small cell lung cancer tumor delineation. *Journal of Nuclear Medicine*, 52(11), 1690–1697. <https://doi.org/10.2967/jnumed.111.092767>
- He, L., Huang, Y., Ma, Z., Liang, C., Huang, X., Cheng, Z., ... Liu, Z. (2017). A CT-based radiomics analysis for clinical staging of non-small cell lung cancer. *Chinese Journal of Radiology (China)*, 51(12), 906–911. <https://doi.org/10.3760/cma.j.issn.1005-1201.2017.12.004>
- He, L., Huang, Y., Yan, L., Zheng, J., Liang, C., & Liu, Z. (2019). Radiomics-based predictive risk score: A scoring system for preoperatively predicting risk of lymph node metastasis in patients with resectable non-small cell lung cancer. *Chinese Journal of Cancer Research*, 31(4), 641–652. <https://doi.org/10.21147/j.issn.1000-9604.2019.04.08>
- Henzler, T. (2017). MTE22.01 Perspectives in Lung Cancer Imaging. *Journal of Thoracic Oncology*, 12(1), S172–S173. <https://doi.org/10.1016/j.jtho.2016.11.154>
- Herth, F. J. F. (2013). Endobronchial ultrasound: First choice for the mediastinum. *Endoscopic Ultrasound*. Spring Media. <https://doi.org/10.4103/2303-9027.121235>
- Huang, Y. Q., Liang, C. H., He, L., Tian, J., Liang, C. S., Chen, X., ... Liu, Z. Y. (2016). Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *Journal of Clinical Oncology*, 34(18), 2157–2164. <https://doi.org/10.1200/JCO.2015.65.9128>
- Huidrom, R., Jina Chanu, Y., & Manglem Singh, K. (2018). Automated Lung Segmentation on Computed Tomography Image for the Diagnosis of Lung Cancer. <https://doi.org/10.13053/CyS-22-3-2526>
- Hylton, D. A., Turner, J., Shargall, Y., Finley, C., Agzarian, J., Yasufuku, K., ... Hanna, W. C. (2018). Ultrasonographic characteristics of lymph nodes as predictors of malignancy during endobronchial ultrasound (EBUS): A systematic review. *Lung*

- Cancer (Amsterdam, Netherlands)*, 126, 97–105.  
<https://doi.org/10.1016/j.lungcan.2018.10.020>
- Hylton, D. A., Turner, S., Kidane, B., Spicer, J., Xie, F., Farrokhyar, F., ... Hanna, W. C. (2019). The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. *The Journal of Thoracic and Cardiovascular Surgery*. <https://doi.org/10.1016/j.jtcvs.2019.10.205>
- Hylton, D. A., Turner, S., Kidane, B., Spicer, J., Xie, F., Farrokhyar, F., ... Hanna, W. C. (2020). The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. *Journal of Thoracic and Cardiovascular Surgery*. <https://doi.org/10.1016/j.jtcvs.2019.10.205>
- Inoue, K., Moriya, E., Suzuki, T., Ohnuki, Y., Sato, T., Kitamura, H., ... Fujii, H. (2011). The usefulness of fully three-dimensional OSEM algorithm on lymph node metastases from lung cancer with 18F-FDG PET/CT. *Annals of Nuclear Medicine*, 25(4), 277–287. <https://doi.org/10.1007/s12149-010-0462-y>
- Jalil, B. A., Yasufuku, K., & Khan, A. M. (2015). Uses, Limitations, and Complications of Endobronchial Ultrasound. *Baylor University Medical Center Proceedings*, 28(3), 325–330. <https://doi.org/10.1080/08998280.2015.11929263>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2000). *An introduction to Statistical Learning. Current medicinal chemistry* (Vol. 7). <https://doi.org/10.1007/978-1-4614-7138-7>
- Jethanandani, A., Lin, T. A., Volpe, S., Elhalawani, H., Mohamed, A. S. R., Yang, P., & Fuller, C. D. (2018, May 14). Exploring applications of radiomics in magnetic resonance imaging of head and neck cancer: A systematic review. *Frontiers in Oncology*. Frontiers Media S.A. <https://doi.org/10.3389/fonc.2018.00131>
- Jha, S., & Topol, E. J. (2016, December 13). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA - Journal of the American Medical Association*. American Medical Association. <https://doi.org/10.1001/jama.2016.17438>
- Jhun, B. W., Um, S. W., Suh, G. Y., Chung, M. P., Kim, H., Kwon, O. J., ... Lee, K. J. (2014). Clinical value of endobronchial ultrasound findings for predicting nodal metastasis in patients with suspected lymphadenopathy: A prospective study. *Journal of Korean Medical Science*, 29(12), 1632–1638. <https://doi.org/10.3346/jkms.2014.29.12.1632>
- Jiang, Y., Yang, M., Wang, S., Li, X., & Sun, Y. (2020, April 1). Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Communications*. John Wiley and Sons Inc. <https://doi.org/10.1002/cac2.12012>
- Jones, S. R., Carley, S., & Harrison, M. (2003). *An introduction to power and sample size estimation. Emerg Med J* (Vol. 20). Retrieved from [www.emjonline.com](http://www.emjonline.com)
- Kalemkerian, G. P., & Schneider, B. J. (2017, February 1). Advances in Small Cell Lung Cancer. *Hematology/Oncology Clinics of North America*. W.B. Saunders. <https://doi.org/10.1016/j.hoc.2016.08.005>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>



- Kottner, J., Audig, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., ... Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. <https://doi.org/10.1016/j.jclinepi.2010.03.002>
- Lecun, Y., Bengio, Y., & Hinton, G. (2015, May 27). Deep learning. *Nature*. Nature Publishing Group. <https://doi.org/10.1038/nature14539>
- Li, H., Zhu, Y., Burnside, E. S., Drukker, K., Hoadley, K. A., Fan, C., ... Giger, M. L. (2016, November). MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, oncotype DX, and PAM50 gene assays. *Radiology*. Radiological Society of North America Inc. <https://doi.org/10.1148/radiol.2016152110>
- Little, A. G., Rusch, V. W., Bonner, J. A., Gaspar, L. E., Green, M. R., Webb, W. R., ... Reed, C. E. (2005a). Patterns of surgical care of lung cancer patients. *Annals of Thoracic Surgery*, 80(6), 2051–2056. <https://doi.org/10.1016/j.athoracsur.2005.06.071>
- Little, A. G., Rusch, V. W., Bonner, J. A., Gaspar, L. E., Green, M. R., Webb, W. R., ... Reed, C. E. (2005b). Patterns of surgical care of lung cancer patients. *Annals of Thoracic Surgery*, 80(6), 2051–2056. <https://doi.org/10.1016/j.athoracsur.2005.06.071>
- Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., ... Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6), e271–e297. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2)
- Liu, Y., Kim, J., Balagurunathan, Y., Hawkins, S., Stringfield, O., Schabath, M. B., ... Gillies, R. J. (2018). Prediction of pathological nodal involvement by CT-based Radiomic features of the primary tumor in patients with clinically node-negative peripheral lung adenocarcinomas. *Medical Physics*, 45(6), 2518–2526. <https://doi.org/10.1002/mp.12901>
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... Whitlock, E. (2016). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Revista Espanola de Nutricion Humana y Dietetica*, 20(2), 148–160. <https://doi.org/10.1186/2046-4053-4-1>
- Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., ... Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*, 162(1), W1–W73. <https://doi.org/10.7326/M14-0698>
- Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer diagnosis using deep learning: A bibliographic review. *Cancers*, 11(9). <https://doi.org/10.3390/cancers11091235>
- Mutasa, S., Sun, S., & Ha, R. (2020). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical Imaging*, 65, 96–99. <https://doi.org/10.1016/j.clinimag.2020.04.025>
- Na, S., Ko, Y., Park, S., Beck, K., Choi, J., & J. Hong. (2018). Annual Congress of the

- European Association of Nuclear Medicine October 13 - 17, 2018 Düsseldorf, Germany. In *European journal of nuclear medicine and molecular imaging* (Vol. 45, pp. 1–844). NLM (Medline). <https://doi.org/10.1007/s00259-018-4148-3>
- Navani, N., Brown, J. M., Nankivell, M., Woolhouse, I., Harrison, R. N., Jeebun, V., ... Janes, S. M. (2012). Suitability of endobronchial ultrasound-guided transbronchial needle aspiration specimens for subtyping and genotyping of non-small cell lung cancer: A multicenter study of 774 patients. *American Journal of Respiratory and Critical Care Medicine*, 185(12), 1316–1322. <https://doi.org/10.1164/rccm.201202-0294OC>
- Navani, N., Nankivell, M., Lawrence, D. R., Lock, S., Makker, H., Baldwin, D. R., ... Janes, S. M. (2015). Lung cancer diagnosis and staging with endobronchial ultrasound-guided transbronchial needle aspiration compared with conventional approaches: An open-label, pragmatic, randomised controlled trial. *The Lancet Respiratory Medicine*, 3(4), 282–289. [https://doi.org/10.1016/S2213-2600\(15\)00029-6](https://doi.org/10.1016/S2213-2600(15)00029-6)
- NICE. (2019). *Lung cancer: diagnosis and management NICE guideline*. Retrieved from [www.nice.org.uk/guidance/ng122](http://www.nice.org.uk/guidance/ng122)
- O’Sullivan, J. W., Banerjee, A., Heneghan, C., & Pluddemann, A. (2018). Verification bias. *Evidence-Based Medicine*, 23(2), 54–55. <https://doi.org/10.1136/bmjebm-2018-110919>
- Ortakoylu, M. G., Iliaz, S., Bahadir, A., Aslan, A., Iliaz, R., Ozgul, M. A., & Urer, H. N. (2015a). Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *Jornal Brasileiro de Pneumologia*, 41(5), 410–414. <https://doi.org/10.1590/S1806-37132015000004493>
- Ortakoylu, M. G., Iliaz, S., Bahadir, A., Aslan, A., Iliaz, R., Ozgul, M. A., & Urer, H. N. (2015b). Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *Jornal Brasileiro de Pneumologia*, 41(5), 410–414. <https://doi.org/10.1590/S1806-37132015000004493>
- Owens, C. A., Peterson, C. B., Tang, C., Koay, E. J., Yu, W., Mackin, D. S., ... Yang, J. (2018). Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer. *PLoS ONE*, 13(10). <https://doi.org/10.1371/journal.pone.0205003>
- Parekh, V. S., & Jacobs, M. A. (2019, March 4). Deep learning and radiomics in precision medicine. *Expert Review of Precision Medicine and Drug Development*. Taylor and Francis Ltd. <https://doi.org/10.1080/23808993.2019.1585805>
- Pham, T. D. (2018). Complementary features for radiomic analysis of malignant and benign mediastinal lymph nodes. In *Proceedings - International Conference on Image Processing, ICIP* (Vol. 2017-September, pp. 3849–3853). IEEE Computer Society. <https://doi.org/10.1109/ICIP.2017.8297003>
- Pham, T. D., Watanabe, Y., Higuchi, M., & Suzuki, H. (2017). Texture Analysis and Synthesis of Malignant and Benign Mediastinal Lymph Nodes in Patients with Lung Cancer on Computed Tomography OPEN. <https://doi.org/10.1038/srep43209>
- Rena, O. (2016). The “N”-factor in non-small cell lung cancer: Staging system and institutional reports. *Journal of Thoracic Disease*. AME Publishing Company.

- <https://doi.org/10.21037/jtd.2016.11.37>
- Rodriguez-Roisin, R. (2000). Toward a consensus definition for COPD exacerbations. *Chest*, 117(5 Suppl 2), 398S-401S. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10843984>
- Sampsonas, F., Kakoullis, L., Lykouras, D., Karkoulias, K., & Spiropoulos, K. (2018). EBUS: Faster, cheaper and most effective in lung cancer staging. *International Journal of Clinical Practice*, 72(2), e13053. <https://doi.org/10.1111/ijcp.13053>
- Schmidt-Hansen, M., Baldwin, D. R., Hasler, E., Zamora, J., Abaira, V., & Roquéí Figuls, M. (2014, November 13). PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer. *Cochrane Database of Systematic Reviews*. John Wiley and Sons Ltd. <https://doi.org/10.1002/14651858.CD009519.pub2>
- Schmidt, R. L., & Factor, R. E. (2013a). Understanding Sources of Bias in Diagnostic Accuracy Studies. *Arch Pathol Lab Med*, 137, 558–565. <https://doi.org/10.5858/arpa.2012-0198-RA>
- Schmidt, R. L., & Factor, R. E. (2013b). Understanding Sources of Bias in Diagnostic Accuracy Studies. *Archives of Pathology & Laboratory Medicine*, 137(4), 558–565. <https://doi.org/10.5858/arpa.2012-0198-RA>
- Shafiek, H., Fiorentino, F., Peralta, A. D., Serra, E., Esteban, B., Martinez, R., ... Cosío, B. G. (2014). Real-Time Prediction of Mediastinal Lymph Node Malignancy by Endobronchial Ultrasound. *Archivos de Bronconeumologia*, 50(6), 228–234. <https://doi.org/10.1016/j.arbr.2014.05.003>
- Silvestri, G. A., Gonzalez, A. V., Jantz, M. A., Margolis, M. L., Gould, M. K., Tanoue, L. T., ... Detterbeck, F. C. (2013). Methods for staging non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5 SUPPL). <https://doi.org/10.1378/chest.12-2355>
- Smith, L., Bryan, S., De, P., Rahal, R., Shaw, A., Turner, D., ... Dixon, M. (2018). Members of the Canadian Cancer Statistics Advisory Committee Project management.
- Smith, L., Cancer Society, C., John, S., Ryan Woods, L., Brenner, D., Bryan, S., ... Dixon, M. (2019). Members of the Canadian Cancer Statistics Advisory Committee Analytic leads.
- Song, Y., Cai, W., Eberl, S., Fulham, M. J., & Feng, D. (2011). Discriminative pathological context detection in thoracic images based on multi-level inference. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6893 LNCS, pp. 191–198). [https://doi.org/10.1007/978-3-642-23626-6\\_24](https://doi.org/10.1007/978-3-642-23626-6_24)
- Song, Y., Cai, W., Kim, J., & Feng, D. D. (2012). A Multistage Discriminative Model for Tumor and Lymph Node Detection in Thoracic Images. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 31(5), 1061. <https://doi.org/10.1109/TMI.2012.2185057>
- Steyerberg, E. W., & Harrell, F. E. (2016, January 1). Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*. Elsevier USA. <https://doi.org/10.1016/j.jclinepi.2015.04.005>

- Tagaya, R., Kurimoto, N., Osada, H., & Kobayashi, A. (n.d.). Automatic Objective Diagnosis of Lymph Nodal Disease by B-Mode Images From Convex-Type Echobronchoscopy\*. *CHEST*, 133, 137–142. <https://doi.org/10.1378/chest.07-1497>
- Tagaya, R., Kurimoto, N., Osada, H., & Kobayashi, A. (2008). Automatic objective diagnosis of lymph nodal disease by B-mode images from convex-type echobronchoscopy. *Chest*, 133(1), 137–142. <https://doi.org/10.1378/chest.07-1497>
- Teoh, E. J., McGowan, D. R., Bradley, K. M., Belcher, E., Black, E., Moore, A., ... Gleeson, F. V. (2016). 18F-FDG PET/CT assessment of histopathologically confirmed mediastinal lymph nodes in non-small cell lung cancer using a penalised likelihood reconstruction. *European Radiology*, 26(11), 4098–4106. <https://doi.org/10.1007/s00330-016-4253-2>
- Toney, L. K., & Vesselle, H. J. (2014). Neural networks for nodal staging of non-small cell lung cancer with FDG PET and CT: Importance of combining uptake values and sizes of nodes and primary tumor. *Radiology*, 270(1), 91–98. <https://doi.org/10.1148/radiol.13122427>
- Topol, E. J. (2019, January 1). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. Nature Publishing Group. <https://doi.org/10.1038/s41591-018-0300-7>
- Traverso, A., Wee, L., Dekker, A., & Gillies, R. (2018). Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology Biology Physics*, 102(4), 1143–1158. <https://doi.org/10.1016/j.ijrobp.2018.05.053>
- Tugwell, P., & Knottnerus, J. A. (2015, March 1). Transferability/generalizability deserves more attention in “retest” studies in Diagnosis and Prognosis. *Journal of Clinical Epidemiology*. Elsevier USA. <https://doi.org/10.1016/j.jclinepi.2015.01.007>
- Usher-Smith, J. A., Sharp, S. J., & Griffin, S. J. (2016). The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ (Online)*, 353. <https://doi.org/10.1136/bmj.i3139>
- Vesselle, H., Turcotte, E., Wiens, L., & Haynor, D. (2003a). *Application of a Neural Network to Improve Nodal Staging Accuracy with 18 F-FDG PET in Non-Small Cell Lung Cancer*. *J Nucl Med* (Vol. 44). Retrieved from [www.gnu.org](http://www.gnu.org)
- Vesselle, H., Turcotte, E., Wiens, L., & Haynor, D. (2003b). Application of a neural network to improve nodal staging accuracy with 18F-FDG PET in non-small cell lung cancer. *Journal of Nuclear Medicine*, 44(12), 1918–1926.
- Wahidi, M. M., Herth, F., Yasufuku, K., Shepherd, R. W., Yarmus, L., Chawla, M., ... Feller-Kopman, D. J. (2016). Technical aspects of endobronchial ultrasound-guided transbronchial needle aspiration CHEST guideline and expert panel report. *Chest*, 149(3), 816–835. <https://doi.org/10.1378/chest.15-1216>
- Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1), 101–110. [https://doi.org/10.1002/\(SICI\)1097-0258\(19980115\)17:1<101::AID-SIM727>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E)
- Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., ... Yu, L. (n.d.). Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-

- small cell lung cancer from 18 F-FDG PET/CT images.  
<https://doi.org/10.1186/s13550-017-0260-9>
- Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., ... Yu, L. (2017). Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. *EJNMMI Research*, 7(1), 11. <https://doi.org/10.1186/s13550-017-0260-9>
- Wang, X., Nan, W., Yan, S., Li, Q., Guo, N., & Guo, Z. (2018). *MA05.11 Radiomics Analysis Using SVM Predicts Mediastinal Lymph Nodes Status of Squamous Cell Lung Cancer by Pre-Treatment Chest CT Scan*. *Journal of Thoracic Oncology* (Vol. 13). <https://doi.org/10.1016/j.jtho.2018.08.357>
- Warfield, S. K., Zou, K. H., & Wells, W. M. (2008). Validation of image segmentation by estimating rater bias and variance. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1874), 2361–2375. <https://doi.org/10.1098/rsta.2008.0040>
- Wnuk, P., Kowalewski, M., Małkowski, B., Bella, M., Dancewicz, M., Szczesny, T., ... Kowalewski, J. (2014). PET-CT derived Artificial Neural Network can predict mediastinal lymph nodes metastases in Non-Small Cell Lung Cancer patients. Preliminary report and scoring model. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging : Official Publication of the Italian Association of Nuclear Medicine (AIMN) [and] the International Association of Radiopharmacology (IAR), [and] Section of the Society Of...* Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/25289632>
- Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018, August 1). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*. Springer Verlag. <https://doi.org/10.1007/s13244-018-0639-9>
- Zhao, W., & Shi, F. (2018). A Method for Lymph Node Segmentation with Scaling Features in a Random Forest Model. *Current Proteomics*, 15(2), 128–134. <https://doi.org/10.2174/1570164614666171030161753>
- Zhong, Y., Yuan, M., Zhang, T., Zhang, Y. D., Li, H., & Yu, T. F. (2018). Radiomics approach to prediction of occult mediastinal lymph node metastasis of lung adenocarcinoma. *American Journal of Roentgenology*, 211(1), 109–113. <https://doi.org/10.2214/AJR.17.19074>
- Zhu, J., Xu, W.-G., Xiao, H., & Zhou, Y. (2019). [Application of a Radiomics Model for Preding Lymph Node Metastasis in Non-small Cell Lung Cancer]. *Sichuan Da Xue Xue Bao. Yi Xue Ban = Journal of Sichuan University. Medical Science Edition*, 50(3), 373–378. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/31631606>
- Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine*, 31(29), 3972–3981. <https://doi.org/10.1002/sim.5466>

## APPENDICES

### Appendix 1. Search Strategy

#### Databases

#### EMBASE (OvidSP Interface)

#### 1974 to 2020 Week 3

#	Search	Results
1	exp machine learning/	185488
2	"neural networks (computer)".mp.	165
3	supervised machine learning/ or unsupervised machine learning/	1774
4	(naive bayes or random forest or boosting or deep learning or machine intelligence or computational intelligence or computer reasoning or convolutional neural network or residual network or variational auto encoder or principal components analysis or k nearest neighbours or linear discriminant analysis or genetic algorithm or regression analysis or LASSO regression or ridge regression or decision tree).ti,ab,kw.	364787
5	(radiomics or learning algorithm or coding algorithm or computer heuristics).ti,ab,kw.	8483
6	(comput* language or comput* prediction or comput* simulation or comput* aided diagnosis).ti,ab,kw.	22320
7	exp algorithms/	292045
8	1 or 2 or 3 or 4 or 5 or 6 or 7	775569
9	(lung? adj3 (cancer* or neoplasm? or tumo?r* or malignan*)).ti,ab,kw.	280813
10	exp respiratory tract neoplasms/ or lung neoplasms/ or "adenocarcinoma of lung"/ or bronchial neoplasms/ or carcinoma, non-small-cell lung/	448973
11	(lung? adj3 (lesion* or mass* or neoplas*)).ti,ab,kw.	27931
12	((mediastin* or thoracic or chest or hilar or interlobar or lobar or segmental or subsegmental) adj2 lymph nod*).ti,ab,kw.	13403
13	9 or 10 or 11 or 12	535550

15	exp Positron-Emission Tomography/	163271
16	image interpretation, computer-assisted/ or exp tomography computed/	38650
17	Endoscopic Ultrasound-Guided Fine Needle Aspiration/	3139
18	((Endobronchial or Endoscopic) adj2 Ultrasound Guided Needle Aspiration).ti,ab,kw.	68
19	((CT or PET or CAT) adj2 scan).ti,ab,kw.	117395
20	((PET or CT or CAT) adj2 image).ti,ab,kw.	12198
21	(electron adj3 tomography).mp.	6984
22	exp tomography emission-computed/	197183
23	14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22	352296
24	8 and 13	15179
25	22 and 23	2022
26	Animal/ not (human/ and animal/)	1065416
27	25 not 26	1972

**MEDLINE (Ovid MEDLINE® and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily)**

**1946 to January 14, 2020**

#	Searches	Results
1	exp Machine Learning/	17590
2	"neural networks (computer)".mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	25284

3	supervised machine learning/ or unsupervised machine learning/	841
4	(naive bayes or random forest or boosting or deep learning or machine intelligence or computational intelligence or computer reasoning or convolutional neural network or residual network or variational auto encoder or principal components analysis or k nearest neighbours or linear discriminant analysis or genetic algorithm or regression analysis or LASSO regression or ridge regression or decision tree).ti,ab,kw,kf.	256113
5	(radiomics or learning algorithm or coding algorithm or computer heuristics).ti,ab,kw,kf.	5947
6	(comput* language or comput* prediction or comput* simulation or comput* aided diagnosis).ti,ab,kw,kf.	18701
7	exp algorithms/	317189
8	1 or 2 or 3 or 4 or 5 or 6 or 7	575635
9	(lung? adj3 (cancer* or neoplasm? or tumo?r* or malignan*)).ti,ab,kw,kf.	191099
10	exp respiratory tract neoplasms/ or lung neoplasms/ or "adenocarcinoma of lung"/ or bronchial neoplasms/ or carcinoma, non-small-cell lung/	286706
11	(lung? adj3 (lesion* or mass* or neoplas*)).ti,ab,kw,kf.	24018
12	((mediastin* or thoracic or chest or hilar or interlobar or lobar or segmental or subsegmental) adj2 lymph nod*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	8518
13	"Adenocarcinoma of Lung"/pa [Pathology]	858
14	Lung Neoplasms/pa [Pathology]	73517
15	Mediastinum/pa [Pathology]	1775
16	9 or 10 or 11 or 12 or 13 or 14 or 15	361661



17	diagnostic imaging/ or image interpretation, computer-assisted/ or radiography/ or radionuclide imaging/ or tomography/ or ultrasonography/ or whole body imaging/	655544
17	exp Positron-Emission Tomography/	59746
19	image interpretation, computer-assisted/ or exp tomography computed/	44684
20	((CT or PET or CAT) adj2 scan).ti,ab,kw,kf.	60494
21	((PET or CT or CAT) adj2 image).ti,ab,kw,kf.	6521
22	(electron adj3 tomography).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	4756
23	exp tomography emission-computed/	111028
24	((Endobronchial or Endoscopic) adj2 Ultrasound Guided Needle Aspiration).ti,ab,kw,kf.	36
25	17 or 18 or 19 or 20 or 21 or 22 or 23 or 24	814692
26	8 and 16	8426
27	25 and 26	1137
28	Animal/ not (human/ and animal/)	4663765
29	27 not 28	1083

## CENTRAL

### Inception to 2020, Week 3

#1 (supervised OR unsupervised) “machine learning” OR naive bayes OR random forest  
OR deep learning OR comput\* (intelligence OR reasoning OR heuristic\* OR language OR

prediction OR simulation OR aided diagnos\*) OR boosting OR radiomics OR “algorithm”  
OR convolutional neural network OR residual network OR variational auto encoder in Title  
Abstract Keyword AND (lung (cancer\* OR (cancer diagnos\*) OR neoplas\* OR lesion\*  
OR mass\* OR tumo?r OR malignan\*)) OR “non small cell lung cancer” OR ((mediastin\*  
OR chest OR thoracic OR interlobar OR lobar OR segmental OR subsegmental) lymph  
node\*) OR medistin\* in Title Abstract Keyword AND “diagnostic imag\*” OR “computer  
assisted” OR ultrasonograph\* OR (positron emission OR comput\*) tomograph\* (imag\*  
OR scan\*) OR (CT OR PET OR CAT) (imag\* OR scan\*) OR ((endobronchial OR  
endoscopic ultrasound) in Title Abstract Keyword (815)

### **Web of Science (All Databases)**

#### **1926 to January 14, 2020**

#1 TS= (supervised machine learning OR unsupervised machine learning) (34,815)

#2 TS= (naive bayes OR random forest OR boosting OR deep learning OR machine  
intelligence OR comput\* intelligence OR comput\* reasoning OR support vector machine  
OR “neural networks (computer)” OR radiomics) (932,786)

#3 (#2 OR #3) (935,866)

#4 TS=(lung (lesion\* OR mass\*)) (158,928)

#5 TS= (respiratory tract neoplasm\* OR lung neoplasm\* OR “adenocarcinoma of the lung”  
OR bronchial neoplasm\* OR carcinoma, non-small cell lung cancer) (540,032)

#6 TS= (lung (lesion\* OR mass\*)) (158,928)

#7 TS=((mediastin\* OR thoracic OR chest OR hilar OR interlobar OR lobar OR segmental OR subsegmental) lymph nod\*)) (36,721)

#8 TS= (mediastin\*[pathology]) (42,384)

#9 (#8 OR #7 OR #6 OR #5 OR #4) (689,340)

#10 TS= ((Positron Emission OR Compute\*) Tomography) (983,448)

#11 TS= ((Endobronchial OR Endoscopic) Ultrasound Guided Needle Aspiration)) (7,371)

#12 TS= ((CT OR PET OR CAT) scan OR imag\*) (568,190)

#13 (#12 OR #11 OR #10) (1,159,472)

#14 (#13 AND #9 AND 3#) (2,063)

#15 TS= (animal\* NOT (human\* AND animal\*)) (12,390,410)

#16 (#14 NOT #15) (2,047)

### *Ongoing Studies*

#### **Clinicaltrials.gov**

#### **Inception to January 14, 2020**

#1 (deep learning OR computer OR neural OR algorithm OR radiomics) AND (EBUS or endobronchial ultrasound OR CT OR PET OR CAT OR tomography OR diagnostic imaging OR lymph nodes) AND (lung cancer) (96)

**Appendix 2. QUADAS-2 Guideline for Risk of Bias Assessment**

<i>Domain</i>	<i>Yes</i>	<i>Unclear</i>	<i>No</i>
<b><i>Patient selection</i></b>			
<b><i>1. Consecutive or random sample enrolled?</i></b>	<i>A consecutive or random sample of patients was enrolled in the study.</i>	<i>It is unclear whether a consecutive or random sample of patients was enrolled in the study.</i>	<i>There was no consecutive or random sample included in the study (e.g. only patients already suspected or confirmed lung malignancy with/ without lymphadenopathy).</i>
<b><i>2. Case control design avoided?</i></b>	<i>There was no case control design.</i>	<i>It is unclear if there was a case control design</i>	<i>There was a case control design</i>
<b><i>3. Inappropriate exclusions avoided?</i></b>	<i>There are no patients inappropriately excluded (e.g. patients with confirmed lung cancer, who will already be operated on)</i>	<i>It is unclear if there was avoidance of inappropriate exclusions</i>	<i>There is inappropriate exclusion of patients (e.g. exclusion of patients with high risk of malignancy)</i>
<b><i>Index test</i></b>			
<b><i>1. Index test results interpreted without knowledge results reference standard?</i></b>	<i>The index test did not have any clinicopathological information added to the algorithm.</i>	<i>It is unclear whether the index test had knowledge of the reference standard.</i>	<i>The index test had clinicopathological information added to the algorithm.</i>

<b>2. Pre-specified threshold?</b>	<i>There was a pre-specified cut-off level.</i>	<i>It is unclear if there was a pre-specified cut-off level</i>	<i>There was no pre-specified cut-off level.</i>
<b>Reference standard</b>			
<b>1. Reference standard likely to correctly classify target condition?</b>	<i>In patients receiving surgery there is adequate histopathological examination of lymph node tissue.</i>	<i>In patients receiving surgery it's unclear how histopathological examination is performed.</i>	<i>In patients receiving surgery histopathological examination is not adequate.</i>
<b>2. Reference standard results interpreted without knowledge results index test?</b>	<i>The outcome assessor of histopathological and follow-up results was not aware of results.</i>	<i>It is not clear if the outcome assessor of histopathological and follow-up results was aware of results.</i>	<i>The outcome assessor of histopathological and follow-up results was aware of results.</i>
<b>Flow and timing</b>			
<b>1. Appropriate interval between index test and reference standard?</b>	<i>Time between radiomic testing and histopathological examination is &lt; 3 months</i>	<i>It is unclear what the time period between reference standard and index test is.</i>	<i>Time between radiomic analysis testing and histopathological exceeds 3 months.</i>
<b>2. All patients received the reference standard?</b>	<i>All patients received surgery, and patients who did not receive surgery have clinical follow-up with oncology or CT surveillance.</i>	<i>It is not clear if the whole sample did receive surgery or follow-up with oncology or CT surveillance.</i>	<i>Only a selected subset of the patients received or surgery or not all patients have clinical follow-up with oncology or CT surveillance.</i>

<p><b>3. Patients received the same reference standard?</b></p>	<p><i>All patients were operated, and histopathological examination of the lymph nodes was performed.</i></p>	<p><i>It is not clear if all patients were operated on and received histopathological examination.</i></p>	<p><i>Not all patients were operated, or histopathological examination was not performed in all patients.</i></p>
<p><b>4. All patients included in the analysis?</b></p>	<p><i>All patients enrolled were included in the analysis</i></p>	<p><i>It is not clear if all patients were included in the analysis.</i></p>	<p><i>Not all patients enrolled were included in the analysis (e.g. patients lost to follow-up)</i></p>

**Appendix 3. Characteristics of Included Studies and Risk of Bias Assessment [ordered by study ID]**

Author	Bella 2013	
Country	Poland	
Participants	Patient population: n = 150 (derivation), n = 26 (validation) Diagnosis of Interest: NSCLC Prevalence: NR Mean Age: NR % Male: NR LNs imaged/ biopsied: n = 467 for derivation, n = 80 for validation Inclusion Criteria: NSCLC patients treated at the study's thoracic surgery department	
Study Design	Retrospective Derivation Cohort & Prospective Validation Cohort	
Imaging Modality	PET/CT	
Index Test	Artificial Neural Networks	
Reference Test	EBUS-TBNA, mediastinoscopy or lymphadenectomy	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>

<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design, but it did not describe its exclusion criteria nor was it clear whether a consecutive sampling was employed for the prospective validation.
<i>Index Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Reference Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Flow and Timing</i>	High Risk	The interval between index test and reference tests was not specified. Given that EBUS-TBNA, mediastinoscopy and surgical resection were used as possible reference tests, the timing intervals could have varied amongst patients because EBUS-TBNA often occurs before mediastinoscopy and surgery.

Author	Ferreira-Junior 2009
Country	Brazil
Participants	<p>Patient population: n = 85                  Diagnosis of Interest: Lung Cancer                  Prevalence: n = 39 patients (46%)                  Mean Age: 67 years                  % Male: 54                  LNs imaged/ biopsied: n = NR                  Inclusion Criteria: Patients with histologically or surgically confirmed lung cancer</p>
Study Design	Retrospective Cohort
Imaging Modality	CT



Index Test	Volumetric Segmentation	
Reference Test	Pathology from biopsy or surgical resection	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design, but it did not describe its exclusion criteria nor was it clear whether a consecutive sampling was employed for the prospective validation.
<i>Index Test</i>	Unclear Risk	The study did not specify whether the trained reader was blinded to the reference tests results while segmenting.
<i>Reference Test</i>	Low Risk	The reference tests were likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	High Risk	The study excluded patients with missing clinical data, and the interval between index test and reference test was not specified. Given that biopsy and surgical resection were both used as possible reference tests, the timing intervals could have varied amongst patients because biopsies are often done before surgery.
Author	Flechsigg 2017	
Country	Germany	

Participants	Patient population: n = 122 Diagnosis of Interest: Lung Cancer Prevalence of Malignancy: n = 73 patients (60%), n = 130 LNs (52%) Median Age: 59 years % Male: 56 LNs imaged/ biopsied: n = 248 Inclusion Criteria: Patients with histologically or surgically confirmed lung cancer whom did not receive neoadjuvant radiation and/or chemotherapy	
Study Design	Retrospective Cohort	
Imaging Modality	PET/CT	
Index Test	Volumetric CT Histogram Analysis with Semi-Automated Segmentation	
Reference Test	Histological pathology via surgical resection with mediastinal lymph node, mediastinoscopy or transbronchial biopsy	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	While the study avoided inappropriate exclusions, consecutive patient enrollment and exclusion criteria were not provided by the authors.
<i>Index Test</i>	Low Risk	Radiologist was blinded to clinical information.
<i>Reference Test</i>	Low Risk	The reference tests were likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.

<i>Flow and Timing</i>	Low Risk	All patients had their imaging conducted within one week of the reference test, which was appropriate. Although multiple ways to obtain histological confirmation were used, all were effective ways to obtain the pathological ground truth.
Author	Gao 2015	
Country	China	
Participants	Patient population: n = 132 Diagnosis of Interest: NSCLC Prevalence: NR Median Age: 61 years % Male: 61 LNs imaged/ biopsied: n = 768 Inclusion Criteria: Patients diagnosed with lung cancer that underwent lobectomy combined with nodal dissection. Did not receive any therapy for their tumour beforehand.	
Study Design	Prospective Cohort	
Imaging Modality	PET/CT	
Index Test	Support Vector Machine	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		

<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Low Risk	Patients were consecutively enrolled and it was not a case-control design. Exclusion criteria was justifiable.
<i>Index Test</i>	Low Risk	Radiologists were blinded to clinical information.
<i>Reference Test</i>	Unclear Risk	Although the reference test is likely to classify the target condition correctly, it is unclear whether the pathologists were blinded to the index test results.
<i>Flow and Timing</i>	Low Risk	All patients had their imaging conducted within one week of the reference test, which was appropriate. They also received the same reference test.

Author	He 2019
Country	China
Participants	Patient population: n = 717 Diagnosis of Interest: NSCLC Prevalence: n = 325 patients (45%) Median Age: 61 years % Male: 59 LNs imaged/ biopsied: NR Inclusion Criteria: Patients that underwent surgical resection with systematic mediastinal lymphadenectomy for primary NSCLC.
Study Design	Retrospective Cohort
Imaging Modality	CT

Index Test	Radiomic-based Predictive Risk Score	
Reference Test	Surgical Resection (Pathological Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	Information on consecutive patient enrollment was not provided by the authors. Moreover, the study excluded patients for unjustified reasons (i.e. patients that were harder to diagnose).
<i>Index Test</i>	Unclear Risk	Although the threshold was pre-specified, there is insufficient information to determine if the radiologists had access to reference test results during segmentation.
<i>Reference Test</i>	Low Risk	The reference tests were likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	Low Risk	Although not all patients were included within analysis due to missing data, this population comprised less than 20% of patients. Remaining patients had their imaging conducted within two weeks of the reference test, which was appropriate. They also received the same reference test.
Author	Inoue 2011	
Country	Japan	

Participants	Patient population: n = 14 Diagnosis of Interest: Lung Cancer Prevalence: n = 100% Mean Age: 65.85 % Male: 86 LNs imaged/ biopsied: n = 23 Inclusion Criteria: Patients that received PET/CT covered by medical insurance, were diagnosed with mediastinal and/or hilum LN metastases and had blood sugar below 150 ml/dl	
Study Design	Retrospective Cohort	
Imaging Modality	3D - Ordered Subset Expectation Maximization (3D-OSEM)	
Index Test	PET/CT with 2D-OSEM and PET/CT with FORE + OSEM	
Reference Test	Histologically confirmed or clinical observation for over a year	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	High Risk	The study did not justify why they excluded 79 patients.
<i>Index Test</i>	Unclear Risk	Insufficient information to determine if the radiologists had access to reference test results while scoring.
<i>Reference Test</i>	High Risk	Although the reference tests assessors did not have knowledge of the index test results, two drastically

		different reference tests were used. Specifically, clinical observation for more than a year is subjective compared to histological confirmation.
<i>Flow and Timing</i>	High Risk	The study excluded patients with missing clinical data; the interval between index test and reference test was not consistent; and patients could have received two completely different reference tests.
Author	Liu 2018	
Country	United States	
Participants	<p>Patient population: n = 187                  Diagnosis of Interest: Peripheral Lung Adenocarcinoma                  Prevalence: n = 34 patients (18.2%)                  Median Age: 59 years                  % Male: 41                  LNs imaged/ biopsied: n = NR                  Inclusion Criteria: underwent lobectomy or pneumonectomy with systematic lymph node dissection of both hilar and medi-astinal lymph nodes; acquisition of preoperative thin-section CT scan and the location of the lung tumor was peripheral (tumor involving subsegmental bronchus or smaller airway);                  (c) clinical N stage was N0.</p>	
Study Design	Retrospective Cohort	
Imaging Modality	CT	
Index Test	Cognition Network Technology	

Reference Test	Surgical Resection (Pathological Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	High Risk	Patients were not consecutively or randomly selected (already confirmed to have peripheral lung adenocarcinoma).
<i>Index Test</i>	Low Risk	Radiologists were blinded to clinical information.
<i>Reference Test</i>	Low Risk	The reference tests were likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	Low Risk	All patients had their imaging conducted within one month of the reference test, which was appropriate. They also received the same reference test.
Author	Na 2018	
Country	South Korea	
Participants	Patient population: n = 468 Diagnosis of Interest: NSCLC Prevalence: n = 157 patients (34%) Mean Age: NR % Male: NR LNs imaged/ biopsied: n = NR Inclusion Criteria: Patients with NSCLC smaller than 3 cm (T1 stage)	



Study Design	Prospective Cohort	
Imaging Modality	PET/CT	
Index Test	Convolutional Neural Network with XGBoost Classifier	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design, but it did not describe its exclusion criteria nor was it clear whether a consecutive sampling was employed.
<i>Index Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Reference Test</i>	Unclear Risk	Although the reference test is likely to classify the target condition correctly, it is unclear whether the pathologists were blinded to the index test results.
<i>Flow and Timing</i>	Unclear Risk	Patients received the same reference test, however, the time between index test and reference test was not specified.
Author	Pham I 2017	

Country	Japan	
Participants	Patient population: n = 148 Diagnosis of Interest: Lung Cancer Prevalence: n = NR Mean Age: NR % Male: 63 LNs imaged/ biopsied: n = NR Inclusion Criteria: Patients with biopsy-proven primary lung malignancy, pathological mediastinal nodal staging, and underwent unenhanced CT	
Study Design	Retrospective Cohort	
Imaging Modality	CT	
Index Test	Unsupervised Neural Network with Gray-Level Co-Occurrence Matrix and Semi-Variogram Features	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	While the study avoided inappropriate exclusions, consecutive patient enrollment and exclusion criteria were not provided by the authors.
<i>Index Test</i>	Low Risk	Index test utilized unsupervised learning, thus minimizing human influence.

<i>Reference Test</i>	Low Risk	The reference test was likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	Unclear Risk	Patients received the same reference test, however, the time between index test and reference test was not specified.
<b>Risk of Bias</b>		
Author	Pham II 2017	
Country	Japan	
Participants	Patient population: n = 148 Diagnosis of Interest: Lung Cancer Prevalence: n = 133 LNs (49.1%) Mean Age: 69.41 years % Male: 63 LNs imaged/ biopsied: n = 271 Inclusion Criteria: Patients with biopsy-proven primary lung malignancy, pathological mediastinal nodal staging, and underwent unenhanced CT	
Study Design	Retrospective Cohort	
Imaging Modality	CT	
Index Test	Support Vector Machine with Gray-Level Co-Occurrence Matrix and Semi-Variogram Features	
Reference Test	Biopsy-Proven (Pathology Results)	
<b>Risk of Bias</b>		

<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	While the study avoided inappropriate exclusions, consecutive patient enrollment and exclusion criteria were not provided by the authors.
<i>Index Test</i>	Low Risk	Thoracic surgeon was blinded to clinical information during segmentation.
<i>Reference Test</i>	Low Risk	The reference test was likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	Low Risk	All patients had their imaging conducted within three months of the reference test, which was appropriate. They also received the same reference test.

Author	Song 2011
Country	Australia
Participants	Patient population: n = 50 Diagnosis of Interest: NSCLC Prevalence: n = 23 patients (46.0%) Mean Age: NR % Male: NR LNs imaged/ biopsied: NR Inclusion Criteria: Patients diagnosed with NSCLC
Study Design	Not Reported

Imaging Modality	PET/CT	
Index Test	Support Vector Machine	
Reference Test	Expert Radiologist Identifying Region of Interest	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design, but it did not describe its exclusion criteria nor was it clear whether a consecutive or random sampling was employed.
<i>Index Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Reference Test</i>	High Risk	Reference test was subjective in its diagnosis of abnormal LNs. Did not specify what classified an abnormal LN nor did it rely on histological confirmation. Was based on the similarities of images.
<i>Flow and Timing</i>	Unclear Risk	Patients received the same reference test, however, the time between index test and reference test was not specified.
Author	Song 2012	
Country	Australia	

Participants	Patient population: n = 85 Diagnosis of Interest: NSCLC Prevalence: n = 23 patients (46%) Mean Age: NR % Male: NR LNs imaged/ biopsied: n = 72 Inclusion Criteria: Patients diagnosed with NSCLC	
Study Design	Retrospective Cohort	
Imaging Modality	PET/CT	
Index Test	Support Vector Machine	
Reference Test	Expert Radiologist Identifying Region of Interest	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	While the study avoided inappropriate exclusions, consecutive patient enrollment and exclusion criteria were not provided by the authors.
<i>Index Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Reference Test</i>	High Risk	Reference test was subjective in its diagnosis of abnormal LNs. Did not specify what classified an abnormal LN nor did it rely on histological confirmation. Was based on the similarities of images.

<i>Flow and Timing</i>	Unclear Risk	Patients received the same reference test, however, the time between index test and reference test was not specified.
Author	Tagaya 2008	
Country	Japan	
Participants	Patient population: n = 91 (n = 66 Lung Cancer) Diagnosis of Interest: Lung Cancer & Sarcoidosis Prevalence: n = 66 (73%) Mean Age: NR % Male: NR LNs imaged/ biopsied: n = 91 (n = 66 Lung Cancer) Inclusion Criteria: Patients undergoing EBUS-TBNA for lung cancer or sarcoidosis	
Study Design	Prospective Cohort	
Imaging Modality	EBUS	
Index Test	Supervised Layered Artificial Neural Networks	
Reference Test	Transbronchial Needle Aspiration (Pathology Results) or Cytological Examination	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>

<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design, but it did not describe its exclusion criteria nor was it clear whether a consecutive or random sampling was employed.
<i>Index Test</i>	Low Risk	Thoracic surgeons were blinded to clinical information during assessments.
<i>Reference Test</i>	Low Risk	Although cytology is not as accurate as the TBNA pathology results for diagnosis, the study specified that the cytology results were confirmed by surgery.
<i>Flow and Timing</i>	High Risk	The interval between index test and reference tests was not specified. Given that TBNA pathology results and cytology were both used as possible reference tests, the timing intervals likely varied amongst patients because cytology often occurs before histological assessment of TBNA samples.

Author	Teoh 2016
Country	England
Participants	<p>Patient population: n = 47                  Diagnosis of Interest: NSCLC                  Prevalence: n = 18 patients (38.3%)                  Mean Age: 69 years                  % Male: 61.7                  LNs imaged/ biopsied: n = 112                  Inclusion Criteria: Patients who underwent PET/CT for staging of NSCLC and had subsequent nodal station histopathological diagnosis</p>
Study Design	Retrospective Cohort



Imaging Modality	PET/CT with Ordered Subset Expectation Maximum Reconstruction	
Index Test	Bayesian Penalized Likelihood Reconstruction	
Reference Test	Surgical Resection or EBUS-TBNA (Pathological Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	While the study avoided inappropriate exclusions, consecutive patient enrollment and exclusion criteria were not provided by the authors.
<i>Index Test</i>	Low Risk	Although the radiologist was not blinded to the nature of the reconstruction, they were blinded to clinical information and a predefined SUV threshold of 2.5 was used to distinguish malignant and benign LNs.
<i>Reference Test</i>	Low Risk	The reference tests were likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	High Risk	The interval between index test and reference tests was not specified. Given that EBUS-TBNA and surgical resection were both used as possible reference tests, the timing intervals could have varied amongst patients because EBUS-TBNA often occurs before surgery.
Author	Toney 2014	

Country	United States	
Participants	Patient population: n = 133 Diagnosis of Interest: NSCLC Prevalence: n = 67 (50.4%) Mean Age: 64.4 years % Male: 64.7 LNs imaged/ biopsied: NR Inclusion Criteria: NSCLC patients with surgically proven nodal status. Could NOT have pleural implants or evidence of stage IV disease.	
Study Design	Prospective Cohort	
Imaging Modality	PET/CT	
Index Test	Supervised Artificial Neural Networks	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design and the exclusion criteria was justifiable, but it was not clear whether a consecutive or random sampling was employed.
<i>Index Test</i>	Low Risk	Expert reader was blinded to surgical pathology.
<i>Reference Test</i>	Low Risk	The reference test was likely to classify the target condition correctly. Moreover, nodal status was performed independent of index test.

<i>Flow and Timing</i>	Unclear Risk	Patients received the same reference test, however, the time between index test and reference test was not specified.
Author	Vesselle 2003	
Country	United States	
Participants	<p>Patient population: n = 133                  Diagnosis of Interest: NSCLC                  Prevalence: n = 67 (50.4%)                  Mean Age (years): N0) 67.7, N1) 67.4, N2) 63.7, N3) 56.0                  % Male: 64.66                  LNs imaged/ biopsied: NR                  Inclusion Criteria: Patients with potentially resectable NSCLC after chest CT and clinical evaluation. Could NOT have type 1 diabetes, stage IV disease, chemotherapy/radiotherapy prior to PET.</p>	
Study Design	Prospective Cohort	
Imaging Modality	PET	
Index Test	Supervised Artificial Neural Networks	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>

<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design and the exclusion criteria was justifiable, but it was not clear whether a consecutive or random sampling was employed.
<i>Index Test</i>	High Risk	Feedback of surgical nodal staging results was provided to the index test assessor.
<i>Reference Test</i>	High Risk	Results of index test were provided to the surgeon prior to confirming mediastinal nodal status.
<i>Flow and Timing</i>	Unclear Risk	Patients received the same reference test, however, the time between index test and reference test was not specified.

Author	Wang 2017
Country	China
Participants	<p>Patient population: n = 168                  Diagnosis of Interest: NSCLC                  Prevalence: n = 1270 LNs (90.91%)                  Median Age: 61 years                  % Male: 54.17                  LNs imaged/ biopsied: n = 1397                  Inclusion Criteria: Patients who had PET/CT within 1 week of surgery</p>
Study Design	Retrospective Cohort
Imaging Modality	PET/CT
Index Tests	<ol style="list-style-type: none"> <li>1. Random Forest</li> <li>2. Support Vector Machine</li> <li>3. Adaptive Boosting</li> <li>4. Back-Propagation Artificial Neural Network</li> </ol>

	5. Convolutional Neural Networks	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	While the study avoided inappropriate exclusions, consecutive patient enrollment and exclusion criteria were not provided by the authors.
<i>Index Test</i>	Unclear Risk	Insufficient information to determine if the radiologists had access to reference test results while segmenting.
<i>Reference Test</i>	Low Risk	The reference test was likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	Low Risk	All patients had their imaging conducted within 1 week of the reference test, which was appropriate. They also received the same reference test.
Author	Wang 2018	
Country	China	
Participants	Patient population: n = 93 Diagnosis of Interest: Squamous Cell Lung Cancer Prevalence: n = 31 with N2 (33.3%)	

	Mean Age: NR % Male: NR LNs imaged/ biopsied: NR Inclusion Criteria: patients with squamous cell lung cancer that underwent pretreatment CT scans	
Study Design	NR	
Imaging Modality	CT	
Index Test	Support Vector Machine	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>
<i>Patient Selection</i>	Unclear Risk	Study did not appear to have a case-control design, but it did not describe its exclusion criteria nor was it clear whether a consecutive sampling was employed for the prospective validation.
<i>Index Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Reference Test</i>	Unclear Risk	Insufficient information to make a judgement.
<i>Flow and Timing</i>	Unclear Risk	Insufficient information to make a judgement.

Author	Zhong 2018	
Country	China	
Participants	<p>Patient population: n = 492                  Diagnosis of Interest: Adenocarcinoma                  Prevalence: n = 78 (15.85%)                  Mean Age: 61.4 years                  % Male: 35.16                  LNs imaged/ biopsied: NR                  Inclusion Criteria: underwent surgical resection and systematic LN dissection (removal of at least three hilar stations and three mediastinal stations), had no enlargement of the hilar or mediastinal LNs, at CT had a clinical diagnosis of no LN metastasis (clinical N0), and had no distant metastasis (M0). Excluded if: IV administration of contrast material, unsatisfactory image quality due to respiratory artifact and surgical resection not performed within 90 days of CT</p>	
Study Design	Retrospective Cohort	
Imaging Modality	CT	
Index Test	Support Vector Machine	
Reference Test	Surgical Resection (Pathology Results)	
<b>Risk of Bias</b>		
<b>Domain</b>	<b>Author's Judgement</b>	<b>Support of Judgement</b>

<i>Patient Selection</i>	High Risk	Although the authors investigated occult disease, which justifies the exclusion of >N0 patients, the rationale for why they solely included these patients with adenocarcinoma was not convincingly explained.
<i>Index Test</i>	Unclear Risk	Insufficient information to determine if the radiologists had access to reference test results while segmenting.
<i>Reference Test</i>	Low Risk	The reference test was likely to classify the target condition correctly. Moreover, the study was retrospective, therefore, the reference test assessors did not have knowledge of the index test results.
<i>Flow and Timing</i>	Low Risk	All patients had their imaging conducted within 90 days of the reference test, which was appropriate. They also received the same reference test.



#### **Appendix 4. Ongoing and Awaiting Classification Studies**

##### *Ongoing*

1. NCT03849040
2. NCT03648151
3. NCT04000620

##### *Awaiting Classification*

1. Genseke 2019
2. He 2017
3. Wnuck 2014
4. Zhao 2018
5. Zhu 2019