J Cancer Clini Trials
Volume 6:3, 2021

Journal of Cancer Clinical Trial

Research Article

Open Access

A Risk Score to Evaluate the Risk of Lung Cancer in People with HIV

Vincent J. Amoruccio*, Dinesh P. Mital and Shankar Srinivasan

Department of Health Informatics, Rutgers, The State University of New Jersey, USA

Abstract

The advent of antiretroviral therapy (ART) changed the prognosis of HIV. People with HIV (PWH) live longer lives but are susceptible to the same age-related disease as people without HIV. Cancer is a leading cause of death for PWH, and lung cancer is the leading cause of cancer-related death. Smoking and increased age are the primary causes of lung cancer in the general population; however, risk factors specific to HIV, such as immunocompetence and respiratory disease, present additional lung cancer risk in PWH. Existing guidance from the National Lung Cancer Screening Trial (NLST) excludes PWH, does not consider HIV-specific risk factors, and misses significant amounts of lung cancer cases in PWH when it is applied. This deficiency has led to increased incidence and mortality from lung cancer and at younger ages and more advanced stages.

This study integrated two longitudinal, multi-city, multi-center cohorts made publicly available by Johns Hopkins University to address this urgent need. Predictive models used logistic regression with a forward selection of demographic, smoking, pulmonary, and immunocompetency variables conditioned on two levels of gender, race (white vs. black/other), and smoking status (smoker vs. non-smoker) to predict lung cancer. The risk score was multiplicative on each subject's gender, race, and smoking status and then summed to create a single risk score stratified using quartiles of predicted risk among those with PWH and lung cancer.

Using an integrated dataset consisted of 12,320, a lung cancer diagnosis occurred in 100 people out of 7,607 with HIV. Ten people with HIV with lung cancer and ten without lung cancer were excluded and reserved for a parallel sub-study making the analytic sample 7,587 HIV-positive men and women. Lung cancer was predicted in the analytical sample using a summed risk score after multiplying risk across predictors across the three stratifications and resulted in sensitivity at 77% and specificity at 60%. In this analytical sample, the existing NLST criteria missed 97% of the lung cancer cases.

This study is the first to demonstrate that traditional, HIV-specific, and respiratory risk factors can develop a risk score to assess lung cancer risk in PWH. Predictive models conditioned on smoking status, gender, and race independently identified risk factors for lung cancer that summed to a single risk score. Traditional risk factors such as age, education, and ethnicity are significant predictors of lung cancer risk. A history of reoccurring pneumonia and respiratory disease and clinical factors that describe immunocompetency are HIV-specific predictors for lung cancer risk in PWH. The results of this study are significantly different compared to the predictors used in the general population and have dramatically improved accuracy.

Keywords: Human Immunodeficiency Virus • HIV • Acquired Immunodeficiency Deficiency Syndrome • AIDS • Lung cancer • Risk score • Prevention • Earlier diagnosis

Introduction

Lung cancer is the leading non-AIDs defining cancer (NADC) and the leading cause of cancer-related death in PWH [1-5]. In the post-ART era, PWH are living longer but are susceptible to age-related illnesses such as cancer. Lung cancer has had the highest number of deaths of any NADC and will remain to be the leading cause of death through 2030 [6]. Lung cancer has a worse prognosis in PWH as compared to the general population. While there is a lack of 5-year overall survival reported in the literature, studies have reported a significantly reduced survival in PWH compared to the general population. In a cohort study of 80 HIV positive men and 507 HIV negative men, a statistically significant difference was reported in 5-year survival between HIV positive and HIV negative men. The study showed a reduced 5-year survival of 9.5% in HIV positive men compared to 19.3% in HIV negative men [7]. There are several factors contributing to poor survival from lung cancer in PWH. PWH are less likely to receive treatments compared to the general population and are typically excluded from clinical trials. Outside of clinical trials, treatment guidelines for lung cancer in PWH do not exist due to a lack of understanding and fear of drug-to-drug interactions between chemotherapies and ART. Additionally, PWH may have more post-operative complications from invasive procedures compared to the general population. These reasons have led to a call for enhanced survival strategies for PWH [8].

In addition to higher mortality, the incidence rate of lung cancer in PWH is higher compared to the general population [2,7] and varies from as low as 2 – 4 and as high as 7 – 10 times that of the general population [8,9]. Some have theorized that the increased incidence of lung cancer in PWH is related to increases in health encounters but this theory lacked significance in a large cohort study [10]. Others have argued that the increased incidence is due to decreases in all-cause mortality. Lung cancer is considered a unique NADC since it not linked to any viral co-infections as are other ADCs and NADCs. The etiology of lung cancer in PWH is not fully understood beyond smoking. PWH have a higher incidence of smoking compared to the general population. The CDC reports that in 2018 13.7% of adults in the general population were smokers compared to 40% of adult PWH smokers. Several studies have reported significantly higher rates of smoking, almost 100%, and it is believed that most PWH are smokers. The National Lung Cancer Screening Trial (NLST) demonstrated a 20% reduction in mortality among "heavy smokers" in the general population through the use of Low-Dose Computed Topography (LDCT) with 94% and 72% sensitivity and specificity. It showed that LDCT was a superior diagnostic tool compared to CXR. As a result, LDCT became the standard-of-care for screening and diagnosing "high risk" people in the general population. It defined "high risk" as men and women between the ages of 55 and 74 who were either heavy smokers (30+ pack-years of smoking) or former smokers who had quit smoking no more than 15 years prior. This showed that the NLST criteria to screen high

*Corresponding Author: Vincent J. Amoruccio, Department of Health Informatics, Rutgers, The State University of New Jersey, USA, Tel: (212) 951-1953; E-mail: vincent.amoruccio@ gmail.com.

Received: April 09, 2021; **Accepted:** April 23, 2021; **Published:** April 30, 2021

risk people in the general population can detect lung cancer at a stage early enough to prevent significant mortality [11]. The same is not true for PWH. The NLST excluded people who had "serious comorbid conditions" that were either competing risks for death or reduced their chances of surviving lung cancer treatment. PWH were excluded [12] since HIV and AIDS were considered both competing risks of death and contributed to a reduced benefit of treatment. In two large cohort studies of HIV positive patients, more than 70% of lung cancer cases were missed in both studies when applying the NLST guidelines for lung cancer screening [1,13]. They concluded that age and pack-years-of-smoking fell outside of the lower limits of inclusion and that research should be conducted to improve criteria for lung cancer screening in PWH. PWH are considered a "unique high-risk group" [3] who could benefit significantly from improved lung cancer screening and diagnosis. PWH are being diagnosed with lung cancer at advanced stages which are more difficult to treat. The age and smoking limitations of the existing screening criteria are serious gaps in the earlier diagnosis of lung and reduction in mortality from lung cancer in PWH.

Smoking is a significant risk factor for lung cancer in PWH. The same is true for the general population however, smoking has been shown to be a larger threat to PWH. It has been reported that for 1 pack-year-of-smoking, PWH have a 9% increased risk of dying compared to the general population. The increased risk of dying from smoking in PWH is greater than the increased risk from HIV itself. It is not widely understood why smoking impacts PWH greater than it does the general population. Some believe that carcinogens found in tobacco products interact with the body differently and there might be unknown interactions between carcinogens and ART. Many have shown other risk factors, aside from the traditional risks in the general population, explain the excess risk for lung cancer in PWH. HIV infection has been shown to be an independent risk factor for lung cancer, after adjusting for smoking [3, 14-17]. The independent association of HIV to lung cancer is in part due to the immunodeficiency caused by three well studied biomarkers: CD4, CD8, and HIV Viremia (HIV RNA). CD4, CD8, and HIV viremia counts have been widely accepted by many, although not all, as independent risk factors for lung cancer. Several pulmonary comorbidities have also been heavily studied and disputed as risk factors for lung cancer in PWH. The two most studied are Chronic Obstructive Pulmonary Disorder (COPD) and bacterial pneumonia. Both have repeatedly shown significant association with increased incidence of lung cancer for PWH [9,10,14]. One study demonstrated that PWH were 63% more likely to develop lung cancer if they had a pulmonary comorbidity such as bacterial pneumonia. It is relatively unknown whether the actual risk comes from pulmonary disease or the inflammation that is caused by the pulmonary disease. Pulmonary inflammation alone has been reported to increase the risk of lung cancer in addition to COPD and pneumonia [17]. The connection between inflammation and other less cited pulmonary disease such as emphysema, asthma, and occupational lung disease have also been reported to increase the risk of lung cancer in PWH.

The absence of screening guidelines and the lack of understanding for risk factors beyond smoking is a significant problem for PWH. Due to the evidence of increased incidence and mortality, there is an urgent need for an "effective means to reduce lung cancer death in PWH" and more research to determine the optimal strategy for screening are needed [4,8,12,17,18]. The purpose of this study was to create a risk score that would predict the risk of lung cancer in PWH using known risk factors with established associations to lung cancer in PWH. It will provide a viable solution to the unmet need for a risk score that can aide in the prevention and earlier diagnosis of lung cancer in PWH [16].

Methodology

Two separate samples from Johns Hopkins University were curated for analysis, the Multicenter AIDS Cohort Study (MACS) and the Women's Interagency HIV Study (WIHS). MACS was a 35-year study of HIV infection in gay and bisexual men and their families (e.g., partners and spouses) that ran from 1984-2019 [19]. The study was conducted in multiple cities in the United States and contained longitudinal biological and behavioral data collected every six months on more than 7,000 men. The cut-off date for the release of the MACS public-use data set (PDS) was September 30, 2017, and covered Visits 1 through 67 for 7,338 subjects. Like the MACS, the WIHS was a multicenter study conducted in multiple cities in the United States that ran from 1993 to 2019 [19]. It contained longitudinal biological and behavioral data on more than 5,000 HIV-positive women and their families. The WIHS PDS release's cut-off date was September 30, 2015, and covered up to visit 42 for 4,982 subjects. Together, the PDS included 12,320 men and women.

The MACS cohort contained variables to identify cancer cases using the International Classification of Disease for Oncology, 3rd edition (ICD-O-3). Topography code C34 (bronchus and lung) identified 65 lung cancer cases (90% primary). The WIHS cohort differed from the MACS study by splitting cancer outcomes into multiple, single visit, and longitudinal datasets. Some datasets used subject-specified text fields to report cancer type and location (e.g., "Right Lung"). Other datasets used either binary indicators (e.g., "Since your last visit have you been diagnosed with lung cancer?") or outcome codes. WIHS did not use ICD-O-3 codes for the reporting of cancers. Lung cancer cases were mined over each set of data and combined to derive one record per subject and identify lung cancer cases. It was impracticable to discern original ("primary") cancer from metastasized cancer in the WIHS data sets. Data mining identified 61 cases of lung cancer in the WIHS cohort.

The MACS and WIHS were normalized as appropriate and integrated to form one dataset. Approximately 62% (N=7,607) were infected with HIV and formed the analytical sample for this study. Among PWH, 100 lung cancer cases were observed over 73,401 person-years compared to 26 lung cancer cases observed over 63,163 person-years in those without HIV (Figure 1). The incidence of lung cancer in the HIV Positive was 136 versus 41 cases per 100,000 person-years in the HIV Negative. The incidence rate ratio, 3.31, is significantly greater than other studies. Twenty subjects, 10 with and 10 without lung cancer, were randomly excluded and reserved from the analytical sample for validation of a web-based, clinical decision support system as part of a parallel sub-study. As a result, the analytical sample for the study became 7,587 men and women with HIV.

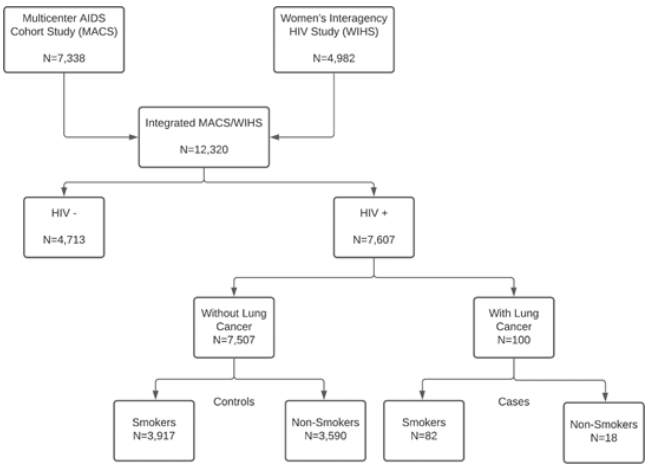

Figure 1. Explanation of analytical sample.

This study assessed traditional and novel risk factors for lung cancer. These include the following demographics, socio-demographic variables, drug-related, smoking, respiratory, and clinical HIV risk factors: Age, Gender, Race, Ethnicity, Body mass index, Annual income, Education, Alcohol use, Marijuana use, Cocaine use, Years smoked, Packs per day, Months since quitting smoking, History of re-occurring pneumonia, History of HIV specific respiratory disease, A diagnosis of AIDS, CD4 cell count, Lowest ("nadir") CD4 cell count, CD8 cell count, CD4/CD8 ratio, Longitudinal fluctuations in CD4 cell count, and HIV RNA viral load.

PWH in this sample with a diagnosis of lung cancer were predominantly smokers who quit smoking in the past 24 months and there were very few 'never-smokers.' As result, smoking status was defined as 'Smoker' if the end-user is either a current smoker or quit smoking within the last 24 months. This definition followed evidence from another study that used a 12-month cut-off [20] instead of a 24-month cut-off for smoking status classification. Twenty four months was selected as the threshold instead of 12 months not only because of the distribution of data, but also due to evidence suggesting that the risk of lung cancer is significantly reduced 24 months after smoking cessation.

Race was condensed into three levels, White, Black, or Other, due to extremely low counts among those with lung cancer. The MACS and WIHS collected five levels for race: White, Black or African American, American Indian or Alaska Native, Asian or Pacific Islander, or Other Specify. Those with lung cancer were predominantly white or black. For analysis, race was collapsed into a binary variable, White versus Black/Other, due to the extremely low counts of "other" races among those with lung cancer.

A lung cancer diagnosis was the study's primary endpoint and was derived as a binary response variable (1= Lung Cancer Diagnosis, 0=No Long Cancer Diagnosis). Logistic regression was used to predict lung cancer from a set of risk factors (predictors). The logistic regression models were stratified into six different multivariable models using two levels of gender, race, and smoking status. These stratifications are due to differences in risk factors by gender, race, and smoking status among PWH who have lung cancer. Gender and racial differences stem from historical, cultural, regional, and socioeconomic disparities in smoking. As a result, this study looked independently at gender (Female vs. Male), race (White vs. Black/Other), and smoking status (Smoker vs. Non-Smoker).

Continuous variables collected longitudinally were analyzed using the last visit on or immediately before lung cancer diagnosis for those with lung cancer or the previous visit in the study for those without lung cancer. For example, suppose a lab collection visit to collect CD4, a continuous variable, occurred on January 1, 2018, another visit occurred on March 1, 2018, and lung cancer diagnosis occurred on February 15, 2018. The visit from January 1, 2018, would be used to analyze CD4 values. Last observation carried forward (LOCF) was used to impute missing data for continuous variables collected over time. Missing values were infrequent for most values, except for HIV viral loads, which had a more significant amount of missing data. Continuous variables were converted into binary indicator variables using the quartiles from those with lung cancer. All continuous risk factors used in logistic regression models were transformed into binary indicator variables. Likewise, categorical variables were converted into binary indicator variables for each category of the variable. The use of indicator variables made it easier to interpret in the predictive models. A total of 51 variables were tested for a significant association with lung cancer for each of the six models, a total of 306 logistic regression models. Risk factors that were statistically significant at 0.1 alpha using the Wald Chi-Square (x^2) test in the bivariable models were selected as candidates for the multivariable models. The p-value cut-off of 0.1 instead of 0.05 was used since the traditional value of 0.05 could erroneously exclude essential predictors [21].

For each of the six logistic regression models, the candidate predictor variables significant at the 0.1 alpha level were added to multivariable models using a forward selection method with an entry criterion at the 0.1 alpha level. In this study, each model was tested using the stepwise selection and backward elimination methods; however, there were no differences in the models compared to the model using forward selection. Due to many candidate predictors, separate multivariable models were performed for demographic variables and another for respiratory and immunocompetency variables. Significant variables from both models were combined for the final multivariable models. Multicollinearity was assessed using a correlation matrix, and all two-by-two correlations greater than 0.5 were evaluated for removal. Additionally, interaction was considered using interaction terms for all possible combinations of variables that remained in the final models. The odds ratios from the six final models served as the probability for lung cancer. The probability of lung cancer was multiplicative for a specific risk factor and accounted for eight (2^3) possible models (e.g., Male, White, Smokers vs. Male, White, Non-Smokers, etc.). A single, patient-specific risk score was the sum of all risk factors and was used to derive qualitative stratifications of risk as low, medium, and high. The risk stratifications were derived from the quartiles of risk in those with lung cancer since the 1st quartile's value was the threshold that maximized sensitivity and specificity. As a result, risk scores in the first quartile were low risk, risk scores in the 2nd and 3rd quartiles were medium risk, and risk scores in the 4th quartile were high risk. All analyses utilized the statistical software SAS, version 9.4.

Results

Risk factors curated from the PDS are described in Table 1 by lung cancer diagnosis to show the differences and similarities between those with lung cancer and those without lung cancer. In this sample, PWH with a diagnosis of lung cancer were significantly different on many risk factors. Specifically, they were different by age, race, ethnicity, BMI, alcohol use, cocaine use, marijuana use, smoking, years smoked, the number of months since smoking cessation, history of respiratory disease, history of reoccurring pneumonia, history of AIDS diagnosis, CD4 count, CD4 stability (fluctuations), CD4/CD8 ratio, and viral load.

Table 1. Descriptive Statistics, Risk Factors for Lung Cancer.

Variable		Lung cancer	No lung cancer	P-Value
Demographics information				
Age		38.63 (8.93) [32.00, 37.50, 44.00]	35.89 (8.57) [30.00, 35.00, 41.00]	0.0015
Gender (Female)	N	48 (48.00)	3856 (51.37)	0.5036
	Y	52 (52.00)	3651 (48.63)	
Race	White	46 (46.00)	3292 (43.85)	0.0369
	Black	44 (44.00)	2723 (36.27)	
	Other	10 (10.00)	1492 (19.87)	
Ethnicity	N	12 (12.00)	1730 (23.05)	0.0090
	Y	88 (88.00)	5777 (76.95)	
(Not Hispanic)				
BMI		23.92 (5.74) [21.46, 23.48, 26.21]	25.67 (6.94) [21.79, 24.10, 27.96]	0.0038
Sociodemographic information				
Annual Income		20285 (19458) [4998, 14998, 27000]	23233 (20542) [4998, 15000, 34998]	0.2331

Education	No Degree	51 (51.00)	3755 (50.02)	0.0763
	Completed High School	29 (29.00)	1535 (20.45)	
	Completed College	14 (14.00)	1396 (18.60)	
	Attended/Completed Graduate School	6 (6.00)	821 (10.94)	
Alcohol use	N	58 (58.00)	3551 (47.32)	0.0335
	Y	42 (42.00)	3954 (52.68)	
Cocaine use	N	26 (26.00)	3178 (42.35)	0.0010
	Y	74 (74.00)	4327 (57.65)	
Marijuana use	N	41 (41.00)	3894 (51.87)	0.0307
	Y	59 (59.00)	3613 (48.13)	
Sociodemographic information				
Smoking Status (Smoker)	N	18 (18.00)	3590 (47.82)	<.0001
	Y	82 (82.00)	3917 (52.18)	
Years smoked		21.17 (9.17) [15.00, 21.00, 28.00]	17.44 (9.40) [10.00, 17.00, 23.00]	0.0002
Packs smoked per day	< 1 per day	42 (51.85)	2264 (56.39)	0.7133
	>= 1 but < 2 per day	28 (34.57)	1244 (30.98)	
	2 or more per day	11 (13.58)	507 (12.63)	
Months since quit		52.13 (60.74) [12.00, 27.00, 79.50]	156.9 (1403) [24.00, 60.00, 99.00]	0.0252
Respiratory disease				
Reoccurring Pneumonia, History of	N	74 (74.00)	6795 (90.52)	<.0001
	Y	26 (26.00)	712 (9.48)	
Respiratory Disease, History of	N	70 (70.00)	6500 (86.59)	<.0001
	Y	30 (30.00)	1007 (13.41)	
Clinical HIV characteristics				
AIDS Diagnosis ever	N	50 (50.00)	4739 (63.13)	0.0069
	Y	50 (50.00)	2768 (36.87)	
CD4 cell counts		294.9 (246.6) [78.00, 250.5, 446.5]	451.3 (355.3) [138.0, 416.0, 674.0]	<.0001
AIDS Diagnosis ever	N	50 (50.00)	4739 (63.13)	0.0069
	Y	50 (50.00)	2768 (36.87)	
CD4 cell counts		294.9 (246.6) [78.00, 250.5, 446.5]	451.3 (355.3) [138.0, 416.0, 674.0]	<.0001
CD4, fluctuations over time	N	15 (15.00)	1959 (26.10)	0.0119
	Y	85 (85.00)	5548 (73.90)	
CD4, Nadir		180.5 (172.6) [42.50, 142.5, 262.0]	245.9 (225.7) [59.00, 201.0, 360.0]	0.0003
CD8 cell counts		733.8 (517.1) [385.0, 600.5, 970.0]	783.7 (492.9) [460.0, 701.0, 1002]	0.3148
CD4/CD8 Ratio		0.45 (0.42) [0.12, 0.35, 0.60]	0.64 (0.56) [0.20, 0.51, 0.92]	<.0001
Viral load greater than 500	N	36 (37.11)	3723 (51.74)	0.0042
	Y	61 (62.89)	3472 (48.26)	

The first step in developing the risk score was determine which risk factors were predictive of lung cancer using a bivariable logistic regression model with a lung cancer diagnosis as the dependent variable (outcome) and the risk factor as the independent variable (predictor). After transforming continuous variables and categorical variables into binary indicator variables, this study tested a total of 51 variables across the six strata in 306 logistic regression models. Predictors with a p-value less than 0.1 were considered candidate risk factors for inclusion in multivariable models.

In the multivariable models, multicollinearity was detected between viral load and CD4 Q1 and CD4/CD8 ratio to CD4 Q1 values. Viral load and CD4 Q1 were removed where applicable to resolve multicollinearity. The chi-square goodness of fit statistic, the C-statistic, quantifies the model's performance. The final predictors and accuracy of each of the final six models are summarized in Table 2. Each model performed well with accuracy ranging from approximately 70% to 90%. The odds ratios for each risk factor in each of the six models are described in Table 3. The odds ratio for each risk factor is multiplied across gender, race, and smoking status and then summed to form the risk score. For example, the 2nd quartile of CD4 increases lung cancer probability by 2.26 for females, 1.90 for males, 2.91 for black or other races, and 2.34 for smokers. In this example, a black female smoker has an increased likelihood of 15.38.

Table 2. Summary of final predict model accuracy.

Model	Predictors of lung cancer	C-Statistic
Females	Age, Q4, Smoking >1 Pack / day, History of reoccurring pneumonia, History of respiratory disease, CD4, Q2, CD4 Q3	82%
Males	High school education, Years smoked Q3, Years smoked Q4, History of Respiratory Disease, CD4 Q2, CD4/CD8 Q1	77%
White	High school education, Years smoked Q4, History of Respiratory Disease, CD4/CD8 Q1	68%
Black/Other	Non-Hispanic, Smoking>1 Pack /day, Years smoked Q3, History of Respiratory Disease, AIDS Diagnosis, CD4 Fluctuation, CD4 Q2, CD4 Q3, CD4/CD8 Q1	88%
Smokers	Non-Hispanic, Years smoked Q3, Years smoked Q4, History of reoccurring pneumonia, History of Respiratory Disease, CD4 Fluctuations, CD4 Q2, CD4 Q3, CD4/CD8 Q1	73%
Non-Smokers	History of Reoccurring pneumonia, Viral load>500	83%

Table 3. Multivariable logistic regression odds ratios.

Model	Gender		Race		Smoking Status	
	Female	Male	White	Black/Other	Smoker	Non-Smoker
Demographics and Socio Demographic Information						
Not Hispanic				2.0183	1.8107	
AGE Q4	1.7316					
High School		3.0883	2.3661			
Smoking Information						
GT 1 PD (Pack/Day)	2.2347			1.7954		
Years Smoked Q3		2.4130		1.7144	2.0738	
Years Smoked Q4		3.3921	2.7124		2.0309	
Respiratory Disease						
Reoccurring Pneumonia, History Of	1.9384				1.7743	4.3792
Respiratory Disease, History Of	2.3726	2.3858	2.1263	2.0962	2.2202	
Clinical HIV Characteristics						
AIDS Ever				2.5451		
CD4 Fluctuations, Over Time				2.7450	1.8771	
CD4 Q2	2.2644	1.8980		2.9081	2.3360	
CD4 Q3	2.7000			4.1657	2.5109	
CD4/CD8 Q1		2.2894	2.1489	1.9665	2.4599	
Viral Load GT 500						9.9413

The risk score's quartiles from the analytical sample of 7,587 HIV positive men and women (Figure 2) determined the risk stratifications (low, medium, and high). The first, second, and third quartiles of predicted lung cancer risk scores were 14.86, 25.20, and 39.1. The first quartile, 14.86, is the value that optimized sensitivity and specificity, 77% and 60%, respectively, with overall accuracy at approximately 70%. Sensitivity is a conditional probability measuring the correct number of lung cancer predictions given a lung cancer diagnosis. It is also known as the true positive rate (TPR). Of the 90 lung cancer cases in the analytical sample, the risk score accurately predicted lung cancer risk for 69 subjects. Specificity is a conditional probability measuring the correct number of non-lung cancer predictions given no lung cancer diagnosis. It is also known as the true negative rate (TNR). Of the 7,497 subjects without a lung cancer diagnosis, the risk score did not predict lung cancer for 4,498. Therefore, the accuracy of the risk score is the total of true positive and true negative risk predictions. Using the quartile values, a risk score less than 14.86 is low-risk, a risk score between 14.86 and 39.1 is medium risk, and any risk score above 39.1 is high risk.

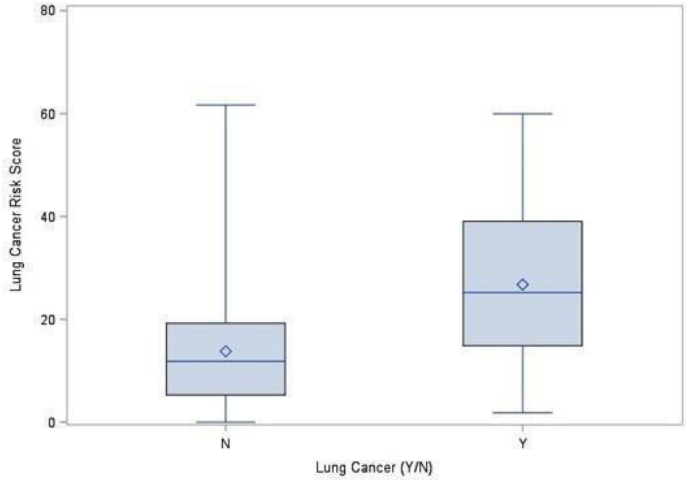

Figure 2. Distribution of predicted lung cancer risk by lung cancer diagnosis.

The sensitivity and specificity for this risk score are lower than the sensitivity and specificity of the NLST criteria (94 and 72%); however, the NLST excluded PWH, and several studies have reported a significant number of PWH missed by the NLST criteria. The same is true for this study. The NLST criteria applied to this study's analytic sample would only recommend lung cancer screening for 3 out of 90 subjects with lung cancer, a sensitivity of 3.3%. This study's risk score recommends screening for 95% more subjects than the NLST. While the NLST has recently changed their guidance to using 20 pack-years of smoking versus 30 pack-years of smoking, there would be no significant changes in NLST criteria' sensitivity applied to this analytical sample.

Discussion

This study is the first to create an HIV-specific risk score to assess lung cancer risk in PWH. Existing studies of PWH have reported significant associations between lung cancer and traditional risk factors in the general population and HIV-specific risk factors in PWH. None have created a tool, such as a risk score, to prevent or diagnose lung cancer earlier. There is a loud call for such a tool, and the risk score developed in this study has not only addressed the need but has demonstrated significantly better results than existing criteria [2,8,22]. PWH have different risks for lung cancer compared to the general population and require different guidelines from those put out by the NLST and adopted by the USPSTF. This study's tool has a true positive rate, or sensitivity close to 80%. In comparison, existing guidelines missed 97% of lung cancer cases, proving that risk factors other than age and smoking are predictive of lung cancer in PWH.

There are differences to findings of risk factors from this study compared to existing studies. Some risk assessment tools used for the general population included risk factors beyond age and smoking, such as body mass index (BMI), education, and alcohol use. They were normalized and integrated into the analytical dataset. While these traditional risk factors were significant in bivariable models, they lost their significance in multivariable models. Since this study is the first to develop a predictive model for lung cancer in PWH using traditional risk factors, it is difficult to determine if this phenomenon is due to inherent differences in risk factors between PWH and the general population or due to the sample itself.

In contrast, there were many similarities to findings of HIV-specific risk factors from this study compared to existing studies. While there is a variation from study to study, an association to increased risk of lung cancer has been shown with a history of AIDS diagnosis, CD4 cell count, the ratio of CD4 to CD8, re-occurring pneumonia, and respiratory disease in samples of HIV positive men and women. The same has been shown in this study. Longitudinal fluctuations of CD4 cell counts have not been previously tested; however, immune reconstitution has been suggested to affect lung cancer risk. It was a significant predictor in this study and contributed to the increased risk of lung cancer in the final models. One difference in HIV-specific risk factors used in this study compared to others is in viral loads. Most studies have reported viral load not being a predictor of lung cancer in PWH; however, viral load was a significant predictor but only in non-smokers. This contradiction is likely due to the small number of non-smokers in the study and the amount of imputed viral load values due to missing data.

There was a significant difference in smoking status and behaviors between those with lung cancer and those without lung cancer. PWH with a lung cancer diagnosis were significantly more likely to be a smoker and have smoked longer than those without lung cancer. There was also a significant difference in the number of months since quitting smoking. This difference shows people with lung cancer, who quit smoking, had stopped fewer months than those without lung cancer. For this reason, there were not enough subjects to form a third smoking status category, and subjects were classified either as a smoker or a non-smoker. The differences in months since quitting smoking should encourage smoking cessation for PWH. This study was also similar to other studies demonstrating that PWH smoke more than the national average of smokers in the US, 53% compared to 20%.

Race was collected as White, Black or African American, American Indian or Alaska Native, Asian or Pacific Islander, and Other (Specify). PWH with a lung cancer diagnosis in this sample were significantly more black and white than any other race. This distribution is consistent with the number of lung cancer cases by sex and race/ethnicity reported by the Center for Disease Control and Prevention (CDC). This shortcoming led to the combination of black and other races into a single category. An exploratory analysis was preformed to evaluate the effect of separating the black race from the 'Other' races. The models predicting lung cancer using 'other' race were questionable due to the small number of patients with lung cancer who were an 'other' race. Despite this problem, the exploratory analysis provides better insight into the actual risk of lung cancer for black PWH. A stratified model of 'Black' separated from 'Other' race showed the addition of cocaine use into the final predictive model. This change suggests illicit drug use and other novel risk factors could play a more significant role in predicting lung cancer risk and should be explored for future research.

Each of the six multivariable models performed well with areas under the curve (AUC) ranging from 0.6842 to 0.8773. While the non-smoker model had a decent AUC, 0.8282, it had the least amount of risk factors predictive of lung cancer. This deficit is likely due to the small number of non-smoking subjects with lung cancer. Blacks or other races who smoke seemed to be most at risk for lung cancer while white, non-smokers appeared to have the least risk of lung cancer, which corroborates historical data findings. The final risk score was a conditional, multiplicative model. The study also assessed summation instead of multiplication for each risk factor across the stratum, but there were no improvements in sensitivity and specificity.

This study's risk score algorithm outperformed the NLST criteria. The USPSTF criteria were assessed, but the only difference between the two is the maximum age, and there was no one in the analytical sample older than 70 years of age. The risk score algorithm developed by this study has a sensitivity of 77% and a specificity of 60%. The NLST criteria applied to the analytical sample missed 97% of those with lung cancer. It would have only recommended screening for 3% of those with lung cancer missing 97% of lung cancers. This discrepancy is because those with lung cancer in the analytical cohort were significantly younger than 50 and had less pack-years of smoking than 30. The NLST has recently modified their criteria to be 20 pack-years of smoking; however, the age has remained and still missed 97% of the analytical sample.

This study's primary strength is the size and similarity to other studies reporting lung cancer risks in PWH both in sample size and incidence rates. In the PDS of 7,607 PWH, a total of 73,401 person-years were observed, and a total of 100 lung cancer cases were identified in PWH. The incidence rate was estimated to be 136 lung cancer cases per 100,000 person-years. In the PDS of HIV negative, 26 lung cancer cases were observed for 63,163 person-years for an incidence rate of 41 lung cancer cases per 100,000 person-years. The incidence rate ratio for the MACS and WIHS data is 3.31. This study's analytical sample has significantly more lung cancer cases per 100,000 person-years than other studies. Sigel et al. reported 204 cases of lung cancer cases per 100,000 person-years in 37,294 PWH with an incidence rate ratio of 1.7 [10], and Marcus et al. reported 66 lung cancer cases per 100,000 person years in 24,768 PWH with an incidence rate ratio of 1.9 [17].

Despite this study's strength, there are a few limitations. First, it was difficult to identify primary lung cancer cases versus secondary lung cancer cases, particularly in the WIHS. Five lung cancer cases were identified as secondary in the MACS; however, the WIHS did not fully and consistently explain primary versus secondary across all outcomes datasets for lung cancer. In the few cases where it was identifiable, four lung cancer cases were secondary. As a result, no lung cancer cases were dropped since they could not be consistently assessed. Second, due to low counts in crucial variables such as race and smoking status, categories were collapsed. Other race categories were collapsed with black race due to low counts, and former smoking status was algorithmically collapsed with either current smoker or non-smoker using the number of months since quitting. Third, the exploratory analysis on separating the 'other' race from the black race suggests novel risk factors might play a more significant role. Having more data to form more strata, such as former smokers and other races, will help understand novel risk factors' effects on lung cancer in PWH. Likewise, novel risk factors were either not collected or not clean enough to be analyzed. This is because the MACS/WIHS were not designed to predict lung cancer risk in PWH. Instead, they report lung cancer outcomes that occur naturally in the sample.

Future work should consider further validation using other cohorts of PWH. The public-use datasets (PDS) from Johns Hopkins University were observational, longitudinal, and generalizable to the larger population but not specifically designed to study lung cancer. The PDS was challenging to normalize, integrate and was incomplete. Many novel risk factors, such as alcohol, cocaine, and marijuana use, were significant in bivariable models but not in multivariable models due to missing data. Likewise, another novel risk factor, e-cigarette use, was collected in one study but not the other making it difficult to infer appropriately. There is an ongoing debate over the robustness and precision of stepwise variable selection procedures such as those used in the study. Some studies are comparing them to explain performance differences. In this study, backward and stepwise selection methods did not result in any differences. Alternative methods such as bootstrapping and least absolute shrinkage and selection operator (LASSO) should be explored to see if different variables explain lung cancer risk or improve model performance. Similarly, different imputation methods, specifically for HIV viral load, should be explored to see its effects in the multivariable models. To improve specificity, other studies specifically designed to predict lung cancer are needed to explore additional risk factors that explain lung cancer risk further that this study has. Retrospective data collection from existing sources is feasible; however, prospective observational studies will be more informative. Likewise, additional studies should increase the number of race categories used in the predictive models and identify more non-smokers. Identifying larger cohorts of non-smokers and never-smokers will be challenging to do but will provide more insight into PWH risk factors that do not smoke.

Conclusion

Clinical advances in anti-retroviral therapies (ART) have significantly improved the prognosis of HIV. People with HIV (PWH) live longer, albeit their life expectancy is significantly worse than the general population. One of the leading causes of death impacting life expectancy is cancer, and lung cancer is the leading cause of cancer-related death. PWH are being diagnosed with and dying from lung cancer at significantly greater rates, at younger ages, and at advanced stages than the general population. There is a lack of treatment guidelines for lung cancer in PWH, but more importantly, there is an absence of criteria to assess the risk of lung cancer in PWH. The NLST and USPSTF guidelines excluded PWH from the clinical trials that formed their criteria. When applied to PWH, they miss significant amounts of lung cancer cases because risk factors far more prominent than smoking and age explain more of the lung cancer risk for PWH. As a result, there are no guidelines appropriate to identify PWH at risk for lung cancer. There remains an urgent and unmet need for a lung cancer risk assessment tool specific to PWH.

This study is the first to address this need and has successfully demonstrated that traditional, HIV specific, and respiratory risk factors can develop a risk score to assess lung cancer risk in PWH. Predictive models conditioned on smoking status, gender, and race independently identified risk factors for lung cancer that summed to a single risk score. Traditional risk factors such as age, education, and ethnicity are significant predictors of lung cancer risk. A history of reoccurring pneumonia and respiratory disease and clinical factors that describe immunocompetency are HIV specific predictors for lung cancer risk in PWH. This list is a stark difference in comparison to the predictors used in the general population.

This study also demonstrated that the risk score could be risk-stratified into low, medium, and high-risk with reasonable accuracy. HIV specific predictors for lung cancer explain a far more significant amount of lung cancer risk than those used in the general population. The near 80% sensitivity outperformed the 3% sensitivity of the NLST criteria when applied to the analytical sample. This difference proves not only that the risk score could be risk-stratified into low, medium, and high-risk but with greater accuracy.

The risk score led to the development of a CDSS using a validated tool accessible via the internet. This web-based CDSS provides a clinician and patient-oriented tool that clearly, and lucidly explains its risk for lung cancer. By allowing both the patient and the clinician to access this risk assessment tool, PWH can prevent lung cancer by understanding known risks and modifying behaviors or undergoing proactive screening for earlier lung cancer diagnosis.

In conclusion, PWH are being diagnosed with and dying from lung cancer at alarming rates compared to the general population. There are no tools for PWH or their clinicians to evaluate their risks for lung cancer and understand how they can prevent lung cancer from occurring. Likewise, there is no means to seek screening to diagnose it early. As the first tool to address these gaps and deficiencies, it will profoundly decrease lung cancer incidence and mortality.

References

1. Makinson, Alain, Laure Tron, Sophie Grabar and Bernard Milleron. "Potential Lung Cancer Screening Outcomes Using Different Age and Smoking Thresholds in the ANRS-CO4 French Hospital Database on HIV Cohort". *HIV Med* 21(2019):180-188.
2. Robbins, Hilary A. "Lung Cancer Screening in People Living with HIV: Modeling to Bridge the Evidence". *Aids* 32 (2018):1369-1371.
3. Sigel, Keith, Juan Wisnivesky, Shahida Shahri and Sheldon T Brow, et al. "Findings in Asymptomatic HIV-Infected Patients Undergoing Chest Computed Tomography Testing: Implications for Lung Cancer Screening". *Aids* 28 (2014):1007-1014.
4. Kong, Chung Yin, Keith Sigel, Steven Criss and Deirdre F. Sheehan, et al. "Benefits and Harms of Lung Cancer Screening in HIV-Infected Individuals with CD4+ Cell Count at Least 500 Cells/MI". *Aids*. 32 (2018):1333-1342.
5. Sigel, Keith, Juan Wisnivesky, Kristina Crothers and Kirsha Gordon, et al. "Immunological and Infectious Risk Factors for Lung Cancer in US Veterans with HIV: A Longitudinal Cohort Study. *Lancet Hiv* 4 (2017):E67-E73.
6. Shiels, Meredith, Jessica Islam, Philip S. Rosenberg and H. Irene Hall et al. "Projected Cancer Incidence Rates and Burden of Incident Cancer Cases in HIV-Infected Adults in the United States Through 2030". *Ann Intern Med*. 168 (2018): 866-873.
7. Marcus, Julia L, Chun Chao, Wendy A Leyden and Lanfang Xu, et al. "Survival among HIV-Infected and HIV-Uninfected Individuals with Common Non-AIDS-Defining Cancers". *Cancer Epidemiol Biomarkers Prev* 24 (2015):1167-73.
8. Makinson, Alain, Vincent Le Moing, Jacques Reynes and Tristan Ferry, et al. "Lung Cancer Screening with Chest Computed Tomography in People Living with HIV: A Review by the Multidisciplinary CANCERVIH Working Group". *J Thorac Oncol* 11(2016): 1644-52.
9. Hou, Wenli, Jun Fu, Yuanyuan Ge and Jian Du, et al. "Incidence and Risk of Lung Cancer in HIV-Infected Patients. *Journal of Cancer Research and Clinical Oncology* 139 (2013): 1781-1794.
10. Sigel, Keith, Juan Wisnivesky, Kirsha Gordon and Robert Dubrow, et al. "HIV as an independent risk factor for incident lung cancer". *Aid* 26 (2012):1017-1025.
11. Aberle, Denise R, Sarah DeMello, Christine D Berg and William C Black, et al. "Results of the Two Incidence Screenings in the National Lung Screening Trial". *N Engl J Med* 369 (2013): 920-31.
12. Hulbert, Alicia, Craig M. Hooker, Jeanne Keruly and Travis Brown, et al. "Prospective CT Screening for Lung Cancer in a High-Risk Population: HIV-Positive Smokers". *J Thorac* Oncol 9 (2014): 752-9.
13. Shcherba M, Hosgood HD and Lin J, et al. Characteristics of HIV Plus Lung Cancer Cases in a Large Clinical Population: Implications for Long Cancer Screening". *Journal of Clinical Oncology* 32 (2014).
14. Kirk Gregory D, Christian Merlo, Peter O' Driscoll and Shruti H Mehta, et al. "HIV Infection is Associated with an Increased Risk for Lung Cancer, Independent of Smoking". *Clinical Infectious Diseases*. 45 (2017): 103-110.
15. Uldrick, Thomas S. "HAART and Lungs: Do HIV Protease Inhibitors Impact Cancer Risk?". *Aids* 29 (2015): 1111-1112.
16. Makinson, Alain, Jacques Reynes. "A Novel Marker of Lung-Cancer Risk in People with HIV". *Lancet Hiv* 4 (2017): E53-E55.
17. Marcus, Julia L, Wendy A Leyden, Chun R Chao, Michael A Horberg, et al. "Immunodeficiency, AIDS-Related Pneumonia, and Risk of Lung Cancer Among HIV-Infected Individuals". *Aids* 32 (2018): 681-681.
18. Triplette, Matthew, Keith M Sigel, Alison Morris and Shahida Shahrir, et al. "Emphysema and Soluble CD14 are Associated with Pulmonary Nodules in HIV-Infected Patients: Implications for Lung Cancer Screening". *Aids* 31 (2017): 1715-1720.
19. A combined cohort study of the MACS Multicenter AIDS Cohort Study and WIHS Women's Interagency HIV Study. Web page. November 20, 2020, 2020. Accessed September 01, 2020, 2020.
20. Spitz, Margaret R, Waun Ki Hong, Christopher I Amos and Xifeng Wu, et al. "A Risk Model for Prediction of Lung Cancer. *J Natl Cancer Inst* 99 (2007): 715-26.
21. Bursac, Zoran, C Heath Gauss, David Keith Williams, David W Hosmer. "Purposeful Selection of Variables in Logistic Regression". *Source Code Biol Med* 17 (2008).
22. Sigel, Keith, Makinson Alain, Thaler Jonathan. "Lung Cancer in Persons with HIV". *Current Opinion in Hiv and Aids* 12(2017): 31-38.