Improving lung cancer outcomes: risk prediction of postoperative readmission and assessing the efficacy of smoking cessation interventions

Ankita Ghatak
Department of Medicine, Division of Experimental Medicine, McGill University, Montreal

April 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

© Ankita Ghatak, 2022

Table of Contents

ABSTRACT

Introduction

Project 1: Lung cancer remains the leading cause of cancer death in Canada and despite advances in technology, screening and treatment, 5-year survival for lung cancer has yet to see significant improvement. Currently, surgical resection is the primary treatment for early-stage lung cancer patients. Studies have shown that post-operative re-admission is a strong predictor of 90-day mortality. Postoperative complications are a negative predictor of overall survival. Current models for perioperative risk assessment provide are dated, developed on general surgical populations, including non cancer cases, and do not use the wealth of clinical data currently available in electronic health records. They have been shown to be largely inaccurate in risk prediction specific to lung cancer patients undergoing resection.

Project 2: Smoking cessation intervention when in conjunction with lung cancer screening has been shown to further reduce lung cancer mortality. However, the efficacy of various smoking cessation services integrated in a lung cancer screening program has yet to be established. The efficacy of smoking cessation phone counselling in this population is unknown. It has been assumed that lung cancer screening participants referred to smoking cessation phone lines are similar to other participants referred by healthcare workers, but this claim has not been validated.

Objectives:

Project 1: We will predict 90-day post operative re-admission in lung cancer patients treated by lobectomy. We will compare the predictive performance of a Random Forests model, which can

handle a large number of predictors, to a traditional logistic regression model. Predictor importance using the feature importance metric from the Random Forests.

Project 2: To assess the effectiveness of smoking cessation by phone counselling in the McGill lung cancer screening study. We will compare smoking cessation rates and smoking behaviours between participants referred by the McGill lung cancer screening program to those referred to the smoking cessation quit line by healthcare workers in the community those who self refer.

Methods:

Project 1: This was a retrospective cohort study of lung cancer patients resected at the MUHC from 2016-2019. Data was extracted from the National Surgical Quality Improvement Program (NSQIP) database, the RI-MUHC Datawarehouse and MUHC Pulmonary Function Testing database. The inclusion criterion was patients resected by lobectomy with a pathologic diagnosis of lung cancer. Patients with bi-lobectomies, segmentectomies, wedge resections were excluded. Those with an index ICU admission prior to lobectomy were excluded. The primary outcome was 90-day postoperative readmission. We used two statistical models to predict 90 day readmission: logistic regression and random forest models. We compared sensitivity, specificity, accuracy, and false positive rate and AUC between models. We assessed random forests feature importance (calculated using Gini impurity). In the logistic regression model, we used age, gender, body mass index, history of severe COPD and Charlson Comorbidity Index as predictors. The random forest model has no limit on the number predictors, and we therefore used all available predictors.

Project 2: We conducted a retrospective cohort study of active smokers referred to Quebec's smoking cessation phoneline. The cohort was defined by three groups: self-referred participants, healthcare worker referred participants from the community and McGill lung cancer screening referred participants. Variables considered include sociodemographic information, smoking history, history of mental health disorder and quit intentions (stage of change, readiness for change, previous use of quit programs, and previous quit attempts). The primary outcome was self-reported 30-day abstinence rates at 6 months. Multivariable logistic regression was used to identify whether group assignment was associated with higher quit rates.

Results:

Project 1: The results of our logistic regression suggested that the most important predictors of our outcome were gender (male) (OR 2.38, CI: 1.20-2.87, $p<0.05$), history of severe COPD (True) (OR 2.1, CI: 1.16-3.2, $p<0.05$) and Charlson comorbidity index (per 1 unit increment) (OR 1.29, CI: 0.9-1.4, $p>0.05$). The Random Forests model ranked features predictors by importance as follows (in descending order): age, white blood cell count, hematocrit, albumin, sodium, creatinine, body mass index, sex, hypertension, Charlson comorbidity index, active smoking within 1 year of surgery and PFT availability. Dyspnea, history of severe COPD and diabetes all had zero feature importance and were ranked last. The random forests model also assessed the relationship between all available predictors and our outcome via the visualization of its decision trees. The random forests also assessed the relationship between all available features, including preoperative laboratory variables, and our outcome via the visualization of its decision trees. The random forests had better model performance based on all assessed metrics, most significantly on the AUC (0.72) compared to the logistic regression (AUC 0.59).

Project 2: Lung cancer screening program referred participants and the lowest six month quit rates amongst all groups (12%, 95% CI: 5-19%). Compared to self-referred participants, lung cancer screening referred participants were much less likely to quit (adjusted OR 0.37; 95% CI 0.17-0.8), whereas healthcare workers referred participants were twice as likely to quit (adjusted OR 2.16; 95% CI 1.3-3.58), despite adjustment for difference in smoking intensity and quit intentions.

Conclusions:

Project 1: Our first study suggests that the random forests model may be a better predictive model than a traditional logistic regression in predicting clinical outcomes such as readmission. Results from the random forests feature importance ranking suggest that using preoperative laboratory data for risk assessment is worth further validation. Random forests may be a useful tool for decision making and have several clinical applications.

Project 2: Our second study suggests that phone counselling alone has very limited benefit in a lung cancer screening program and lung cancer screening referred participants differ significantly from those who are otherwise referred by healthcare workers.

RÉSUMÉ

Introduction

Projet 1 : Le cancer du poumon demeure la principale cause de décès par cancer au Canada et, malgré les progrès de la technologie, du dépistage et du traitement, le taux de survie à 5 ans pour le cancer du poumon ne s'est pas encore beaucoup amélioré. Actuellement, la résection chirurgicale est le principal traitement des patients atteints d'un cancer du poumon à un stade précoce. Des études ont montré que la réadmission postopératoire est un facteur prédictif important de la mortalité à 90 jours. Les complications postopératoires sont un facteur prédictif négatif de la survie globale. Les modèles actuels d'évaluation du risque péri-opératoire sont dépassés, développés sur des populations chirurgicales générales, y compris des cas non cancéreux, et n'utilisent pas la richesse des données cliniques actuellement disponibles dans les dossiers médicaux électroniques. Il a été démontré qu'ils sont largement inexacts dans la prédiction du risque spécifique aux patients atteints de cancer du poumon et subissant une résection.

Projet 2 : Il a été démontré que l'intervention de désaccoutumance au tabac, lorsqu'elle est associée au dépistage du cancer du poumon, réduit davantage la mortalité due au cancer du poumon. Cependant, l'efficacité de divers services de désaccoutumance au tabac intégrés dans un programme de dépistage du cancer du poumon n'a pas encore été établie. On ne connaît pas l'efficacité du conseil téléphonique pour le sevrage tabagique dans cette population. Il a été supposé que les participants au dépistage du cancer du poumon orientés vers les lignes téléphoniques d'aide à l'arrêt du tabac sont similaires aux autres participants orientés par les travailleurs de la santé, mais cette affirmation n'a pas été validée.

Objectifs :

Projet 1 : Nous avons prédit la taux de réadmission postopératoire à 90 jours chez les patients atteints de cancer du poumon traités par lobectomie. Nous avons comparerés la performance prédictive d'un modèle Random Forests, qui peut gérer un grand nombre de prédicteurs, à un modèle de régression logistique traditionnel. L'importance des prédicteurs en utilisant la métrique de l'importance des caractéristiques des forêts aléatoires.

Projet 2 : Évaluer l'efficacité de l'abandon du tabac par des conseils téléphoniques dans le cadre de l'étude de McGill sur le dépistage du cancer du poumon. Nous comparerons les taux d'abandon du tabac et les comportements tabagiques entre les participants orientés par le programme de dépistage du cancer du poumon de McGill et ceux orientés vers la ligne d'aide à l'abandon du tabac par des travailleurs de la santé dans la communauté, et a ceux auto-référés.

Méthodes :

Projet 1 : Il s'agissait d'une étude de cohorte rétrospective de patients atteints de cancer du poumon ayant subi une résection au CUSM de 2016 à 2019. Les données ont été extraites de la base de données du Programme national d'amélioration de la qualité des interventions chirurgicales (NSQIP), de l'entrepôt de données RI-CUSM et de la base de données des tests de fonction pulmonaire du CUSM. Le critère d'inclusion était les patients réséqués par lobectomie avec un diagnostic de cancer du poumon. Les patients ayant subi une bilobectomie, une segmentectomie ou une résection en coin ont été exclus. Les patients ayant été admis dans une unité de soins intensifs avant la lobectomie ont été exclus. Le résultat primaire était la réadmission postopératoire à 90 jours. Nous avons comparé la sensibilité, la spécificité, la

précision, le taux de faux positifs et l'AUC dans les modèles de régression logistique et de forêts aléatoires. Nous avons également évalué l'importance des caractéristiques des forêts aléatoires (calculée en utilisant l'impureté de Gini). Dans le modèle de régression logistique, nous avons utilisé l'âge, le sexe, l'IMC, les antécédents de BPCO sévère et l'indice de comorbidité de Charlson. Le modèle de forêt aléatoire n'a pas de limite quant au nombre de prédicteurs et a utilisé tous les prédicteurs disponibles.

Projet 2 : Nous avons mené une étude de cohorte rétrospective des fumeurs actifs référés à la ligne téléphonique de désaccoutumance au tabac du Québec. La cohorte a été définie par trois groupes : les participants référés par eux-mêmes, les participants référés par des travailleurs de la santé de la communauté et les participants référés par le dépistage du cancer du poumon de McGill. Les variables prises en compte comprennent les informations sociodémographiques, les antécédents de tabagisme, les antécédents de troubles mentaux et les intentions d'abandon (stade de changement, volonté de changement, utilisation antérieure de programmes d'abandon et tentatives d'abandon antérieures). Le résultat primaire était le taux d'abstinence à 30 jours autodéclaré à 6 mois. Une régression logistique multivariable a été utilisée pour déterminer si l'affectation à un groupe était associée à des taux d'abandon plus élevés.

Résultats :

Projet 1 : Les résultats de notre régression logistique ont suggéré que les prédicteurs les plus importants de notre résultat étaient le sexe (masculin) (OR 2,38, IC : 1,20-2,87, $p<0,05$), les antécédents de BPCO sévère (Vrai) (OR 2,1, IC : 1,16-3,2, $p<0,05$) et l'indice de comorbidité de Charlson par incrément de 1 unité (OR 1,29, IC : 0,9-1,4, $p>0,05$). Le modèle Random Forests a

classé les caractéristiques prédictives par importance comme suit (par ordre décroissant) : âge, numération leucocytaire, hématocrite, albumine, sodium, créatinine ; IMC, sexe (masculin), hypertension, indice de comorbidité de Charlson, fumeur actif dans l'année précédant l'intervention et disponibilité de l'examen physique. La dyspnée, les antécédents de BPCO sévère et le diabète n'avaient aucune importance et ont été classés en dernier. Le modèle Random Forests a également évalué la relation entre tous les prédicteurs disponibles et notre résultat via la visualisation de ses arbres de décision. Le modèle Random Forests a également évalué la relation entre toutes les caractéristiques disponibles, y compris les variables de laboratoire préopératoires, et notre résultat via la visualisation de ses arbres de décision. Les forêts aléatoires ont eu une meilleure performance de modèle sur la base de tous les paramètres évalués, plus particulièrement sur l'AUC (0,72) par rapport à la régression logistique (0,59).

Projet 2 : Les participants référés par le programme de dépistage du cancer du poumon présentaient les taux d'abandon à six mois les plus faibles de tous les groupes (12 %, IC : 5-19 %). Comparativement aux participants qui se sont adressés à eux, les participants adressés au programme de dépistage du cancer du poumon étaient beaucoup moins susceptibles de cesser de fumer (OR ajusté 0,37 ; IC à 95 % 0,17-0,8), tandis que les participants adressés par des travailleurs de la santé étaient deux fois plus susceptibles de cesser de fumer (OR ajusté 2,16 ; IC à 95 % 1,3-3,58), malgré l'ajustement pour tenir compte de la différence dans l'intensité du tabagisme et le renoncement au tabac.

Conclusions :

Projet 1 : Notre première étude suggère que le modèle des forêts aléatoires peut être un meilleur modèle prédictif qu'une régression logistique traditionnelle pour prédire des résultats cliniques tels que la réadmission. Les résultats du classement par importance des caractéristiques des forêts aléatoires suggèrent que l'utilisation des données de laboratoire préopératoires pour l'évaluation des risques mérite une validation supplémentaire. Les forêts aléatoires peuvent être un outil utile pour la prise de décision et ont plusieurs applications cliniques.

Projet 2 : Notre deuxième étude suggère que le conseil téléphonique seul a un avantage très limité dans un programme de dépistage du cancer du poumon et que les participants orientés vers le dépistage du cancer du poumon diffèrent significativement de ceux qui sont orientés par d'autres travailleurs de la santé.

ACKNOWLEDGEMENTS

I am honoured to have been a Kuok Fellowship recipient for the year 2020 and 2021. I would like to express my sincere gratitude to the Rossy Cancer Network and the Gerald Bronfman Department of Oncology for enabling me to pursue the research I've always wanted to do.

There are many people I need to thank for their support. I would like to start by thanking my supervisor Dr Nicole Ezer. Firstly, I would like to thank her for giving me the opportunity to work on such a unique project. It has changed the course of my research career and indeed my life, for the better. I am grateful to have been chosen. Dr Ezer has been a constant source of knowledge and guidance throughout these years. I am fortunate to have had a supervisor who cared so deeply about my work and gave me so much of her time. She promptly responded to my questions, spent time guiding me and reviewing my work, even with her many duties as a clinician during the COVID-19 pandemic. Dr Ezer supported me fiercely through the challenges created by the pandemic with regards to data accessibility. This project would never have come to completion without her determination. When I first began this research project, the methods involved were entirely new to me. Nevertheless, Dr Ezer always encouraged me and guided me. I would like to thank her for believing in me so deeply – I could not have gone through this learning curve without it. Dr Ezer truly embodies the merits of an excellent researcher and has taught me how to understand and adapt to current research needs. Watching her work and working with her has contributed tremendously to my growth as a researcher. I would like to sincerely thank her for this – I am very lucky to call her my mentor. This thesis would not be possible without her support, help, and guidance. Her careful editing contributed greatly to the

production of this thesis, and I would once again like to mention how much I appreciate her efforts.

I would also like to thank my co-supervisor Dr Benjamin Smith for his constructive comments on this thesis.

My sincere thanks to Dr Andrea Benedetti for her guidance regarding the statistical methods used this thesis. She has been an incredible source of expertise and has guided me immensely throughout my graduate studies. I would also like to thank the bioinformatics team at the MUHC for their help with data management.

A very special thanks to Dr Jean Bourbeau, for his invaluable comments and support throughout my graduate degree. He greatly influenced the direction of this study and has helped me realize its potential. I would like to thank him and his team members for inviting me to their team meetings and giving me opportunities to present my work.

I would like to thank my thesis committee members who provided me insightful suggestions on my work as well as my academic advisor, Dr Andrew Bateman facilitating my committee meetings.

My heartfelt thanks to my dearest friends, Patricia Stamatelos and Catalina Karam for supporting me through all my personal struggles and always holding space for me through all these years.

They have profoundly shaped my post secondary career and life, and I am blessed to call them my friends.

I must express my earnest gratitude to my partner Adam Levine for always believing me, supporting me and being a such a source of hope and happiness. I am a better researcher and a better person because of him.

Finally, I would like to express my most profound gratitude to my parents. It is difficult to put into words how much they have supported me throughout my life and graduate studies. All my achievements, past, present, and future, are made possible because of them. My deepest thanks and love to them for their unwavering support, their sacrifices, and their endless love.

PREFACE AND CONTRIBUTION OF AUTHORS

One of the major goals of the projects constituting this Master thesis was contribute towards the improvement of lung cancer outcomes. This thesis complies with the Graduate and Postdoctoral Studies' guidelines and general requirements of a manuscript-based (article-based) Master's theses at McGill University. This thesis consists of two manuscripts that address important research topics related to lung cancer.

This thesis contains **six chapters**:

Chapter 1 provides a comprehensive literature review on lung cancer incidence and mortality, staging, screening, survival, smoking and risk assessment. It provides thorough explanations of the machine learning methods used in the preparation of this study.

Chapter 2 introduces the thesis rationale, hypothesis, and objectives of the 2 projects.

Chapter 3 to 5 include the two manuscripts, which constitute my thesis. Chapter 3 is the project on lung cancer (manuscript 1) that assesses the predictive performance of a random forests model compared to a logistic regression model for prediction of post operative readmission. Chapter 4 is the bridging chapter which highlights the importance of a holistic approach to lung cancer care to improve outcomes most effectively. Chapter 5 is the project on smoking cessation which assesses the effectiveness of phone counselling as a smoking cessation intervention in the McGill Lung Cancer screening program.

Chapter 6 summarises and discusses the overall findings and provides the final conclusions

Chapter 7 provides references for all parts of the thesis excluding the two manuscripts

My thesis supervisor, Nicole Ezer contributed instrumentally to every stage of research. From the conception and design of the projects to analysis and discussion of the results, and thoughtful manuscript revisions. Andrea Benedetti provided fundamental insight on analysis and the statistical methods used in this thesis. Daniel Jiminez and Cedric Roux of the bioinformatics team at the MUHC assisted in storing our data securely and cleaning the pulmonary function testing data.

All chapters were written and completed by Ankita Ghatak. Nicole Ezer reviewed and edited the text in this thesis. Benjamin Smith also reviewed the text in this thesis.

Chapter 3 (Manuscript 1) is still ongoing. Nicole Ezer contributed to the conception, planning and design of the study. Andrea Benedetti contributed to design of the statistical methods used in the study. Ankita Ghatak is the first author. Nicole Ezer is the senior author. Figures and tables are embedded in the manuscript.

Chapter 5 (Manuscript 2) is formatted and written according to the respective peer-reviewed journal's specifications. The manuscript in chapter 3 was submitted to the Canadian Respiratory Journal on September 09, 2021 and is currently under review. Note, a revised version of manuscript 2 was accepted to the Canadian Respiratory Journal on All authors made substantive intellectual contributions to the development of the study. Ankita Ghatak is the first author and submitting author. Nicole Ezer is the senior author and corresponding author. All authors

approved the version of the submitted manuscript, as it appears in this thesis. Figures and tables are embedded in the manuscript. Note: a revised version of manuscript 2 was accepted to the Canadian Respiratory Journal on March 21, 2022. All authors made substantive intellectual contributions to the development of the study.

CHAPTER 1: BACKGROUND

1.1 Lung Cancer Epidemiology: Incidence and Mortality

Globally, lung cancer has been the most common diagnosed cancer for the last several decades. In both the United States and Canada, it is the leading cause of cancer death across both sexes[1,2] . In Canada, it is the most diagnosed cancer and the leading cause of cancer death, accounting for 25% of all cancer deaths [2]. In 2020, lung cancer caused more deaths than colorectal, pancreatic and breast cancers combined[3]. The high mortality reflects both its high incidence and low survival[1]. This trend applies to Quebec as well. In Quebec, lung cancer is the leading cause of cancer death in women and men. Quebec has one of the highest age standardized incidence and mortality rates for lung cancer, of all Canadian provinces. As per Canadian Cancer Society estimates, 30% of all new lung cancer diagnoses occurred in Quebec (8900 out of 29,400) and 31% of all lung cancer deaths in Canada occurred in Quebec (6600 out of 21,000)[4].

1.2 Lung Cancer Survival

1.2.1 Staging

Note: There are two main types of lung cancer: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). They progress differently and are treated in different ways. NSCLC accounts for 85% of all cases and is the focus of this thesis[5]. Clinically, NSCLC are categorized using the TNM (tumour-nodes-metastasis) staging system (Figure 1)[6]. Broadly, the stages are defined as follows: ‘local’ is confined to the primary sites, ‘regional’ has spread to the regional lymph nodes, and ‘distant’ is cancer that has metastasized[6].

TABLE 5
Treatment of Lung Cancer According to Stage

STAGE	*PRIMARY TREATMENT*	*ADJUVANT THERAPY*	*FIVE-YEAR SURVIVAL RATE (%)*
Non–small cell carcinoma			
I	Resection	Chemotherapy	60 to 70
II	Resection	Chemotherapy with or without radiotherapy	40 to 50
IIIA (resectable)	Resection with or without preoperative chemotherapy	Chemotherapy with or without radiotherapy	15 to 30
IIIA (unresectable) or IIIB (involvement of contralateral or supraclavicular lymph nodes)	Chemotherapy with concurrent or subsequent radiotherapy	None	10 to 20
IIIB (pleural effusion) or IV	Chemotherapy or resection of primary brain metastasis and primary T1 tumor	None	10 to 15 (two-year survival)
Small cell carcinoma			
Limited disease	Chemotherapy with concurrent radiotherapy	None	15 to 25
Extensive disease	Chemotherapy	None	< 5

Figure 1. Staging of Lung Cancer. Adapted from Spira A, Ettinger DS. [6]

Analysis of SEER 18 (2017-2018) data showed that only 16 % of cases were diagnosed at the localized stage, while 22 % were diagnosed at the regional stage and 57 % were diagnosed at the distant stage (2003–2009, SEER data)[7]. Although improvements are expected due to implementation of screening, the data still suggests that the majority of patients still present with advanced stage disease at the time of diagnosis. Looking at survival rates by stage illustrates why stage at diagnosis is of major concern. We see that that 5 year survival for cases diagnosed at the localized stage is 56.3% compared to just 4.7% at the distant stage. Given that the majority cases

of cases (57%) were diagnosed the distant stage, stage at time of diagnosis is of major concern, and early detection is critical to improving survival rates[1,8].

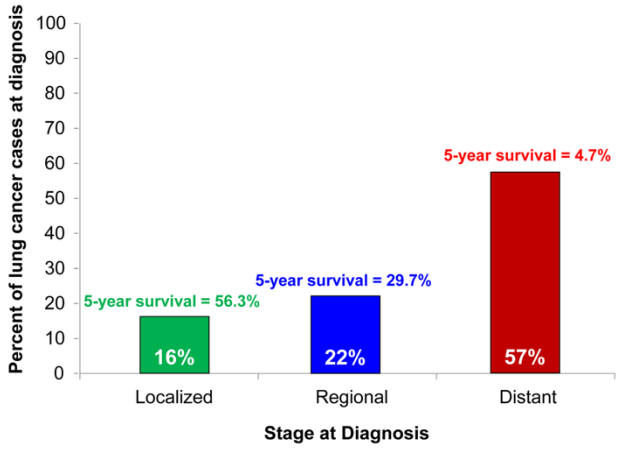


Figure 2. Percent of lung cancer cases at diagnosis and 5-year relative survival by stage. Figure 6 shows the percentage of lung cancer cases diagnosed in the U.S. by stage and their respective 5-year survival rates using data from SEER 18 (https://seer.cancer.gov/statfacts/html/lungb.html). "Localized" is confined to the primary sites, "regional" has spread to the regional lymph nodes, and "distant" is a cancer that has metastasized. "Unknown", which accounts for 4% of diagnoses and has an 8.2% 5-year survival, is not shown. Adapted from Schabath MB, Cote ML. [7]

1.2.2 Survival Statistics

Although most cancer types have seen significant improvements in 5-year survival rates, survival for lung cancer has seen little improvement. Common cancer types such as prostate, breast and colorectal now have relatively higher survival rates and significantly more improvement in 5-year survival when compared to lung cancer (Figure 3)[9]. This lack of improvement can be attributed to the fact that at the time of diagnosis, most patients already have advanced stage disease, at which point survival rates are very low. Age adjusted incidence based mortality per 100,000 for lung cancer was 39 in 2001 - and has scarcely improved since then, at just 25 in 2016 (Figure 4)[10]. Observed 5-year survival in 2013 for all stages was 22.1% and is 21.7% based

on 2017 data (Figure 5)[9]. In Canada, the 5 year survival rate for lung cancer (all stages combined) is only 19%, with late stage 5 year survival being even lower, at less than 15%[11].

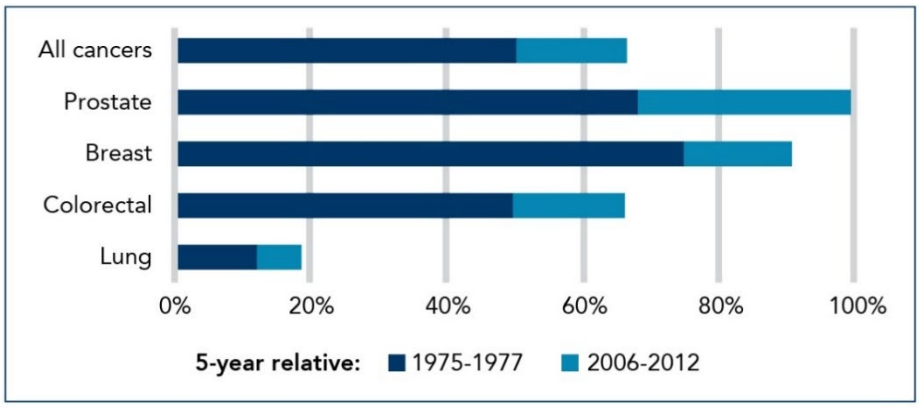


Figure 3. 5 year survival rates for the most common cancers in the United States. Adapted from SEER Database, Cancer Stat Facts, Cancer of the Lung and Bronchus.[9]

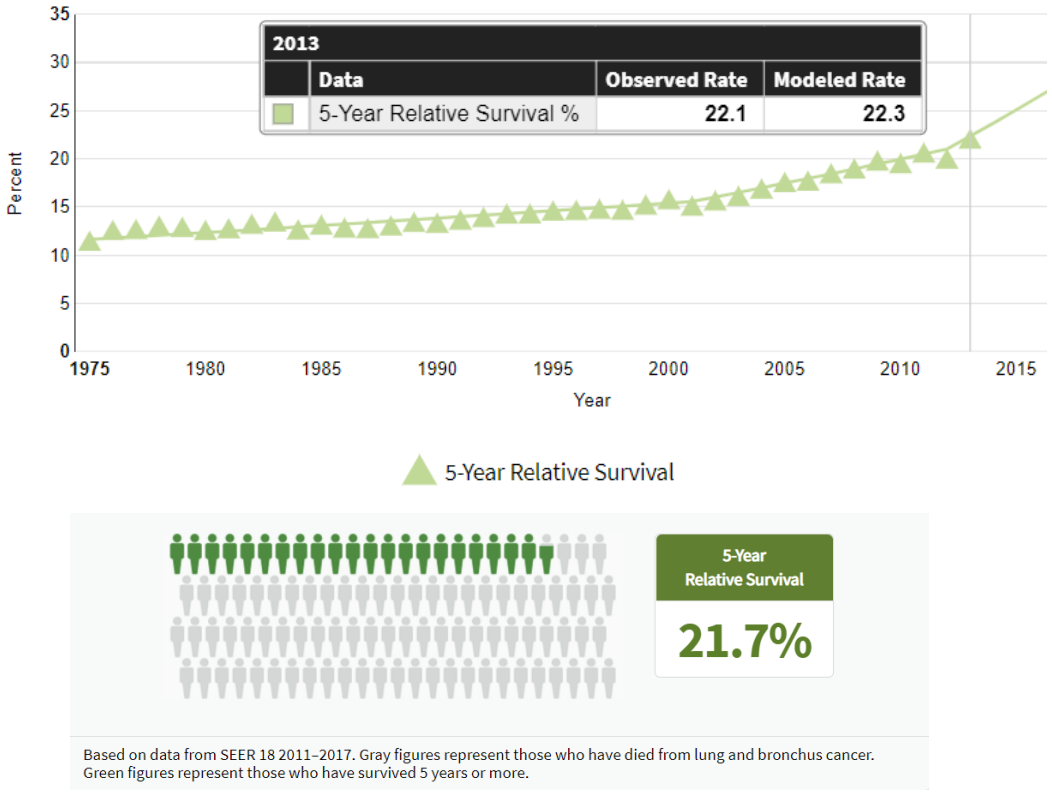


Figure 4. (A) SEER 9 5-Year Relative Survival Percent from 1975–2013, All Races, Both Sexes. (B) 5 year survival based on SEER 18 2011-2017. Adapted from SEER Database, Cancer Stat Facts, Cancer of the Lung and Bronchus.[9]

1.3 Lung Cancer Screening

Screening for lung cancer provides an opportunity to detect cases at an early stage and offer curative resection. The first of many large screening trials to show the significant mortality benefit of lung cancer screening was the National Lung Cancer Screening Trial (NLST). The NLST was conducted on 53, 454 patients between 55-75 years, current or former smokers (quit within 15 years), 30 pack years and no significant competing risks on mortality. The NLST found that annual low dose CT (LDCT) screening provided a 20% reduction in lung cancer mortality, and 7% reduction in overall mortality as compared to the control arm (Figure 6)[12]. Subsequent trials such as the Dutch-Belgian Lung Cancer Screening Trial (NELSON) have shown similar results (24% reduction in lung cancer specific mortality in men and 33% reduction in women with LDCT screening)[13]. The significant mortality benefit of lung cancer screening has been widely established.

Table 2. Results of Three Rounds of Screening.*

Screening Round	Low-Dose CT				Chest Radiography			
	Total No. Screened	Positive Result	Clinically Significant Abnormality Not Suspicious for Lung Cancer	No or Minor Abnormality	Total No. Screened	Positive Result	Clinically Significant Abnormality Not Suspicious for Lung Cancer	No or Minor Abnormality
			no. (% of screened)				*no. (% of screened)*	
T0	26,309	7191 (27.3)	2695 (10.2)	16,423 (62.4)	26,035	2387 (9.2)	785 (3.0)	22,863 (87.8)
T1	24,715	6901 (27.9)	1519 (6.1)	16,295 (65.9)	24,089	1482 (6.2)	429 (1.8)	22,178 (92.1)
T2	24,102	4054 (16.8)	1408 (5.8)	18,640 (77.3)	23,346	1174 (5.0)	361 (1.5)	21,811 (93.4)

* The screenings were performed at 1-year intervals, with the first screening (T0) performed soon after the time of randomization. Results of screening tests that were technically inadequate (7 in the low-dose CT group and 26 in the radiography group, across the three screening rounds) are not included in this table. A screening test with low-dose CT was considered to be positive if it revealed a nodule at least 4 mm in any diameter or other abnormalities that were suspicious for lung cancer. A screening test with chest radiography was considered to be positive if it revealed a nodule or mass of any size or other abnormalities suspicious for lung cancer.

Figure 5. National Lung Cancer Screening Trial results. Adapted from Aberle DR, et al.[12]

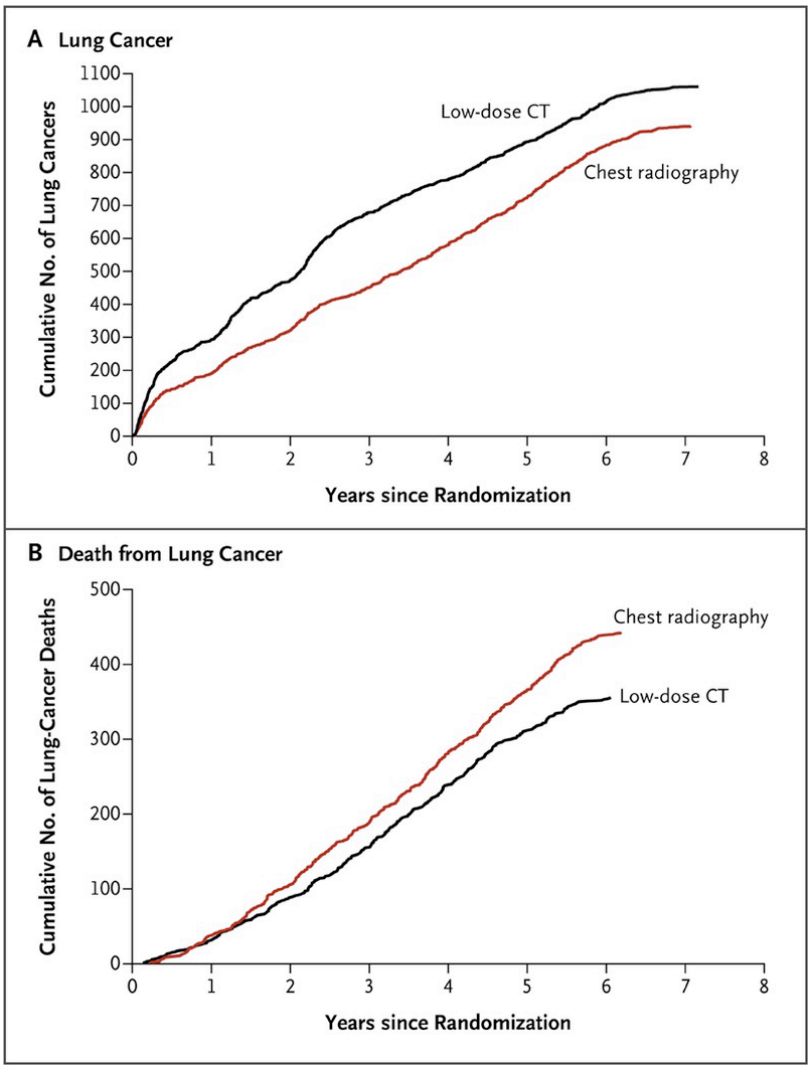


Figure 6. Cumulative Numbers of Lung Cancers and of Deaths from Lung Cancer in the NLST. The number of lung cancers (Panel A) includes lung cancers that were diagnosed from the date of randomization through December 31, 2009. The number of deaths from lung cancer (Panel B) includes deaths that occurred from the date of randomization through January 15, 2009. Adapted from Aberle DR, et al. [12]

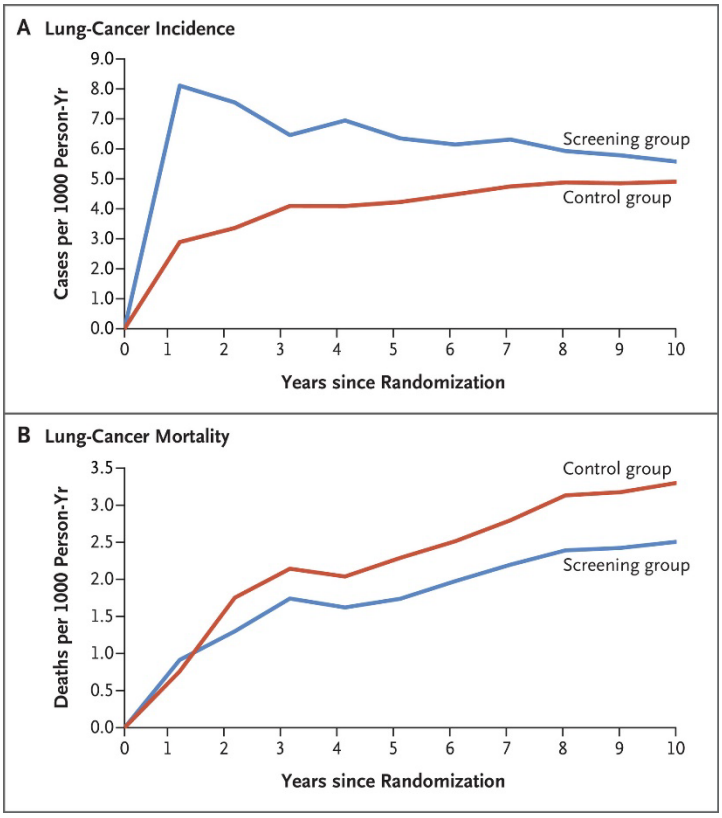


Figure 7. Lung-Cancer Incidence and Lung-Cancer Mortality among Male Participants in NELSON. Panel A shows the cumulative lung-cancer incidence (per 1000 person-years) according to follow-up year since randomization. Panel B shows the cumulative lung-cancer mortality (per 1000 person-years) according to follow-up year since randomization. Cause of death (with known date of lung-cancer diagnosis) was defined by the cause-of-death committee, if available, or by vital-statistics registries. Adapted from de Koning HJ, et al.[13]

1.4 Smoking and Smoking Cessation

1.4.1 Smoking and Lung Cancer

The 1964 landmark report released by the Surgeon General of the US Public Health Service detailed the association between lung cancer and cigarette smoking, showing that smokers had a 9-10 fold greater risk of lung cancer than non-smokers[14]. Since then, this association has been unequivocally established, with studies showing that tobacco smoking is the most important and

prevalent lung cancer risk factor[15]. It has been shown that although only 15% of smokers develop lung cancer, between 80-90% of lung cancer diagnoses can be ascribed to cigarette smoking in the United States[7]. In fact, the *risk* of lung cancer in smokers may be up to 20-fold greater than that of a never smoker[7]. Additionally, lung cancer risk has been shown to be related to smoking intensity (measured by heaviness of smoking index which is calculated based on the number of cigarettes smoked per day, number of years smoked, how long after waking up do you smoke)[15] . There are various lung cancer risk models available on the web that provide risk assessment, most commonly referenced is the Prostate, Lung, Colorectal, and Ovarian (PLCO) model, which has been validated in several countries including the United States, Canada, Germany, and Australia[16]. The PLCOm2012 defines risk as the probability of a diagnosis of lung cancer in 6 years and includes the following predictors: age, level of education, body-mass index (BMI), family history of lung cancer, chronic obstructive pulmonary disease (COPD), chest radiography in the previous 3 years, smoking status (current smoker vs. former smoker), history of cigarette smoking in pack-years, duration of smoking, and quit time (the number of years since the person quit smoking)[16] [17].

1.4.2 Smoking Cessation

Lung cancer screening presents an opportunity to implement smoking cessation services at a teachable moment for individuals engaging with a screening program[18]. The National Lung Cancer Screening Trial (NLST) showed that seven years of smoking cessation alone led to a 20% reduction in lung cancer mortality for in participants in the control group – which is equivalent to the mortality reduction from lung cancer screening itself[19]. Seven years of smoking cessation in the LDCT screening arm led to an additional reduction in mortality of about 10%

(Figure 9)[19]. In former smokers, each additional year of smoking abstinence resulted in a 6% decrease in the risk of lung cancer death (HR, 0.94; 95% CI, 0.92–0.96) (Figure 10), which increased to 9% in those screened with LDCT versus 3% in those who were not[19].

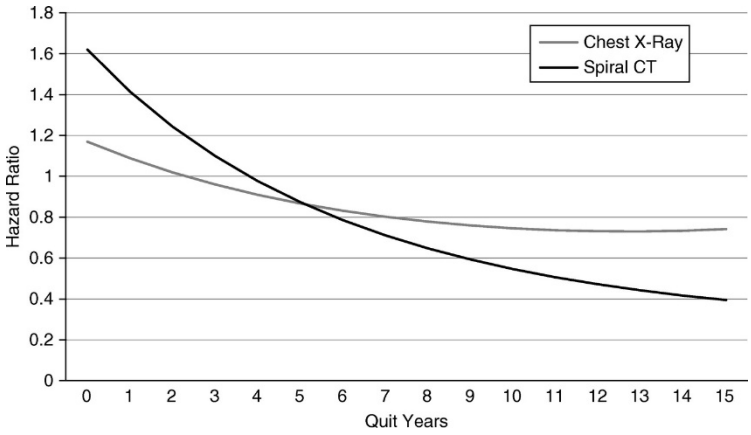


Figure 8. Adjusted hazard ratios for lung cancer–specific mortality; quit-years by screening arm for former smokers. CT = computed tomography. Adapted from Tanner NT, et al. [19]

Table 3. Hazard Ratios and 95% CI for Years since Quitting Smoking and Lung Cancer–Specific Mortality in Former Smokers (1- and 5-yr Increments) and by Screening Group, Adjusted for Demographics

Variable	All Former Smokers			LDCT Scan			Chest Radiograph		
	Hazard Ratio	95% Wald CI	*P* Value	Hazard Ratio	95% Wald CI	*P* Value	Hazard Ratio	95% Wald CI	*P* Value
Centered age (by 5 yr)	1.67	(1.52–1.83)	<0.0001	1.67	(1.45–1.92)	<0.0001	1.67	(1.50–1.87)	<0.0001
Years since quitting (by 1 yr)	0.94	(0.92–0.96)	<0.0001	0.91	(0.87–0.95)	<0.0001	0.97	(0.95–1.00)	0.0782
Years since quitting (by 5 yr)	0.75	(0.67–0.83)	<0.0001	0.62	(0.51–0.76)	<0.0001	0.88	(0.76–1.02)	0.0782

Figure 9. Hazard Ratios and 95% CI for Years since Quitting Smoking and Lung Cancer-Specific Mortality in Former Smokers (1- and 5-yr increments) and by Screening Group, Adjusted for Demographics. Adapted from Tanner NT, et al.[19]

However, studies have shown that lung cancer screening by itself does not result in smoking cessation. Danish Lung Cancer Screening Trial (DLCST) showed that 12-month smoking cessation rates (no smoking cessation intervention) were not significantly different between the screening and control arm (11% and 10%)[20]. Thus, smoking cessation intervention is necessary. While it has been shown that Lung Cancer Screening patients are more motivated to quit than general population [21], it has also been shown that they are more dependent on cigarettes. A trial by Rajewski et al showed the mean Fagerstrom score was 6.1 in former smokers undergoing screening compared to 4.4-4.6 in the general US smoking population[22]. As a result, there arises the notion that smokers participating in lung cancer screening programs have different smoking behaviour compared to non-screening populations. This has been supported by the finding that most smoking cessation interventions, across types, have lower efficacy in screening populations compared to the general population. A 2019 meta-analysis of 9 smoking cessation intervention trials showed that in contrast to general population, telephone counseling was not statistically significantly in lung cancer screening participants (OR 1.38, 95% CI 1.19-1.61 (Figure 11)[23]. The only Canadian study was published by Tremblay et al., with active smokers participating in lung cancer screening randomized to phone counselling smoking cessation intervention (n = 171) or control arm (n = 174), showed no statistical difference in 30-day smoking abstinence rate at 12 months between the groups (14.0% in intervention arm, 12.6% in control arm, p=0.7)[24].Thus, the consensus on the effectiveness of phone counselling as a smoking cessation intervention is still inconclusive.

	Electronic/Web-Based			In-Person Counseling			Pharmacotherapy			Telephone Counseling		
	n	OR (95% CI)	I^2 (%)*	n	OR (95% CI)	I^2 (%)*	n	OR (95% CI)	I^2 (%)*	n	OR (95% CI)	I^2 (%)*
Overall at 6-Months	25	1.14 (1.03-1.25)	57.2	20	1.46 (1.25-1.70)	24.7	25	1.53 (1.33-1.77)	73.7	9	1.21 (0.98-1.50)	74.4

Figure 10. Odds or smoking cessation from random-effects meta-analysis of trials with smokers potentially eligible for lung cancer screening based on 7-day point prevalence of abstinence at 6-months and 12-months by primary intervention type. Adapted from Cadham CJ, et al. [23]

1.5 Lung Cancer Treatment and Risk

1.5.1 Surgical Resection and Post Operative Readmission

Surgical resection is the standard primary treatment for early-stage lung cancer (stages I and II) and is associated with the best long-term survival, with 5-year survival rates of 60-80% for stage I and 30-50% for stage II[25–27] . 30-day post operative readmission is a strong negative predictor of 90-day mortality [28,29]. A 2014 a large-scale study conducted on, 11,432 lung cancer resection patients from the SEER-Medicare database, found that 90-day mortality was six times higher among patients who were readmitted at least once, and that 30 day post operative readmission was the largest contributor to predicting mortality (OR 5.79, p<.001)[28]. The study found that most readmissions were directly related to postoperative complications, which are huge negative predictor of survival[30]. Analysis of 1992-2002 SEER Medicare data showed a 30-day postoperative readmission rate of 15% in this population – a figure that has not changed in decades[28]. The 2014 study by Hu et al., used a hierarchical generalized regression model to predict 30 day post operative readmission, using demographics, comorbidities, socioeconomic factors, hospital provider and diagnoses and reported an AUC of just 0.604 (Figure 12)[28]. Post operative readmission has generally been difficult to predict due to the presence of many correlated variables and factors that cannot be captured by conventional variables. Existing models rarely achieve AUC great than 0.6. [28]

Variable	OR	CI	p-value	F test
Procedure Type			0.018	2.4
Pneumonectomy	1.21	0.85 – 1.72		
Chest wall resection with lung	1.15	0.62 – 2.13		
Lobectomy/bilobectomy	0.82	0.66 – 1.01		
Segmentectomy	0.96	0.72 – 1.27		
Wedge resection	0.78	0.60 – 1.01		
VATS lobectomy	0.74	0.58 – 0.95		
VATS segmentectomy	0.69	0.49 – 0.98		
VATS wedge resection	REF			
Surgical Year			0.584	0.7
2007	1.08	0.78 – 1.49		
2008	1.17	0.85 – 1.61		
2009	1.14	0.83 – 1.57		
2010	REF			
Age			0.025	2.8
85+	1.47	1.11 – 1.94		
80-84	1.22	1.00 – 1.48		
75-79	1.21	1.03 – 1.44		
70-74	1.07	0.91 – 1.26		
65-69	REF			
Gender			0.049	3.9
Female	0.88	0.78 – 1.00		
Male	REF			
Race			0.190	1.6
Asian	0.80	0.52 – 1.24		
Black	0.75	0.57 – 1.01		
Other	0.88	0.60 – 1.28		
White	REF			
Comorbidity				
Induction Chemoradiation	1.52	1.19 – 1.93	<0.001	11.5
Acute myocardial infarction	1.25	1.01 – 1.50	0.015	5.9
Congestive Heart Failure	1.56	1.32 – 1.83	<0.001	27.9
Peripheral Vascular Disease	1.14	0.98 – 1.34	0.093	2.8
Cerebrovascular Disease	1.18	0.99 – 1.42	0.070	3.3
COPD	1.47	1.29 – 1.67	<0.001	34.3
Diabetes	1.15	1.01 – 1.31	0.035	4.5
Renal Failure	1.25	1.04 – 1.51	0.018	5.6
Married	0.96	0.85 – 1.09	0.530	0.4
Regional Median Income			0.570	0.7
Q1 (lowest income)	0.89	0.67 – 1.18		
Q2	0.95	0.74 – 1.22		
Q3	1.06	0.85 – 1.32		
Q4	1.05	0.86 – 1.29		
Q5 (highest income)	REF			
Regional Population Density			0.032	2.7
Q1 (least dense)	0.92	0.75 – 1.13		
Q2	1.15	0.95 – 1.39		
Q3	1.14	0.94 – 1.38		
Q4	1.24	1.03 – 1.50		
Q5 (most dense)	REF			
Regional High School Non-Graduation			0.067	2.2
Q1 (fewest non-graduates)	0.79	0.60 – 1.03		
Q2	0.73	0.57 – 0.93		
Q3	0.91	0.73 – 1.13		
Q4	0.97	0.79 – 1.18		
Q5 (most non-graduates)	REF			

Figure 11. Clinical risk factors for 30-day readmission in lung cancer patients treated by resection, based on 1992-2002 SEER-Medicare data. Model C-statistic: 0.604. Adapted from Hu Y, et al. [28]

1.6 Risk and Risk Scores

1.6.1 Risk Assessment and Risk Scores Overview

Traditionally, clinicians integrate a variety of clinical data and risk prediction scores to evaluate patients for lung resection. However, these scores come with significant limitations as they are built on older data and do not necessarily reflect modern surgical techniques, additionally they were not created solely based on data from oncology patients, where the benefits of surgical resection often outweigh risk.

The American College of Chest Physicians clinical practice guidelines *Physiologic Evaluation of Lung Cancer Patients Being Considered for Resection Surgery* (3rd edition, 2013) recommend using the Revised Cardiac Index (RCRI) or American College of Surgeons National Surgical Quality Improvement (ACS NSQIP) surgical risk score (for cardiovascular evaluation) and pulmonary function testing (predicted post operative FEV 1 and DLCO) for risk evaluation prior to lung resection (Figure 12)[31]. Alternatively, there exists the thoracoscore, a risk prediction model designed to predict 30-day in hospital mortality for all thoracic surgery patients[32].

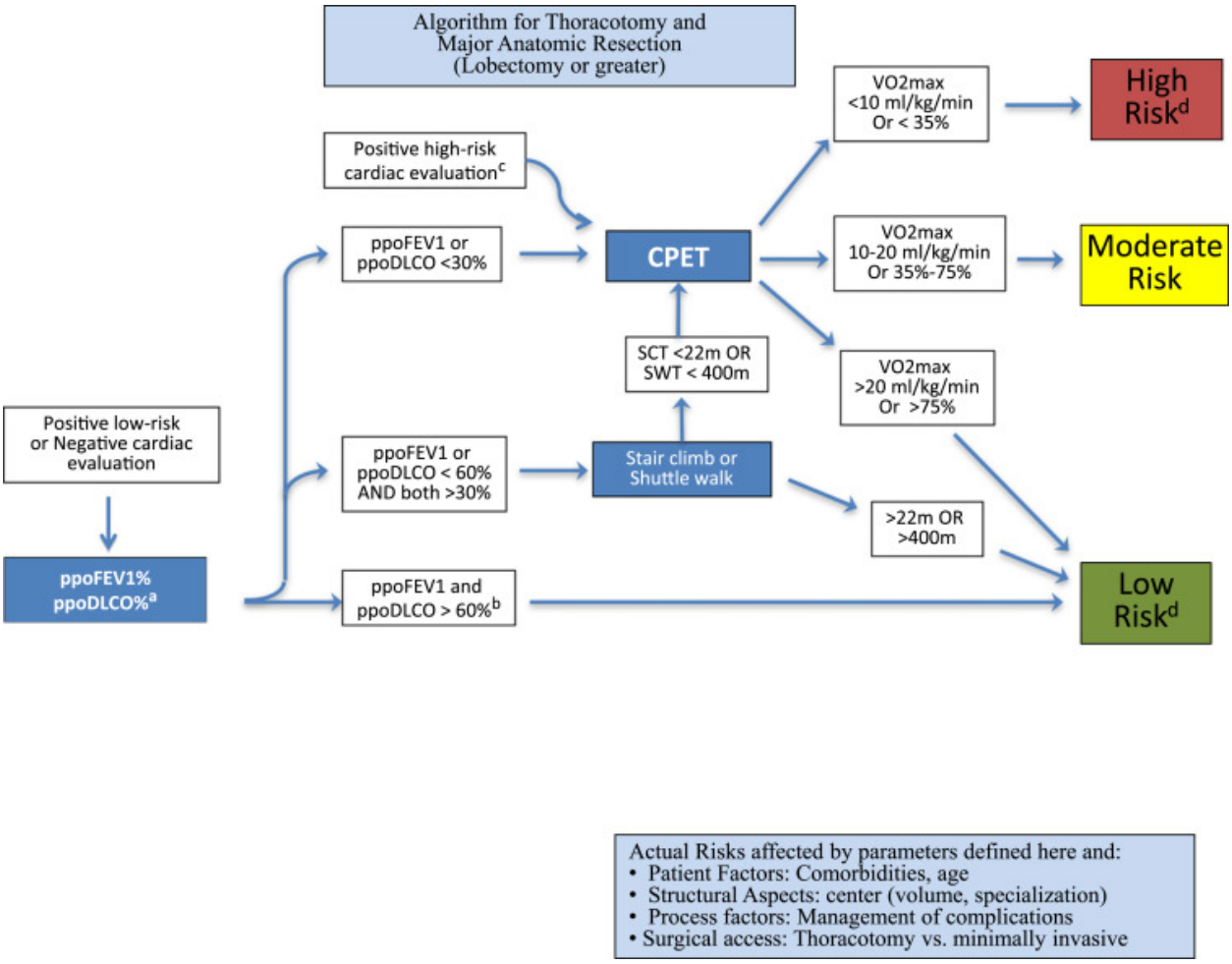


Figure 12. Algorithm for pulmonary preoperative assessment of patients requiring lung resection. ppoDLCO = predicted postoperative diffusing capacity for carbon monoxide; ppoDLCO% = percent predicted postoperative diffusing capacity for carbon monoxide; ppoFEV1 = predicted postoperative FEV1; ppoFEV1% = percent predicted postoperative FEV1; SCT = stair climb test; SWT = shuttle walk test; VO2max = maximal oxygen consumption. Adapted from American College of Chest Physicians clinical practice guidelines for the physiologic evaluation of lung cancer patients being considered for resection surgery, 3rd edition, 2013. Adapted from Brunelli A, et al. [31]

1.6.2 Revised Cardiac Risk Index (RCRI)

The Revised Cardiac Risk Index (RCRI) was developed in 1999 on 3000 patients undergoing major noncardiac surgery and predicts risk of major cardiac outcomes (myocardial infarction, pulmonary edema, ventricular fibrillation, primary cardiac arrest or complete heart block)

(AUC: 0.81). However, the RCRI has several limitations in that it was developed from a general surgical population with only a small number of thoracic patients (346/2893), and also does not provide an estimate of all-cause mortality[33]. Validation in lung resection patients in predicting major cardiac complications has shown the RCRI to have poor performance with AUC 0.59[34]. The RCRI was recalibrated in a lung resection population (oncologic and non-oncologic), to create the thoracic RCRI (ThRCRI), however studies have shown even the ThRCRI shows poor discriminative ability for predicting post operative cardiac complications in a lung cancer population undergoing resection (AUC 0.57). [34,35] The American College of Surgeons (ACS) risk calculator was designed to predict 18 different post operative outcomes, including various post operative complications, readmission, and death within 30 days of surgery[36]. It was developed on the general surgical population, is not externally validated, and has shown poor risk stratification for pulmonary resection[37].

Revised cardiac risk index, adapted from Lee[11]

Lee index		
1. High-risk surgical procedures	Intraperitoneal	
	Intrathoracic	
	Suprainguinal vascular	
2. History of ischaemic heart disease	History of myocardial infarction	
	History of abnormal exercise ECG	
	Current complaint of chest pain considered secondary to myocardial ischaemia	
	Use of nitrate therapy	
	ECG with pathological Q-waves	
3. History of congestive heart failure	History of congestive heart failure	
	Pulmonary oedema	
	Paroxysmal nocturnal dyspnoea	
	Bilateral rales or S3 gallop	
	Chest radiograph showing pulmonary vascular redistribution	
4. History of cerebrovascular disease	History of transient ischaemic attack or stroke	
5. Preoperative treatment with insulin		
6. Preoperative serum creatinine >2.0 mg/dL		
Risk of major cardiac event (each risk factor is assigned one point)		
Points	Class	Risk (%)
0	I	0.4
1	II	0.9
2	III	6.6
3 or more	IV	11

Figure 13. Revised Cardiac Risk Index (RCRI) risk factors and risk score calculation. Adapted from Pannell LMK, et al. [33,38]

1.6.3 Pulmonary Function Testing

FEV1 and DLCO are specific pulmonary function measures that are used to quantify severity of lung disease. In COPD, a disease that affects many smokers presenting for lung resection, they are used to quantify disease severity. Mild COPD is >80% FEV1 % predicted, moderate disease

is 50-80%, severe is 30-50% and very severe COPD is considered when FEV1 % predicted drops below 30%[39]. To estimate the severity of COPD expected after pulmonary resection clinicians can calculate the post operative predictive FEV1 (ppoFEV1) and DLCO (ppoDLCO) to predict a patient pulmonary status after resection. This value is also dependent on which lobes are removed as each bronchopulmonary segment is assumed to contribute equally to lung function. There are in total 19 segments: 3 in right upper lobe, 2 in the right middle lobe, 5 in the right lower lobe; 4 in the left upper lobe, 5 in the left lower lobe. For example, a right middle lobe has less contribution to pulmonary function because there are fewer segments removed.[40] However, recent data shows that observed post operative FEV1 and DLCO are often far lower than those predicted. A 2007 study by Brunelli et al., showed that at 1 month post discharge, observed postoperative FEV1 was 11% lower than ppoFEV1 ($p<0.0001$) and observed post operative DLCO was 12% lower than ppoDLCO ($p<0.0001$) in 180 lung cancer patients treated by lobectomy. [41]

1.6.4 Thoracoscore

Thoracoscore predicts the risk of 30 day in hospital mortality using logistic regression. It was developed on patients receiving thoracic surgery (both oncologic and non-oncologic surgical indications, N=10,122) and modeled the following pre-operative factors: age, sex, American Society of Anesthesiologists score, dyspnea score, performance class, diagnosis group, procedure class, priority of surgery and comorbidity score (0, 1-2, 3+) defined as: smoking, history of cancer, COPD, arterial hypertension, heart disease, diabetes mellitus, peripheral vascular disease, obesity, alcoholism (Figure 14). [32]

TABLE 3. Prediction of risk of in-hospital mortality

Variable	Value	Code	β coefficient
Age (y)	<55	0	
	55-65	1	0.7679
	≥65	2	1.0073
Sex	Female	0	
	Male	1	0.4505
American Society of Anesthesiologists score	≤2	0	
	≥3	1	0.6057
Performance status classification	≤2	0	
	≥3	1	0.689
Dyspnea score	≤2	0	
	≥3	1	0.9075
Priority of surgery	Elective	0	
	Urgent or emergency	1	0.8443
Procedure class	Other	0	
	Pneumonectomy	1	1.2176
Diagnosis group	Benign	0	
	Malignant	1	1.2423
Comorbidity score	0	0	
	≤2	1	0.7447
	≥3	2	0.9065
Constant	—		−7.3737

Figure 14. Thoracoscore Variables. Adapted from Falcoz PE, et al.[32]

Although at first glance it has relatively good predictive performance, (AUC= 0.86) it comes with several key limitations. It only models 30 day in hospital mortality and includes only conventional predictors (Figure 13)[32]. Furthermore, it was developed on heterogenous population undergoing all thoracic procedures and is not specific to lung cancer resection patients. [32] Consequently, validation of the thoracoscore in lung cancer resection patients has shown significantly reduced model performance (8 % thoracoscore predicted mortality vs. 4.5% observed in-hospital, AUC=0.44). [42]

1.6.5 Traditional Risk Assessment - Conclusions

None of the traditional models were developed exclusively on patients presenting for lung cancer resection limiting their utility in this population. They do not use granular data available in

electronic health record (EHR). Furthermore, much of the validation and performance data is prior to the implementation of more modern surgical techniques, and more recent evaluation has shown they inaccurately predict risk.[43,44] Thus, traditional methods of risk assessment come with significant limitations and need to be updated.

1.7 Supervised Machine Learning Methods

1.7.1 Logistic Regression

Logistic regression is widely regarded as the statistical model of choice for modelling the relationship between a binary outcome and two or more predictors (independent variables). Logistic regression assumes a linear relationship between the logarithm of the odds of a positive outcome (defined as Y=1) and the predictors, which are expressed as the sum of the product of each predictor and its coefficient (Figure 15). Coefficients are calculated from the data and describe the contribution of its predictor towards the outcome when all other predictors are controlled for. Predictors may be categorical or continuous. The odds ratio, obtained by taking the exponential of the regression coefficient for a given predictor, represents the effect of said predictor on the likelihood of a positive outcome (i.e., the likelihood of Y=1). The odds ratio, in combination with its respective p-value, is often used to determine whether or not a given predictor is a risk factor for a given outcome. [45,46]

$$\ln\left[\frac{P(Y)}{1-P(Y)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Figure 15. Logistic regression formula. Where Y is the binary outcome, P(Y) represents the probability of a positive outcome, i.e., the probability of Y=1. $X_1, X_2 \ldots X_k$ are the predictor variables and β_0 is the intercept, and $\beta_1, \beta_2 \ldots \beta_k$ are the model coefficients.

Introducing additional predictors to a model generally produces a model that overfits the data– this is because greater the number of predictors added, more the model begins to fit the idiosyncrasies in the data. Thus, it is crucial to select an appropriate number of predictors to avoid model overfitting and create a generalizable model. When developing prediction models for binary outcomes, an established rule of thumb for is the 10 event per parameter (10 EPP) rule: regardless of sample size, to have at least 10 positive outcome events for each predictor parameter, i.e., for each beta term in the model equation. The 10 EPP rule was developed based on simulation studies conducted in the 1990's, however several studies since have shown that the required events per parameter is far more context specific. The required events per parameter depends not only on the number of positive outcome events relative to the number of parameters, but also on the total sample size, the proportion of the positive outcome events and the expected predictive performance of the model. [47] Riley et al, proposes several updated methods to customize sample size requirements to specific datasets while minimizing the potential for overfitting and producing accurate estimates of parameters. Their pmsampsize package in R allows implementation of their methods to any given dataset and can determine a more suitable number of predictors based on the specific features of the dataset as listed above. [47,48]

1.7.2 Random Forests

1.7.2.1 Decision Trees and Gini Impurity

Random Forests is an ensemble learning technique that uses a combination of different decision trees, i.e., a 'forest'[49]. A decision tree 'splits' the dataset based on the features. Decision learning

involves modelling features (predictors) from training data and subsequently applying it to testing data to evaluate the model. [50]

Decision trees are built based on 'nodes'. A node is a splitting point - the data is split on a feature at a given node based on feature importance, i.e., how well the feature predicts the outcome. The Root Node contains the feature that best predicts the outcome, and is at the top of the tree. The Internal Nodes contain further features in a hierarchy based on feature importance and the final node, the leaf node, contains the outcome. Thus, decision trees perform classification using feature importance[49–51] (Figure 16).

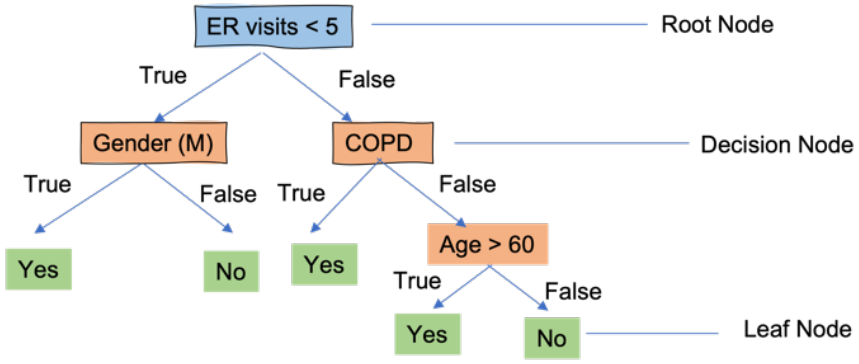


Figure 16. The structure of a decision tree.

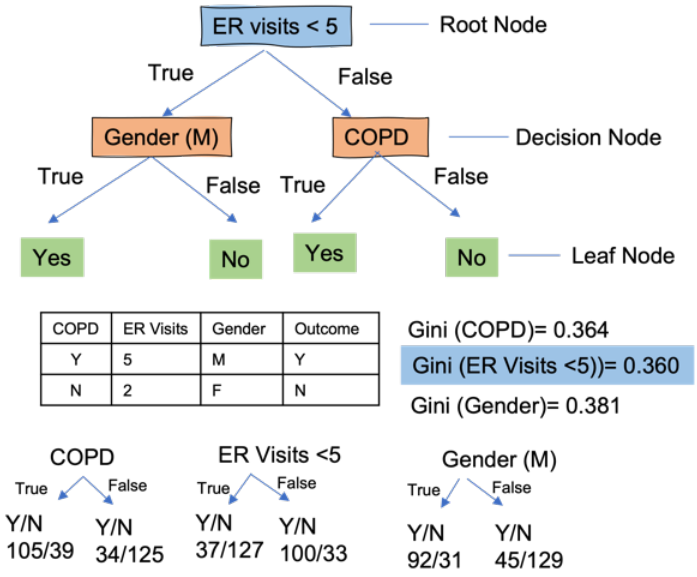
For classification, decision trees calculate feature importance via various metrics. Most commonly used is Gini impurity. Gini impurity is a measure of 'impurity', wherein 'pure' is a 100% split, i.e., a 100 True / 0 False classification, and 'impure' is any split that is not 100% 'pure'. [49] For a given leaf, Gini impurity is calculated as 1 minus the sum of the square probability of the condition being yes and the square of the probability of the condition being no. The total Gini impurity in a node is then calculated as the weighted average of the Gini impurity of its leaf nodes (Figure 17A, Figure 17B).

$$GI = 1 - \sum_{i=1}^{n} (p)^2$$

$$GI = 1 - \left[(P_{(+)})^2 + (P_{(-)})^2 \right]$$

Total Gini impurity in a node= weighted average of Gini impurity of its leaf nodes

Figure 17A. Gini impurity calculation. P (+) represents the probability of "yes" or of the condition being true, P (-) represents the probability of "no" or of the condition being false.



For ER visits < 5 $=$ True, Gini impurity $= 1 - \left[\left(\frac{37}{37+127}\right)^2 + \left(\frac{127}{37+127}\right)^2\right] = 0.349$

For ER visits < 5 $=$ False, Gini impurity $= 1 - \left[\left(\frac{100}{100+33}\right)^2 + \left(\frac{33}{100+33}\right)^2\right] = 0.373$

For the node ER visits < 5, Total Gini impurity $= \left(\frac{37+127}{37+127+100+33} \times 0.349\right) + \left(\frac{100+33}{37+127+100+33} \times 0.373\right)$

$= 0.360$

Figure 17B. Gini impurity calculation

Essentially Gini impurity measures how inaccurately a feature classifies an outcome - the lower the inaccuracy, i.e., the lower the Gini impurity, the more 'important' the feature. Thus, when comparing Gini impurities of different features for splitting, the feature with the lower Gini

impurity is selected[49]. In figure 17B, this is determined to be the feature ER Visits <5. Decision trees operate on if-else logic and are relatively simple to interpret for clinicians[50,51].

Note: when decision trees are visualized using scikit-learn in Python, they typically display the following information for each node: the feature, Gini impurity, the total number of samples in that node as a percentage of the samples in the previous node, the percentage of samples in each class after bootstrapping (labeled as 'value') and the predicted outcome or 'class'.

1.7.2.2 Ensemble Learning: Bagging

Ensemble learning is a technique used to overcome the issue of overfitting. Overfitting is the notion that the model in its effort for high accuracy starts to pick up 'noise' in the data, i.e., model features that are specific to the training data and not generalizable. Decision trees when used alone are prone to overfitting[52]. This is due to the amount of specificity decision trees provide – the deeper the tree gets, the stricter the rules, the smaller the subsets that meet those rules resulting in a model that is not generalizable. Overfitting is a particularly huge challenge in large clinical datasets where there are large number of correlated features. By using ensemble learning we can overcome the issue of overfitting and for our dataset, it is a highly relevant tool[53].

Ensemble learning uses a combination of different models. In ensemble learning, each model is built using different training sets via bootstrap sampling. Averaging the outcomes from all the models results in the final outcome – a process called aggregation. Together, this technique is referred to 'bagging' – short for **B**ootstrap **Agg**regation[54,55].

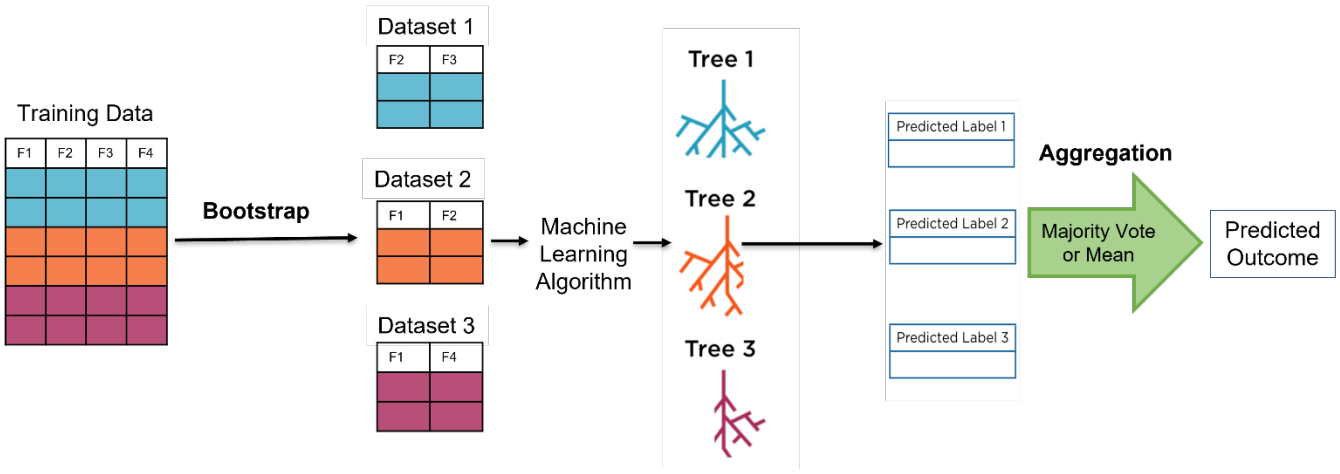


Figure 18. Bootstrap Aggregation (bagging)

Although bagging introduces some randomness into the model, decision trees built by bagging alone are still prone to overfitting. This is because when bootstrapping, all features are utilized, i.e., all correlated features are still used to build the trees, creating correlated trees, and resulting in overfitting.

1.7.2.3 Random Forests: an Improvement on Ensemble Learning

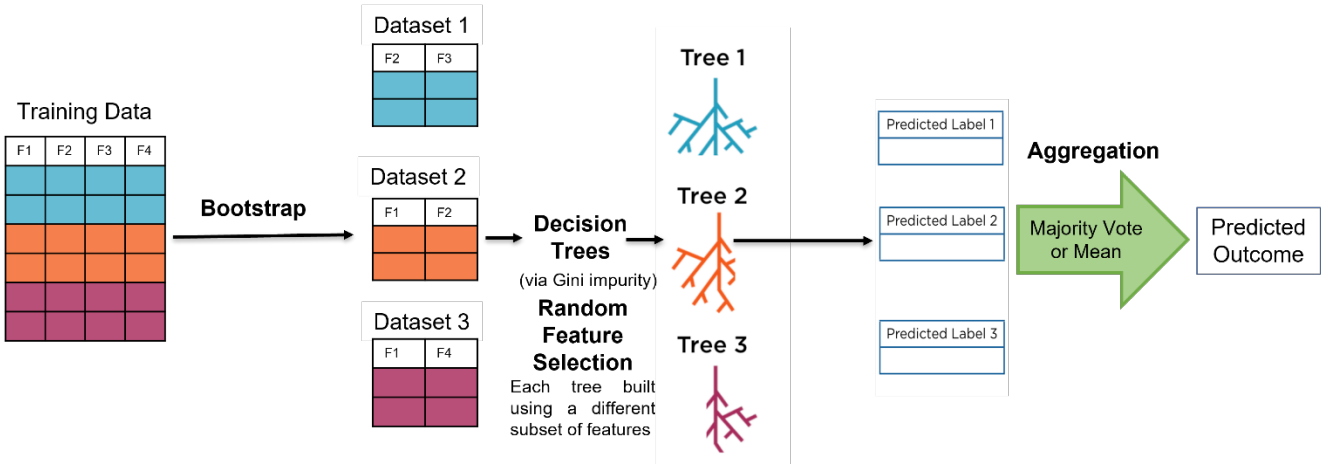


Figure 19. Random Forests

Random Forests solves the problem of overfitting with bagging by using random feature selection: instead of using all the features in each bootstrap sample, it takes a random sample of features. It fits the model and aggregates the outcomes. The idea is that the added randomness from each tree using a different subset of features and the subsequent aggregation of results can cancel out the effect of correlated features and make the model robust to noise[49]. Thus, by introducing randomness at two junctures (bagging and random feature selection), random forests addresses the issue of variable correlation – a significant challenge in clinical settings[49,53,55].

1.7.2.4 Random Forests for Feature Selection

Random Forests allows for intuitive calculation of feature importance[49]. This allows random forests models to be relatively interpretable compared to other supervised machine learning algorithms. It also allows for the use of random forests as a tool for feature selection.

There exist various measures of feature importance. The two most common are Gini impurity, which is the implicit measure in random forests, and Permutation Importance[49]. When using Gini impurity to rank features by importance, the model takes the mean Gini impurity across the forest for each feature – the lower the mean Gini impurity, the higher the feature importance. However, it is important to note that using Gini impurity to rank features by importance can be highly biased towards continuous variables and categorical variables with many levels[49,53].

1.7.3 Measures of Model Performance

Accuracy is defined as the number of classifications a model correctly predicts (true positives, true negatives) divided by the total number of predictions made (true positives, true negatives,

false positives, false positives). Sensitivity is the true positive rate. Specificity is the true negative rate.

$$\text{Sensitivity} = \text{Recall(Yes)} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \text{Recall(No)} = \frac{TN}{TN + FP}$$

Figure 20. Measures of model performance: sensitivity and specificity. TP=True Positive. TN=True Negative. FP=False Positive. FN=False Negative.

AUC is the area under the receiver operating characteristic (ROC) curve. The ROC curve is the curve between sensitivity (true positive rate) and false positive rates (1-specificity). Thus, the AUC is a measure of the ability of a model to discriminate between classes. When the AUC of a model is 1, the model can perfectly discriminate between positive and negative classes. When the AUC of a model is 0.5, the model either is predicting a random class or constant class for all the data. The higher the AUC, the better the performance of the model at discriminating between classes. Models that are clinically useful tend to have a AUC in the range of 0.75-0.9. Positive and negative predictive values are influenced by underlying disease prevalence. Sensitivity, specificity and therefore AUC are independent of underlying disease prevalence. [56]

1.7.4 Imputation Methods

The reality of clinical data mining is that there is often significant missing data. Imputation of missing data can be conducted in various ways, the most robust of which are multiple imputation by chained equations (MICE) and multiple imputation. The only necessary requirement for imputation is that the data be missing at random (MAR)[48].

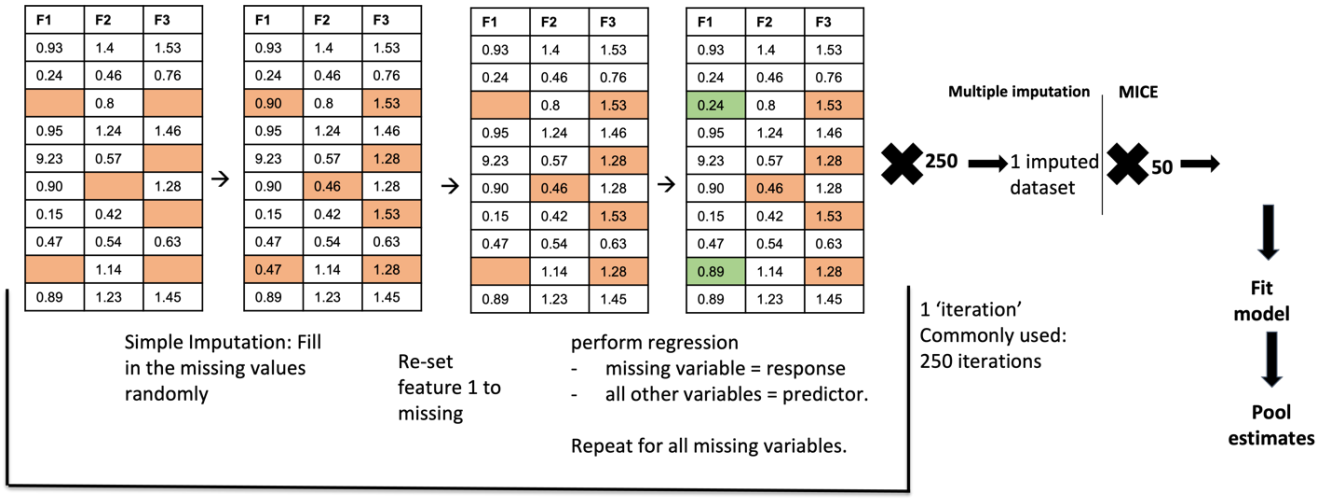


Figure 21. Imputation methods: (1) multiple imputation and (2) multiple imputation by chained equations (MICE)

(1) Multiple imputation starts with a simple imputation, i.e., filling in all the missing values randomly, in order to produce a filled in dataset. (2) It then takes a feature, resets the values to missing, and conducts a regression using all other features in the dataset to predict the missing one. It then repeats this for every variable, producing a full dataset filled by regression. The entire process thus far is one iteration. (3) This iteration is then repeated many times, resulting in a dataset filled by conducting many regressions. This imputed dataset is then ready for analysis. This is the extent of multiple imputation. MICE adds several more steps. (4) With MICE, instead of using the first imputed dataset obtained in step 3, steps 1-3 are repeated many times in order to get many different imputed datasets. (5) Your model of choice is then fitted on each imputed dataset. (6) The estimate from every model is then pooled and averaged to get the final result. [57–59]

CHAPTER 2: Rationale, Hypothesis and Objectives

2.1 Rationale:

2.1.1 Project 1: Lung Cancer Project

Big data in a clinical context can include EHR data, preoperative laboratory data (e.g., hemoglobin, creatinine, and albumin), genomics data and even radiomics data (from CT scans). Given the limitations associated with traditional risk prediction models, such large clinical data sets present an opportunity to capture a level granularity not previously utilized, and thereby improve risk prediction.

Hypothesis: We hypothesized that a Random Forest model using large clinical data may have superior predictive performance when compared to a multivariate logistic regression with traditional epidemiological data, in predicting 90-day post operative readmission in lung cancer patients.

2.1.2 Project 2: Smoking Cessation Project

Current studies assessing the effectiveness of phone counselling as a smoking cessation intervention in lung cancer screening groups use very small sample sizes and are largely inconclusive. Additionally, there is no data on smoking cessation rates lung cancer screening groups as compared to in healthcare worker referred groups.

Hypothesis: we hypothesized that phone counselling does not contribute significantly to increasing smoking cessation rates in a population of patients referred by a lung cancer screening

program, compared to individuals who self-refer or those referred by community health care workers.

2.2 Objectives

2.2.1 Project 1: Lung Cancer Project

To integrate multiple sources of large clinical data to better predict 90-day post operative readmission in early-stage lung cancer patients resected by lobectomy at the McGill University Health Centre (MUHC) and further improve risk stratification pre-operatively compared to traditional risk estimation models.

2.2.2 Project 2: Smoking Cessation Project

To assess the effectiveness of telephone counselling for smoking cessation in lung cancer screening eligible participants as compared to healthcare worker referred and self-referred participants.

CHAPTER 3: Manuscript 1 "Improving lung cancer outcomes: risk prediction of postoperative readmission following lobectomy in lung cancer"

Title: Improving lung cancer outcomes: risk prediction of 90-day postoperative readmission following lobectomy in lung cancer

Ankita Ghatak[1] , Benjamin Smith MD MSc [2, 4] , Andrea Benedetti PhD[2,3], Nicole Ezer MD MPH[1,2,3,4]

1. Division of Experimental Medicine, Department of Medicine, McGill University, Montreal, Quebec, Canada
2. Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada Center for Outcomes Research and Evaluation, McGill University Health Centre, Montreal, Quebec, Canada
3. Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Quebec, Canada
4. Department of Medicine, Division of Respiratory Medicine, McGill University Health Centre, Montreal, Quebec, Canada

Abstract

Introduction: Post-operative readmission affects patient care and is a strong negative predictor of survival. Traditional models used to assess pre-operative risk come with significant limitations with regards to accuracy and their application to lung cancer patients undergoing resection. The objective of this study was to use large clinical data to assess the predictive accuracy of Random Forests as compared to a traditional model such as logistic regression that uses only conventional

variables in predict 90-day post-operative readmission in lung cancer patients resected by lobectomy.

Methods: Data was extracted for lung cancer patients undergoing resection by lobectomy from 2016-2019 from the NSQIP database at the Montreal General Hospital, the Pulmonary Function Data database at the RI-MUHC and EHR data from the RI-MUHC Datawarehouse. Datasets included, demographic information, comorbidities, pre-operative laboratory results, pulmonary function data and admission records. Our primary outcome was 90-day post-operative re-admission. The predictive power of a Random Forests model using all available variables was compared to a logistic regression model with only demographics and comorbidities.

Results: The results of our logistic regression suggested that the most important predictors of our outcome were gender (male) (OR 2.38, CI: 1.20-2.87, $p<0.05$), history of severe COPD (True) (OR 2.1, CI: 1.16-3.2, $p<0.05$) and Charlson comorbidity index (per 1 unit increment) (OR 1.29, CI: 0.9-1.4, $p>0.05$). The Random Forests model ranked features predictors by importance as follows (in descending order): age, white blood cell count, hematocrit, albumin, sodium, creatinine, body mass index, sex, hypertension, Charlson comorbidity index, active smoking within 1 year of surgery and PFT availability. Dyspnea, history of severe COPD and diabetes all had zero feature importance and were ranked last. The random forests model also assessed the relationship between all available predictors and our outcome via the visualization of its decision trees. The random forests also assessed the relationship between all available features, including preoperative laboratory variables, and our outcome via the visualization of its decision trees. The

random forests had better model performance based on all assessed metrics, most significantly on the AUC (0.72) compared to the logistic regression (AUC 0.59).

Conclusions: Random forests shows better predictive performance than logistic regression and may improve preoperative risk stratification in lung cancer patients. Preoperative laboratory tests may be valuable for risk assessment and should be further validated. A random forests model may have several clinical decision-making applications.

Introduction

Lung cancer is the leading cause of cancer death in Canada, in both men and women, exceeding breast, colon and pancreatic cancer combined[1]. Currently, lung resection is the primary treatment for early-stage lung cancer and is associated with the best long-term survival. Post operative readmission is often directly related to post operative complications, which occur in over 40% of patients and are a huge negative predictor of survival[2]. A 2014 study with 11,432 lung cancer resection patients from the SEER-Medicare database, showed 30-day post operative readmission to be associated with a six-fold increase in 90-day mortality – outranking all commonly reported preoperative predictors[3]. Additionally, this study showed a 12.8% 30-day post operative readmission rate, suggesting that the risk of post operative readmission in this population has not significantly changed in this population in over 15 years[3]**.** As such, accurate assessment of perioperative risk is crucial. Currently pre-existing clinical risk factors thought to be associated with readmission are COPD, congestive heart failure, induction chemoradiation, recent myocardial infarction, renal failure, and age[3,4]. While 30-day postoperative readmission is directly related to post-surgical complications, 90-day postoperative readmission is reflective of underlying patient characteristics. These have a stronger association with 1 year survival compared to immediate post operative complications. We chose to assess 90-day re-admission in order to capture patients beyond the immediate post operative period and thereby identify patients who may be candidates for prehabilitation prior to lobectomy. Prehabilitation has been shown to significantly improve post operative outcomes. However, in such a resource limited setting, it cannot be offered to all patients and implementation must be optimized by selecting candidates who would benefit most.

Although existing perioperative risk evaluation metrics exist, they are severely limited in their application to lung cancer patients. Risk scores such as the Revised Cardiac Risk Index (RCRI), post operative predicted FEV1, post operative predicted DLCO and thoracoscore were not developed exclusively developed on lung cancer patients. RCRI predicts risk of major cardiac complications after surgery and thoracoscore predicts post-surgical 30-day in hospital mortality. Although they appear to show good performance (RCRI AUC: 0.80 , thoracoscore AUC: 0.85) [5,6], recent studies have shown that when applied to a lung cancer specific population, they exhibit highly inaccurate risk prediction (RCRI AUC: 0.59 , thoracoscore AUC: 0.65)[7,8]. These existing risk scores were developed exclusively on conventionally available epidemiological variables and fail to take advantage of EHR data or the types of large clinical datasets that have become accessible over the last 5 years. Currently, a risk adjusted readmission metric for pulmonary resection does not exist. NSQIP ACS reports risk of 30-day readmission for all thoracic procedures to be 11.9% and development of a risk adjusted readmission metric for coronary artery bypass surgery is in progress – but there is no such effort towards development of one for pulmonary resection[3]. Presently, the highest quality study on readmission following resection is a 2014 study by Hu et al., using SEER Medicare data from 2006-2011 and patients with NSCLC of any stage. It proposes a hierarchical generalized regression model with demographics, comorbidities, socioeconomic factors, readmitting facility, and diagnosis to predict 30-day postoperative re-admission following lung resection. The model achieved an AUC of only 0.604[3]. There is therefore a great need to improve lung cancer specific risk stratification to improve post operative lung cancer outcomes.

We performed a retrospective cohort analysis to assess the predictive performance of Random Forests as compared to logistic regression in predicting 90-day postoperative readmission in lung cancer patients treated by lobectomy. Our secondary objective was to assess the contribution of predictive variables as reported by Random Forests feature importance metrics[9].

Methods

Databases

This study was performed using three datasets: a subset of National Surgical Quality Improvement (NSQIP) database from the Montreal General Hospital, Pulmonary Function Testing (PFT) data from the McGill University Health Centre (MUHC) and electronic health record (EHR) data containing admission records from the MUHC Datawarehouse. All data cleaning, mining, and linking was performed using Python. [10,11] Ethics approval was obtained from the MUHC research ethics board (REB number 2021-6528).

Patient Selection and Outcomes

The inclusion criterion was lung cancer patients treated by lobectomy at the MUHC between 2016 and 2019. Patients treated by bilobectomy, segmentectomy, pneumonectomy, wedge resection and patients whose index admission was to ICU, were excluded. Our primary outcome was 90-day post operative readmission. This was extracted from admission records and was defined as at least one readmission to ward or emergency room (ER) within 90 days of discharge.

Available Predictors

Continuous variables included were age, body mass index (BMI), sodium, creatinine, albumin, White Blood Cell (WBC) count, hematocrit and Charlson comorbidity index. Charlson comorbidity index was calculated from NSQIP and was extracted using the R package 'comorbidity' [12]. Categorical variables included were gender (Male/Female), active smoker within 1 year of surgery (Yes/No), dyspnea (Yes/No), history of severe COPD (Yes/No), hypertension (Yes/No), diabetes (Yes/No) and PFT availability (Yes/No). PFT availability was determined based on availability of PFT test results at the MUHC. The MUHC is a quaternary care center and receives a high volume of referrals of patients who had PFTs done at outside hospitals. Consequently, PFT datapoints were not available for a large subset of participants.

Statistical Analysis

All statistical analyses were conducted in Python.

BMI, Sodium, Creatinine, WBC, and Hematocrit all had missing values (missingness <60%) and were imputed by multiple imputation with 250 iterations. [13,14]

In both models, the data was split into training (321/535, 60%) and testing (214/535, 40%) sets[13,15]. For both models, the training data was balanced by assigning the classes (admitted/not readmitted) weights inversely proportional to their respective frequencies[13,16].

Logistic regression was performed with the following variables: age, gender, history of severe COPD, BMI and Charlson comorbidity index. As opposed to the 10 event per variable rule, we

followed the recommendations of Riley et al., and used 5 degrees of freedom for our sample size[17]. The logistic regression was conducted on balanced data using bagging: the training data was sampled by bootstrapping with replacement with 250 iterations, and the results were then aggregated and averaged across predictors[13,18].

Random Forests was executed using the following variables: Age, BMI, Sodium, Creatinine, Albumin, WBC, Hematocrit, Charlson Comorbidity Index, Gender, Active smoker within 1 year of surgery, Dyspnea, History of severe COPD, Hypertension, PFT availability, Diabetes. The model was set to aggregate results across 150 trees and have a minimum of 50 samples per leaf. Feature importance was assessed using Gini impurity and predictors were ranked by importance. [13,19–22] Various individual decision trees were visualized[13,20,23]. Models were assessed using the following metrics: accuracy, sensitivity, specificity, false positive rate, and AUC. ROC curves (testing) for both models were visualized. [13,24(p3)]

Results

Population

There was a total of 535 patients with early-stage lung cancer treated by lobectomy. The overall readmission rate within 90 days of discharge was 21% (111/535) (Table 1). The overall mean age was 67 years (SD 9.2) but was higher in those who were readmitted (69.6 years, SD 8.4) compared to those who were not (66.9 years, SD 9.3) ($p<0.05$). There was a higher number of female patients overall (60.2%) but a greater proportion of readmitted patients were male (53.2%) ($p<0.05$). Most patients were not active smokers within 1 year of surgery both overall (55.9%) and amongst those readmitted (58.6%) ($p>0.05$). Overall and across both admitted and

non-readmitted patients, the majority did not have dyspnea, history of severe COPD or diabetes (Table 1). A greater proportion of readmitted patients had hypertension (59.5%, $p<0.05$) as compared to those not readmitted. Overall only half of patients had hypertension (49.5%). Overall most patients did not have PFTs available (63.6%). Overall mean BMI was 26.8 (SD 5.5), but was higher in readmitted patients (27.2, SD 5.8) compared to patients who were not readmitted (26.7, SD 5.4, $p>0.05$). The mean Charlson comorbidity index was 2.0 (SD 1.0) but was higher in patients who were readmitted (2.2, SD 1.1, $p<0.05$) as compared to those not readmitted. Sodium (overall mean 138.4, SD 2.4), creatinine (overall mean 0.9, SD 0.8) and hematocrit (overall mean 40.5, SD 3.8) were roughly the same across both classes ($p>0.05$). Overall mean albumin was 4.2 (SD 0.3) but was lower in those who were readmitted (4.1, SD 0.3) as compared to those who were not (4.2, SD 0.3) ($p=0.001$). Overall mean WBC level was 8.4 (SD 3.2) but was higher in those who were readmitted (9.3, SD 5.5) compared to those who were not (8.2, SD 2.2) ($p<0.05$). (Table 1).

Variable		Total N=535	Not Readmitted N=424 (79%)	Readmitted N=111 (21%)	P-Value
Age, mean (SD)		67.4 (9.2)	66.9 (9.3)	69.6 (8.4)	0.004
Gender, n (%)	Female	322 (60.2)	270 (63.7)	52 (46.8)	0.002
	Male	213 (39.8)	154 (36.3)	59 (53.2)	
Smoker, n (%)	False	299 (55.9)	234 (55.2)	65 (58.6)	0.597
	True	236 (44.1)	190 (44.8)	46 (41.4)	
Dyspnea, n (%)	No	456 (85.2)	367 (86.6)	89 (80.2)	0.125
	Yes	79 (14.8)	57 (13.4)	22 (19.8)	
History of Severe	False	422 (78.9)	346 (81.6)	76 (68.5)	0.004
COPD, n (%)	True	113 (21.1)	78 (18.4)	35 (31.5)	
Hypertension, n (%)	False	270 (50.5)	225 (53.1)	45 (40.5)	0.025
	True	265 (49.5)	199 (46.9)	66 (59.5)	
PFT Availability, n (%)	PFT Not Available	340 (63.6)	282 (66.5)	58 (52.3)	0.008
	PFT Available	195 (36.4)	142 (33.5)	53 (47.7)	
Diabetes, n (%)	No	504 (94.2)	401 (94.6)	103 (92.8)	0.626
	Yes	31 (5.8)	23 (5.4)	8 (7.2)	
BMI, mean (SD)		26.8 (5.5)	26.7 (5.4)	27.2 (5.8)	0.418

Sodium, mean (SD)	138.4 (2.4)	138.4 (2.4)	138.4 (2.3)	0.983
Creatinine, mean (SD)	0.9 (0.6)	0.9 (0.7)	0.9 (0.2)	0.881
Albumin, mean (SD)	4.2 (0.3)	4.2 (0.3)	4.1 (0.3)	0.001
WBC, mean (SD)	8.4 (3.2)	8.2 (2.2)	9.3 (5.5)	0.041
Hematocrit, mean (SD)	40.5 (3.8)	40.5 (3.7)	40.7 (4.2)	0.546
Charlson Comorbidity Index, mean (SD)	2.0 (1.0)	2.0 (0.9)	2.2 (1.1)	0.034

Table 1. Summary of available predictors.[25]

Logistic Regression

In our logistic regression, gender (OR 2.38, $p < 0.05$) and history of severe COPD (OR 2.1, $p<0.05$) had the most significant contribution to predicting 90 day post operative readmission. Other predictors in order of significance were Charlson comorbidity index (per unit increment) (OR 1.29, $p>0.05$, Age (OR 1.04, $p<0.05$) and BMI (OR 1.01, $p>0.05$) (Table 2).

Variables	**Odds Ratio**	**CI**	**P-Value**
Age	1.04	1-1.05	0.041
Charlson Comorbidity Index	1.29	0.9-1.41	0.301
Gender (Male)	2.38	1.2-2.87	0.006
History of Severe COPD (True)	2.1	1.16-3.2	0.011
BMI	1.01	0.97-1.05	0.702

Table 2. Logistic Regression results. Outcome: 90-day postoperative readmission. N=535. Degrees of freedom=5.

Random Forests

The random forests ranked all predictors by feature importance calculated by Gini impurity. Note that feature importance is a relative measure. In our random forests, age (0.21), WBC (0.15), hematocrit (0.13), albumin (0.12), and sodium (0.11) were determined to be the five most important, i.e., least inaccurate, predictors for prediction of 90-day post operative readmission (relative to all other predictors in our model). In order of descending importance, they were

followed by creatinine (0.09), BMI and gender (0.06), hypertension (0.03) and Charlson Comorbidity index, active smoker within 1 year of surgery, PFT availability which all had equal feature importance (0.01). Dyspnea, history of severe COPD and diabetes all had zero feature importance, meaning they had no importance in prediction of the outcome, and were ranked last (Table 3, Figure 1).

Feature	**Importance**
Age	0.21
WBC	0.15
Hematocrit	0.13
Albumin	0.12
Sodium	0.11
Creatinine	0.09
BMI	0.06
Gender (Male)	0.06
Hypertension	0.03
Charlson Comorbidity Index	0.01
Active Smoker Within 1 Year of Surgery	0.01
PFT Availability	0.01
Dyspnea	0
History of Severe COPD	0
Diabetes	0

Table 3. Random Forests Feature importance ranking in order of descending importance. Feature importance was calculated by Gini impurity, wherein the lower the Gini impurity in a feature, the higher importance it has, and the higher it is ranked.

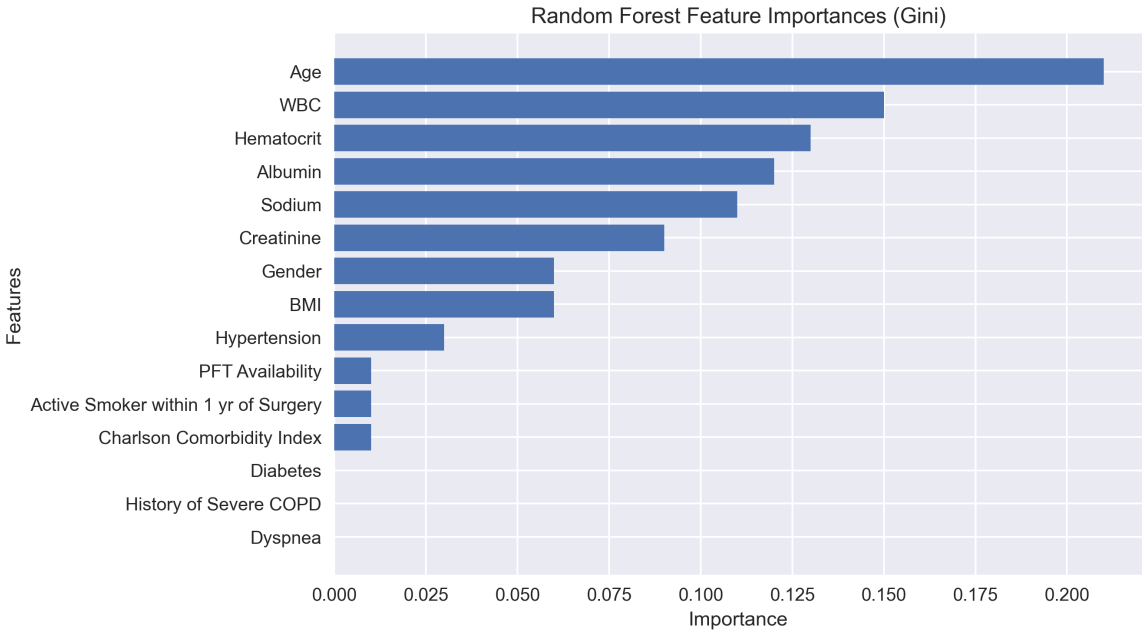


Figure 1. Random forests features ranked by feature importance (via Gini impurity).

Of the 150 trees in the random forests model, a select few were visualized (Figure 2). As visualized they are highly interpretable for clinicians as they can see what decisions are made at each node, and in what order they are classified.

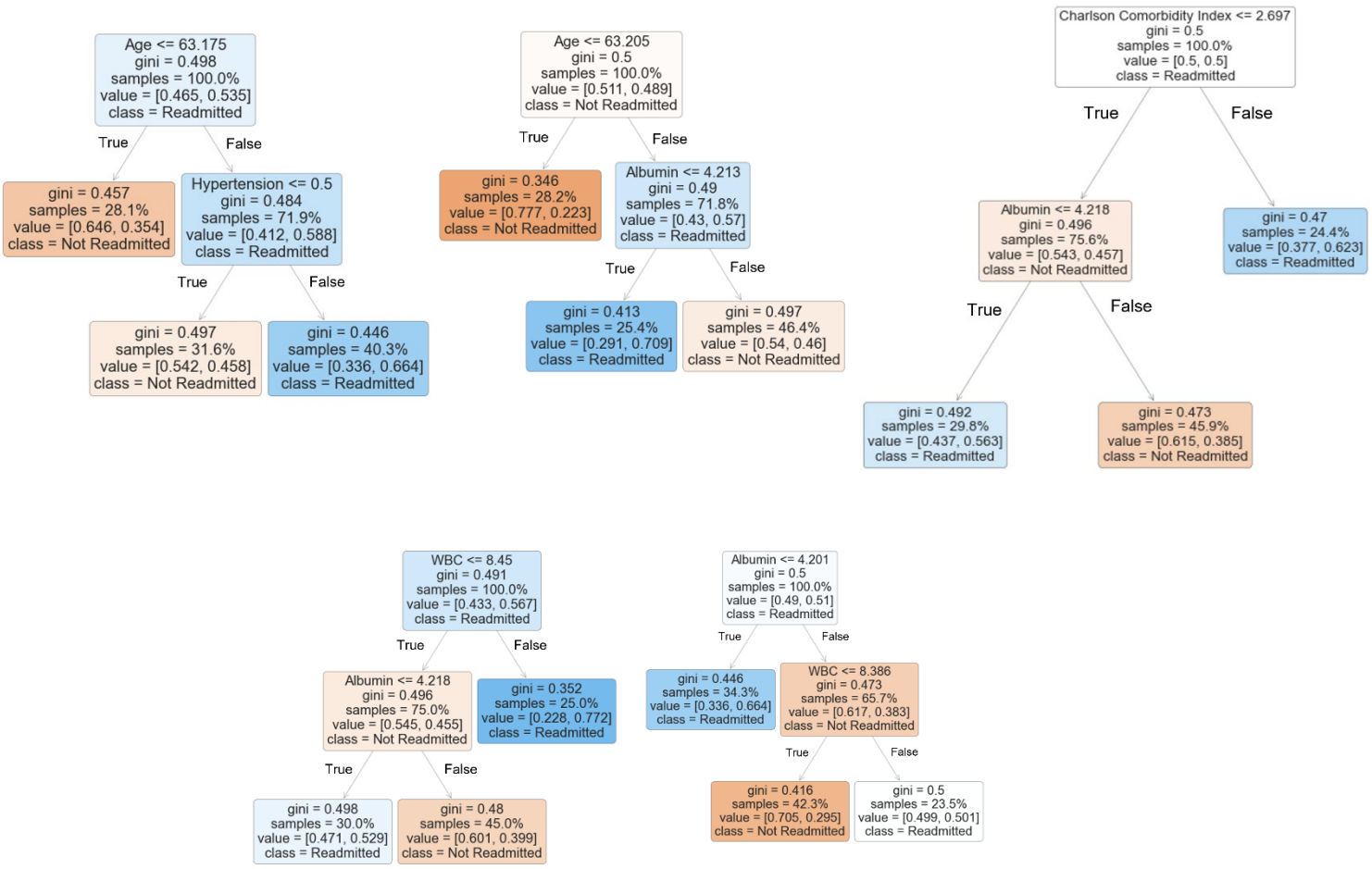


Figure 2. Individual decision trees visualized from the random forest. Left to right: Tree 1, Tree 2, Tree 3, Tree 4, Tree 5.

Comparison of Logistic Regression and Random Forest models

The random forests model had better performance on all assessed metrics (70.1% accuracy, 63.5% sensitivity, 72.2% specificity, 27.8% false positive rate, 0.77 training AUC) as compared to the logistic regression (68.5% accuracy, 49.1% sensitivity, 69.8% specificity, 31.1% false positive rate, 0.7 training AUC) (Table 4). The random forests model had a testing AUC of 0.72 – significantly greater than the logistic regression testing AUC of 0.59 (Table 4, Figure 3). The

difference between training and testing AUC was significantly lower in the random forests (0.05) compared to the logistic regression (0.11) (Table 4).

Scores	Random Forests	Logistic Regression
Accuracy (%)	70.1	68.5
Sensitivity (%)	63.5	49.1
Specificity (%)	72.2	68.9
False Positive Rate (%)	27.8	31.1
AUC (Training)	0.77	0.7
AUC (Testing)	0.72	0.59

Table 4. Model Performance Metrics

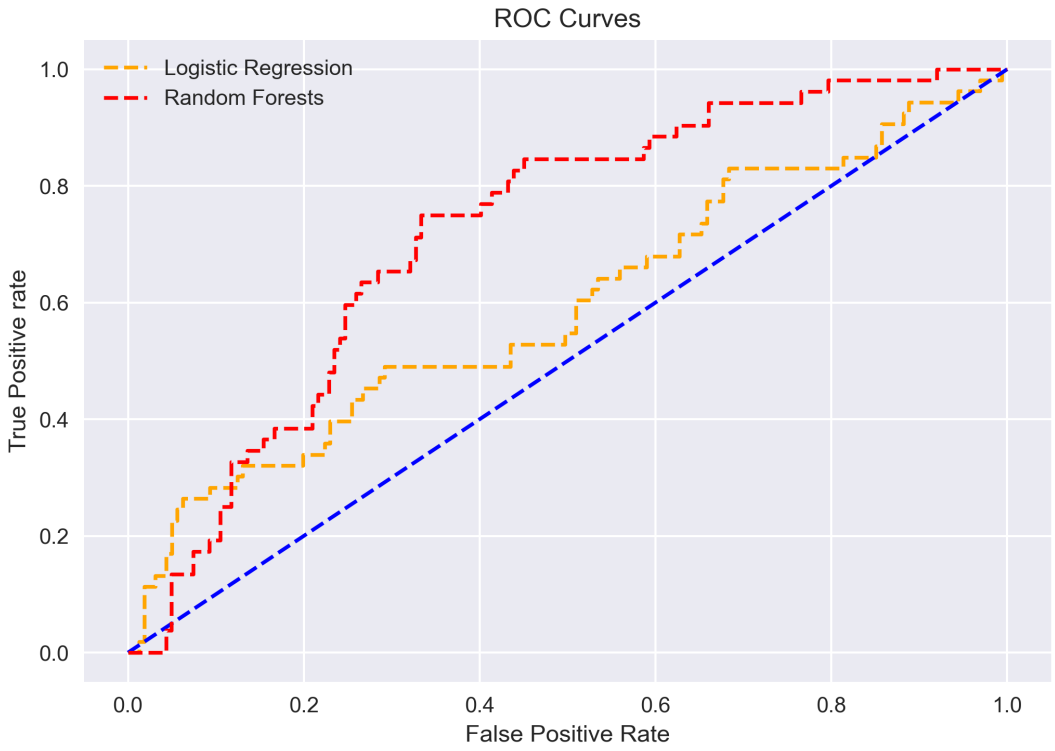


Figure 3. ROC curves for model performance on training data. Random Forests AUC = 0.72. Logistic Regression AUC=0.59.

Discussion

The Random forests had significantly better predictive performance for predicting 90-day readmission as assessed by all commonly reported performance metrics when compared to a logistic regression model. The logistic regression AUC of 0.59 is at par with existing literature on prediction of postoperative readmission [3]. As such, the random forests with AUC of 0.72 is a significant improvement, showing predictive performance that has not yet been shown with existing models.

Generalizability is crucial in the application of predictive models to clinical settings - models must be applicable to external data with good accuracy. However, overfitting is an important concern in clinical settings where there may be a high number of correlated predictors. Existing risk prediction models such as the RCRI and thoracoscore have shown poor generalizability and fail when applied to external data[7,8]. The logistic regression model in our study fits this narrative, and despite being specific to a lung cancer population, shows a huge difference between training (AUC 0.7) and testing performance (AUC 0.59) suggesting that the model is highly overfitted. Comparatively, the random forests shows only a minimal decrease between training AUC (0.77) and testing AUC (0.72), indicating that it is far less overfit, and more generalizable.

Additionally**,** it is also important to note that unlike logistic regression, random forests feature importance is calculated relative to all other predictors in the model without controlling for them. Compared to traditional statistical modelling, this provides a more accurate and realistic assessment when presented with a large set of predictors, which is the reality of any clinical decision-making process.

When predictors are included in a logistic regression model, the traditional teaching is that the number of predictors is limited by the degrees of freedom allowed in a model. This introduces clinical biases into selecting which features are included and may not identify important new relationships in the data. In a random forests model, there is no selection of which features are included in the model, with all available features available entered. This allows data scientists to identify previously unknown relationship between predictors and outcomes. Random forests can take thousands to millions of features while remaining robust to overfitting and maintaining good predictive performance.[9] This allows for inclusion of both a greater number of predictors and many more types of predictors. Random forests is thus able to take advantage of the wealth of clinical data that has become available over the last decade and has the potential to determine predictor importance towards clinical outcomes with a greater degree of granularity than is currently known. In our model, the random forests feature importance ranking showed several strong associations between preoperative laboratory data and 90 day post operative readmission. Our model suggests that the importance of WBC and hematocrit levels as predictors for readmission may outrank conventionally important predictors such as COPD severity. Despite the limitations of our study and model, this suggests that using preoperative laboratory data as predictors for risk assessment may be worth validation. Random forests models are not limited to numerical features alone and can also handle imagining – which could provide insight that cannot be determined by traditional statistical models.

The intuitive ability to calculate feature importance and visualize decision trees makes random forests highly interpretable. It thus has the potential for many applications in clinical decision

making. One possible novel application is in the context of clinical intervention. Feature importance and decision trees combined can identify which clinically modifiable variables are most important towards predicting an outcome, and thus identifying points of clinical intervention. In our model, we see that albumin, which is a marker of nutrition, is an important predictor for 90 day post operative readmission. Were this claim to be validated, it could serve as point for clinical intervention, wherein change in the patient's nutritional status may lead to better clinical outcomes, such as reduced readmission. Along those lines, random forests may help select the ideal candidates for prehabilitation for lung cancer resection. Although recent guidelines highlight the importance of prehabilitation in reducing postoperative complications, there is a severe lack of data on how to implement it effectively [26]. Ideally all patients would be sent for prehabilitation, but the reality of any healthcare system is that there are limited resources – patients who would benefit the most should be prioritized for the intervention to be cost effective. Random forests has the potential to serve as an effective tool in analyzing a huge number of preoperative variable and determining which factors make the best prehabilitation candidates.

Limitations

Our study comes with several limitations. Although random forests are known to be effective even with small sample sizes, our sample size (N=535) is still relatively small. Further validation which a larger sample is crucial to determining the validity of our model. Several of our continuous variables, particularly the preoperative labs had missing data. Missing data was imputed by multiple imputation under the assumption that it was missing at random. Unfortunately, pulmonary function testing was surprisingly not available for 62% of participants

as they had their lung function testing outside the MUHC. We decided to include the variable as "PFT available" or "PFT not available" in order to capture a patient who would have had his entire investigation at the MUHC, by specialized teams in respirology and thoracic surgery, versus a patient who only had surgery at the MUHC and had pulmonary function testing at a smaller hospital. Future work should identify the difference in the various data points available in the PFT such as RV/TLC ratio, FEV1 and Mean Inspiratory capacity which is more susceptible to improvement with inspiratory muscle training. Several variables from NSQIP are not clearly defined as they are collected by registrars without standardized definitions. For example neither dyspnea nor history of severe COPD were defined by standardized clinical scales such as the Medical Research Council (MRC) dyspnea scale or the GOLD system[27,28]. Data collection and quality is an issue that is at the core of predictive modelling – model utility depends greatly on data quality and is a fundamental. As previously noted, performance in external datasets is of foremost importance when assessing predictive models. Although our data was split into training and testing sets, our model was not externally validated.

Conclusions

The random forests model exhibits improved predictive performance and lower overfitting as compared to logistic regression and may be better suited to the prediction of complex clinical outcomes such as 90-day readmission and may improve risk stratification in lung cancer patients. Random forests feature importance and decision trees have several possible clinical applications and can be a valuable tool for clinical decision making.

References

1. Facts & Figures 2020 Reports Largest One-year Drop in Cancer Mortality. Accessed February 11, 2022. https://www.cancer.org/latest-news/facts-and-figures-2020.html

2. Andalib A, Ramana-Kumar AV, Bartlett G, Franco EL, Ferri LE. Influence of postoperative infectious complications on long-term survival of lung cancer patients: a population-based cohort study. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2013;8(5):554-561. doi:10.1097/JTO.0b013e3182862e7e

3. Hu Y, McMurry TL, Stukenborg GJ, Kozower BD. Readmission Predicts 90-Day Mortality Following Esophagectomy: Analysis of SEER-Medicare Outcomes. *J Thorac Cardiovasc Surg*. 2015;150(5):1254-1260. doi:10.1016/j.jtcvs.2015.08.071

4. Bouabdallah I, Pauly V, Viprey M, et al. Unplanned readmission and survival after video-assisted thoracic surgery and open thoracotomy in patients with non-small-cell lung cancer: a 12-month nationwide cohort study. *Eur J Cardio-Thorac Surg Off J Eur Assoc Cardio-Thorac Surg*. 2021;59(5):987-995. doi:10.1093/ejcts/ezaa421

5. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. 1999;100(10):1043-1049. doi:10.1161/01.cir.100.10.1043

6. Falcoz PE, Conti M, Brouchet L, et al. The Thoracic Surgery Scoring System (Thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *J Thorac Cardiovasc Surg*. 2007;133(2):325-332. doi:10.1016/j.jtcvs.2006.09.020

7. Wotton R, Marshall A, Kerr A, et al. Does the revised cardiac risk index predict cardiac complications following elective lung resection? *J Cardiothorac Surg*. 2013;8:220. doi:10.1186/1749-8090-8-220

8. Qadri SSA, Jarvis M, Ariyaratnam P, et al. Could Thoracoscore predict postoperative mortality in patients undergoing pneumonectomy? *Eur J Cardiothorac Surg*. 2014;45(5):864-869. doi:10.1093/ejcts/ezt517

9. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

10. McKinney W. Data Structures for Statistical Computing in Python. *Proc 9th Python Sci Conf*. Published online 2010:56-61. doi:10.25080/Majora-92bf1922-00a

11. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

12. Gasparini A. *The {comorbidity} Package: Computing Comorbidity Scores*.; 2022. Accessed February 11, 2022. https://github.com/ellessenne/comorbidity

13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.

14. sklearn.impute.IterativeImputer. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/generated/sklearn.impute.IterativeImputer.html

15. sklearn.model_selection.train_test_split. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/generated/sklearn.model_selection.train_test_split.html

16. sklearn.utils.class_weight.compute_class_weight. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

17. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441

18. sklearn.ensemble.BaggingClassifier. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

19. sklearn.ensemble.RandomForestClassifier. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

20. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng*. 2007;9(3):90-95. doi:10.1109/MCSE.2007.55

21. Waskom ML. seaborn: statistical data visualization. *J Open Source Softw*. 2021;6(60):3021. doi:10.21105/joss.03021

22. Feature importances with a forest of trees. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/auto_examples/ensemble/plot_forest_importances.html

23. sklearn.tree.plot_tree. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/generated/sklearn.tree.plot_tree.html

24. 3.3. Metrics and scoring: quantifying the quality of predictions. scikit-learn. Accessed February 11, 2022. https://scikit-learn/stable/modules/model_evaluation.html

25. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1(1):26-31. doi:10.1093/jamiaopen/ooy012

26. Mahendran K, Naidu B. Prehabilitation in lung cancer resection—are we any closer to the ideal program? *J Thorac Dis*. 2020;12(4):1628-1631. doi:10.21037/jtd.2020.02.15

27. Stenton C. The MRC breathlessness scale. *Occup Med Oxf Engl*. 2008;58(3):226-227. doi:10.1093/occmed/kqm162

28. Global Strategy for the Diagnosis, Management and prevention of Chronic, Obstructive Pulmonary Disease, 2022 Report. http://goldcopd.org/

CHAPTER 4: Bridging Chapter

Lung cancer remains the leading cause of cancer death in Canada[2]. Quebec specifically has one of the highest age standardized mortality rates in the country[4]. Lung cancer survival rates have not seen significant improvement in years, and significantly lags behind that of other common cancers including breast, prostate and colorectal[3]. As such there is an overall drive to improve survival rates. Improving lung cancer outcomes is complicated and multifaceted, and thus requires varied approaches in order to ensure optimal outcomes. Screening and early diagnoses, accurate risk assessment and early intervention are all key strategies to this end. Manuscript 1 addresses the need to improve perioperative risk stratification for early-stage lung cancer and offer machine learning, particularly random forests models, as a potential solution to the current lack of accuracy in this regard. Smoking cessation is an important tool for clinicians, public health providers and program leaders have to improve health outcomes. Smoking cessation is beneficial in reducing post operative complications after a lung cancer resection, reducing cardiac complications and improving wound healing.[43,44] Smoking cessation delivery is varied in the health care system. Integrating smoking cessation upstream, at a point where patients are presenting for lung cancer screening, and before they are even diagnosed with lung cancer, may provide the largest benefit if done well. Smoking abstinence combined with lung cancer screening has shown to provide additional mortality of 10%[19]. If implemented effectively, smoking cessation programs may have a huge impact in reducing lung cancer mortality. Yet there is an acute lack of data assessing the efficacy of various smoking cessation services. Consequently, the next chapter (manuscript 2) assesses the effectiveness of smoking cessation by phone counselling in the McGill Lung Cancer Screening Trial.

CHAPTER 5: Manuscript 2 "Smoking cessation by phone counselling in a lung cancer screening program: a retrospective comparative cohort study"

Smoking cessation by phone counselling in a lung cancer screening program: a retrospective comparative cohort study

Ankita Ghatak[1] , Sean Gilman MD[4] , Siobhan Carney BScN[4], Anne V Gonzalez MD MSc[1,4], Andrea Benedetti PhD[2,3], Nicole Ezer MD MPH[1,2,3,4]

5. Division of Experimental Medicine, Department of Medicine, McGill University, Montreal, Quebec, Canada
6. Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada Center for Outcomes Research and Evaluation, McGill University Health Centre, Montreal, Quebec, Canada
7. Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Quebec, Canada
8. Department of Medicine, Division of Respiratory Medicine, McGill University Health Centre, Montreal, Quebec, Canada

Corresponding author:

Nicole Ezer, MD, MPH, FRCPC
Assistant Professor of Medicine
nicole.ezer@mcgill.ca

Office 3F.52
Centre for Outcomes Research and Evaluation
5252 De Maisonneuve Boulevard
Montréal, Québec Canada
H4A 3S9

Abstract word count: 293

Abstract

Introduction: Smoking cessation integration within lung cancer screening programs is challenging. Currently phone counselling is available across Canada for individuals referred by health care workers, and by self-referral. We compared quit rates after phone counselling intervention between participants who self-refer, those referred by healthcare workers, and those referred by a lung cancer screening program.

Methods: This is a retrospective cohort study of participants referred to provincial smoking cessation quit-line in contemporaneous cohorts: self-referred participants, healthcare worker referred, and those referred by a lung cancer screening program if they were still actively smoking at the time of first contact. Baseline covariates (sociodemographic information, smoking history, and history of mental health disorder) and quit intentions (stage of change, readiness for change, previous use of quit programs, and previous quit attempts) were compared among the three cohorts. Our primary outcome was defined as self-reported 30-day abstinence rates at 6 months. Multivariable logistic regression was used to identify whether group assignment was associated with higher quit rates.

Results : Participants referred by a lung cancer screening program had low quit rates (12%, CI: 5-19%) at six months despite the use of phone counselling. Compared to patients who were self-referred to the smoking cessation phone help line, individuals referred by a lung cancer screening program were much less likely to quit (adjusted OR 0.37; 95% CI 0.17-0.8), whereas those referred by healthcare workers were twice as likely to quit (adjusted OR 2.16 (1.3-3.58) even after adjustment for differences in smoking intensity and quit intentions.

Conclusions: Phone counselling alone has very limited benefit in a lung cancer screening program. Participants differ significantly from those who are otherwise referred by healthcare workers. Screening programs should tailor smoking cessation interventions to differences in quit intentions and higher nicotine dependence.

Introduction

Across Canada multiple provinces are implementing lung cancer screening following results from multiple large trials showing a mortality benefit of low dose CT screening [1,2,3,4] and cost-effectiveness in Canada[5]. Given the significant mortality benefit of smoking cessation in combination with lung cancer screening[6], as well as the high proportion of active smokers in screening programs, there is interest using engagement in screening as a teachable moment to promote smoking cessation. In a secondary analysis of the National Lung Screening trial the 7-year smoking abstinence in the control arm (i.e., who underwent chest X-ray) was equivalent to a 20% reduction in lung cancer-specific mortality[6]. The authors note that this reduction is equivalent to the mortality benefit of three annual CT screening rounds. Combined abstinence and CT screening was associated with an almost twofold increase in benefit, resulting in a 38% reduction in lung cancer death, HR: 0.62 (95% CI: 0.51–0.76). Smoking cessation provides an additional 10% mortality benefit when combined with LDCT screening[1,6].

Participation in a lung cancer screening program without a smoking cessation intervention has been shown not to increase quit rates[3]. In the Cancer Care Ontario pilot lung cancer screening program, in person counselling is offered to participants[7]. However, in the context of reduced health care resources since the onset of the COVID-19 pandemic, many programs may only be able to integrate smoking cessation services into existing phone counselling programs because of lack of manpower. Studies assessing phone counseling in patients screened for lung cancer as part of clinical trials are limited [8 9 10]. Little information is available on the comparative effectiveness of programs between participants referred by health care workers (nurses, community pharmacists and physicians), participants who are self-referred, and those referred by a lung cancer screening program. There is a need for evidence based data

for provincial lung cancer screening programs as they deploy resources to increase smoking cessation rates among participants. Additionally, understanding how lung cancer screening referred participants differ from those traditionally referred to smoking quit-lines will allow for interventions to be tailored to address these differences.

We performed a retrospective cohort study to assess the effectiveness of telephone counselling for smoking cessation in lung cancer screening eligible participants from the McGill Lung Cancer Screening Pilot program to participants referred by healthcare workers, and those who self-refer to the Quebec's smoking cessation helpline (Ligne J'arrete).

Materials and Methods

Participants were matched by date and referral status in a 2:2:1 ratio. They were identified as self-referred, those referred by healthcare workers in the context of usual clinical care by nurses, pharmacists, or their family doctors, and those referred by the McGill Lung Cancer Screening Pilot program between 2019-2020. Participants who called the lung cancer screening program and identified as active smokers at time of first contact with the program were referred to the smoking cessation quit-line. Participants who engaged with the lung cancer screening program were those who had at a minimum one phone counseling intervention with the quit-line, we excluded participants who had quit smoking by the time they were contacted by the quit program. The lung cancer screening program followed the Quebec's Institut National d'Excellence en Santé et Services Sociaux recommendation to determine lung cancer screening eligibility with a 6-year lung cancer risk greater than or equal to 2% using the PLCO m2012 risk prediction model[5]**.** Participants who were not eligible for screening based on PLCO and/or age

but initiated contact with the program and identified as active smokers were referred to the quit-line and were included in our analysis.

Baseline sociodemographic information such as age and sex, highest educational attainment (less than high school, some training after high school, high school graduate, college graduate, postgraduate) were collected in all participants. History of mental health disorder was defined as (anxiety, bipolar, clinical depression, seasonal depression, pathological gambling, schizophrenia, , eating disorder, border personality disorder, drug or alcohol use) was collected by the smoking cessation phone line by self-report.

To quantify smoking we collected time to first cigarette (within the first 5 minutes after waking up, between 6 and 30 minutes after waking up, between 31 and 60 minutes after waking up, more than 60 minutes), number of cigarettes used per day (at baseline) and heaviness of smoking index. Heaviness of smoking index uses a 6 point scale and combines data on baseline cigarette use per day and time to first cigarette (reference). It was used to compare nicotine dependence in the three groups[11]. Participants stage of change was categorized using the transtheoretical model of behavioral stages of change to categorize them as precontemplation (no thoughts of quitting), contemplation (thinking about quitting) , preparation (planning to quit in the next 30 days), action (quitting successfully for up to six months), maintenance (no smoking for more than six months)[12]. Previous quit attempts, previous use of pharmacological intervention (nicotine patches, gum, etc.) and previous use of quit-lines were defined as “yes/no”. Readiness for change, both importance and confidence in quitting, were measured on a 10-point scale, with 1 being very low and 10 being very high.

The primary outcome was self-reported 30-day abstinence rates at 6 months after first contact with the phone quit-line. Baseline sociodemographic information, history of mental health disorders and smoking data and heaviness of smoking index were compared between the three groups using a chi squared test, one-way ANOVA or Kruskal-Wallis test where appropriate. Quit intentions were compared in the three groups using a chi-squared test.

Missing data was imputed using Multiple Imputation by Chained Equations (MICE) using 250 iterations and pooled 50 imputed datasets to get the final dataset[13]. A multivariate logistic regression was developed to determine the impact of group allocation using the following variables of interest: age, gender (male, female), education (collapsed into less than high school, more than high school), time to first cigarette (collapsed into within 5 minutes after waking up, more than 5 minutes after waking up), cigarette use per day (at baseline), and group (self-referred, healthcare worker referred, McGill Lung Cancer Screening Pilot program referred).

All data was cleaned and analyzed using Python [14] and R[13]. The study was approved by the Research Ethics Board of the McGill University Health Center.

Results and Discussion

Results

A total of 417 active smokers were included in the study, 176 (42%) were self-referred arm, 165 (40%) were referred by health care workers (family doctors, nurses, pharmacists), and 76 (18%) were referred by the lung cancer screening program. Three individuals were excluded from the study as they had quit smoking after contacting the lung cancer screening program and prior to being contacted by the smoking cessation quit line. As expected, mean age was highest

in the lung cancer screening referred group (63 years, standard deviation (SD) 6), and was younger in the self-referred (53 years, SD 15) and healthcare worker referred (49 years, SD 13) groups ($p<0.001$). The lung cancer screening referred group had lower educational attainment, with the majority of participants (26.3%) having a less than a high school education ($p<0.001$), those referred by healthcare workers had higher educational attainment (32.1% college graduates and 20.6% postgraduates).

There was no substantial difference in nicotine dependence between the three groups. Overall, the majority of participants smoked within 5 minutes of waking up (48.7%) or within 6 and 30 minutes of waking up (24.2%). This trend was the same in the three groups: self-referred (55.1% within 5 minutes of waking up, and 19.3% within 6 and 30 minutes of waking up), healthcare worker referred (46.7% and 23.6%), and lung cancer screening referred (38.2% and 36.8%) ($p<0.001$). The mean number of cigarettes used per day was highest amongst those referred by the lung cancer screening program (median 20 per day; IQR 13-25), followed by the self-referred group (median 18 per day; IQR 10-25) and the healthcare worker referred group (median 16 per day, IQR 10-25), although the difference was not statistically significant between the three groups ($p>0.05$). Additionally, participants had on average moderate nicotine dependence defined using the heaviness of index. Despite the higher number of cigarettes used per day in the lung cancer screening group, the heaviness of smoking index was similar in the three groups ($p>0.05$) (Table 1).

Stage of change was significantly different in the three groups ($p <0.001$) (Table 2). Among the participants referred by the lung cancer screening program the majority were in the first three stages of change – precontemplation (15.8%), contemplation (27.6%) and preparation (43.4%). By comparison, very few self-referred and healthcare worker referred participants were

in the precontemplation (5.7% and 4.8% respectively) or contemplation stages (8.5% and 10.3% respectively). Among participants referred by healthcare workers, most were in the action stage (48.5%) compared to only 9.2% among lung cancer screening referred participants. Participants' quit histories differed significantly between groups. Considerably more lung cancer screening referred participants had had previous quit attempts (36.8%) as compared to self-referred and healthcare worker referred participants (11.9% and 10.9% respectively) ($p<0.001$). A larger proportion of lung cancer screening referred participants reported previous use of pharmacological therapy (30.3%) as compared to self-referred and healthcare worker referred participants (24.4% and 12.7%) ($p<0.01$). Across all groups, the majority of participants reported previous use of quit-lines, with the highest being amongst the healthcare worker referred group (94.5%), followed by the lung cancer screening referred group (92.1%) and the self-referred group (84.1%) ($p<0.01$). Notably, a higher proportion of lung cancer screening referred participants reported mental health disorders (47.4%) as compared to both self-referred and healthcare worker referred participants (39.2% and 28.5% respectively) ($p<0.05$). Both readiness for change measures differed significantly between the three groups. Approximately half of all lung cancer screening referred participants rated their readiness for change – importance in quitting as 5 (19.7%) or 6 (30.3%) - the lowest scores reported in the study – as compared to only a small percentage of self-referred (7.4% and 9.7%) and healthcare worker referred participants (2.4% and 7.9%). As such, a significantly lower proportion of lung cancer screening referred participants rated their importance in quitting as 10 (23.7%) compared to self-referred (57.4%) and healthcare worker referred participants (63%) ($p<0.001$). Similarly, very few lung cancer screening referred participants rated their readiness for change – confidence in quitting as

10 (2.6%) compared to self-referred (7.4%) and healthcare worker referred participants ($p<0.001$) (Table 2).

Overall, 30-day abstinence at 6 months was 30% among all participants (Table 3). Six month quit rates were the lowest amongst participants referred by the lung cancer screening program (12%, CI: 5-19), and highest amongst participants referred by healthcare workers (42%, CI: 35-50) ($p <0.001$). After adjustment for sex, age, education (less than high school or high school or more), baseline cigarette use per day and time to first cigarette (less than or more than 5 minutes from waking up), participants who were referred by healthcare workers were almost twice as likely to quit than those who were self-referred (adjusted OR: 2.12, 95% CI: 1.29-3.51); whereas those participants who were referred by the lung cancer screening program were significantly less likely to quit, even after adjustment (adjusted OR: 0.34, 95% CI: 0.15-0.76) (Table 4).

Discussion

Combining smoking cessation with lung cancer screening by low dose CTs has been shown to be associated with a 38% reduction in death from lung cancer[6]. Although it is evident that smoking cessation should be incorporated into screening programs[15], there is limited evidence on how best to integrate these services. Across studies of participants screened for lung cancer, quit rates with no smoking cessation intervention range from 7-23%[16]. Our 6 month quit rate of 12% (CI: 5-19%) among individuals screened for lung cancer is comparable to similar studies with no smoking cessation intervention[4]. Our quit rates are unchanged even after adjustment for age, smoking intensity and education. Our results show that overall, phone

counselling as a smoking cessation intervention had no additional benefit among participants referred by a lung cancer screening program. This is of significant concern given the significant cost and resources needed to absorb the added volume to smoking cessation quit lines if they systematically will be targeting participants in screening programs.

The prevalent view is that participants referred by a lung cancer screening program to a phone quit-line are similar to participants referred by healthcare workers. However our results demonstrate this is definitely not the case. Participants demonstrate key differences in quit intentions and readiness for change, despite similar use of quit programs in the past. Most notably, a referral by a health care worker outside of a lung cancer screening program is likely a sign that an individual is in the "action" stage of change, whereas lung cancer screening referred participants were more likely to be in pre-contemplation or contemplation stages of change.

Some studies suggest that findings after lung cancer screening low dose CT is performed would help tailor the intervention to encourage cessation using personalized results of the screening study. However a recently published Canadian randomized control trial of a telephone based smoking cessation intervention demonstrated incorporating lung cancer screening results did not result in increased 12-month cessation rates versus written information alone in unselected smokers undergoing lung cancer screening[9]. To optimize cessation interventions in this population behavioral counselling combined with pharmacotherapy are more promising than telephone counselling alone. Cessation rates have been demonstrated to be up to 57% with these strategies in the first six months in clinical trials[17] [18] [19] [20] . Beneficial effects decline after a year and participants increasingly relapse with passage of time, and follow-up sessions might be required to maintain treatment effects[18]. Internet-based interventions such as computer-tailored

cessation advice or a list of internet resources has not shown to be beneficial over standard written information material[21 22].

Our study is limited by our short follow up time of 6 months, and the fact that smoking cessation was not confirmed biochemically. Nevertheless, verbal assessment alone is likely to overestimate the effectiveness of the intervention. Additionally, we used multiple imputation to deal with missing data and under the assumption that data was missing at random. Notably, a complete case analysis showed similar numbers also before and after adjustment, with the lung cancer screening referred group still being significantly less likely to quit, especially compared to the healthcare worker referred group, thus supporting our results.

Conclusions

These findings, along with those from another Canadian randomized clinical trial of smoking cessation integration into a lung cancer screening trial[9], have important implications for lung cancer screening programs across Canada. They suggest options other than phone counseling such as multi-modality interventions with in person motivational interviewing and pharmacotherapy, are more likely to demonstrate clinical effectiveness for lung cancer screening participants.

Tables

	Self-Referred n=176	Healthcare Worker Referred n=165	Lung Cancer Screening Referred n=76	Total n=417	P-Value
Age, mean (SD)	53 (15)	49 (13)	63 (6)	53 (14)	<0.001
Age, median [Q1, Q3]	57 [39,65]	48 [39,59]	63 [59,67]	57 [40,64]	
Sex, n (%)					
F	90 (51)	108 (65.5)	33 (43.4)	231 (55.4)	0.002
M	86 (48.9)	57 (34.5)	43 (56.6)	186 (44.6)	
Education, n (%)					
Less than high school	55 (31.2)	15 (9.1)	20 (26.3)	90 (21.6)	<0.001
Some training after high school	17 (9.7)	24 (14.5)	17 (22.4)	58 (13.9)	
High School Graduate	44 (25.0)	39 (23.6)	12 (15.8)	95 (22.8)	
College Graduate	37 (21.0)	53 (32.1)	9 (11.8)	99 (23.7)	
Postgraduate	23 (13.1)	34 (20.6)	18 (23.7)	75 (18.0)	
Baseline Cigarette Use Per Day, mean (SD)	18 (10)	18 (11)	20 (9)	19 (11)	0.471
Baseline Cigarette Use Per Day, median [Q1, Q3]	18 [10,25]	16 [10,25]	20 [13,25]	18 [10,25]	0.283
Time to First Cigarette, n (%)					
Within the first 5 minutes after waking up	97 (55.1)	77 (46.7)	29 (38.2)	203 (48.7)	<0.001
Between 6 and 30 minutes after waking up	34 (19.3)	39 (23.6)	28 (36.8)	101 (24.2)	
Between 31 and 60 minutes after waking up	14 (8.0)	14 (8.5)	15 (19.7)	43 (10.3)	
More than 60 minutes after waking up	31 (17.6)	35 (21.2)	4 (5.3)	70 (16.8)	
Heaviness of Smoking Index, mean (SD)	3 (2)	3 (2)	3 (1)	3 (2)	0.415
Heaviness of Smoking Index, median [Q1,Q3]	4 [2,5]	3 [2,5]	3 [3,4]	3 [2,5]	0.652

Table 1. Baseline Demographics

	Self Referred	Healthcare Worker Referred	Lung Cancer Screening Referred	Overall	P-Value
	n=176	n=165	n=76	n=417	
Stage Of Change, n (%)					
Precontemplation	10 (5.7)	8 (4.8)	12 (15.8)	30 (7.2)	<0.001
Contemplation	15 (8.5)	17 (10.3)	21 (27.6)	53 (12.7)	
Preparation	100 (56.8)	59 (35.8)	33 (43.4)	192 (46.0)	
Action	44 (25.0)	80 (48.5)	7 (9.2)	131 (31.4)	
Maintenance	7 (4.0)	1 (0.6)	3 (3.9)	11 (2.6)	
Previous Quit Attempts, n (%)					
No	21 (11.9)	18 (10.9)	28 (36.8)	67 (16.1)	<0.001
Yes	155 (88.1)	147 (89.1)	48 (63.2)	350 (83.9)	
Previous Use of Pharmacological Therapy, n (%)					
No	43 (24.4)	21 (12.7)	23 (30.3)	87 (20.9)	0.002
Yes	133 (75.6)	144 (87.3)	53 (69.7)	330 (79.1)	
Previous Use of Quit-lines, n (%)					
No	28 (15.9)	9 (5.5)	6 (7.9)	43 (10.3)	0.005
Yes	148 (84.1)	156 (94.5)	70 (92.1)	374 (89.7)	
Mental Health, n (%)					
No	107 (60.8)	118 (71.5)	40 (52.6)	265 (63.5)	0.011
Yes	69 (39.2)	47 (28.5)	36 (47.4)	152 (36.5)	
Readiness for Change - Importance in Quitting, n (%)					
10	101 (57.4)	104 (63.0)	18 (23.7)	223 (53.5)	<0.001
9	7 (4.0)	20 (12.1)	7 (9.2)	34 (8.2)	
8	27 (15.3)	17 (10.3)	4 (5.3)	48 (11.5)	
7	11 (6.2)	7 (4.2)	9 (11.8)	27 (6.5)	
6	17 (9.7)	13 (7.9)	23 (30.3)	53 (12.7)	
5	13 (7.4)	4 (2.4)	15 (19.7)	32 (7.7)	
Readiness for Change - Confidence in Quitting, n (%)					
10	13 (7.4)	27 (16.4)	2 (2.6)	42 (10.1)	<0.001
9	16 (9.1)	13 (7.9)	3 (3.9)	32 (7.7)	
8	46 (26.1)	49 (29.7)	13 (17.1)	108 (25.9)	
7	27 (15.3)	23 (13.9)	11 (14.5)	61 (14.6)	
6	10 (5.7)	19 (11.5)	1 (1.3)	30 (7.2)	
5	15 (8.5)	12 (7.3)	3 (3.9)	30 (7.2)	
4	3 (1.7)	3 (1.8)	8 (10.5)	14 (3.4)	

	3	38 (21.6)	16 (9.7)	24 (31.6)	78 (18.7)
	2	8 (4.5)	3 (1.8)	11 (14.5)	22 (5.3)

Table 2. Quit Intentions

	Self Referred n=176	**Healthcare Worker Referred** n=165	**Lung Cancer Screening Referred** n=76	**Overall** n=417	**P-Value**
Smoking Status, n (%, 95% CI)					
Smoker	129 (73%, 95% CI 67-80)	95 (58%, 95% CI 50-65)	67 (88%, 95% CI 81-95)	291 (70%, 95% CI 65-74)	<0.001
Quitter	47 (27%, 95% CI 20-33)	70 (42%, 95% CI 35-50)	9 (12%, 95% CI 5-19)	126 (30%, 95% CI 26-35)	

Table 3. Smoking Status of participants at 6 months. Quitter is defined as self reported 30 day abstinence rates at 6 months.

Group	**Unadjusted**	**Adjusted***
Healthcare Worker Referred	2.02 (1.29-3.20)	2.12 (1.29-3.51)
Lung Cancer Screening Referred	0.37(0.16-0.77)	0.34 (0.15-0.76)

Table 4. Logistic Regression with imputed values. Reference group is Control 1 (Self-Referred). *Adjusted for Sex, Education, Age, Time to First Cigarette (categorical), Baseline Cigarette Use per Day (continuous).

Data Availability Anonymized data is stored on the RedCap Database of the Research Institute of the McGill University Health Center. Request for data can be sent to the principal investigator Dr N Ezer and will be addressed on a case by case basis.

Conflicts of Interest None declared. NE works as a consultant for the Programme Quebecois de Cancerologie lung cancer screening demonstration project.

Funding Statement This project was funded by the Association des Pneumologues du Quebec.

Acknowledgments We would like to thank the Quebec Ligne J'arrete team for collaborating on data extraction.

Supplementary Materials none

References

1. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409. doi:10.1056/NEJMoa1102873

2. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med*. 2020;382(6):503-513. doi:10.1056/NEJMoa1911793

3. Pedersen JH, Tønnesen P, Ashraf H. Smoking cessation and lung cancer screening. *Ann Transl Med*. 2016;4(8):157. doi:10.21037/atm.2016.03.54

4. Pastorino U, Silva M, Sestini S, et al. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Ann Oncol Off J Eur Soc Med Oncol*. 2019;30(7):1162-1169. doi:10.1093/annonc/mdz117

5. Tammemagi MC, Schmidt H, Martel S, et al. Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a

single-arm, prospective study. *Lancet Oncol*. 2017;18(11):1523-1531. doi:10.1016/S1470-2045(17)30597-1

6. Tanner NT, Kanodra NM, Gebregziabher M, et al. The Association between Smoking Abstinence and Mortality in the National Lung Screening Trial. *Am J Respir Crit Care Med*. 2016;193(5):534-541. doi:10.1164/rccm.201507-1420OC

7. Evans WK, Truscott R, Cameron E, et al. Implementing smoking cessation within cancer treatment centres and potential economic impacts. *Transl Lung Cancer Res*. 2019;8(Suppl 1):S11-S20. doi:10.21037/tlcr.2019.05.09

8. Cadham CJ, Jayasekera JC, Advani SM, et al. Smoking cessation interventions for potential use in the lung cancer screening setting: A systematic review and meta-analysis. *Lung Cancer Amst Neth*. 2019;135:205-216. doi:10.1016/j.lungcan.2019.06.024

9. Tremblay A, Taghizadeh N, Huang J, et al. A Randomized Controlled Study of Integrated Smoking Cessation in a Lung Cancer Screening Program. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2019;14(9):1528-1537. doi:10.1016/j.jtho.2019.04.024

10. Cahill K, Lancaster T, Green N. Stage-based interventions for smoking cessation. *Cochrane Database Syst Rev*. 2010;(11):CD004492. doi:10.1002/14651858.CD004492.pub4

11. Heatherton TF, Kozlowski LT, Frecker RC, Rickert W, Robinson J. Measuring the heaviness of smoking: using self-reported time to the first cigarette of the day and number of cigarettes smoked per day. *Br J Addict*. 1989;84(7):791-799. doi:10.1111/j.1360-0443.1989.tb03059.x

12. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. *J Consult Clin Psychol*. 1983;51(3):390-395. doi:10.1037//0022-006x.51.3.390

13. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw Artic*. 2011;45(3)(1548-7660):1--67. doi:10.18637/jss.v045.i03

14. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1(1):26-31. doi:10.1093/jamiaopen/ooy012

15. Fucito LM, Czabafy S, Hendricks PS, Kotsen C, Richardson D, Toll BA. Pairing smoking-cessation services with lung cancer screening: A clinical guideline from the Association for the Treatment of Tobacco Use and Dependence and the Society for Research on Nicotine and Tobacco. *Cancer*. 2016;122(8):1150-1159. doi:10.1002/cncr.29926

16. Moldovanu D, de Koning HJ, van der Aalst CM. Lung cancer screening and smoking cessation efforts. *Transl Lung Cancer Res*. 2021;10(2):1099-1109. doi:10.21037/tlcr-20-899

17. Pistelli F, Aquilini F, Falaschi F, et al. Smoking Cessation in the ITALUNG Lung Cancer Screening: What Does "Teachable Moment" Mean? *Nicotine Tob Res Off J Soc Res Nicotine Tob*. 2020;22(9):1484-1491. doi:10.1093/ntr/ntz148

18. Pozzi P, Munarini E, Bravi F, et al. A combined smoking cessation intervention within a lung cancer screening trial: a pilot observational study. *Tumori*. 2015;101(3):306-311. doi:10.5301/tj.5000282

19. Lucchiari C, Masiero M, Mazzocco K, et al. Benefits of e-cigarettes in smoking reduction and in pulmonary health among chronic smokers undergoing a lung cancer screening program at 6 months. *Addict Behav*. 2020;103:106222. doi:10.1016/j.addbeh.2019.106222

20. Filippo L, Principe R, Cesario A, et al. Smoking cessation intervention within the framework of a lung cancer screening program: preliminary results and clinical perspectives from the "Cosmos-II" Trial. *Lung*. 2015;193(1):147-149. doi:10.1007/s00408-014-9661-y

21. van der Aalst CM, de Koning HJ, van den Bergh KAM, Willemsen MC, van Klaveren RJ. The effectiveness of a computer-tailored smoking cessation intervention for participants in lung cancer screening: a randomised controlled trial. *Lung Cancer Amst Neth*. 2012;76(2):204-210. doi:10.1016/j.lungcan.2011.10.006

22. Clark MM, Cox LS, Jett JR, et al. Effectiveness of smoking cessation self-help materials in a lung cancer screening population. *Lung Cancer Amst Neth*. 2004;44(1):13-21. doi:10.1016/j.lungcan.2003.10.001

CHAPTER 6: Summary of findings and final conclusions

Despite lung cancer being the most commonly diagnosed cancer, and the biggest cause of cancer death worldwide, only 5% of global research funding is directed to lung cancer research[60]. There is a distinct lack of research several aspects of lung cancer care – risk stratification, particularly in the context of readmission, and how best to implement effective smoking cessation interventions, are particularly neglected.

The primary goal of this Master's thesis was to contribute towards improvement of lung cancer outcomes. Our studies addressed two key aspects of lung cancer care where there is a distinct lack of data: (1) improving perioperative risk stratification for early-stage lung cancer patients and (2) determining which smoking cessation strategies are legitimately effective in a population of patients presenting for lung cancer screening.

First, Manuscript 1 attempts to address the notable lack of data surrounding readmission risk for pulmonary resection. It showed that machine learning model such as random forests showed significantly better predictive performance in predicting 90-day readmission following pulmonary resection as compared to a traditional statistical model like logistic regression. The random forests model was far less overfitted, indicating that it may have far greater generalizability. This addresses the two key failures of existing readmission risk prediction models: poor performance and poor generalizability. Besides these primary strengths, an additional strength of our study was the inclusion of preoperative laboratory data. Currently there is no such literature using large clinical data such as EHR records, preoperative laboratory data and PFT data for risk prediction. Our model shows that using large clinical data with machine

learning models can help significantly improve risk stratification for prediction of complex clinical outcomes such as readmission and have widespread applications.

The explanatory ability of a model is greatly different from its predictive ability and as manuscript 1 demonstrates, the utility of random forests lies in its predictive capabilities. So while logistic regression can be thought of as an explanatory modelling technique capable of determining effect size, i.e., the odds ratio, random forests is a purely predictive algorithm capable of producing highly accurate predictive models. Although random forests cannot quantify risk directly, using the right combination of predictors as determined by its feature importance metric, it can create powerful decision trees that have incredibly high predictive accuracy.

Note that the goal of this manuscript was to pinpoint the most important underlying patient characteristics predictive of poorer outcomes. To this end we used 90-day post operative readmission only, an outcome we determined to be most reflective of underlying patient characteristics based on our dataset. Determinants of surgical complications could be better captured by a composite outcome of mortality and readmission. Also of note is the exclusion of patients undergoing pneumonectomy, including those who were planned for a lobectomy but converted to a pneumonectomy. We expect this number to be less than 1% of cases, given our total rate of pneumonectomy was 3% (22/731 total patients), and as such do not expect this to significantly bias our outcome.

As discussed, our model determined that albumin – a nutritional marker may be an important predictor. In fact, our model deemed increased WBC count – which can be an indication of

ongoing infection or recurrent COPD exacerbation – to be the most important predictor of 90-day post operative readmission. This may be of clinical significance and provide an additional opportunity for intervention.

Manuscript 2 showed that phone counselling as a smoking cessation intervention for participants enrolling in lung cancer screening has very limited efficacy. In addition to the significantly larger sample sized relative to existing studies, an important strength of this study was the comparison of lung cancer screening referred participants, to participants referred by other healthcare workers in the community. However, our study is limited by the fact that we did not adjust for characteristics such as stage of change or previous quit attempts despite them differing significantly between the three groups. This is due to the substantial amount of missingness in those variables, which as previously noted, is a significant limitation of our study. Our study showed that lung cancer screening referred participants have significantly different smoking behaviours from those who are otherwise referred by healthcare workers.

Future Directions

This thesis establishes random forests as potentially valuable tool for risk assessment in lung cancer patients treated by surgical resection. However, several further investigations must be conducted to confirm the utility of our model. A test of statistical significance must be conducted to validate the superior predictive performance of random forests. Analysis must be conducted on a greater sample size (>1000 patients), perhaps even at a hospital wide level to confirm our findings. Our model must also be externally validated using data outside the MUHC, from hospitals in and outside of Quebec. There are several substantial ways in which it can be improved, each with the possibility to hugely improve risk stratification. Adding key PFT values

(FEV 1, FVC, DLCO) to the model could add significant granularity to risk stratification. Given our results, addition of other preoperative labs such as prealbumin, hemoglobin, etc. could also be further investigated. There is also a huge potential to integrate even further large scale clinical data such as genomic data and imaging data from CT scans. Imaging data may be of particular use as computerized image analysis approaches such as radiomics can be used to assess sarcopenia, emphysema, and cardiac disease severity. With the appropriate development – high quality big data and proper validation – a machine learning model such as random forests may prove to be a useful clinical decision making tool.

The second project in this thesis ascertains the need for alternative smoking cessation interventions in participants presenting for lung cancer screening. Their quit intentions and smoking behaviours differ significantly from the traditional individuals engaging with a phone quit line (self referred individuals and individuals referred by health care workers in the community). Smoking cessation interventions should be tailored to the needs of this population to be successful.

CHAPTER 7: References

1. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. In: Ahmad A, Gadgeel S, eds. *Lung Cancer and Personalized Medicine: Current Knowledge and Therapies*. Advances in Experimental Medicine and Biology. Springer International Publishing; 2016:1-19. doi:10.1007/978-3-319-24223-1_1

2. Lee S. Lung cancer statistics. Canadian Cancer Society. Accessed February 11, 2022. https://cancer.ca/en/cancer-information/cancer-types/lung/statistics

3. Facts & Figures 2020 Reports Largest One-year Drop in Cancer Mortality. Accessed February 11, 2022. https://www.cancer.org/latest-news/facts-and-figures-2020.html

4. Lung cancer. Accessed February 11, 2022. https://www.quebec.ca/en/health/health-issues/cancer/lung-cancer

5. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non–Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clin Proc Mayo Clin*. 2008;83(5):584-594.

6. Spira A, Ettinger DS. Multidisciplinary Management of Lung Cancer. *N Engl J Med*. 2004;350(4):379-392. doi:10.1056/NEJMra035536

7. Schabath MB, Cote ML. Cancer Progress and Priorities: Lung Cancer. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2019;28(10):1563-1579. doi:10.1158/1055-9965.EPI-19-0221

8. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol*. 2017;7(9):170070. doi:10.1098/rsob.170070

9. Cancer of the Lung and Bronchus - Cancer Stat Facts. SEER. Accessed February 11, 2022. https://seer.cancer.gov/statfacts/html/lungb.html

10. Howlader N, Forjaz G, Mooradian MJ, et al. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *N Engl J Med*. 2020;383(7):640-649. doi:10.1056/NEJMoa1916623

11. Lee S. Survival statistics for non–small cell lung cancer. Canadian Cancer Society. Accessed February 11, 2022. https://cancer.ca/en/cancer-information/cancer-types/lung/prognosis-and-survival/non-small-cell-lung-cancer-survival-statistics

12. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409. doi:10.1056/NEJMoa1102873

13. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med*. 2020;382(6):503-513. doi:10.1056/NEJMoa1911793

14. Health USSGAC on S and. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. U.S. Department of Health, Education, and Welfare, Public Health Service; 1964.

15. Schwartz AG, Cote ML. Epidemiology of Lung Cancer. *Adv Exp Med Biol*. 2016;893:21-41. doi:10.1007/978-3-319-24223-1_2

16. Tammemägi MC, Katki HA, Hocking WG, et al. Selection Criteria for Lung-Cancer Screening. *N Engl J Med*. 2013;368(8):728-736. doi:10.1056/NEJMoa1211776

17. Tammemägi MC, Church TR, Hocking WG, et al. Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. *PLoS Med*. 2014;11(12):e1001764. doi:10.1371/journal.pmed.1001764

18. Pistelli F, Aquilini F, Falaschi F, et al. Smoking Cessation in the ITALUNG Lung Cancer Screening: What Does "Teachable Moment" Mean? *Nicotine Tob Res Off J Soc Res Nicotine Tob*. 2020;22(9):1484-1491. doi:10.1093/ntr/ntz148

19. Tanner NT, Kanodra NM, Gebregziabher M, et al. The Association between Smoking Abstinence and Mortality in the National Lung Screening Trial. *Am J Respir Crit Care Med*. 2016;193(5):534-541. doi:10.1164/rccm.201507-1420OC

20. Pedersen JH, Tønnesen P, Ashraf H. Smoking cessation and lung cancer screening. *Ann Transl Med*. 2016;4(8):9-9. doi:10.21037/atm.2016.03.54

21. Hahn EJ, Rayens MK, Hopenhayn C, Christian WJ. Perceived risk and interest in screening for lung cancer among current and former smokers. *Res Nurs Health*. 2006;29(4):359-370. doi:10.1002/nur.20132

22. Rojewski AM, Tanner NT, Dai L, et al. Tobacco Dependence Predicts Higher Lung Cancer and Mortality Rates and Lower Rates of Smoking Cessation in the National Lung Screening Trial. *Chest*. 2018;154(1):110-118. doi:10.1016/j.chest.2018.04.016

23. Cadham CJ, Jayasekera JC, Advani SM, et al. Smoking cessation interventions for potential use in the lung cancer screening setting: A systematic review and meta-analysis. *Lung Cancer*. 2019;135:205-216. doi:10.1016/j.lungcan.2019.06.024

24. Tremblay A, Taghizadeh N, Huang J, et al. A Randomized Controlled Study of Integrated Smoking Cessation in a Lung Cancer Screening Program. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2019;14(9):1528-1537. doi:10.1016/j.jtho.2019.04.024

25. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology: non-small cell lung cancer, version 1.2022. https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf

26. Howington JA, Blum MG, Chang AC, Balekian AA, Murthy SC. Treatment of Stage I and II Non-small Cell Lung Cancer: Diagnosis and Management of Lung Cancer, 3rd ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2013;143(5, Supplement):e278S-e313S. doi:10.1378/chest.12-2359

27. Ginsberg RJ, Rubinstein LV. Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. Lung Cancer Study Group. *Ann Thorac Surg*. 1995;60(3):615-622; discussion 622-623. doi:10.1016/0003-4975(95)00537-u

28. Hu Y, McMurry TL, Stukenborg GJ, Kozower BD. Readmission Predicts 90-Day Mortality Following Esophagectomy: Analysis of SEER-Medicare Outcomes. *J Thorac Cardiovasc Surg*. 2015;150(5):1254-1260. doi:10.1016/j.jtcvs.2015.08.071

29. Bouabdallah I, Pauly V, Viprey M, et al. Unplanned readmission and survival after video-assisted thoracic surgery and open thoracotomy in patients with non-small-cell lung cancer: a 12-month nationwide cohort study. *Eur J Cardio-Thorac Surg Off J Eur Assoc Cardio-Thorac Surg*. 2021;59(5):987-995. doi:10.1093/ejcts/ezaa421

30. Andalib A, Ramana-Kumar AV, Bartlett G, Franco EL, Ferri LE. Influence of postoperative infectious complications on long-term survival of lung cancer patients: a population-based cohort study. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2013;8(5):554-561. doi:10.1097/JTO.0b013e3182862e7e

31. Brunelli A, Kim AW, Berger KI, Addrizzo-Harris DJ. Physiologic Evaluation of the Patient With Lung Cancer Being Considered for Resectional Surgery: Diagnosis and Management of Lung Cancer, 3rd ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2013;143(5, Supplement):e166S-e190S. doi:10.1378/chest.12-2395

32. Falcoz PE, Conti M, Brouchet L, et al. The Thoracic Surgery Scoring System (Thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *J Thorac Cardiovasc Surg*. 2007;133(2):325-332. doi:10.1016/j.jtcvs.2006.09.020

33. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. 1999;100(10):1043-1049. doi:10.1161/01.cir.100.10.1043

34. Wotton R, Marshall A, Kerr A, et al. Does the revised cardiac risk index predict cardiac complications following elective lung resection? *J Cardiothorac Surg*. 2013;8:220. doi:10.1186/1749-8090-8-220

35. Brunelli A, Varela G, Salati M, et al. Recalibration of the Revised Cardiac Risk Index in Lung Resection Candidates. *Ann Thorac Surg*. 2010;90(1):199-203. doi:10.1016/j.athoracsur.2010.03.042

36. Bilimoria KY, Liu Y, Paruch JL, et al. Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aid and Informed Consent Tool for Patients and Surgeons. *J Am Coll Surg*. 2013;217(5):833-842.e3. doi:10.1016/j.jamcollsurg.2013.07.385

37. Samson P, Robinson CG, Bradley J, et al. The National Surgical Quality Improvement Program risk calculator does not adequately stratify risk for patients with clinical stage I non-small cell lung cancer. *J Thorac Cardiovasc Surg*. 2016;151(3):697-705.e1. doi:10.1016/j.jtcvs.2015.08.058

38. Pannell LMK, Reyes EM, Underwood SR. Cardiac risk assessment before non-cardiac surgery. *Eur Heart J - Cardiovasc Imaging*. 2013;14(4):316-322. doi:10.1093/ehjci/jes288

39. Global Strategy for the Diagnosis, Management and prevention of Chronic, Obstructive Pulmonary Disease, 2022 Report. http://goldcopd.org/

40. Kearney DJ, Lee TH, Reilly JJ, DeCamp MM, Sugarbaker DJ. Assessment of Operative Risk in Patients Undergoing Lung Resection: Importance of Predicted Pulmonary Function. *Chest*. 1994;105(3):753-759. doi:10.1378/chest.105.3.753

41. Brunelli A, Refai M, Salati M, Xiumé F, Sabbatini A. Predicted versus observed FEV1 and DLCO after major lung resection: a prospective evaluation at different postoperative periods. *Ann Thorac Surg*. 2007;83(3):1134-1139. doi:10.1016/j.athoracsur.2006.11.062

42. Qadri SSA, Jarvis M, Ariyaratnam P, et al. Could Thoracoscore predict postoperative mortality in patients undergoing pneumonectomy? *Eur J Cardiothorac Surg*. 2014;45(5):864-869. doi:10.1093/ejcts/ezt517

43. Lugg ST, Agostini PJ, Tikka T, et al. Long-term impact of developing a postoperative pulmonary complication after lung surgery. *Thorax*. 2016;71(2):171-176. doi:10.1136/thoraxjnl-2015-207697

44. Mills E, Eyawo O, Lockhart I, Kelly S, Wu P, Ebbert JO. Smoking Cessation Reduces Postoperative Complications: A Systematic Review and Meta-analysis. *Am J Med*. 2011;124(2):144-154.e8. doi:10.1016/j.amjmed.2010.09.013

45. Hosmer, D.W., Jr., Lemeshow, S. and Sturdivant, R.X. Introduction to the Logistic Regression Model. In: *Applied Logistic Regression (Eds D.W. Hosmer, S. Lemeshow and R.X. Sturdivant)*. ; 2013. https://doi.org/10.1002/9781118548387.ch1

46. Tolles J, Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*. 2016;316(5):533-534. doi:10.1001/jama.2016.7653

47. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441

48. Ensor J, Martin EC, Riley RD. *Pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model*.; 2022. Accessed February 15, 2022. https://CRAN.R-project.org/package=pmsampsize

49. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

50. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81-106. doi:10.1007/BF00116251

51. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification And Regression Trees*. Routledge; 2017. doi:10.1201/9781315139470

52. Kotsiantis SB. Decision trees: a recent overview. *Artif Intell Rev*. 2013;39(4):261-283. doi:10.1007/s10462-011-9272-4

53. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform*. 2013;14(3):315-326. doi:10.1093/bib/bbs034

54. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. doi:10.1007/BF00058655

55. Altman N, Krzywinski M. Ensemble methods: bagging and random forests. *Nat Methods*. 2017;14(10):933-934. doi:10.1038/nmeth.4438

56. Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. *Crit Care*. 2004;8(6):508-512. doi:10.1186/cc3000

57. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw Artic*. 2011;45(3)(1548-7660):1--67. doi:10.18637/jss.v045.i03

58. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40-49. doi:10.1002/mpr.329

59. Wilson S. The MICE Algorithm. Published June 9, 2021. Accessed February 11, 2022. https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html