M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

**Title Heading**:

Clinical Utility of Lymph Node Features during EBUS

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

# PROSPECTIVE DEVELOPMENT AND VALIDATION OF A MALIGNANCY SCORING SYSTEM DURING ENDOBRONCHIAL ULTRASOUND EVALUATION OF MEDIASTINAL LYMPH NODES FOR LUNG AND ESOPHAGEAL CANCER

By DANIELLE ALEXANDRIA HYLTON, BSc.

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

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

McMaster University MASTER OF SCIENCE (2018) Hamilton, Ontario (Health Research Methodology)

TITLE: Prospective Development and Validation of a Malignancy Scoring System During Endobronchial Ultrasound Evaluation of Mediastinal Lymph Nodes for Lung and Esophageal Cancer.

Author: Danielle Alexandria Hylton, BSc.

Supervisor: Dr. Waël C. Hanna, MD., MBA.

Pages: 86 (Preliminary pages =16, Text = 70)

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

**Lay Abstract**:

During lymph node staging for lung and esophageal cancer, specific features of lymph nodes can be seen. Using diagnostic tools these features can be used to predict whether a lymph node is cancerous or benign. However, many of these diagnostic tools are inaccurate or unreliable. To address this, this thesis aimed to develop a novel diagnostic tool based on lymph node features seen during staging procedures and determine its clinical usefulness and application to the wider lung and esophageal cancer population. This thesis also aimed to use improved methods to develop this diagnostic tool such that patient and clinician experiences would be significantly improved. The results of this thesis may contribute to a reduction in the number of repeat procedures required for patients undergoing staging prior to their treatment for lung and esophageal cancers.

**Abstract:**

Background: At the time of endobronchial ultrasound (EBUS) staging, ultrasonographic features can be used to predict mediastinal lymph node (LN) malignancy. Predictive tools have been developed, however they have not gained widespread use due to lack of research demonstrating validity and reliability. We sought to develop a novel predictive tool, the Canada Score, capable of predicting malignancy and potentially guide LN biopsy decision making.

Methods: We prospectively analyzed the ultrasonographic features of LNs from patients with NSCLC. Ultrasonographic features were identified by a single experienced endoscopist, this data was used to develop the Canada Score. Pathological specimens were used as the gold standard for determination of malignancy. Videos were then circulated to endoscopists across Canada, who were also asked to identify ultrasonographic features for each LN. Hosmer-Lemeshow test, logistic regression, receiver operator characteristic (ROC) curve, and Gwet's AC1 analyses were used to test the performance, discriminatory capacity, and inter-rater reliability of the Canada Score.

Results: A total of 300 LNs from 140 patients were analyzed by 12 endoscopists across 7 Canadian centres. Backwards elimination was used to create a

multivariate model. Hosmer-Lemeshow test and ROC curves indicated the model was well-calibrated ($chi^2$=11.86, p=0.1567) with good discriminatory power (c-statistic= 0.72 ±0.042, 95%CI: 0.64-0.80). Beta-coefficients were used to create a simplified score out of four. Evaluation of the tool showed that LNs scoring 3 or 4 had odds ratios of 15.17 (p<0.0001) and 50.56 (p=0.001), respectively for predicting malignancy. A score of 4/4 was associated with 99.59% specificity and a positive likelihood ratio of 22.78. Inter-rater reliability for a score ≥ 3 was 0.81 ± 0.02 (95%CI: 0.77-0.85).

Conclusions: The Canada Score shows excellent performance in identifying malignant LN at the time of EBUS. A cut-off of ≥ 3 has the potential to inform decision-making regarding biopsy or repeat/mediastinoscopy if the initial results are inconclusive.

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

**Acknowledgements**:

I would like to thank my thesis supervisor, Dr. Waël C. Hanna, for all his support, guidance, and encouragement during my graduate studies.

I would also like to thank Dr. Forough Farrokhyar and Dr. Feng Xie for their support during my graduate studies.

I also wish to thank Christine Fahim and Yogita Patel for their insightful and knowledgeable advice and encouragement throughout this process.

**Table of Contents**:

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

**List of Tables and Figures**

**TABLES:**

**FIGURES**:

**List of Abbreviations and Symbols**:

| | |
|---|---|
| AUC | Area under the curve |
| CHS | Central hilar structure |
| CI | Confidence interval |
| CP | Convex probe |
| CT | Computed tomography |
| EBUS | Endobronchial ultrasound |
| EBUS-TBNA | Endobronchial ultrasound transbronchial needle aspiration |
| EUS | Endoscopic ultrasound |
| HR | Hazard ratio |
| IASLC | International Association for the Study of Lung Cancer |
| IRA | Inter-rater agreement |
| IRR | Inter-rater reliability |
| LN | Lymph node |
| LR | Likelihood ratio |
| NPV | Negative predictive value |
| NSCLC | Non-small cell lung cancer |
| OR | Odds ratio |
| PET | Positron emission tomography |
| PPV | Positive predictive value |

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

| | |
|---|---|
| ROC | Receiver operator characteristic |
| RR | Risk ratio |
| SD | Standard deviation |
| SUV | Standardized uptake value |

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

**Declaration of Academic Achievement**:

Danielle A. Hylton was involved in the conception and design of the study, collection of the data, performed the statistical analyses, and was primary author of the subsequent manuscripts published based on this thesis work.

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

Drs. Forough Farrokhyar and Feng Xie were involved in the design of the study and review of the statistical analysis.

## Chapter 1: Introduction

### 1.1 Lung and Esophageal Cancer – Canadian Context:

Lung cancer is the most common cancer amongst Canadian males and females, accounting for 14 percent of all cancers. In 2017, 26% (n=21,100) of the cancer-related deaths in Canada were attributed to lung cancer, this was the highest percentage of all cancers. Projections indicated that in 2017, the number of people expected to die of lung cancer would be higher than that of pancreatic, colorectal, and breast cancer combined (n=21,100 vs. n=19,200) (Canadian Cancer Society (CCS), 2017).

Incidence rates for lung cancer by age group follow an upward trend, implying that the likelihood of developing lung cancer increases with age. After age 50, the incidence of lung cancer for males and females is 46%. Despite the relatively high incidence rates after age 50, trends indicate that lung cancer incidence is decreasing. This decrease has been attributed to reductions in tobacco use during the 1970s and 1980s and is expected to continue for as long as tobacco use continues to decline. The ten-year prevalence rates, which identify the length of time individuals live with lung cancer, can be broken down into three divisions: 0 to 2 years, 3 to 5 years, and 6 to 10 years. The numbers of prevalent lung cancer cases are, 48% (n=18,755), 28% (n=11,165), and 24% (n=9,430), for 0 to 2 years, 3 to 5 years, and 6 to 10 years, respectively. There are significantly fewer

individuals living with lung cancer after years 0 to 2 post-diagnosis, suggesting that lung cancer often has a poor prognosis (CCS, 2017).

Within the western hemisphere, esophageal cancer represents one of the most rapidly growing incidence rates and lowest survival rates of any cancer type (Otterstatter, et al., 2012). In 2010, the estimated number of new esophageal cancer cases for Canadian men and women was 1,700 (male = 73.53%, n=1,250) (Canadian Institute for Health Information (CIHI), 2011). Within the Canadian population, the five-year survival rate for esophageal cancer is the second lowest amongst all cancers at 14%. Comparatively, the incidence and prevalence of lung cancer is higher than that of esophageal cancer. However, patients diagnosed with esophageal cancer experience higher rates of mortality.

Despite the apparent differences in incidence and prevalence between lung and esophageal cancer, the way in which these cancers are diagnosed are similar. Thus, the development of novel diagnostic methods may positively impact the quality of life and care received from both cancer populations.

## 1.2 Mediastinal Detection & Investigations:

Methods used to detect lung and esophageal cancer tumors include computed tomography (CT), chest radiography, and sputum cytology (Gelberg, et al., 2014). Once a cancerous tumor is detected, the primary concern shifts to detecting the presence of cancerous mediastinal lymph nodes via staging

procedures. The staging system most often used for non-small cell lung cancer (NSCLC) and esophageal cancer is the American Joint Committee on Cancer (AJCC) TNM system. This system is based on three key pieces of information: the main tumor size (T), cancer spread to nearby lymph nodes (hilar and mediastinal) (N), and cancer spread to distant sites (metastasis) (M). Numbers after the T, N, and M provide details about how advanced a patient's cancer is, higher numbers indicate a more advanced case (CCS, 2017).

Staging the mediastinal and hilar lymph nodes is crucial to identifying the optimal treatment option for patients. The presence of cancer in hilar or mediastinal lymph nodes indicates that chemotherapy or radiotherapy are available treatment options. Whereas when cancer is absent from the hilar or mediastinal lymph nodes, patients qualify for surgical resection, which is often curative. There are two methods to diagnostically assess lymph nodes for lung and esophageal cancer: invasive or minimally invasive. The invasive approach, also called surgical staging, refers to cervical mediastinoscopy. This method is considered the gold-standard method for hilar and mediastinal lymph node biopsy (Gelberg, et al., 2014). A mediastinoscopy is performed in an operating room under general anesthesia and provides access to the upper and lower paratracheal (2R, 2L, 4R, and 4L) and subcarinal (7) lymph nodes (Gelberg, et al., 2014). Systematic reviews report a median sensitivity of 78% and negative predictive

value (NPV) of 91% for surgical staging (Silvestri et al., 2013). The endobronchial ultrasound (EBUS) guided transbronchial needle aspiration (EBUS-TBNA) is an alternative to the cervical mediastinoscopy. The EBUS is minimally invasive, does not require general anesthesia, and is completed on an out-patient basis. The procedure is completed using specialized endoscopes that allow for real-time ultrasound visualization of the lymph nodes during biopsy. Compared to the mediastinoscopy, EBUS provides access to upper (2R, 2L) and lower paratracheal (4R, 4L), subcarinal (7), hilar (10R, 10L), and interlobar (11R, 11L) lymph nodes. Additionally, meta-analyses have indicated that the median sensitivity for EBUS procedures is 89% and the NPV is 91% (Silvestri, et al., 2013; Gelberg, et al., 2014). Another minimally invasive procedure for lymph node investigation is the endoscopic ultrasound (EUS) guided fine needle aspiration, which operates similarly to the EBUS. The advantages offered by EUS over EBUS are technical including: higher quality image, wider field of ultrasound image, and less air artefact (Gelberg, et al., 2014). The EUS is also able to reach lymph nodes that the EBUS and mediastinoscopy cannot: subaortic (5), paraesophogeal (8), pulmonary ligament (9) lymph nodes. The EUS offers a sensitivity of 89% and NPV of 86% (Gelberg, et al., 2014). Standard of care often involves clinicians combining EBUS and EUS during a staging procedure. This enables access to nearly all the lymph nodes in the mediastinum. When EBUS and

EUS are combined the sensitivity and NPV have been reported as 91% and 86%, respectively (Silvestri et al., 2013; Gelberg, et al., 2014). In terms of diagnostic capability, the EBUS or combined EBUS/EUS outperform the traditional cervical mediastinoscopy.

Despite the mediastinoscopy being considered the gold-standard approach to mediastinal lymph node staging, more clinicians prefer the minimally invasive approach offered by EBUS. In addition to offering greater sensitivity, higher NPV, and improved lymph node access, the EBUS provides the opportunity for clinicians to assess the ultrasonographic features of lymph nodes in real-time prior to needle biopsy.

## 1.3 Characteristics of Mediastinal Lymph Nodes:

There is extensive literature discussing the distinct characteristics of lymph nodes visualized via ultrasound. In recent years, certain characteristics have been evaluated and associated with malignant infiltration or benign status. Furthermore, the evaluation of these lymph node characteristics has been proven clinically useful in breast cancers, thoracic malignancies, thyroid cancer, and cervical lymph node metastases relating to head and neck cancers (Harris et al., 2015; Akissue de Camargo Texeria, et al., 2017; Sun et al., 2017). Several studies

have assessed the ultrasonographic features of mediastinal and hilar lymph nodes for patients with lung and esophageal cancer specifically.

Fujiwara et al. (2010) completed one of the first studies identifying the ultrasonographic features that predict malignancy in mediastinal and hilar lymph nodes. Using a retrospective study design a total of 1,061 lymph nodes were evaluated from 487 patients. The following characteristics were assessed: (1) size (short axis) less than 10 millimetres (mm) or greater than 10 mm, (2) shape (oval or round), (3) margin status (well-defined or ill-defined), (4) echogenicity (homogeneous or heterogeneous), (5) presence or absence of a central hilar structure, and (6) presence or absence of a coagulation (central) necrosis sign. These characteristics were evaluated using strict definitions to be described in detail in Chapter 2.

In the Fujiwara et al. (2010) study a complete analysis of the abovementioned ultrasonographic features was completed to determine their predictive capability. Results of a multivariate analysis indicated that round shape, distinct margin, heterogeneous echogenicity, and presence of coagulation necrosis sign were independent predictive factors for malignant lymph nodes. Conclusions supported the theory that ultrasonographic features may be helpful in the prediction of metastatic lymph nodes during EBUS procedures.

The Fujiwara study lead to a series of studies aiming to determine the clinical utility of a predictive scoring system that could be used to detect malignancy in mediastinal or hilar lymph nodes during EBUS. The most recent was conducted by Shafiek et al. (2014) and used the malignancy criteria identified by Schmidt-Bindert et al. (2012) and definitions described by Fujiwara et al. (2010). A 6-point score was developed and proven internally valid. Results indicated that a score greater than or equal to 5 had a sensitivity and specificity for detecting malignant lymph nodes of 78% and 86%, respectively (c-index = 0.852; 95% confidence interval (CI) = 0.743-0.928, p=0.0001) (Shafiek et al., 2014). The authors concluded that the score may assist clinicians when making decisions regarding which lymph nodes to biopsy and be particularly helpful during decision making after an EBUS procedure has been determined inconclusive. Despite the positive results of the study, there is a need for an external validity assessment before this score can be applied to the lung and esophageal cancer population. Furthermore, the methods used to develop this predictive scoring tool and others are not methodologically rigorous, often do not include robust statistical analyses, and do not reflect the methodology used within the literature to develop clinically useful predictive tools.

1.4 Rationale for Ultrasonographic Feature Usage:

The importance of ultrasonographic assessment with malignancy criteria during EBUS-TBNA becomes obvious when biopsy results are inconclusive or insufficient for cytological interpretation. EBUS-TBNA samples are deemed inconclusive for pathological diagnosis in as high as 42.14% of cases (Ortakoylu, et al., 2015). In such situations, either the EBUS-TBNA procedure needs to be repeated or the patient must undergo a mediastinoscopy (Jalil et al., 2015). A reported 29.85% of patients with inconclusive EBUS-TBNA staging are referred to mediastinoscopy, and clinical guidelines mandate repeat biopsy when EBUS-TBNA results are inconclusive (Ortakoylu et al., 2015; National Institute for Health and Care Excellence, 2011). As such, the correct identification of benign or malignant ultrasonographic features at the time of EBUS staging, has the potential to reduce the number of repeat EBUS-TBNA and/or mediastinoscopy after initial inconclusive EBUS results. Despite the potential clinical utility, ultrasonographic features are not frequently reported or used by interventional respirologists or thoracic surgeons. This can be explained by the lack of high quality scientific evidence confirming the reliability and validity of using ultrasonographic features to predict mediastinal lymph node malignancy. The development and formal validation of a novel predictive tool can address this knowledge gap.

## 1.5 Objectives:

Clinical prediction tools reliably estimate the probability of an outcome based on clinical features observed during diagnostic test or treatment modality. Ultrasonographic features of mediastinal and hilar lymph nodes are examples of clinical features that can be used to predict diagnostic outcomes. Despite these features being proven to be predictive of malignant infiltration, widespread use of these features in clinical practice is limited. Few studies have assessed how to apply these features in clinical practice. Of the few studies that have looked at the clinical applications of these features, several predictive tools have been developed with the intention of enabling clinicians to be able to reliably classify lymph nodes as malignant or benign without a need for repeat staging. However, many of these predictive tools have either not been developed using rigorous methodology or do not use robust statistics. This suggests that the true clinical applicability of these predictive tools may not be reflected in the reported utility.

The primary objective of this thesis was therefore to develop a predictive tool based on ultrasonographic features using robust statistics and rigorous methodology that is accurate, reliable, and capable of being used in a clinical setting. The ultrasonographic features identified by Fujiwara et al. and Shafiek et al. were used to prospectively assess lymph nodes being biopsied by endobronchial ultrasound procedures. This information was then used to

determine which ultrasonographic features are significant predictors of malignancy and develop a novel predictive tool based on these features called the Canada Score. The secondary objectives for this thesis were to externally validate the Shafiek et al. predictive tool, and compare the performances of the Shafiek et al. tool with the newly developed predictive tool described in objective one.

## Chapter 2: Ultrasonographic Features - A Summary of the Pertinent Literature

Within the literature relating to ultrasonographic features seen during endobronchial ultrasound procedures there are several features that are commonly reported as being predictive of malignant infiltration. These features include: shape, echogenicity, margin status, central necrosis, short axis length, and central hilar structure. The importance of these features followed by their incorporation into diagnostic tools will be described below.

### 2.1 Shape:

Possible lymph node (LN) shapes included round, oval, and triangular. Shape was determined in several studies by calculating the ratio of the long axis to short axis of the LN (measured by the individual operating the endoscope) (Shafiek et al., 2014; Wang et al., 2016; Fujiwara et al., 2010). A ratio less than 1.5 would attribute a round shape to the LN. A ratio equal or greater than 1.5 would attribute an oval shape to the LN (Fujiwara et al., 2010). Triangular shape was determined if three distinct arms could be seen by the operator, otherwise, the LN was considered round/oval (Gogia et al., 2015). Shape was also determined via subjective assessment of the long and short axes during EBUS procedures and agreed upon by participating bronchoscopists (Wang-Memoli et al., 2011). Round shape was commonly associated with LN malignancy. One study, by Jhun et al.

(2014) determined that round shape was a significant predictor of malignancy during univariate regression analysis and not statistically significant during multivariate regression analysis (Jhun et al., 2014). Schmidt-Bindert et al. (2012) concluded that any LN shape other than round contributed to a benign LN. Triangular shape, which was only evaluated in the study by Gogia et al. (2015), was also found to be a strong predictor of benign LN status. The presence of a triangular shaped LN, compared to round or oval shape, was associated with a specificity of 98.90% and an odds ratio (OR) of 17.40 (95% CI: 7.0-43.1) for benign diagnosis after log-binomial regression analysis was completed (Gogia et al., 2015).

2.2 Echogenicity:

Echogenicity is divided into two possible categories: heterogeneous or homogeneous, referring to the grayscale texture of the LN being imaged during EBUS (Fujiwara et al., 2010). Heterogeneous echogenicity was frequently a significant predictor of malignancy, and homogeneous echogenicity a predictor of benign LN. Evison et al. (2015) reported that heterogeneous echogenicity was the strongest predictor of LN malignancy, and it further proved to be the only significant ultrasonographic predictor during multivariate analysis (OR = 48, 95% CI: 8-282, $p<0.001$) (Evison et al., 2015). Jhun et al. (2014) reported similar results, demonstrating that only absence of CHS and heterogeneous echogenicity (OR = 3.1, 95% CI: 1.4-6.7, $p=0.005$) remained significant predictors of malignancy after

analyzing a model including five ultrasonographic features. In contrast, echogenicity was found not to be a statistically significant predictor of malignancy by Wang-Memoli et al. after logistic regression analysis (Wang-Memoli et al., 2011).

2.3 Margin Status:

Margin status is dichotomously categorized as well-defined (predictor of malignant LN) or ill-defined (predictor of benign LN) (Fujiwara et al., 2010). Margins are considered well-defined if the majority (>50%) of the LN border is hyperechoic. The following studies found well-defined margins to not be predictive of malignant disease: Ayub et al. (2016), Wang-Memoli et al. (2011), and Evison et al. (2015). These studies reported odds ratios for LNs with well-defined margins compared to those with ill-defined margins as 0.3 (95% CI: 0.1-0.9, p=0.11), 0.98 (95% CI: 0.58-1.66, p=0.93), and 1.2 (95% CI: 0.6-2.6, p=0.57), respectively. Gogia et al. (2015) also assessed margins and reported no significant correlation between ill-defined margins and pathologic results. However, analyses showed that the absence of well-defined margins had a 92.60% specificity for predicting benign LNs (Gogia et al., 2015). Schmid-Bindert et al. (2012) were unable to conclude whether well-defined or ill-defined margins could predict LN pathology. Though, they reported that the presence of ill-defined margins, in association with several other ultrasonographic features, most likely predicted a benign LN (Schmid-Bindert

et al., 2012). Conversely, Fujiwara et al. (2010) reported well-defined margins compared to ill-defined margins as strong predictors of malignancy, with a hazard ratio of 3.05 (95% CI: 1.61-5.75, p=0.0006) on logistic regression.

2.4 Central Necrosis:

Central necrosis, which is thought to be a distinct sign of malignancy is defined as the presence of a centrally located hypoechoic structure within a LN. Alici et al. (2016) assessed central necrosis in benign and malignant LNs, and found it to be present in significantly more pathologically confirmed malignant LNs than benign LNs ($p<0.001$). Central necrosis presence (compared to central necrosis absence) was found to be a statistically significant predictor of malignancy by Fujiwara et al. (2010) with a hazard ratio (HR) of 5.64 (95% CI: 3.40-9.38) ($p<0.001$) on logistic regression.

2.5 Short Axis Length:

A short axis length equal to or greater than 10 mm is thought to be associated with malignant LNs (Fujiwara et al., 2010; Shafiek et al., 2014). Gogia et al. (2015) confirmed short axis length less than 10 mm (versus short axis length greater than 10 mm) was an independent predictor of benign LNs in a multivariate regression model (Risk Ratio = 1.31, 95% CI: 1.107-1.549, p=0.002). In the analysis by Schmid-Bindert et al. (2012), short axis length less than 10 mm was

predictive of benign LNs and had a negative predictive value (NPV) of 82.50%. Shafiek et al. (2014) determined that the presence of short axis lengths greater than 10 mm was significantly different between biopsy proven malignant and benign LNs (p=0.023). Fujiwara et al. (2010) concluded that the presence of short axis greater than 10 mm, round shape, well-defined margins, heterogeneous echogenicity, absent CHS, and present central necrosis within a LN increased the likelihood of malignancy. However, their logistic regression analysis indicated that short axis length (greater than 10 mm versus less than 10 mm) alone was not an independent predictor of malignancy (HR = 1.34, 95% CI: 0.882-2.03, p=0.171) (Fujiwara et al., 2010).

2.6 Central Hilar Structure:

The presence of a central hilar structure (CHS) is predictive of a benign LN, and its absence is predictive of malignancy (Fujiwara et al., 2010). Ayub et al. (2016) determined that the presence of a CHS was independently predictive of benign LN during univariate analysis (p=0.03). Gogia et al. (2015) reported the risk ratio of a LN being benign as 2.22 (95% CI: 1.876-2.621, p<0.001) when the central hilar structure was present. In predicting malignant LNs, Fujiwara et al. (2010) found that the absence of a CHS resulted in 89.70% sensitivity and 92.90% NPV. However, their logistic regression analysis did not identify absence of CHS as independently predictive of malignancy (Fujiwara et al., 2010). Shafiek et al.

(2014) reported a statistically significant difference (p=0.0012) between malignant and benign LNs with respect to CHS absence. In their study, absent CHS was also associated with 99.00% (95% CI: 94.8-99.8%) sensitivity and 90.00% (95% CI: 54.1-99.5%) NPV for the prediction of malignant disease (Shafiek et al., 2014).

## 2.7 Development of Predictive Tools:

Four predictive tools have been developed either partially or completely based on the absence or presence of certain ultrasonographic features. These tools are described below.

### 2.7.1 The Alici et al. Algorithm:

Alici et al. conducted a retrospective analysis of the ultrasonographic features of 1051 LNs and used the diagnostic results to develop an algorithm for LN sampling during EBUS. This algorithm was then internally validated using a subset of the study population. The sensitivity, specificity, positive predictive value (PPV), NPV, and diagnostic accuracy were 100.00%, 51.20%, 50.60%, 100.00%, and 67.50%, respectively. Authors concluded that the algorithm did not provide any suggestions to clinicians with respect to systematic N3-N2-N1 sampling and that their proposed algorithm should be externally validated prospectively (Alici et al., 2016).

2.7.2 The Schmid-Bindert et al. Score:

Schmid-Bindert et al. (2012) retrospectively analyzed 281 LNs from 145 patients to develop a score to predict malignancy. The score was based on the presence of six ultrasonographic features that were shown to be predictive of malignancy. The authors stratified the ultrasonographic features such that the presence of 3-6 features was considered high-risk for malignancy and the presence of 1-2 features was considered low-risk. Results of this stratification showed that the odds ratio for malignancy was 15.50 (95% CI: 3.63-66.17) when 3 or more features were present. Results of individual ultrasonographic features showed that the single best criterion for predicting malignancy was heterogeneous echogenicity, which was present in 85% of the malignant LNs (Schmid-Bindert, et al., 2014).

2.7.3 The Evison et al. Risk Stratification Model:

Evison et al. (2015) retrospectively analyzed 329 LNs that were split into a derivation set (n=196) and validation set (n=133) with the intention of developing a risk stratification model to be used during EBUS procedures. The risk stratification model was developed based on the natural logs of the odds ratios for echogenicity, standardized uptake value (SUV), and lymph SUV percentage. The previously mentioned features were the only statistically significant covariates after multivariate logistic regression analyses. When applied to the validation set, the

risk stratification model had a NPV, PPV, sensitivity, and specificity of 97.90%, 64.90%, 92.30%, and 87.80%, respectively (Evison et al., 2015).

### 2.7.4 The Shafiek et al. Score:

Shafiek et al. developed a score-based tool based on research by Schmid-Bindert et al. (2012). Development of this tool included a full retrospective analysis of six ultrasonographic features from 208 LNs (n=141 patients) and prospective internal validation of the tool with 65 LNs (n=39 patients). Based on Schmid-Bindert et al. (2012) results, the score was modified to give heterogeneous echogenicity and absent CHS higher weights than the other 4 features, to reflect their strong predictive capability of malignancy. Results indicated that a cumulative score $\geq 5$ had a sensitivity and specificity of 78.00% and 86.00%, respectively in detecting malignant LNs. The area under the curve (AUC) was 0.85 (95% CI: 0.74-0.93, p=0.0001) and the PPV and NPV was 75.00% and 88.00%, respectively (Shafiek et al., 2014).

2.7.5 A Need for Further Investigation:

The summary of literature above highlights the lack of consensus within the literature regarding which features are associated with malignancy or benign lymph node status. This is a concern when considering the validity and clinical application of these ultrasonographic features. There is a need for further investigation using correct methodologies to identify the true predictors of lymph node malignancy and improve subsequent development of predictive diagnostic tools.

## Chapter 3: Methods

### 3.1 Primary Research Question:

In patients with confirmed or suspected lung and esophageal cancer undergoing mediastinal and hilar lymph node assessment via endobronchial ultrasound, can the ultrasonographic features identified (central hilar structure, shape, margin status, echogenicity, small axis length, and central necrosis) be used to develop a clinical predictive tool capable of accurately and reliably detecting malignant infiltration?

### 3.2 Secondary Research Question:

In patients with lung and esophageal cancer undergoing mediastinal and hilar lymph node assessment via endobronchial ultrasound is the novel clinical predictive tool compared with the Shafiek et al. tool more accurate in predicting malignant infiltration?

### 3.3 Hypothesis:

1) Based on the ultrasonographic features identified by Fujiwara et al. (2010) a clinical tool, called the Canada Score, can be developed that is capable of accurately and reliably predicting malignancy in mediastinal and hilar lymph nodes.

2) The Canada Score will outperform the Shafiek et al. predictive tool with respect to validity and reliability.

This thesis is divided into two parts. Part one focuses on data collection and development of the novel predictive tool (the Canada Score) and part two focuses on the external reliability assessment of the predictive tool and comparison to the Shafiek et al. (2014) tool.

## 3.4 Part One: Identification of Ultrasonographic Features & Predictive Tool Development

### 3.4.1 Patient Selection & Study Design:

A prospective study design with a predefined protocol was used to collect ultrasonographic lymph node feature information from consecutive patients that underwent endobronchial ultrasound staging for mediastinal investigation of suspected or confirmed lung or esophageal cancer lymph node infiltration. Potential participants were identified based on the following inclusion criteria: patient must be diagnosed with confirmed or suspected lung or esophageal cancer and be undergoing EBUS diagnosis/staging. There were no exclusion criteria for patients to prevent limiting the study population. Informed consent was obtained from each patient prior to the procedure. Each EBUS procedure was video recorded and saved to a secure external hard drive. Final pathology reports were

collected for each assessed and biopsied lymph node. Lymph node stations were identified using the International Association for the Study of Lung Cancer (IASLC) Lymph Node Mapping Nomenclature (El-Sherief, et al., 2014). The lymph nodes selected for biopsy were left to the operating surgeon's discretion. This part of the study was conducted between August 2016 and September 2017 at a designated thoracic cancer surgery centre. This study was approved by the Hamilton Integrated Research Ethics Board (HIREB) prior to study initiation.

### 3.4.2 Controlling for Bias:

Several methods were used to control biases during the study. When completing diagnostic studies spectrum bias is major concern as it can impact the generalizability of the results. Spectrum bias refers to the phenomenon where a diagnostic tests performance may vary in different clinical settings (Schmidt & Factor, 2013). This tends to occur when the study population being investigated does not accurately reflect the clinically relevant general population (Schmidt & Factor, 2013). When the severity of the disease in the study population differs significantly from the general population it may influence the sensitivity of the diagnostic test being evaluated or developed. To control for this bias, the literature suggests designing high quality studies where the diagnostic tool is evaluated against a study population that reflects the full breadth of disease severity (Schmidt & Factor, 2013). With respect to our study, the inclusion and exclusion

criteria were developed to be intentionally broad. Participants included within the study had cancer severity ranging from stage one to four, encompassing the full range of clinical staging for lung and esophageal cancer.

Diagnostic review bias occurs when the interpretation of the gold standard reference test, histopathological interpretation, is made with knowledge of the results from the diagnostic test that is under investigation. Similar to diagnostic review bias, incorporation bias occurs when the results of the diagnostic test being studied are being used to make a final diagnosis (Schmidt & Factor, 2013). To control for these biases, pathologists were blinded to the ultrasonographic feature information for each lymph node included in the analyses.

To prevent any biases related to improper identification of ultrasonographic features each clinician participating in the reliability assessment portion of the study (part two) was required to successfully complete the Ultrasonographic Feature Education Program. This online learning module was specifically developed for this study and teaches clinicians how to correctly identify ultrasonographic features.

Within the literature, when inter-rater reliability relating to identifying ultrasonographic features was assessed a maximum of two raters were used. Using two raters from the same institution to evaluate inter-rater reliability may introduce bias, as these clinicians would likely have similar experiences with

respect to patients, endobronchial ultrasound procedures, and training. To minimize this possibility, clinicians from across Canada were contacted and invited to participate. In total, 12 clinicians from seven different institutions participated in identifying ultrasonographic features. The variation in clinical and patient experience made the study less prone to bias and improved the generalizability of the results.

### 3.4.3 Sample Size & Recruitment Strategy:

The sample size was calculated with the lymph node as the unit of analysis rather than the patient, since each lymph node finding is independent from another within the same individual. Assuming 90.00% sensitivity and specificity for EBUS and a population of 1000 newly diagnosed lung and esophageal cancer patients in five centers each year, 300 LNs would provide enough precision to achieve confidence intervals of ± 3.00% for diagnostic properties of post-EBUS score (sensitivity and specificity). A 95.00% confidence interval z-score (1.96), 0.05 accuracy level, a prevalence of 0.50, and 90.00% for sensitivity and specificity were used to determine an appropriate sample size. The calculation was completed using standard formulae (Jones et al., 2003). It was determined that a sample equivalent to 300 LNs would enable diagnostic statistics to be calculated accurately. Most patients have at least 3 lymph node specimens biopsied, it was estimated that approximately 100 patients would be needed to achieve this sample size.

Patient recruitment began after research ethics approval was obtained. At St. Joseph's Healthcare Hamilton (SJHH), prior to study initiation the operating surgeon involved in this study completed six EBUS procedures on average each week. Assuming a recruitment rate of 60.00-80.00%, we anticipated enrolling 100 patients within 6-7 months. This recruitment rate was based on current clinical trial rates of recruitment at SJHH. However, we conservatively estimated our recruitment timeline to reflect a pragmatic approach. The patients study involvement ended at completion of the EBUS procedure.

### 3.4.4 EBUS-TBNA Procedure:

All EBUS procedures for the purposes of lymph node sampling, feature detection, and video collection were completed by the same thoracic surgeon at SJHH. Prior to the EBUS procedure, a conventional flexible bronchoscopic examination of the tracheobronchial tree was completed. Each EBUS procedure was completed using an endoscope and digital ultrasound scanner. The procedure was performed via the oral route with the patient under deep sedation. Convex probe (CP) EBUS was initially used to identify the lymph nodes and their surrounding blood vessels, the IASLC lymph node map was used to classify lymph nodes (El-Sherief et al., 2014). The dimensions (axes lengths) of the lymph nodes were measured using frozen ultrasound images. A needle was then inserted

into the working channel of the bronchoscope, it was then used to puncture the lymph node being visualized via EBUS guidance. The aspirated cellular material was spread onto glass slides, fixed, and air-dried. The dried slides were evaluated immediately by an on-site cytopathologist to determine whether the cellular aspirate was adequate for pathological analysis.

### 3.4.5 Analysis of Ultrasonographic Features:

All ultrasound videos were evaluated to determine the presence or absence of the five ultrasonographic features described by Shafiek et al. (2014) (CHS, small axis length, echogenicity, shape, and margin status) and coagulation (central) necrosis. The following definitions were used for the malignant ultrasonographic features being identified during EBUS procedures:

1. Round shape: defined as a ratio of the long and short axis less than 1.5.
2. Well-defined margins: distinguished by a majority echogenic line delimiting the lymph node.
3. Heterogeneous echogenicity: presence of non-uniform echogenic patterning.
4. Absence of a CHS: missing flat, central, echogenic structure in the lymph node.
5. Small axis > 10 mm: presence of a small axis length greater then 10 mm (1 cm).

6. Presence of central necrosis: presence of a central hypoechoic structure in the lymph node.

All ultrasonographic feature identification was completed during EBUS procedures using a standardized form. All lymph nodes included in the study were assessed using the gold standard histopathological interpretation approach. For part two of the study during the reliability assessment raters watched the lymph node videos and identified ultrasonographic features using an online version of the standardized form.

### 3.4.6 Statistical Analyses:

#### 3.4.6.1 Diagnostic Statistics:

Diagnostic statistics were calculated using receiver operator characteristic (ROC) curves. The sensitivity, specificity, negative likelihood ratio, and positive likelihood ratio were calculated for the novel predictive tool. Data was presented as means ± standard deviations (SD), median (range), or as a number (percentage) as appropriate. 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 lymph nodes. All statistical tests used two-sided hypotheses with p values less than 0.05 considered statistically significant. Stata 15 (StataCorp, College Station, Texas, USA) was used for all statistical analyses.

3.4.6.2 Regression Analysis:

Univariate analyses via binary logistic regression was completed using all the ultrasonographic features reported above (shape, small axis length, margin status, echogenicity, central hilar structure, and central necrosis) as covariates and final pathology (benign or malignant) as the dependent variable. Each covariate was treated dichotomously. The univariate analyses provided insight as to which ultrasonographic features should be included in the multivariate regression based on statistical significance ($p < 0.05$). Backwards elimination automatic variable selection with an elimination point of p-value > 0.05 was used to assist in developing the multivariate model. Backwards elimination starts with the full model followed by dropping the least significant variable, this process continues until all the remaining variables are statistically significant. A stepwise automatic variable selection analysis was also completed. The stepwise procedure allows for movement of variables in either forward or backward directions, dropping or adding variables at each step. A significance level of $0.05 \geq$ was used as the removal point, and a significance value of 0.04 was used as the addition point. Two automatic variable selection processes were completed to see the possible multivariate models, therefore allowing the most statistically significant model to be selected. The covariates included during model development were the same covariates assessed during the univariate analyses described previously.

3.4.6.3 Model Calibration and Discrimination:

The Hosmer-Lemeshow test was used to evaluate the model's calibration in combination with a calibration plot. Calibration refers to the level of agreement between predicted and observed outcomes. A well calibrated model is especially desired for predictive modeling as it provides information on the likelihood of predicted outcomes being either over or underestimated (Han et al., 2016). A p-value greater than 0.05 after completing the Hosmer-Lemeshow test suggests that the model is well calibrated. For a calibration plot, when the future predicted probabilities agree perfectly with the observed probabilities, the plotted line follows the 45-degree bisecting line indicating perfect calibration (Han et al., 2016).

Receiver operator characteristic (ROC) curves and c-statistics generated after binary logistic regression were used to evaluate model discrimination. Model discrimination reflects the ability of a prediction model to differentiate between two outcome classes, for example benign and malignant lymph nodes. The concordance statistic (c-index), also referred to as the area under the curve (AUC), is used to measure the level of discrimination, it can be interpreted as the probability that a patient with an outcome is given a higher probability of the outcome by the predictive model than a randomly selected patient without the

outcome (Han et al., 2016). A c-statistic equal to 0.5 reflects a model without any discriminatory ability, 1.0 reflects perfect discriminatory ability.

3.4.6.4 Criticisms of Calibration and Discrimination Statistical Procedures:

Although both model calibration and discrimination are important steps during the development of a predictive model, there are criticisms for both procedures. For the Hosmer-Lemeshow test, the sample size and p-value have an inverse relationship. Implying that when used, the Hosmer-Lemeshow test is nearly always significant (an indicator of poor model calibration) for large samples (Bertolini et al., 2000). Despite this criticism, the Hosmer-Lemeshow test remains an important method for assessing model calibration. As for ROC curves and c-statistics, although widely used their interpretation is not directly clinically relevant. Furthermore, the predicted values can differ significantly from the observed values even when the c-statistic is 1.0 (Bertolini et al., 2000). The criticisms of the above-mentioned statistics explain why the use of one statistical method to assess a predictive model's calibration and discrimination is ill-advised. The use of multiple statistical methods with results that agree with one another provides additional confidence that the predictive model is truly well-calibrated and with high discriminatory capability.

3.4.6.5 Development of a Novel Predictive Tool: The Canada Score:

Several other predictive diagnostic tools that aimed to predict malignancy in mediastinal lymph nodes based on ultrasonographic features, have been described in chapter 2. There are three main concerns with the previously developed tools: 1) the published research does not report any attempts to calibrate their models using any statistical measures, 2) weighting of the most important ultrasonographic features via beta coefficient values was often not completed, and 3) prospective data collection was often not possible. To address these issues, development of the Canada Score incorporated the previously mentioned criticisms to create a novel predictive tool. Using formulae from Han et al. (2016) the smallest beta coefficient from the multivariate model (CHS ß = 0.85) was used as a base constant. Each beta coefficient was then divided. The resulting quotient from dividing the betas for CHS, small axis length, margin status, and central necrosis by CHS ß provided the score value. The results are described in Table 5. Based on calculations, each covariate was allotted a score of 1 resulting in a maximum score of 4 for each lymph node. Using the beta coefficients to derive the score ensured that covariates with higher betas and therefore higher predictive capability were reflected in the score. Despite each ultrasonographic feature being scored either a zero or one, this method ensured that values were not assigned arbitrarily and reflected statistical significance.

## 3.5 Part Two: Inter-Rater Reliability Assessment for the Predictive Tool

### 3.5.1 Patient Selection & Study Design:

The second part of this study included inter-rater reliability assessment of the Canada Score and formal external validity assessment of the Shafiek et al. tool. Thoracic surgeons and respirologists (referred to as raters) from across Canada were asked to participate in identifying ultrasonographic features for each of the 300 lymph nodes recorded and analysed in part one of the study. These raters were blinded to the final pathology results of each lymph node they assessed and any patient information. Prior to participating in this part of the study all raters were required to complete an online education module (described below). The education module was developed for the purposes of this study to ensure consistency in ultrasonographic feature identification. For ease and simplicity, a specialized online survey tool was developed using the Research Electronic Data Capture (REDCap) software to disseminate the lymph node videos to raters. Using REDCap, raters were sent one survey, which included 10 LN videos, at a time. After watching each LN video, raters were asked to identify ultrasonographic features. The intention of this part of the study was to formally assess the reliability related to identifying ultrasonographic features across different clinicians.

3.5.2 Education Program:

The Ultrasonographic Feature Education Program was developed to teach clinicians how to correctly identify ultrasonographic features during EBUS. The education program was completely online and made use of interactive and repetitious learning theories to accomplish mastery over a short time period. The design of the module included a pre-test and post-test along with ten additional immersive practical experiences in between. During the practical experience portion of the program participants viewed videos of lymph nodes being imaged via EBUS and were interactively guided through learning how to identify each ultrasonographic feature. Those who did not successfully complete the module were ineligible to participate as raters.

3.5.3 Statistical Analyses:

The inter-observer reliability was calculated using the standard definitions of Gwet's AC1 (Blood & Spratt, 2007). The Gwet's AC1 alpha is an alternative to the commonly used Cohen's Kappa approach for determining inter-rater reliability. The literature reports several criticisms with using Cohen's Kappa including: inability to calculate coefficients over more than two raters and low coefficient values despite there being high levels of agreement (also referred to as Kappa's Paradox) (Blood & Spratt, 2007). Gwet's AC1 is capable of comparing

the level of agreement between multiple raters using a categorical rating system (Blood & Spratt, 2007). Therefore, Gwet's AC1 was deemed most appropriate for reliability analyses. Raw agreement percentage, standard errors, and 95% confidence intervals were calculated for each Gwet's AC1 coefficient.

For external validation of the Shafiek et al. predictive tool, diagnostic statistics were calculated using ROC curves and logistic regression. The sensitivity, specificity, negative likelihood ratio, and positive likelihood ratio were calculated. Data was presented as means ± standard deviations (SD), median (range), or as a number (percentage) as appropriate. All statistical tests used two-sided hypotheses with p-values less than 0.05 considered statistically significant. Stata 15 (StataCorp, College Station, Texas, USA) was used for all statistical analyses.

# Chapter 4: Results

## 4.1 Part One: Results of Ultrasonographic Feature Assessment and Predictive Tool Development

### 4.1.1 Demographic Data:

In total, 300 lymph nodes from 140 patients with suspected or confirmed lung or esophageal cancer were included in this study. The average age of the participants was 67.92 ± 10.64 years, with 54.30% (n=76) being male (Table 1). Standard of care mandates diagnostic imaging prior to mediastinal lymph node assessment and/or surgical resection. As such, 99.29% of patients participating in this study received either a chest computed tomography (CT) or positron emission tomography (PET) scan prior to their EBUS procedure. On average, 2.14 ± 0.95 lymph nodes were sampled per patient. Of the 300 lymph nodes sampled, the most commonly biopsied were those at station 7 (n=131, 43.67%) and station 4R (n=85, 28.33%).

After pathological assessment, 55.00% (n=77) of the patients had confirmed lung cancer with adenocarcinoma (42.86%, n=33) being the most common histological type. A total of 32 (22.86%) patients were confirmed to have esophageal cancer with adenocarcinoma also being the most common histological type (87.50%, n=28). Benign disease was confirmed in 31 patients (22.14%). With respect to individual lymph nodes, definitive diagnosis of

malignant disease was obtained in 18.00% of biopsied LNs compared to 82.00% that were benign.

4.1.2 Ultrasonographic Features:

Analysis of the ultrasonographic features present in each biopsied LN are shown in Table 2. Malignant and non-malignant lymph nodes were compared using Pearson's chi square, results were statistically significant ($p<0.05$) for each ultrasonographic feature assessed except echogenicity and shape which had chi square values (p-value) of 0.002 ($p=0.966$) and 0.67 ($p=0.415$), respectively. This implied that the presence or absence of central hilar structure ($p<0.0001$), margin status ($p<0.0001$), small axis length ($p=0.001$), and central necrosis ($p=0.001$) were dependent upon the malignancy status of a lymph node. To further assess the relationship between these ultrasonographic features and lymph node malignancy status a binary logistic regression analyses was completed.

4.1.3 Multivariate Model Development:

The univariate binary logistic regression analyses for each ultrasonographic feature is reported in Table 3. The results generated from the univariate analyses replicate those completed during the Pearson's chi square test: all the ultrasonographic features, except for echogenicity and shape produced significant p-values ($p<0.05$). The results from the univariate analysis suggested

that the multivariate model should include CHS, margin, small axis length, and central necrosis. Both automatic variable selection processes, backwards elimination and stepwise, included CHS, margin, and small axis length as variables in the multivariate model (each had p-values $< 0.05$). Central necrosis was statistically significant during univariate analysis; however, it became largely non-significant ($p=0.096$) during multivariate analysis. The literature suggests that the presence of central necrosis is a significant predictor of malignant infiltration. Therefore, central necrosis was included in the final multivariate model.

Shape and echogenicity were determined to be non-statistically significant during univariate analysis and both automatic modelling procedures. Aside from statistical reasoning for removing these from the multivariate model, both shape and echogenicity are highly subjective and operator-dependent. This likely accounts for why they were not statistically significant, but also provides clinical reasoning to remove them from the final model. These features were unlikely to be consistently nor reliably assessed amongst clinicians, thus their inclusion in the model would not provide any added clinical value.

4.1.4 Model Calibration and Discrimination:

The resulting Hosmer-Lemeshow chi square value was 11.86 and coincided with a p-value equal to 0.16. The high p-value suggests that there is not enough evidence to state that the model does not fit the data well. Figure 1 illustrates a calibration plot, which is a visual representation of the Hosmer-Lemeshow test. Subjectively, the model is well calibrated however there may be areas of overestimation with lymph nodes pathologically confirmed to be malignant. This overestimation is illustrated by the diversion of the observed outcomes from the predicted outcomes (45-degree angle line). Figure 2 depicts the receiver operator characteristic (ROC) curve for the multivariate model. The c-statistic was 0.72 ±0.04 (95%CI: 0.64-0.80) indicating the model has good discriminatory capability.

4.1.5 Evaluation of the Canada Score:

A logistic regression was completed to evaluate the Canada Score's performance (Table 6). Logistic regression revealed that a score of one out of four was not statistically significant in predicting lymph node malignancy (p=0.68). However, scores greater than or equal to two were statistically significant. A ROC curve analysis showed that the Canada Score has a c-statistic of 0.73 ± 0.04 (95%CI:0.65-0.81) (Figure 3). Table 7 summarizes the sensitivity, specificity,

percent correctly classified, and positive and negative ratios for each possible score using the Canada Score tool. A lymph node scoring four had a specificity of 99.59% and a positive likelihood ratio of 22.78. Figure 4 shows the probability of lymph node malignancy (y-axis) plotted against the Canada Score values (x-axis), this was used to estimate an appropriate malignancy-cut off score. A steep increase in malignancy likelihood is seen when the score equals three, and was therefore selected as the cut-off point for malignancy. This suggests that any lymph node scoring three or higher should be considered malignant.

## 4.2 Part Two: Formal Reliability Assessment of the Canada Score and External Validation of the Shafiek et al. Tool

### 4.2.1 Canada Score: Formal Reliability Assessment:

Gwet's AC1 coefficient was used to assess the level of agreement amongst clinicians identifying each ultrasonographic feature and application of the Canada Score. Gwet's AC1 coefficients for the individual ultrasonographic features ranged from 0.25 ± 0.03 (95%CI: 0.18 - 0.31) for echogenicity and 0.77 ± 0.02 (95%CI: 0.72 - 0.82) with central necrosis (Table 8). The agreement between clinicians (n=12) on the raw Canada Score values was 0.29 ± 0.02 (95%CI: 0.25 - 0.33) (Table 9). However, each of the 12 clinicians did not review all 300 lymph nodes. When reduced to the three clinicians that reviewed the complete sample

(n=300), the reliability statistics improved to 0.74 ± 0.02 (95%CI: 0.70 - 0.79) for agreement on the raw Canada Score values (Table 9). As described above, a score of three was considered the cut-off for malignancy. The reliability for whether clinicians (n=12) could agree when a lymph node should be considered benign or malignant (based on the 3-point cut-off) was 0.81 ± 0.02 (95%CI: 0.77 - 0.85) (Table 10).

4.2.2 Comparison Between Previous Models and the Canada Score:

The Shafiek et al. model to predict mediastinal lymph node malignancy based on ultrasonographic features was externally validated based on the collected data. The Shafiek et al. multivariate model had similar results to initial univariate modeling, in that echogenicity and shape were not statistically significant (Table 11). However, these features remained in the final Shafiek et al. model. Logistic regression analysis indicated that scores greater then 4.5 were significant predictors of lymph node malignancy (Table 12). The c-statistic for this model was 0.77 ± 0.04 (95%CI: 0.70-0.85) (Figure 5). Comparison of the ROC curves (Shafiek et al. model and Canada Score) revealed that despite objective differences in c-statistics, statistically this difference was not significant ($Chi(1)^2$=1.89, p=0.17) (Figure 6).

## Chapter 5: Discussion

### 5.1 External Validation of the Shafiek et al. Score:

Based on the ultrasonographic feature data collected for each lymph node (n=300), external validation of the Shafiek et al. (2014) predictive tool was possible. The Shafiek et al. (2014) score included five features: CHS (absent vs. present), small axis length (≥ 10 mm vs. < 10 mm), margin status (well-defined vs. ill-defined), shape (round vs. non-round), and echogenicity (heterogeneous vs. homogeneous). Three of the five features were categorically scored as either zero or one, where a score of one was associated with malignancy. Heterogeneous echogenicity and absent CHS were scored as 1.5 and their dichotomous counterparts as zero. The difference in scoring values between the ultrasonographic features was done to address the suspected importance of CHS absence and heterogeneous echogenicity as strong predictors of malignancy. A logistic regression analysis was not reported by Shafiek et al., however ROC analysis of the retrospectively collected data indicated that a combined score greater than or equal to five had a sensitivity of 73.30% (c-index= 0.738, 95% CI: 0.673-0.796). Prospective validation of the Shafiek et al. (2014) tool completed by the authors had a reported c-index equivalent to 0.85 (95% CI: 0.74-0.93). The authors concluded that a lymph node scoring five or higher was most accurate for predicting malignancy.

Our external validation of the Shafiek et al. score using logistic regression analyses produced a c-index of 0.77 (95% CI: 0.69-0.85). Unlike in the Shafiek et al. (2014) study, our logistic regression analysis indicated that a score of 4.5 was the best indicator of LN malignancy. A LN scoring 4.5 had a specificity, sensitivity, positive likelihood ratio, and odds ratio of 98.37%, 16.67%, 10.25, and 18.45 ($p \leq 0.0001$), respectively. Comparatively, a score equal to five was associated with a specificity, sensitivity, positive likelihood ratio, and odds ratio of 37.40%, 75.93%, 1.21, and 5.22 ($p=0.015$), respectively (Table 13).

A Pearson chi square comparison of the c-indexes for the externally validated Shafiek et al. (2014) tool and the Canada Score was completed (Figure 6). Despite the differences in c-indexes, there is no significant difference between both ROC curves ($chi^2(1) = 1.89$, $p=0.17$). This implies that both predictive tools would perform similarly with respect to predicting LN malignancy. However, we argue that the lack of rigorous methodology used to develop the Shafiek et al. (2014) tool limits its true clinical applicability.

## 5.2 The Importance of the Canada Score & Criticism of Previous Predictive Tools:

The Canada Score was intended to be novel in that the methods used to develop this predictive tool were rigorous and therefore different than those previously published. To the best of our knowledge, four other predictive scores exist that either partially use or completely focus on using ultrasonographic features to predict malignancy. Each of these studies collected their data retrospectively, and only one included a prospective internal validation portion. Of the four predictive tools three, Alici et al. (2016), Schmid-Bindert et al. (2012), and Shafiek et al. (2014), did not make use of beta coefficients based on logistic regression modelling to develop the scoring for each ultrasonographic feature. Instead, arbitrary values were used to score the absence or presence of each ultrasonographic feature. The problem with this approach is that it neglects the relative significance of each ultrasonographic feature determined during logistic regression.

For example, in the Schmid-Bindert et al. (2012) study a simplistic approach was used where the presence of any ultrasonographic feature associated with malignancy was given a score of one with a maximum score of six for each LN. The authors then stratified the scores, where scores between one and two were deemed low risk and any score three or higher was high risk. Risk

stratification was developed based on ROC curve analyses and predicted probabilities. An exact logit model was used to determine which ultrasonographic features were statistically significant, however this information was only used to select variables to be included in the predictive model. There was no use of the beta coefficients to derive scoring values for each included feature.

The predictive tool developed by Shafiek et al. was based on the Schmid-Bindert et al. (2012) study. This predictive tool included the same ultrasonographic features, except colour power Doppler index was removed. Another key difference was the weighting applied to each ultrasonographic feature. Instead of the absence or presence of each feature being scored as one, certain features were given more weight to therefore influence the sum score for the lymph node more. The Schmid-Bindert et al. (2012) paper suggested that heterogeneous echogenicity and absent central hilar structure were the features most predictive of malignancy. Therefore, heterogeneous echogenicity and absent CHS were both scored 1.5, whereas the other features were scored as zero or one, depending on their absence or presence. The result was a summative score out of six, with a score equal to five or higher being highly suggestive of LN malignancy. The Shafiek et al. (2014) approach is not arbitrary in comparison to the Schmid-Bindert et al. (2012) score; however, it does not use methodologically sound approaches to derive these values. In the above-mentioned studies, the lack

of beta-coefficient derived weighting may impact validity of the results. Without basing the scoring values on the beta-coefficients it is possible that CHS and echogenicity should be weighted more than 1.5, or equally possible that they should be weighted less.

The Alici et al. (2016) algorithm intended to distinguish between benign and malignant lymph nodes, however it did not include any logistic regression analyses, model calibration, or discrimination assessment. Development of the algorithm was based on chi-square analyses, diagnostic statistics, and later validated against a subset of the sample. The lack of robust statistical analyses used to develop this algorithm may negatively impact the validity and external applicability of these results.

Alternatively, the predictive score developed by Evison et al. (2015) included robust logistic regression analyses to develop their 5-point score. The initial intention for this tool was to incorporate both ultrasonographic features and characteristics specific to lymph nodes visualized during PET scan imaging (SUV and Lymph SUV percent) into a tool used to predict malignancy. After multivariate logistic regression analysis only one (of the five) ultrasonographic features analyzed was statistically significant, echogenicity. The authors used the beta-coefficients derived from the multivariate model logistic regression to determine the scoring and weight for each included covariate. The result was a

predictive tool including only one ultrasonographic feature (echogenicity) and requiring all lymph nodes to have been imaged via PET scan for the score to be applied. Based on the beta-coefficients, the presence of heterogeneous echogenicity and a lymph SUV percentage greater than 60 equaled a score of two, all other categories equaled scores of zero or one. With this tool, a lymph node could score a maximum 5 points. Any lymph node scoring $\leq 1$ or $\geq 2$ was considered low-risk or high-risk for malignancy, respectively. The development of this predictive tool included more rigorous methodology and statistics than the previously described studies, however the lack of included ultrasonographic features and requirement for PET scans limits the applicability of this tool. For this study, majority of the lymph nodes analyzed were enlarged or fluoro-deoxy-glucose (FDG) avid. Therefore, risk stratification based on this tool can only be applied to enlarged or FDG avid LNs as it has yet to be proven in small, PET-negative LNs.

Several tools have been developed to evaluate the malignancy status of lymph nodes seen during EBUS procedures. However, there is inconsistency between these studies with respect to the methodology being applied. To develop a predictive tool a model must be developed using regression analyses, model calibration must be assessed, and the discriminatory ability of the model must be determined (Han et al., 2016). Each of these steps must be completed prior to

using beta-coefficients to develop the scoring weights and values for the included covariates. To the best of our knowledge, the Canada Score is the only predictive tool to incorporate each of these steps.

5.3 Inter-Rater Reliability of the Canada Score:

For predictive tools to be used in clinical settings, it is critically important that the clinicians using the tool are consistently able to identify the ultrasonographic features correctly. Of the four existing ultrasonographic feature predictive tools, the study completed by Alici et al (2016). was the only one that did not report a formal inter-reliability assessment. The Schmid-Bindert et al. (2012) tool reported raw agreement for each ultrasonographic feature assessed. The lowest reported raw agreement value was 0.80 (95% CI: 0.75-0.84) for colour power Doppler index. The authors also completed chance-corrected agreement (kappa) and chance-independent agreement, however neither were reported. Within the literature, reporting raw percentage agreement as a measure of IRR has been rejected and widely considered to be inadequate (Hallgren, 2012). The largest criticism against reporting raw percentage agreement is that it does not correct for agreements expected by chance (Hallgren, 2012). The result is a value that likely overestimates the true level of agreement amongst raters. Shafiek et al. (2014) used the interval-to-interval method, a statistical procedure similar to calculating raw percent agreement to assess IRR. The authors used an inter-rater

agreement (IRA) level of 80.00% as an indicator of a positive agreement, the lowest reported IRA was 81.60% for hyperechogenic density in the LN interior. As with the IRR results reported in the Schmid-Bindert et al. (2012) study, the interval-to-interval method has the same limitations and likely resulted in inflated IRA values. The Evison et al. (2015) study assessed IRR using the Cohen's kappa statistic. The lowest reported kappa value was 0.40 (95% CI: 0.13-0.67) for margin status. The Cohen's kappa statistic is capable of testing whether the level of agreement seen between binary ratings exceed chance, however there are limitations. Cohen's kappa is a statistic influenced by trait prevalence and base-rates which make these statistics often incomparable across different studies or populations (Thompson & Walter, 1988; Feinstein & Cicchetti, 1990).

Compared to the IRAs reported by Schmid-Bindert et al. (2012), Shafiek et al. (2014), and Evison et al. (2015), our results are significantly lower. This may be a result of the statistics used to assess IRA for those studies compared to Gwet's AC1. The lowest and highest reported IRAs for ultrasonographic features were 0.25 (95% CI: 0.18-0.31) and 0.77 (95% CI: 0.72-0.82) for echogenicity and central necrosis, respectively. The low level of agreement may also be related to the relative limited experience clinicians have with identifying ultrasonographic features. The use of ultrasonographic features during EBUS procedures has been studied, however it has yet to be widely adopted in clinical practice nor taught

during medical education. Further education and training in proper identification of ultrasonographic features during the formative years of medical education is required to improve IRR. Unlike previous published studies, we assessed the reliability of the ultrasonographic features and of the predictive tool. Using a cut-off of three as an indicator for LN malignancy, clinician-raters (n=12) had a Gwet's AC1 value of 0.81 (95% CI: 0.77-0.85). This indicates that when using the Canada Score clinicians are often capable of agreeing when a LN has the characteristics present that likely suggest malignancy.

## 5.4 Clinical Implications of the Results:

Clinically, the results of the Canada Score suggest that the malignancy status of LNs can be predicted based on the presence of several ultrasonographic features. Statistical analyses revealed that a score of three out of four on the Canada Score correctly classified 84.67% of the lymph nodes and had a specificity of 96.34%. The high specificity associated with a Canada Score equal to three suggests that the tool is highly capable of detecting the presence of malignant LNs. Positive likelihood ratios (LR) provide an indication of how much more likely it is for a test to give a positive result compared to an individual without disease (Moosapour et al., 2011). It is accepted that the further away a positive likelihood ratio is from one, the more valuable it is towards making a clinical diagnosis. A positive LR between five and 10 and greater than 10 are

considered to have moderate and large effects on increasing the probability of disease, respectively (Moosapour et al., 2011). Using the Canada Score a LN scoring 3 had a positive LR of 8.60 and a LN scoring 4 had a positive LR of 22.78 (Table 7). These values indicate that LN scoring three or four is 8.60 and 22.78 times more likely to be malignant compared to a LN without those scores, respectively. The pathological information generated from identifying ultrasonographic features and using the Canada Score may have a profound impact in the event of insufficient biopsy results. Clinicians can confidently determine whether a repeat EBUS procedure needs to be completed, for example, in the case of a LN score $\geq 3$.

## 5.5 Limitations, Next Steps, & Future Endeavours:

This study is not without limitations. The first limitation is that our study did not assess the ultrasonographic features of noncancerous adenopathy. Only patients with confirmed or suspected NSCLC or esophageal cancer were included. Lymphadenopathy is not only present in patients with cancer. Mediastinal lymph nodes in patients with sarcoidosis, certain autoimmune disease, and tuberculosis can also exhibit similar ultrasonographic features presented above (Alici et al., 2016; Fujiwara et al., 2010). In this study we have isolated the ultrasonographic features predictive of malignancy and developed a predictive tool without

considering benign causes of lymphadenopathy, thus the tool developed is more powerful for comparing the ultrasonographic features of malignant LNs and normal LNs. Future studies should include benign causes of lymphadenopathy during analysis to understand which features are more likely associated with malignancy, benign lymphadenopathy, or normal LNs. Secondly, some of the ultrasonographic features assessed are subjective. For example, the small axis measurement is dependent on how the endobronchial ultrasound endoscope is maneuvered by the endoscopist. This may impact the ability for the external validity and reliability of this feature to be properly assessed. Future endeavors related to this research should include formal external validation of the Canada Score. Application of the tool in different health care institutions with other segments of the lung and esophageal cancer population would enable widespread generalization. In conclusion, the use of ultrasonographic features can accurately predict the malignancy status of mediastinal lymph nodes during EBUS procedures. The Canada Score was developed following sound methodology and if used in clinical settings may reduce the number of required repeat EBUS procedures.

**Works Cited:**

Akissue de Camargo Teixeria P, Chala LF, Shimizu C, et al. (2017). Axillary Lymph Node Sonographic Features and Breast Tumor Characteristics as Predictors of Malignancy: A Nomogram to Predict Risk. *Ultrasound Med Biol*, 43(9):1837-1845. http:// doi: 10.1016/j.ultrasmedbio.2017.05.003.

Alici I, Demirci N, Yilmaz A, Karakaya J, & Ozaydin E. (2016). The sonographic features of malignant mediastinal lymph nodes and a proposal for an algorithmic approach for sampling during endobronchial ultrasound. *Clin Respir J*, 10, 606-613.

Ayub I, Mohan A, Madan K, et al. (2016). Identification of specific EBUS sonographic characteristics for predicting benign mediastinal lymph nodes. *Clin. Respir. J*, 1-10.

Bertolini G, D'Amico R, Nardi D, et al. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidemiol Biostat*, 5(4):251-3.

Blood E & Spratt KF. (2007). *Disagreement on Agreement: Two Alternative Agreement Coefficients*. Retrieved from: http://www2.sas.com/proceedings/forum2007/186-2007.pdf (January 2018)

Canadian Cancer Society's Advisory Committee on Cancer Statistics. (2017) *Canadian Cancer Statistics 2017*. Toronto, ON. Retrieved from: cancer.ca/Canadian-Cancer- Statistics-2017-EN.pdf (December 2017)

Canadian Institute for Health Information. (2011). *Surgery for Pancreatic and Esophageal Cancer in Canada: Hospital Experience and Care Centralization*. Retrieved from: https://secure.cihi.ca/estore/productFamily.htm?pf=PFC1655&lang=en&media=0 (May 2018)

El-Sherief AH, Lau CT, Wu CC, et al. (2014). International Association for the Study of Lung Cancer (IASLC) Lymph Node Map: Radiologic Review with CT Illustration. *RadioGraphics*, 34(6): 1681-91.

Evison M, Morris J, Martin J, et al. (2015). Nodal Staging in Lung Cancer: A Risk Stratification Model for Lymph Nodes Classified as Negative by EBUS-TBNA. *J Thorac Oncol*, 10, 126-133.

Feinstein AR & Cicchetti DV. (1990) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543-9.

Fujiwara T, Yasufuku K, Nakajima T, et al. (2010). The Utility of Sonographic Features During Endobronchial Ultrasound Guided Transbronchial Needle Aspiration for Lymph Node Staging in Patients with Lung Cancer:

A Standard Endobronchial Ultrasound Image Classification System. *Chest*, 138(3), 641-647.

Gelberg J, Grondin S, Tremblay A. (2014). Mediastinal staging for lung cancer. *Can Respir J,* 21(3):159-161.

Gogia P, Insaf T, McNulty W, et al. (2015). Endobronchial ultrasound: morphological predictors of benign disease. *ERJ Open Res*, 1-8.

Hallgren KA. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*, 8(1):23-4.

Han K, Song K, & Choi B. (2016). How to Develop, Validate, and Compare Clinical Prediction Models Involving Radiological Parameters: Study Design and Statistical Methods. *Korean J Radiol*, 17(3).

Harris K, Modi K, Kumar A, & Dhillon SS. (2015, July). Endobronchial ultrasound-guided transbronchial needle aspiration of pulmonary artery tumors: A systematic review (with video). *Endosc Ultrasound*, *4*(3), 191-197. doi:10.4103/2303-9027.162996 PMCID: PMC4568630 PMID: 26374576

Jalil B, Yasufuku K, Khan A. (2015). Uses, limitations, and complications of endobronchial ultrasound. *Proc (Bayl Univ Med Cent),* 28(3), 325-330.

Jhun B, Um S, Suh G, et al. (2014). Clinical Value of Endobronchial Ultrasound Findings for Predicting Nodal Metastasis in Patients with Suspected

Lymphadenopathy: A Prospective Study. *J Korean Med Sci*, 29, 1632-1638. http://dx.doi.org/10.3346/jkms.2014.29.12.1632

Jones SR, Carley S, & Harrison M. (2003). An introduction to power and sample size estimation. *Emerg Med J*, 20:453-8.

Moosapour H, Raza M, Rambod M, & Soltani A. (2011). Conceptualization of category-oriented likelihood ratio: a useful tool for clinical diagnostic reasoning. *BMC Med Educ*, 11:94.

National Institute for Health and Care Excellence. (2011). *Lung cancer: diagnosis and management*. Retrieved from: https://www.nice.org.uk/guidance/cg121/chapter/1-Guidance#diagnosis-and-staging

Ortakoylu M, Iliaz S, Bahadir A. (2015). Diagnostic value of endobronchial ultrasound-guided transbronchial needle aspiration in various lung diseases. *J. bras. pneumol*, 41(5), 410-414.

Otterstatter MC, Brierley JD, De P et al. (2012). Esophageal cancer in Canada: Trends according to morphology and anatomical location. *Can J Gastroenterol*, 26(10):723-728.

Schmid-Bindert G, Jiang H, Kahler G, et al. (2012). Predicting malignancy in mediastinal lymph nodes by endobronchial ultrasound: a new ultrasound scoring system. *Respirology*, 17,1190-8.

Schmidt RL & Factor RE. (2013). Understanding Sources of Bias in Diagnostic Accuracy Studies. *Archives of Pathology & Laboratory Medicine*, 137(4), 558-65.

Shafiek H, Fiorentino F, Peralta AD, et al. (2014). Real-time prediction of mediastinal lymph node malignancy by endobronchial ultrasound. *Arch Bronconeumol, 50(6):228-234*.

Silvestri GA, Gonzales AV, Jantz MA, et al. Methods of staging non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd edn. American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2013;143(Suppl):e211S–e250S. doi: 10.1378/chest.12-2355.

Sun YS, Lyu HJ, Zhao YR, et al. (2017). Risk factors for central neck lymph node metastases of papillary thyroid carcinoma. *Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi, 52*(6):421-25.

Thompson WD & Walter SD. (1988) Kappa and the concept of independent errors. *Journal of Clinical Epidemiology*, 41: 969-70.

Wang-Memoli J, El-Bayoumi E, Pastis N, et al. (2011). Using Endobronchial Ultrasound Features to Predict Lymph Node Metastasis in Patients with Lung Cancer. *Chest*, 140(6), 1550-1556.

M.Sc. Thesis – D. Hylton; McMaster University – Health Research Methodology.

Wongpakaran N, Wongpakaran T, Wedding D, & Gwet K. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:61.

## Appendix 1: Tables and Figures

**Table 1. Patient baseline demographics and pathological diagnosis of biopsied and scored lymph nodes**

| | **Population Size [Patients] (n= 140) Sample Size = 300 LNs** |
|---|---|
| **Age (years) [mean ± SD]** | 67.92 ± 10.64 |
| **Males: n (%)/ females: n (%)** | 76 (54.30%)/ 64 (45.70%) |
| | |
| **Pre-planned imaging studies completed** | |
| **MRI, n (%)** | 27 (19.3%) |
| **Head CT, n (%)** | 10 (7.1%) |
| **Chest CT or PET, n (%)** | 139 (99.29%) |
| | |
| **Average Number of LNs Scored/Biopsied per Patient** | 2.14 ± 0.95 |
| | |
| **Scored and biopsied LNs (n=300)** | |
| **7, n (%)** | 131 (43.67%) |
| **4R, n (%)** | 85 (28.33%) |
| **4L, n (%)** | 54 (18.00%) |
| **10, n (%)** | 13 (4.33%) |
| **11, n (%)** | 6 (2.00%) |
| **Other (1, 2R, 2L, 12), n (%)** | 10 (3.33%) |
| | |
| **Pathology diagnosis: malignant cases** | |
| ***Primary lung cancer*** | n= 77 (55.00%) |
| **Adenocarcinoma, n (%)** | 33 (42.86%) |
| **Squamous cell carcinoma, n (%)** | 25 (32.47%) |
| **Other, n (%)** | 16 (20.78%) |
| ***Primary esophageal cancer*** | n= 32 (22.86%) |
| **Adenocarcinoma, n (%)** | 28 (87.50%) |
| **Squamous cell carcinoma, n (%)** | 3 (9.38%) |
| **Other, n (%)** | 1 (3.12%) |

| | |
|---|---|
| **Pathology Diagnosis: benign cases** | 31 (22.14%) |
| | |
| **Pathological Diagnosis: Lymph Nodes** | |
| **Malignant, n (%)** | n= 54 (18.00%) |
| **Benign, n (%)** | n= 246 (82.00%) |

LN = lymph node
SD = standard deviation
MRI = magnetic resonance imaging
CT = computed tomography
PET = positron emission tomography

**Table 2. Ultrasonographic Feature Presence in Malignant and Benign Lymph Nodes**

| Ultrasonographic Feature | Malignant (Pathologically determined) | Benign (Pathologically determined) | Pearson's Chi Square Value | P-Value |
|---|---|---|---|---|
| **Central Hilar Structure** | | | | |
| **Absence** | 39 | 142 | 15.92 | <0.0001 |
| **Presence** | 15 | 104 | | |
| **Echogenicity** | | | | |
| **Heterogeneous** | 23 | 104 | 0.0018 | 0.966 |
| **Homogeneous** | 31 | 142 | | |
| **Margins** | | | | |
| **Well-defined** | 24 | 37 | 23.6336 | <0.0001 |
| **Ill-defined** | 30 | 209 | | |
| **Small Axis Length** | | | | |
| **≥ 10 mm** | 28 | 177 | 11.4670 | 0.001 |
| **< 10 mm** | 26 | 69 | | |
| **Shape** | | | | |
| **Round** | 27 | 138 | 0.6652 | 0.415 |
| **Non-round** | 27 | 108 | | |
| **Central Necrosis** | | | | |
| **Presence** | 9 | 10 | 11.8534 | 0.001 |
| **Absence** | 45 | 236 | | |
| **Sample size = 300 LNs** | | | | |

LN = Lymph Node

**Table 3. Univariate Analyses for Ultrasonographic Features with Logistic Regression**

| Ultrasonographic Features | Odds Ratio | 95% Confidence Interval | Std. Error | P-Value |
|---|---|---|---|---|
| **Central Hilar Structure** | 3.55 | 1.86-6.78 | 1.17 | <0.0001 |
| **Echogenicity** | 1.01 | 0.56-1.84 | 0.31 | 0.966 |
| **Margins** | 4.52 | 2.38-8.57 | 1.48 | <0.0001 |
| **Small Axis Length** | 2.76 | 1.51-5.04 | 0.85 | 0.001 |
| **Shape** | 1.28 | 0.71-2.30 | 0.38 | 0.415 |
| **Central Necrosis** | 4.72 | 1.82-12.27 | 2.30 | 0.001 |

**Table 4. Multivariate Analyses for Ultrasonographic Features with Logistic Regression**

| Ultrasonographic Features | Odds Ratio | 95% Confidence Interval | Std. Error | P-Value |
|---|---|---|---|---|
| **Central Hilar Structure (Absence vs. Presence)** | 2.34 | 1.14-4.81 | 0.86 | 0.021 |
| **Small Axis Length (≥ 10 mm vs. < 10mm)** | 2.49 | 1.29-4.80 | 0.83 | 0.006 |
| **Margin (Well-defined vs. Ill-defined)** | 2.95 | 1.42-6.13 | 1.10 | 0.004 |
| **Central Necrosis (Presence vs. Absence)** | 2.51 | 0.85-7.39 | 1.38 | 0.096 |
| **Constant** | 0.06 | 0.03-0.12 | 0.02 | <0.0001 |

**Table 5. Canada Score Development**

| Covariates | ß | Categories | Ref-erence Value | ß (W – $W_{ref}$) | Points = ß (W – $W_{ref}$)/ $ß_{CHS}$ | Allotted Points |
|---|---|---|---|---|---|---|
| **Central Hilar Structure** | 0.85 | Absent | 0 ($W_{ref}$)* | 0.85(1-0) | 0.85/0.85 | 1 |
| | | Present | 1 (W) | | | 0 |
| **Small Axis Length** | 1.08 | ≥ 10 mm | 0 ($W_{ref}$)* | 1.08(1-0) | 1.08/0.85 | 1 |
| | | < 10 mm | 1(W) | | | 0 |
| **Margin** | 0.91 | Well-defined | 0 ($W_{ref}$)* | 0.91(1-0) | 0.91/0.85 | 1 |
| | | Ill-defined | 1(W) | | | 0 |
| **Central Necrosis** | 0.92 | Absent | 0 ($W_{ref}$)* | 0.92(1-0) | 0.92/0.85 | 1 |
| | | Present | 1(W) | | | 0 |

* = Reference category
ß = Beta coefficient

**Table 6. Canada Score Logistic Regression**

| Canada Score Values | Odds Ratio | 95% Confidence Interval | Std. Error | P-Value |
|---|---|---|---|---|
| **1 (.vs 0)** | 1.21 | 0.49-3.01 | 0.56 | 0.680 |
| **2 (.vs 0)** | 3.52 | 1.44-8.57 | 1.60 | 0.006 |
| **3 (.vs 0)** | 15.17 | 4.92-46.79 | 8.72 | <0.0001 |
| **4 (.vs 0)** | 50.56 | 5.31-481.39 | 58.13 | 0.001 |
| **Constant** | 0.10 | 0.05-0.20 | 0.03 | <0.0001 |

**Table 7. Canada Score Values Diagnostic Statistics**

| Canada Score Value | Sensitivity (%) | Specificity (%) | Percent Correctly Classified | Positive Likelihood Ratio | Negative Likelihood Ratio |
|---|---|---|---|---|---|
| **0** | 100.00 (99.99-100) | 0.00 (0.00-0.00) | 18.00 | 1.00 | N/A |
| **1** | 83.33 (79.08-87.52) | 36.99 (31.53-42.45) | 45.33 | 1.32 (0.94-1.70) | 0.45 (0.39-0.51) |
| **2** | 61.11(55.59-66.62) | 77.64 (72.93-82.35) | 74.67 | 2.73 (2.22-3.23) | 0.51 (0.26-0.76) |
| **3** | 31.48(26.22-36.74) | 96.34 (94.21-98.46) | 84.67 | 8.60 (5.43-11.77) | 0.71 (0.42-1.00) |
| **4** | 9.26(5.98-12.54) | 99.59 (98.87-100) | 83.33 | 22.78 (18.03-27.53) | 0.91 (0.58-1.24) |

**Table 8. Reliability Assessment for Ultrasonographic Features**

| Ultrasonographic Feature | Percent Agreement | Gwet's AC1 Value (± SD) | 95% Confidence Interval |
|---|---|---|---|
| **CHS** | 73.34% | 0.49 ± 0.04 | 0.42 - 0.56 |
| **Echogenicity** | 62.17% | 0.25 ± 0.03 | 0.18 - 0.31 |
| **Margins** | 62.62% | 0.29 ± 0.04 | 0.22 - 0.36 |
| **Small Axis Length** | Not analyzed | | |
| **Shape** | Not analyzed | | |
| **Central Necrosis** | 81.73% | 0.77 ± 0.02 | 0.72 - 0.82 |

SD = standard deviation

**Table 9. Reliability Assessment for the Canada Score: Three Rater & Twelve Rater Comparison**

| | Percent Agreement | Gwet's AC1 Value (± SD) | 95% Confidence Interval |
|---|---|---|---|
| **3 Rater Comparison (n=900)** | 76.56% | 0.74 ± 0.02 | 0.70 - 0.79 |
| **12 Rater Comparison** | 41.70% | 0.29 ± 0.02 | 0.25 - 0.33 |

SD = standard deviation

**Table 10. Reliability Assessment for Agreement between Clinician-Raters Using Canada Score 3 as Malignancy Cut Off**

| | Percent Agreement | Gwet's AC1 Value (± SD) | 95% Confidence Interval |
|---|---|---|---|
| **3 Rater Comparison (n=900)** | 84.44% | 0.80 ± 0.02 | 0.75 - 0.85 |
| **12 Rater Comparison** | 85.36% | 0.81 ± 0.02 | 0.77 - 0.85 |

SD = standard deviation

**Table 11. Shafiek et al. Multivariate Logistic Regression Model**

| Ultrasonographic Features | Odds Ratio | 95% Confidence Interval | Std. Error | P-Value |
|---|---|---|---|---|
| **Central Hilar Structure (Absence vs. Presence)** | 2.37 | 1.15-4.89 | 0.88 | 0.020 |
| **Small Axis Length (≥ 10 mm vs. < 10mm)** | 2.68 | 1.38-5.20 | 0.91 | 0.004 |
| **Margin (Well-defined vs. Ill-defined)** | 3.26 | 1.59-6.68 | 1.19 | 0.001 |

| | | | | |
|---|---|---|---|---|
| **Shape (Round vs. Non-Round)** | 1.21 | 0.63-2.31 | 0.40 | 0.568 |
| **Echogenicity (Heterogeneous vs. Homogeneous)** | 1.00 | 0.52-1.95 | 0.34 | 0.992 |
| **Constant** | 0.06 | 0.03-0.12 | 0.02 | <0.0001 |

**Table 12. Shafiek et al. Model Score Predictive Capability**

| **Shafiek et al. Model Score** | **Odds Ratio** | **Std. Error** | **P-Value** | **95% Confidence Interval** |
|---|---|---|---|---|
| **1** | 1.45 | 0.79 | 0.640 | 0.31 - 6.74 |
| **2** | 0.68 | 1.14 | 0.739 | 0.07 - 6.43 |
| **2.5** | 1.09 | 0.64 | 0.890 | 0.31 - 3.85 |
| **3** | 4.10 | 0.85 | 0.100 | 0.77 - 21.74 |
| **3.5** | 2.34 | 0.59 | 0.150 | 0.73 - 7.51 |
| **4** | 0.51 | 1.13 | 0.560 | 0.06 - 4.73 |
| **4.5** | 18.45 | 0.77 | <0.0001 | 4.12 - 82.66 |
| **5** | 5.22 | 0.68 | 0.015 | 1.38 - 19.66 |
| **6** | 49.20 | 1.18 | 0.001 | 4.88 - 496.48 |

**Table 13. Shafiek et al. Score Performance & Diagnostic Statistics**

| Canada Score Value | Sensitivity (%) | Specificity (%) | Percent Correctly Classified | Positive Likelihood Ratio | Negative Likelihood Ratio |
|---|---|---|---|---|---|
| **0** | 100.00 | 0.00 | 18.00 | 1.00 | N/A |
| **1** | 90.74 | 16.67 | 30.00 | 1.09 | 0.56 |
| **2** | 85.19 | 23.58 | 34.67 | 1.11 | 0.63 |
| **2.5** | 46.30 | 65.85 | 62.33 | 1.36 | 0.82 |
| **3** | 83.33 | 28.46 | 38.33 | 1.16 | 0.59 |
| **3.5** | 35.19 | 84.15 | 75.33 | 2.22 | 0.77 |
| **4** | 77.78 | 30.89 | 39.33 | 1.13 | 0.72 |
| **4.5** | 16.67 | 98.37 | 83.67 | 10.25 | 0.85 |
| **5** | 75.93 | 37.40 | 44.33 | 1.21 | 0.64 |
| **6** | 62.96 | 41.87 | 45.67 | 1.08 | 0.88 |

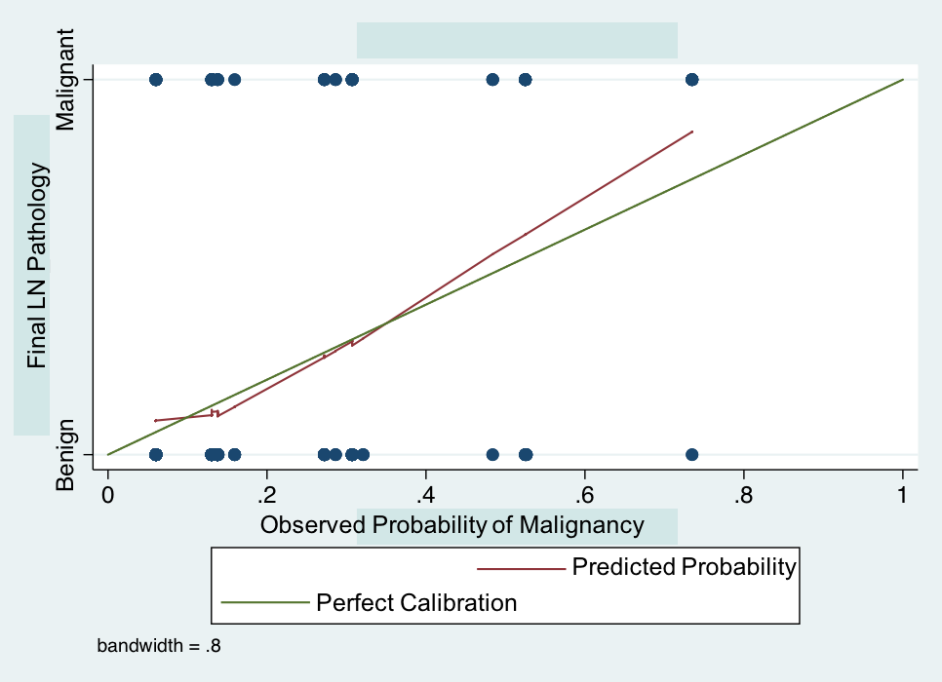


**Figure 1. Canada Score multivariate model calibration plot**

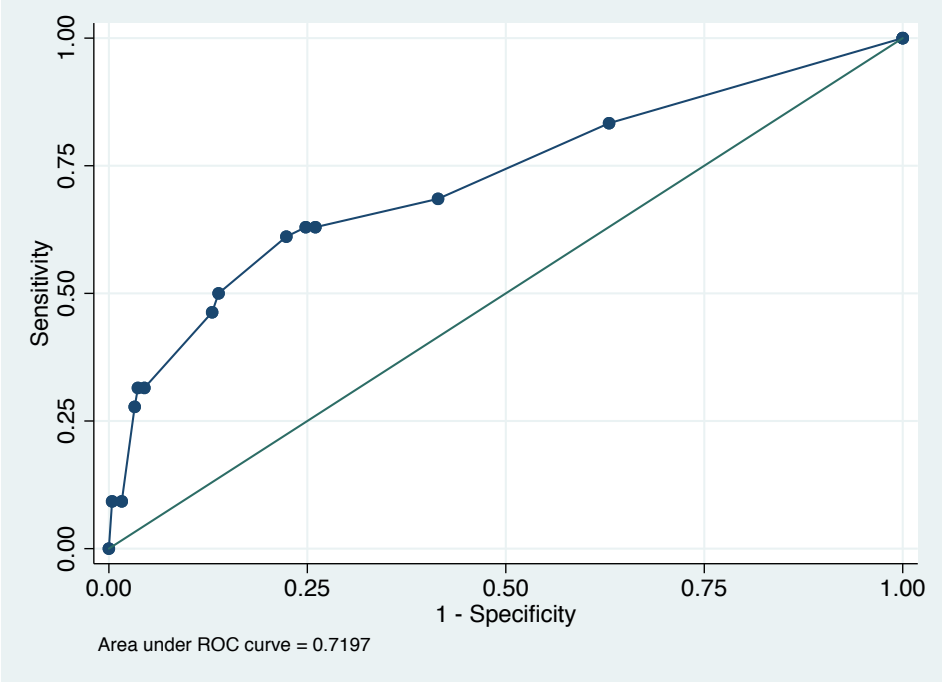


**Figure 2. Canada Score multivariate model receiver operator characteristic curve (c-index = 0.72, std. error = 0.04, 95% CI= 0.64-0.80)**

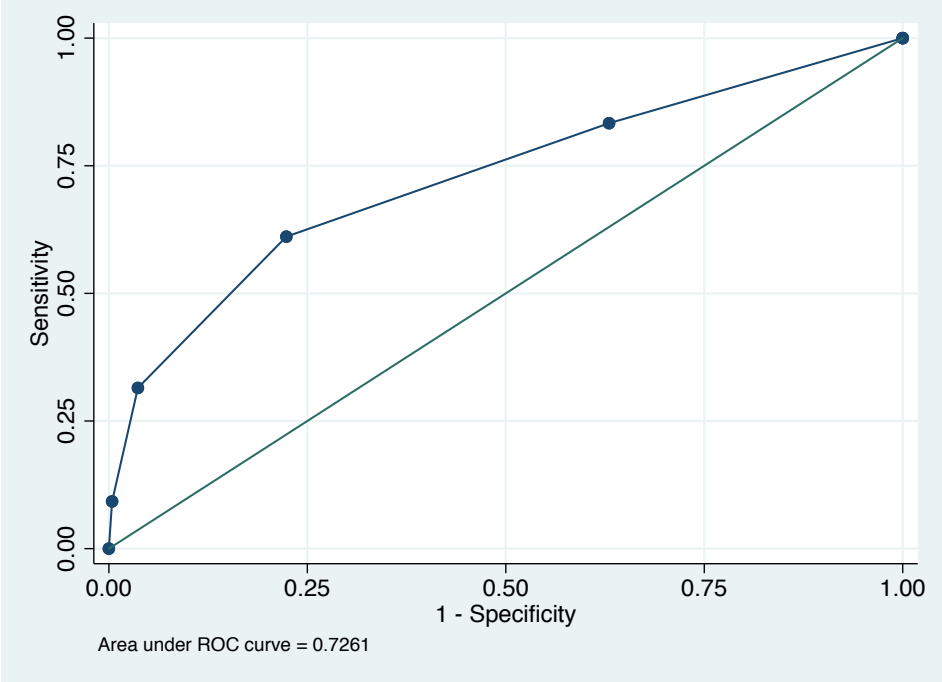


**Figure 3. Canada Score receiver operator characteristic curve (c-index = 0.73, std. error = 0.04, 95% CI = 0.65-0.81)**

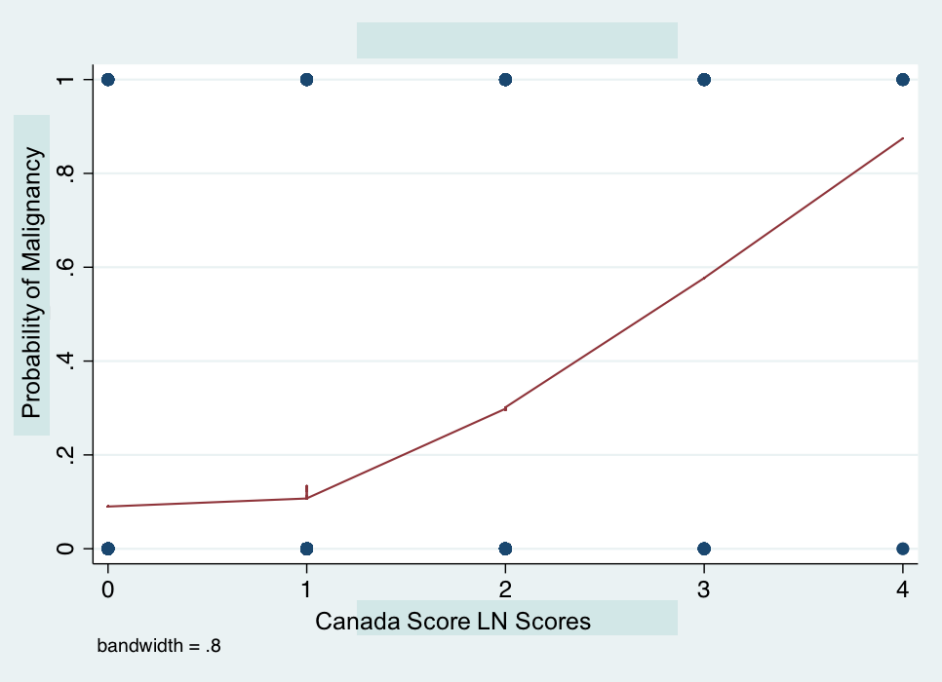


**Figure 4. Malignancy Cut-off Determination for the Canada Score**

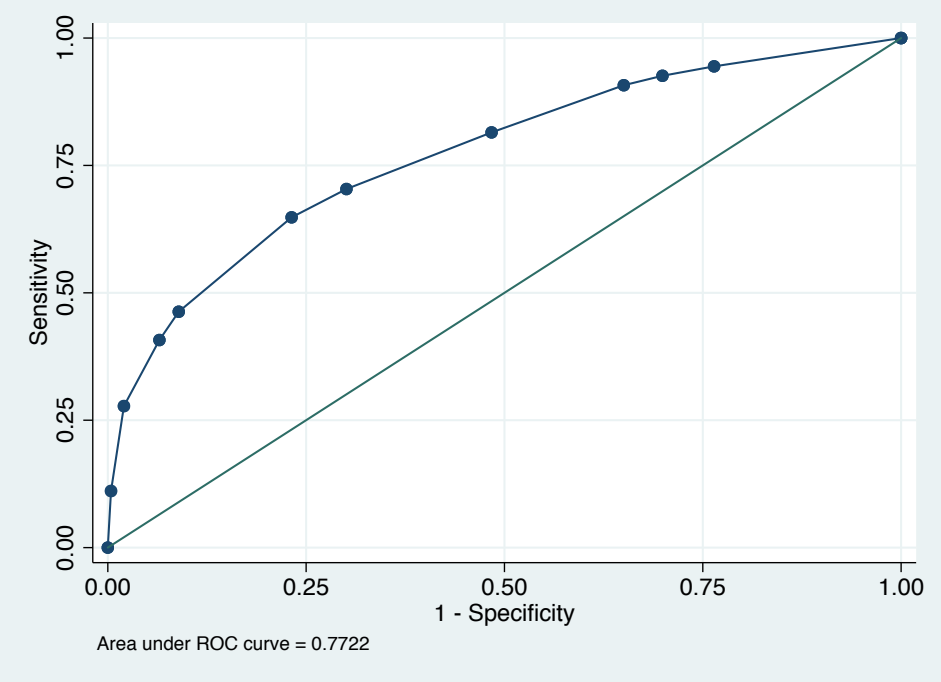


**Figure 5. ROC curve for the Shafiek et al. tool (c-index = 0.77, std error = 0.04, 95% CI: 0.70-0.85)**

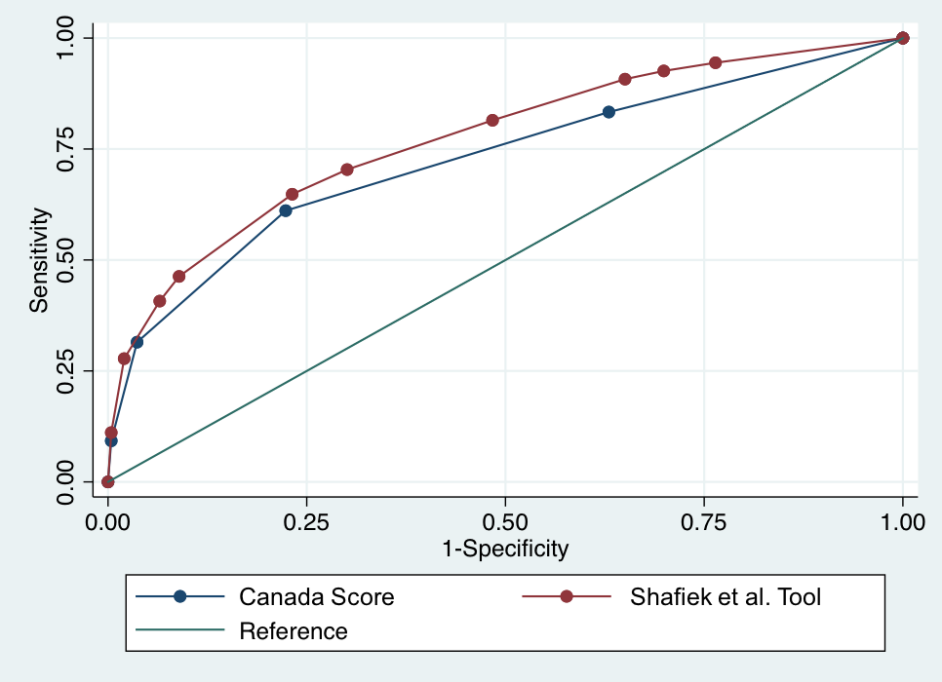


**Figure 6. Chi-square comparison of the ROC curves of the Shafiek et al. tool (c-index= 0.77) and the Canada Score (c-index= 0.73)**