DEVELOPMENT OF A CLINICAL DECISION SUPPORT SYSTEM TO ASSESS
THE RISK OF LUNG CANCER FOR PREVENTION AND EARLIER
DIAGNOSIS IN PEOPLE WITH HIV

By

Vincent James Amoruccio, MS, MA

A Dissertation Submitted
In partial fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics
Rutgers, the State University of New Jersey
School of Health Professions

May, 2021
Table of Contents

LIST OF TABLES .................................................................................................................... V
LIST OF FIGURES .................................................................................................................. VI
ABSTRACT ............................................................................................................................... XI
ACKNOWLEDGEMENTS .......................................................................................................... XIII
CHAPTER 1 .................................................................................................................................. 14
INTRODUCTION ....................................................................................................................... 14
1.1 INTRODUCTION TO THE PROBLEM ............................................................................... 14
1.2 BACKGROUND AND STATEMENT OF THE PROBLEM .................................................. 16
1.3 SIGNIFICANCE OF THE STUDY ....................................................................................... 22
1.4 RESEARCH GOALS AND OBJECTIVES ........................................................................... 26
1.5 RESEARCH HYPOTHESES ............................................................................................... 27
CHAPTER 2 .................................................................................................................................. 28
DISEASE CONDITIONS AND MANAGEMENT ............................................................................. 28
2. 1 LUNG CANCER ................................................................................................................... 28
   2.1.1 BIOLOGY OF CANCER ............................................................................................... 28
   2.1.2 HISTOPATHOLOGY OF LUNG CANCER .................................................................... 30
   2.1.3 SIGNS AND SYMPTOMS .......................................................................................... 33
   2.1.4 RISK FACTORS ......................................................................................................... 33
   2.1.5 TREATMENTS ............................................................................................................. 35
   2.1.6 SCREENING AND DIAGNOSIS .................................................................................. 37
2.2 HIV ..................................................................................................................................... 38
   2.2.1 BACKGROUND ON VIRUSES .................................................................................. 38
   2.2.2 ETIOLOGY OF HIV .................................................................................................... 39
   2.2.3 STAGES OF HIV, SIGNS, AND SYMPTOMS ............................................................... 43
   2.2.4 RISK FACTORS ......................................................................................................... 44
   2.2.5 DIAGNOSIS ............................................................................................................... 45
   2.2.6 ANTI-RETROVIRAL THERAPY (ART) ......................................................................... 46
   2.2.7 COMORBIDITIES, CANCER, AND LUNG CANCER IN PWH ................................. 47
2.3. CLINICAL DECISION SUPPORT SYSTEMS ............................................. 48

2.3.1 OVERVIEW OF CDSS .................................................................. 48
2.3.2 CDSS AS AN EXPERT SYSTEM USING EXSYS CORVID® .............. 50

CHAPTER 3 ......................................................................................... 52

LITERATURE REVIEW ........................................................................ 52

3.1 LUNG CANCER EPIDEMIOLOGY ...................................................... 52

3.1.1 INCIDENCE AND MORTALITY .................................................... 52
3.1.2 RELEVANT FINDINGS IN LUNG CANCER ................................. 54
3.1.3 CHALLENGES IN SCREENING AND EARLY DETECTION ............ 57
3.1.4 THE ROLE OF CDSS FOR LUNG CANCER SCREENING ............. 60

3.2 HIV EPIDEMIOLOGY .................................................................. 63

3.2.1 INCIDENCE AND PREVALENCE OF HIV ................................. 64
3.2.2 MORTALITY FROM HIV AND LIFE EXPECTANCY .................... 67
3.2.3 CHANGES IN COMORBIDITIES ............................................... 69

3.3 HIV-ASSOCIATED LUNG CANCER ................................................. 71

3.3.1 PRE-ART VS POST-ART AND EPIDEMIOLOGIC CHANGES AFFECTING CANCER .................................................. 71
3.3.2 DIFFERENCES IN LUNG CANCER PREVALENCE AMONG PWH .... 72
3.3.3 DIFFERENCES IN SMOKING PREVALENCE AMONG PWH .......... 76
3.3.4 WORSE PROGNOSIS OF LUNG CANCER AMONG PWH ............. 77
3.3.5 RISK FACTORS FOR LUNG CANCER SPECIFIC TO PWH .......... 79

3.3.5.1 Smoking ............................................................................. 79
3.3.5.2 HIV and Immunosuppression risk factors ................................ 80
3.3.5.3 Chronic Inflammation and Pulmonary Infections .................... 83

3.3.6 DISPARITIES IN TREATING PWH FOR LUNG CANCER .......... 86
3.3.7 CHALLENGES IN SCREENING, EARLY DETECTION, AND PREVENTION ........................................................................... 88
3.3.8 THE ABSENCE OF LUNG CANCER RISK ASSESSMENT TOOLS FOR PWH .................................... 90

CHAPTER 4 ......................................................................................... 92

RESEARCH METHODOLOGY .................................................................. 92

4.1 INTRODUCTION ........................................................................... 92

4.2 RESEARCH DESIGN .................................................................... 93

4.2.1 DATA ANALYSIS .................................................................... 94
4.3 CDSS

4.3.1 CORVID VARIABLES

4.3.2 FORWARD AND BACKWARD CHAINING LOGIC BLOCKS

4.3.2.1 Screening Logic Block

4.3.2.2 Smoking Status Logic Block

4.3.2.3 Smoking Information Logic Block

4.3.2.4 Demographics Logic Block

4.3.2.5 Respiratory Disease Logic Block

4.3.2.6 Clinical HIV Characteristics Logic Block

4.3.2.7 Risk Score Logic Block

4.3.3 COMMAND BLOCK

4.3.4 RUNNING THE CORVID CDSS

4.4 TESTING AND VALIDATION

4.5 LIMITATIONS

CHAPTER 5

RESULTS

5.1 OVERVIEW

5.2 DATA ANALYZED

5.2.1 PRIMARY ANALYSIS

5.2.2 EXPLORATORY ANALYSIS

5.3 CDSS SCREEN SHOTS

5.3.1 SCREENING

5.3.2 DEMOGRAPHICS

5.3.3 SMOKING STATUS

5.3.4 SMOKING INFORMATION

5.3.5 DEMOGRAPHICS CONTINUED

5.3.6 RESPIRATORY DISEASE

5.3.7 CLINICAL HIV CHARACTERISTICS

5.3.8 RESULTS SCREEN

5.4 PATIENT NARRATIVES

5.5 RESULTS CONCLUSION
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 6</td>
<td>221</td>
</tr>
<tr>
<td>SUMMARY AND CONCLUSION</td>
<td>221</td>
</tr>
<tr>
<td>6.1 SUMMARY</td>
<td>221</td>
</tr>
<tr>
<td>6.2 DISCUSSION</td>
<td>223</td>
</tr>
<tr>
<td>6.3 FUTURE DIRECTION</td>
<td>229</td>
</tr>
<tr>
<td>6.4 CONCLUSION</td>
<td>230</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>233</td>
</tr>
</tbody>
</table>
List of Tables

Table 3.1: Incidence Rates of HIV + compared to HIV ............................................................ 73
Table 4.1: Confidence Variables using Odds of Lung Cancer from Multivariable Logistic Regression .................................................................................................................. 98
Table 4.2: Variables used in the Screening Logic Block .................................................................. 109
Table 4.3: Variables used in the Smoking Status Logic Block ...................................................... 115
Table 4.4: Variables used in Smoking Information Logic Block ..................................................... 120
Table 4.5: Variables used in the Demographic Logic Block ............................................................. 128
Table 4.6: Variables used in the Respiratory Disease Logic Block ................................................. 133
Table 4.7: Variables used in the Clinical HIV Characteristics Logic Block ..................................... 142
Table 4.8: Variables used in the Risk Score Logic Block ............................................................... 154
Table 5.1: Descriptive Statistics, MACS/WIHS risk factors by lung cancer diagnosis ............... 166
Table 5.2: Explanation of categorized versus collected Race ......................................................... 168
Table 5.3: Bivariable logistic regression results, Candidate risk factors for multivariable models .......................................................................................................................... 169
Table 5.4: Summary of Final Predictive Model Accuracy ............................................................... 171
Table 5.5: Multiplicative Risk Scores for Each Risk Factor by Model for Females ....................... 175
Table 5.6: Multiplicative Risk Scores for Each Risk Factor by Model for Males ............................. 176
Table 5.7: Descriptive Statistics for False Positives ................................................................. 177
Table 5.8: Exploratory Analysis - Bivariable Logistic Regression Results, Race Risk Factors ................................................................................................................................. 180
Table 5.9: Exploratory Race Analysis Odds of Lung Cancer from Multivariable Logistic Regression .......................................................................................................................... 182
List of Figures

Figure 2.1: The Respiratory System .............................................................................. 31
Figure 2.2: The Human Immunodeficiency Virus (HIV) ............................................ 40
Figure 2.3: The HIV Replication Process .................................................................... 41
Figure 2.4: A general model for a knowledge-based CDSS ........................................ 49
Figure 3.1: Number of Cases by Sex and Race/Ethnicity – US – 2017 (CDC) ............ 53
Figure 3.2: Number of Deaths by Sex and Race/Ethnicity – US – 2017 (CDC) ......... 53
Figure 3.3: People with Diagnosed HIV in the US and Dependent Areas by Age, 2016 ...... 65
Figure 3.4: Adults and Adolescents Living with Diagnosed HIV Infection, by Sex and Race/Ethnicity, Year-end 2018—United States ................................................................. 66
Figure 3.5: Rates of Diagnoses of HIV Infection among Adults and Adolescents by Sex and Race/Ethnicity, 2018—United States .................................................................................. 66
Figure 3.6: Percentages of Diagnoses of HIV Infection among Adults and Adolescent by Transmission Category, 2018 – United States ................................................................. 67
Figure 3.7: Stage 3 (AIDS) Classifications, Deaths, and Persons Living with Diagnosed HIV Infection Ever Classified as Stage 3 (AIDS), 1985–2018—United States ......................... 69
Figure 4.1: CDSS Workflow ....................................................................................... 100
Figure 4.2: Development for process for the CDSS ................................................... 102
Figure 4.3: Example Logic Block in Tree Diagram ..................................................... 107
Figure 4.4: Screening Logic Block .............................................................................. 109
Figure 4.5: Smoking Status Logic Block ...................................................................... 113
Figure 4.6: Months Quit Backward Chaining Logic Block .......................................... 114
Figure 4.7: Smoking Message Logic Block ................................................................. 114
Figure 4.8: Smoking Information Logic Block ............................................................ 119
Figure 4.9: Years Smoked Backward Chained Logic Block ................................................................. 119
Figure 4.10: Demographics Logic Block .......................................................................................... 127
Figure 4.11: Respiratory Disease Logic Block .................................................................................. 132
Figure 4.12: Clinical HIV Characteristics Logic Block ................................................................. 140
Figure 4.13: CD4 Quartiles Backward Chained Logic Block ......................................................... 141
Figure 4.14: CD4/CD8 Ratio Backward Chained Logic Block ....................................................... 141
Figure 4.15: Risk Score Logic Block ............................................................................................... 153
Figure 4.16: Risk Level Backward Chained Logic Block ............................................................... 156
Figure 4.17: Command Block ....................................................................................................... 158
Figure 4.18: Corvid Servlet Runtime ............................................................................................ 160
Figure 5.1: MACS/WIHS Analytical Sample .................................................................................... 165
Figure 5.2: ROC for Final Logistic Regression Modeling Females ............................................... 172
Figure 5.3: ROC for Final Logistic Regression Modeling Males ..................................................... 172
Figure 5.4: ROC for Final Logistic Regression Modeling White Race ......................................... 173
Figure 5.5: ROC for Final Logistic Regression Modeling Black or Other Race ........................... 173
Figure 5.6: ROC for Final Logistic Regression Modeling Smokers .............................................. 174
Figure 5.7: ROC for Final Logistic Regression Modeling Non-Smokers ...................................... 174
Figure 5.8: Distribution of Predicted Lung Cancer Risk by Lung Cancer Diagnosis – Box Plots .......................................................................................................................... 179
Figure 5.9: CDSS Name and Brand for Integration ......................................................................... 184
Figure 5.10: Workflow - Screening ............................................................................................... 186
Figure 5.11: CDSS Screenshot: Main Screen ................................................................................ 186
Figure 5.12: CDSS Screenshot: Screening Criteria Not Met – Does not want to continue .......... 186
Figure 5.13: CDSS Screenshot: HIV Positive ............................................................................... 187
Figure 5.14: CDSS Screenshot: Screening Criteria Not Met – HIV Negative...............187
Figure 5.15: CDSS Screenshot: Currently have Lung Cancer .................................187
Figure 5.16: CDSS Screenshot: Screening Criteria Not Met – Current Lung Cancer .......187
Figure 5.17: CDSS Screenshot: Ever had lung cancer ..............................................187
Figure 5.18: CDSS Screenshot: Screening Criteria Not Met – Previous Lung Cancer ......188
Figure 5.19: Workflow - Demographics.................................................................189
Figure 5.20: CDSS Screenshot: Gender .................................................................189
Figure 5.21: CDSS Screenshot: Race .....................................................................189
Figure 5.22: Workflow - Smoking Status ...............................................................190
Figure 5.23: CDSS Screenshot: Currently smoking cigarettes ...............................190
Figure 5.24: CDSS Screenshot: Ever smoked cigarettes ........................................191
Figure 5.25: CDSS Screenshot: Months since smoking cessation ............................191
Figure 5.26: Workflow - Smoking Information......................................................192
Figure 5.27: CDSS Screenshot: Packs smoked per day ........................................192
Figure 5.28: CDSS Screenshot: Years Smoked......................................................192
Figure 5.29: Workflow - Demographics (Continued) ............................................193
Figure 5.30: CDSS Screenshot: Age.....................................................................193
Figure 5.31: CDSS Screenshot: Education.............................................................193
Figure 5.32: CDSS Screenshot: Ethnicity...............................................................193
Figure 5.33: Workflow - Respiratory Disease ......................................................194
Figure 5.34: CDSS Screenshot: History of Respiratory Disease ..............................194
Figure 5.35: CDSS Screenshot: History of Reoccurring Pneumonia .......................194
Figure 5.36: Workflow - Clinical HIV Characteristics ...........................................195
Figure 5.37: CDSS Screenshot: Viral Load .............................................................196
Figure 5.38: CDSS Screenshot: CD4 value .................................................................196
Figure 5.39: CDSS Screenshot: CD8 value .................................................................196
Figure 5.40: CDSS Screenshot: CD4 Fluctuations ......................................................196
Figure 5.41: CDSS Screenshot: History of AIDS Diagnosis ........................................196
Figure 5.42: Workflow - Results ..............................................................................198
Figure 5.43: CDSS Screenshot: Results Screen .........................................................198
Figure 5.44: Results Screen from Patient Narrative 1 – Subject M-1017 ..................199
Figure 5.45: Results Screen from Patient Narrative 2 – Subject W-2547 .................200
Figure 5.46: Results Screen from Patient Narrative 3 – Subject M-1040 .................201
Figure 5.47: Results Screen from Patient Narrative 4 – Subject W-2825 .................202
Figure 5.48: Results Screen from Patient Narrative 5 – Subject M-1154 .................203
Figure 5.49: Results Screen from Patient Narrative 6 – Subject W-3274 .................204
Figure 5.50: Results Screen from Patient Narrative 7 – Subject M-1797 .................205
Figure 5.51: Results Screen from Patient Narrative 8 – Subject W-3472 .................206
Figure 5.52: Results Screen from Patient Narrative 9 – Subject M-3682 .................207
Figure 5.53: Results Screen from Patient Narrative 10 – Subject W-4017 ..............208
Figure 5.54: Results Screen from Patient Narrative 11 – Subject M-6107 ...............209
Figure 5.55: Results Screen from Patient Narrative 12 – Subject W-4745 ..............210
Figure 5.56: Results Screen from Patient Narrative 13 – Subject M-6732 ...............211
Figure 5.57: Results Screen from Patient Narrative 14 – Subject W-9445 ..............212
Figure 5.58: Results Screen from Patient Narrative 15 – Subject M-7025 ...............213
Figure 5.59: Results Screen from Patient Narrative 16 – Subject W-9896 ..............214
Figure 5.60: Results Screen from Patient Narrative 17 – Subject M-7170 ...............215
Figure 5.61: Results Screen from Patient Narrative 18 – Subject W-9929 ...............216
Figure 5.62: Results Screen from Patient Narrative 19 – Subject M-7867 ........................................217

Figure 5.63: Results Screen from Patient Narrative 20 – Subject M-9272 .................................218

Figure 5.64: Summary of Validation Results..................................................................................219
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. Still, they do not explain the significant increase in lung cancer incidence in PWH. Risk factors specific to HIV, such as immunocompetence and respiratory disease, present more lung cancer risk in PWH. Existing guidance for lung cancer prevention exclude PWH and do not consider HIV specific risk factors. This deficiency has led to increased incidence and mortality from lung cancer and at younger ages and more advanced stages.

This study addresses the urgent and unmet need to develop a lung cancer risk assessment tool for prevention and earlier diagnosis in PWH. Using an integrated set of public-use data of 7,607 HIV-positive men and women from Johns Hopkins University, knowledge-based risk factors specific to PWH were formed using logistic regression models predicting lung cancer. Twenty-one HIV-specific risk factors were identified from 51 candidates using bivariant logistic regression models conditioned on gender, race, and smoking status. Similarly, stratified multivariable logistic regression models determined 14 lung cancer predictors for PWH. A multiplicate risk score was created for each risk factor
and then summed to create a single risk score. Risk stratifications for low, medium and high risk were identified using the quartiles of risk.

The risk score in the knowledge-base had a sensitivity of 77%. It led to the developing of the inference engine and a web-based Clinical Decision Support System (CDSS) for PWH and their clinicians. A random sample of 20 patients reserved for validating the CDSS had a sensitivity of 80%, while criteria from existing guidelines for the general population missed between 97 and 100% of lung cancer cases. These results demonstrate that a risk assessment tool, such as this CDSS, predicts lung cancer in PWH better than existing guidance and can prevent or diagnose it earlier.
ACKNOWLEDGEMENTS

This research is dedicated to all the men and women infected with HIV suffering from or who have died from lung cancer. I hope that they are forever seen as a unique population that is never excluded from science again. I hope that this work leads to the broad adoption of a risk assessment tool to prevent lung cancer in people with HIV and diagnose it earlier so that they may live the life they deserve.

The completion of this research would not have been possible if it were not for the support of several individuals at Rutgers University. First and foremost, I would like to thank Dr. Dinesh Mital, the Chair of my Dissertation Committee. Thank you for serving as my Chair but more importantly, thank you for becoming my mentor. I could not have completed this work without your support and guidance.

Second, I would like to thank both Dr. Shankar Srinivasan and Dr. Riddhi Vyas for serving as my dissertation committee members. Your feedback and guidance through this part of my journey significantly contributed to my success.

Finally, I would like to thank my family. Thank you to my mother and father for the support and love I needed to succeed in life. And thank you to my spouse, who is also my best friend, for enduring the challenges created from being a doctoral student while working full-time, sacrificing time with me, but, most importantly, for your love.
CHAPTER 1

INTRODUCTION

1.1 Introduction to the Problem

People with HIV (PWH) live longer due to clinical advances in anti-retroviral therapies (ART). In the pre-ART era before 1996, PWH died quickly from cancers rare in the general population such as Kaposi’s Sarcoma. In the post-ART period, the prevalence of morbidities common in older ages in the general population has increased among PWH due to their increased life expectancy. Lung cancer became the 2nd most common cancer among PWH in 2020 and will remain the 2nd most common cancer among PWH by 2030; however, lung cancer is the leading cause of cancer-related death among PWH.

There are significant differences in the incidence and mortality of lung cancer in PWH compared to the general population. They are diagnosed at younger ages and with more advanced stages. This is due in part to smoking, as it is in the general population. However, other risk factors explain more of the risk than smoking. Smoking is more prevalent among PWH compared to the general population. Some studies have reported that 100% of PWH are smokers, but when models control smoking, the risk of lung cancer remains elevated. This has led to the belief that smoking does not explain all of the risks. Some hypothesize that HIV exacerbates the carcinogenic effects of lung cancer. In contrast, others have speculated that immunologic factors specific to HIV, and perhaps HIV itself are
risk factors that increase the incidence and mortality of lung cancer in PWH. The differences in risks, frequency, and death have heightened awareness for lung cancer screening in PWH.

Existing guidelines for screening lung cancer in the general population are new. The National Lung Cancer Screening Trial (NLST) is a landmark trial demonstrating a 20% reduction in mortality from annual low-dose computed tomography (LDCT). The NLST used a population of people 50 and 74 years of age who were either current smokers with 30 pack-years of smoking, former smokers who ceased smoking no more than 15 years ago, and excluded people not in ‘good health.’ Despite concerns about high false-positive rates, enough evidence demonstrated the benefits of annual LDCT outweigh the risks. As a result, the US Preventative Services Task Force (USPSTF) adopted the NLST recommendations making it the first guideline for lung cancer screening in the general population.

Other organizations have adopted the NLST and USPSTF criteria for lung cancer screening. The results of these criteria are not generalizable to PWH. Several studies have tested the NLST and USPSTF criteria against cohorts and registries of PWH and have found significant proportions of missed cases among PWH. This is partly due to the younger age and fewer pack-years of smoking among PWH but also because PWH have known risk factors that increase the risk and mortality of lung cancer not considered in developing the NLST and USPSTF guidelines. The existing set of guidelines are under question in the general population. Due to the newness of lung cancer screening guidelines in the general population, there are a limited number of risk-based clinical decision support (CDS) tools for assessing lung cancer risk in the general population. While they utilize age and smoking as risk factors, they include other known risk factors for lung cancer from the general population. The same is not valid for PWH.
There are neither risk assessment tools nor CDSS for lung cancer in PWH. Several studies have indicated that it is critical, if not urgent, to identify PWH at high risk for lung cancer to encourage screening. Smoking cessation and early administration of ART are the most straightforward and immediate precautions to follow until the development of screening guidelines for PWH. This study aims to develop a CDSS to assess lung cancer risk in PWH to prevent lung cancer, diagnose them earlier, and reduce their risk and mortality from lung cancer.

1.2 Background and Statement of the Problem

Lung cancer is the leading non-AIDs defining cancer (NADC) and the leading cause of cancer-related death in PWH\textsuperscript{1-5}. In the pre-ART era, PWH died from cancers only seen in people with immunodeficiency such as AIDS. AIDS-defining cancers (ADC) such as Kaposi’s Sarcoma, Non-Hodgkin Lymphoma, and Cervical Cancer took the lives of PWH in as little as 1 year. 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 ADC or NADC in PWH and will remain to be the leading cause of death through 2030\textsuperscript{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, Marcus et al. reported a statistically significant difference 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\textsuperscript{7}. There are several factors contributing to poor survival from lung cancer.
in PWH. The primary reason is that PWH are less likely to receive treatments compared to the general population. PWH are typically excluded from and underrepresented in clinical trials. Only two clinical trials have been reported in the literature and only 20% of ongoing clinical trials include them. 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 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.

In addition to higher mortality, the incidence rate of lung cancer in PWH is higher compared to the general population and varies from to 2 – 4 times that of the general population. The standardized incidence ratio (SIR) has been used by many to further explain the worse incidence of lung cancer in PWH. The SIR for lung cancer in PWH is the number of observed cases of lung cancer in PWH divided by the expected number of lung cancer cases for the general population. In general, the SIR for lung cancer in PWH is significantly higher than in the general population. This is true globally however, the range varies for different parts of the world, race, age and HIV transmission groups. 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. 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 NLST criteria demonstrated a 20% reduction in mortality among “heavy smokers” in the general population through the use of LDCT. It showed that LDCT was a superior diagnostic tool compared to CXR. 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. Other domestic and international organizations used the NLST criteria to establish similar recommendations for lung cancer screening. The USPSTF uses the same criteria as the NLST but extends the maximum age to 80 instead of 74. In the general population, the NLST showed that annual LDCT to be highly sensitive with 94% detection at year 1. More importantly it showed that almost half, 47%, of the detected lung cancers were early stage. This showed that the NLST criteria to screen high risk people in the general population can detect lung cancer at a stage early enough to prevent significant mortality. The same is not true for PWH. The NLST and the USPSTF 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 since HIV and AIDS were considered both competing risks of death and contributed to a reduced benefit of treatment. The USPSTF lists these exclusions as part of its current clinical considerations. Some have tested the NLST and USPSTF criteria against samples of PWH and have reported that the current guidelines miss significant amounts of lung cancer cases in PWH. 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 and USPSTF guidelines for lung cancer screening. 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” 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 purpose of existing screening guidelines is to diagnose lung cancer at a stage early enough to be treated thereby reducing mortality. It is problematic PWH were excluded from the studies deriving the guidelines and PWH fail to receive the benefit of earlier diagnosis from the guidelines. There is a dearth of information on the benefits of LDCT in PWH. Studies have reported its benefit however the full set of benefits and harms still needs further exploration. 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. 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. CD4 cell count is the most extensively investigated risk factor for lung cancer in PWH. There has been debate over the timing of and exposure of CD4 levels as well as the varying thresholds which affect risk. Some believe cumulative exposure to low CD4 cells counts is more significant than recent CD4 cell counts\textsuperscript{10}. The most common thresholds appear to be CD4 < 200, 200 ≤ CD4 < 500, and CD4 > 500 with debate over whether 200 or 500 should be used as a comparator for risk. The CD8 biomarker alone has not been heavily studied however the ratio of CD4/CD8 has been reported by several as a significant risk factor\textsuperscript{5}. HIV Viremia is also widely researched and disputed. Cumulative HIV viremia has been shown to be a significant predictor but it has been shown to have an inverse relationship on lung cancer compared to CD4 cell count\textsuperscript{16,10}. 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\textsuperscript{9,10,15}. 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\textsuperscript{18}. 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. Unsurprisingly, there is an absence of
clinical decision support systems (CDSS) which could detect lung cancer earlier in PWH. The American Health Information Management Association (AHIMA) defines CDSS as a computer system that provides “clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care”. Simply put, CDSS is a computer system to assist in clinical decision making. It is sometimes referred to as an electronic-CDSS (eCDSS) given its integration with advanced electronics and web-based technology such as electronic health records (EHR). There is a vast amount of information supporting the use of CDSS for the treatment of cancers in the general population however there is a lack of information about the screening of cancers in the general population. Nonetheless, there is a sufficient amount of evidence to support the use of CDSS in the screening of cancer. In a systematic review of 11 studies of the role of CDSS in cancer screening, CDSS was shown to improve diagnosis including shorter times to diagnosis\(^{19}\). There is discernable absence of evidence similar to this for cancers associated with HIV such as lung cancer. 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\(^{4,8,12,18,20}\).

The purpose of this study will be to determine risk factors to create a risk score to identify PWH who are at high risk for lung cancer. The risk score will be used to develop and validate a CDSS for prevention and earlier lung cancer diagnosis in PWH. For a CDSS to work, there must be a knowledge base of scientific information to drive the mechanism for decision making. There is sufficient evidence to suggest risk factors which describe the excess risk of lung cancer not explained by smoking alone. This study will address the unmet
need for a risk score that can aide in the prevention and earlier diagnosis of lung cancer in PWH.  

1.3 Significance of the Study

The prognosis for lung cancer in the general population is poor however it is significantly worse in PWH. PWH have an increased risk of lung cancer and mortality from lung cancer compared to the general population of people without HIV (HIV negative). Several studies have demonstrated this using the general population as a control group (controlled studies). Non-controlled studies have shown the incidence of lung cancer to be as much as 5 times the general population while controlled studies have shown the incidence of lung cancer in PWH to be at least double the general population. One such controlled study is a large cohort study of lung cancer in 37,294 HIV positive and 75,750 HIV negative male veterans with a median age of 45 followed for a median of 5.8 and 7.3 person years between 1997 and 2008. In this study the incidence of lung cancer for HIV positive men was nearly two times that of HIV negative men. Sigel et al. reported 204 cases of lung cancer per 100,000 person years in HIV positive men compared to 119 cases of lung cancer per 100,000 person years in HIV negative men; an unadjusted incidence rate ratio (IRR) of 1.710. A larger controlled study used subjects from Kaiser Permanente’s Northern and Southern health care system. From 1996 to 2011, 24,768 HIV positive and 257,600 HIV negative men and women with a mean age of 40 for 4.9 and 5.8 person-years per subject. Marcus et al. reported 66 cases of lung cancer per 100,000 person years in HIV positive individuals and 33 cases of lung cancer per 100,000 person years in HIV negative individuals; an unadjusted IRR of 1.918. In a smaller but prospective study of 2,495 injection drug users followed for
25,708 person-years between 1988 and 2006, PWH had a 2.3 increased risk for developing lung cancer. A systematic review of literature summarized by Hou et al. reported the highest SIR of 27.32 from a retrospective cohort study between 1997 and 2007. Controlled studies have also reported a significant increase in mortality from lung cancer. One such study is a large controlled cohort study of 5,065 men and women by Brock et al. They demonstrated a significant association between mortality and advanced stage in 92 HIV positive individuals compared to 4,973 HIV negative individuals. Overall survival in both groups was poor however, mortality was almost 60% higher among the HIV positive. Of those who died, 94% of the HIV positive group had advanced stage lung cancer compared to 68% of the HIV negative group. Notably, 69% of the HIV positive group had metastatic lung cancer compared to 47% in the HIV negative. In this study, the two groups also differed significantly by age with 62% of the HIV positive group being less than 50 years-of-age compared to 12% of the HIV negative group. The higher mortality, advanced stage, and young age led Brock et al. to conclude PWH not only have higher mortality but also delayed diagnosis. This supports additional claims that lung cancer in PWH progresses faster and more aggressively than lung cancer in the general population.

It is well established in the general population that smoking is the primary factor for genetic mutations in cellular division and explains most of the risk of lung cancer. PWH have a higher prevalence of smoking and increased risk of lung cancer compared to the general population. The prevalence of smoking varies by sample and has been reported by many to be almost 100% prevalent. In the large controlled cohort study by Sigel et al., a significant difference in smoking between HIV positive and negative men was demonstrated. The adjusted IRR comparing HIV positive men to HIV negative men was 6.3 and 3.0 for current and former smokers, respectively. This demonstrates a significantly greater risk for
lung cancer among HIV positive smokers. Marcus et al. reported similar results in their large cohort study with as significant difference in smoking status between HIV positive and negative individuals. When smoking was adjusted in each of the large controlled studies by Sigel, Marcus, and Shiels et al., there was still an increased risk of lung cancer in the HIV positive groups compared to the HIV negative group. Others have demonstrated similar findings in non-controlled studies. The evidence suggests that other factors are contributing to the risk of lung cancer not explained by smoking. Risk factors for lung cancer in PWH include the traditional risk factors seen in the general population (age, gender, and race/ethnicity) but also include other factors specific to HIV infection. Low CD4 and high CD8 values measured as a ratio (CD4/CD8) and HIV viremia (HIV RNA) have been demonstrated by many as risk factors for lung cancer in PWH. Pulmonary infections, specifically COPD and bacterial pneumonia, have also been demonstrated by many as risk factors. Sigel et al. showed an increased risk in lung cancer remains in the HIV positive group compared to HIV negative group after traditional risk factors and pulmonary factors are adjusted for. They demonstrated increases in significantly different IRR when stratified by age, race/ethnicity, smoking, COPD, and previous bacterial pneumonia. Marcus et al. showed similar results but concluded previous pneumonia was not a risk factor and should not be adjusted for.

The research demonstrates that there are significant differences in the presentation of lung cancer in PWH compared to the general population. PWH have a poorer prognosis with higher incidence and mortality. Existing screening guidelines such as those from the NLST and USPSTF exclude PWH. Several studies have demonstrated that existing guidelines miss substantial amounts of lung cancer when applied to PWH. Shcherba et al. retrospectively demonstrated that nearly 73% of lung cancers were missed in a cohort of 90
HIV positive men and women because the age and smoking limits were too high. Similarly, Cioaia et al. retrospectively demonstrated that only 17% of their cohort met the existing criteria for screening. Makinson et al. also showed the NLST and USPSTF screening criteria missed more than 69% of lung cancer cases and further demonstrated how reducing the screening limits would increase the number of lung cancers detected. In their study, reducing the smoking threshold could have increased the number of cases detected by as much as 14% and reducing age limits could have increased the number of cases detected by 2 to 3 times. This research supports the claim for improved and enhanced guidance for lung cancer screening in PWH.

There is an absence of CDSS for the screening and diagnosis of lung cancer in PWH. This unmet need has been highlighted by many investigators in their research. Makinson et al. indicated screening for lung cancer in PWH is not only feasible but a mechanism that could lead to earlier diagnosis and reduced mortality. They suggested a “lung cancer risk score, including age, smoking, COPD, bacterial pneumonia, and CD4/CD8 ratio could be assessed and validated”. Sigel et al. established HIV is independently associated with lung cancer. Sigel et al. also suggested that traditional risk factors, immunocompetency factors, and pulmonary factors could be used to “risk-stratify patients with HIV for targeted lung cancer screening”. Lung cancer will continue to afflict PWH at higher rates for at least the next 10 years. Using cancer trends from 2000 to 2012, Shiels et al. projects that overall incidence of lung cancer will decline marginally however, it will remain to be an NADC with the highest incidence rate secondary to prostate cancer. They also project a substantial shift in the age demographics of PWH due to their increased aging and survival from ART. The largest proportion of PWH will be between 45 to 64-years of age, an increase from 39.4% to 47%. Given the evidence of late-stage lung cancer at younger ages below 55, there will
continue to be a substantial risk for lung cancer in PWH. This study seeks to address this unmet need. It will identify traditional, immunocompetent, and pulmonary factors that influence the risk of lung cancer in PWH. It will use those factors to create a lung cancer risk score which can risk-stratify PWH and be used in a CDSS to prevent and diagnose lung cancer earlier.

1.4 Research Goals and Objectives

There are three primary goals of this study. The first is to develop a risk score that will predict the risk of lung cancer in PWH. The risk factors for lung cancer among PWH are different from those for the general population; however, large, comparative studies have proven their significance. This study's second primary goal is to use the risk score to stratify PWH by risk for lung cancer. Multiple, patient-specific variables will identify PWH at low, medium, and high risk for cancer. This risk score will be a stark contrast to the recommendation for screening in the general population. This study's third primary goal is to develop an expert CDSS using the new risk score. Neither a risk score nor a CDS tool exists for lung cancer in PWH. A risk score and a CDS will allow PWH to assess and take the necessary steps for prevention and treatment at a sufficiently early stage. This tool would be the first of its kind to address this urgent and unmet need.

This study's other goals include facilitating shared-decision making between PWH and their clinicians and contributing to the sparse research on lung cancer in PWH. There are disparities in the prevention and treatment of lung cancer, contributing to higher lung cancer incidence and mortality. These disparities are partly due to the lack of knowledge clinicians have with lung cancer among PWH and lung cancer risk factors in a complex
population such as PWH. This tool will be easily accessible and allow PWH to assess their own risk for lung cancer and initiate discussions with their clinicians. The heightened shared-decision making will bring awareness to risks, benefits, and harms of screening and help both the clinician and PWH take the appropriate next step, ultimately leading to prevention and earlier diagnosis. This study's final goal is to contribute significantly to the lack of evidence in lung cancer among PWH. There is a distinct gap in evidence describing the challenges of lung cancer screening in PWH and a persistent and loud call to action coming from the work. This study will be the first to address this urgent and unmet need.

1.5 Research Hypotheses

This study will contribute to the gaps in the research and address the unmet needs highlighted in the field. It seeks to test the following hypotheses:

1. A risk-score will be developed using traditional, HIV, and pulmonary risk factors.

2. The risk score will be risk-stratified and accurately diagnose lung cancer in PWH with sufficient sensitivity and specificity.

3. A CDSS will be developed using the clinician and patient-oriented risk-score providing the appropriate amount of knowledge to allow for shared-decision making, prevention, and earlier diagnosis of lung cancer in PWH.
2.1 Lung Cancer

2.1.1 Biology of Cancer

Cancer is the name for a collection of diseases characterized by uncontrolled cell growth. All living organisms consist of one or more cells, the basic unit of life that arise from pre-existing cells in a controlled manner. The cells' main components are the cell ("plasma") membrane, nucleus, cytoplasm, and organelles. Eukaryote organisms, such as humans, contain cells with a nucleus, while prokaryote organisms, such as bacteria, contain cells that lack nuclei. Trillions of cells make up the human body, and all but one type, the red blood cell, have a nucleus.

In humans, a membrane surrounds the cell's nucleus that protects and encapsulates its content. The nucleus contains deoxyribonucleic acid ("DNA") with two primary functions to store genetic information and coordinate cell activities. DNA is organized into 22 autosomal chromosome pairs and one pair of sex cells and is considered the chemical blueprint for all cell activities. DNA is a double helix of four chemical bases ("nucleotides") – adenine, thymine, guanine, and cytosine. Sections of DNA found in each chromosome make up genes, which are codes that tell the cells to produce proteins through a complicated process called gene expression. Gene expression transcribes DNA into ribonucleic acid
(RNA) and then translates RNA into proteins. Some viruses, such as HIV, lack DNA but instead have RNA, which is reverse transcribed into DNA to infect the nucleus of their target cells. The gene's length and sequence determine the type and function of the protein. Additionally, proteins derived from genes control somatic and germ cell division, mitosis and meiosis, as well as cell death.

Cancer begins in cells, and cellular DNA mutations cause cancer. There are many causes for DNA mutations, including inheritance, although unlikely, and environmental exposure to carcinogens such as smoking, asbestos, or radon. Cells continuously replenish in a controlled process that kills old, worn out, or damaged cells and creates the correct number of new cells. Four genes regulate the cell division and death process, maintaining homeostasis of cells in the body. Oncogenes tell the cells when to divide while tumor suppressor genes tell the cell when not to divide. Suicide genes control programmed cell death (apoptosis), while DNA repair genes provide instructions to cells about when and how to repair itself. Mutations to oncogenes and tumor suppressor genes create uncontrolled cell growth, while mutations to suicide genes and DNA repair genes inhibit cells from dying and repairing themselves. Mutations to any of these four genes can lead to the formation of benign or malignant tumors.

Benign tumors are usually encapsulated and do not spread into nearby tissues. They are not harmless but relatively easier to remove and less likely to spread and grow back. The same is not valid for malignant tumors. Malignant tumor cells can remain in their primary tissue or primary organ. Still, they can also break off their tumor and invasively spread to nearby tissues and organs, becoming invasive cancer, or they can enter the immune system and spread to blood vessels. When malignant tumors spread to distant tissues and organs,
cancer has metastasized and becomes secondary cancer. For example, breast cancer cells found in the stomach are considered breast cancer, not stomach cancer.

There are five main types of cancer. Sarcomas are cancers of the bone and soft tissues such as muscle, cartilage, fat, fibrous tissues, and blood and lymph vessels. Leukemia is cancer that affects bone marrow and blood, which leads to mutations of blood cells. Lymphomas are cancers that begin in T and B cells, lymphocytes. Melanoma is a cancer of special skin cells known as melanocytes that pigment the skin. Finally, carcinomas, the most common cancer type, are formed by epithelial cells in the skin or tissues that line the body's outside and inside surfaces, particularly internal organs. There are several types of carcinomas. The two most common are adenocarcinoma and squamous cell carcinoma. Adenocarcinoma is a type of carcinoma that starts in epithelial cells of mucus-producing glandular organs such as breast, prostate, and lung cancer. Like adenocarcinoma, squamous cell carcinoma found on epithelial cells affects those cells in the outer surface of skin cells and those that line organs such as the stomach, kidneys, and lungs.

Cancer is the second leading cause of death in the US, preceded by heart disease according to 2017 surveillance data from the CDC. It is also one of the leading causes of death globally. According to the WHO, lung cancer is among the top 10 global causes of death. In the US, lung cancer will have the second highest incidence and be the leading cause of cancer-related death in 2020. Lung cancer is the most common cancer in the world and the most fatal.

2.1.2. Histopathology of Lung Cancer

The lungs are part of the respiratory system, which works in tandem with the cardiovascular system to supply oxygen to cells and relieve carbon dioxide in cells. The respiratory system consists of many parts (Figure 2.1). Air enters and leaves the body
through the nasal cavity or the oral cavity and flows to the lungs through the pharynx, larynx, and trachea. The trachea's air space is supported and kept open by cartilage and surrounded by epithelial cells (respiratory epithelium), smooth muscle, and connective tissue (lamina propria). The trachea branches into two primary bronchi (right and left); however, the right lung bronchus branches into three lobes and the left bronchus branches into two lobes. Each branch of the bronchus and each lobe branch into a series of tree-like bronchioles and ends with hundreds of millions of tiny, air-filled sacs, known as alveoli. Bronchioles and alveoli do not have cartilage. Small amounts of connective tissues, the protein elastin, and capillaries surround each alveolus; however, the alveolus serves as the conduit for gas exchange between the blood and lungs.

Figure 2.1: The Respiratory System

Each branch of the two primary bronchi forms the lungs and is surrounded by a double membrane pleura. One layer of the membrane surrounds the lungs. In contrast, the other membrane, separated by pleural fluid, surrounds the thoracic cavity made mostly of the rib cage muscle and connective tissue known as fascia. The diaphragm is the primary muscle that contracts the lungs and works with the ribs' intercostal muscles.
The uncontrolled growth of cells in lung cells and tissues leads to lung cancer. As in much other cancer, abnormal lung cells do not function properly and grow into tumors that interfere with lung function. Lung cancer can be classified as either small-cell lung cancer (SCLC) or non-small-cell lung cancer (NSCLC). The incidence of both is 15% and 85% respectively. SCLC is deadlier than NSCLC due to its ability to metastasize quickly. It has a 6-month median survival and 33% 1-year overall survival. NSCLC can be classified as adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and NSCLC not otherwise specified (NOS). Adenocarcinoma is the most common and manifests itself in lung cells that secrete substances. Squamous cell carcinoma is the second most common lung cancer, is usually associated with current or former smokers, and usually manifests itself near the bronchi in squamous cells. Large cell carcinoma is a rapidly metastasizing cancer and can be found anywhere in the lungs.

The stages of lung cancer explain the invasiveness levels and range from stage I, II, IIIA, IIIB, or IV. If the tumor is local to the lungs and has not spread to any lymph nodes, it is considered stage I. When cancer in the lungs has spread to nearby lymph nodes, it is considered stage II. Stage I and II lung cancer are also called limited stages since they are contained inside the chest and primarily one lung. Once lung cancer has spread to the middle of the chest or the other lung, it is considered stage III; however, there are two types of stage III lung cancer. If lung cancer has been contained to lymph nodes on the same side of the chest as it started, it is considered stage IIIA. In contrast, if the lung cancer has spread to lymph nodes on the other side of the chest, it is considered stage IIIB. When lung cancer has spread to both lungs, lung fluid, or other body tissue and organs, it is considered stage IV, the most advanced stage, also called an extensive stage.
2.1.3. **Signs and Symptoms**

Signs are effects of a disease that can be observed physically, such as a temperature or a lump. Symptoms are effects of a disease that can only be experienced and explained by the patient, such as pain. The signs and symptoms vary by type of cancer and are noticeable or shared by most, but not all. Signs include a lump, unintended weight loss or gain, skin changes, new or exacerbated cough, hoarseness, fevers, night sweats, and unexplained bleeding or bruising. Most cancer symptoms also include fatigue, indigestion, and muscle or joint pain. The absence or presence of any of these signs and symptoms does not indicate cancer since many other diseases or conditions could also cause these signs and symptoms. Lung cancer is difficult to diagnose since its signs and symptoms are related to several lung diseases such as bronchitis, tuberculosis, asthma, and chronic obstructive pulmonary disorder (COPD). Despite this, the most significant sign is a new, persistent, or worsening cough, dysphonia, or chest pain. Other signs and symptoms include dyspnea, hemoptysis, tachypnea, changes in color or volume of mucus, and blood in the mucus.

2.1.4. **Risk Factors**

Mutations cause uncontrolled cell growth in genes in the cell, which influence the process of cell division. Mutations to cell division genes can be genetically inherited, although this only partially explains 5-10% of cancer. Mutations acquired through lifestyle and environmental factors explain a more significant part of cancers. Understanding the effect of environmental and lifestyle factors on cancer can lead to preventative measures.

Many biological factors increase lung cancer risk, specifically, age, gender, and race. Currently, being a male, older than 50, White, or African American are associated with higher risks of lung cancer than those who are female, younger than 50, and not White or
African. Age is a significant risk factor since more than half of new lung cancer cases occur in 55-74-year-old men and women, and 37% occur in those older than 75\textsuperscript{25}.

The most significant and recognized risk factor for lung cancer is smoking\textsuperscript{25, 30}. More than 4,000 chemicals make up cigarettes, including nicotine. At least 50 of these chemicals are carcinogenic and cause lung cancer in approximately 80-90% of lung cancer cases, or 1 in 6 smokers\textsuperscript{26, 29-31}. The risk of lung cancer is proportionate to the intensity and duration of smoking, the 'age at onset' for smoking, and the lifetime duration of smoking. While cigarette smoking is the most critical and recognized risk factor, so too is environmental tobacco exposure (ETS), also known as "secondhand smoke." More than 80% of lung cancer occurs in current and former smokers; however, people who have never smoked also develop lung cancer, suggesting that environmental and familial exposures and genetic susceptibility increase lung cancer risk.

Environmental risk factors can explain part of the risk of lung cancer in never-smokers. At the same time, they further increase the risk of smokers. When encountered in an occupational setting (e.g., painters, coal miners, etc.), exposure to these risks is called "occupational exposure." After smoking, radon is the second most critical toxin and a risk factor for lung cancer. Occupational exposures have included arsenic, chromium, and ionizing radiation; however, asbestos is the most crucial occupational risk factor. Domestic fuel smoke is also known as "household air pollution," is an environmental risk factor for lung cancer.

Certain viruses such as the Human Papilloma Virus (HPV), Epstein-Barr Virus (EBV), Tuberculosis (TB), and Helicobacter pylori (H-Pylori) cause gene mutations and increase the risk of several cancers. HPV is a known risk factor in breast cancer and cervical
cancer in most populations and has been linked infrequently to lung cancer risk, as is HPV and EBV, which are also not common in lung cancer.

TB increased the risk of lung cancer in regional epidemiological studies; however, pulmonary inflammation is more likely the mechanism of increased risk. Inflammation chronically caused by pulmonary diseases, such as TB, emphysema, chronic bronchitis, and pneumonia, has been shown to increase lung cancer risk and presents a set of risk factors on its own. Emphysema and chronic bronchitis are pulmonary diseases classified as COPD, and COPD is generally considered a risk factor for lung cancer.

There is no consensus on genome-wide associations to lung cancer; however, associations are possible with some common tumor markers, sequences of nucleotides in DNA, which code for specific proteins. These proteins are expressed in higher amounts when cells are cancerous and found in or around tumors and other body parts. The three most common tumor markers for lung cancer are epidermal growth factor (EGFR), the human epidermal growth factor (HER2), and KRAS. Other less common tumor markers include ALK, ROS 1, BRAF, and RET. EGFR and HER2 occur most commonly in 40-60% of never-smokers, while KRAS is found more often in smokers and associated with poor prognosis.

2.1.5. Treatments

There are many different treatment types for lung cancer, and the rationale for each varies by stage. Treatments can be performed alone or in combination and usually include surgery, chemotherapy, radiation, targeted therapies, and immunotherapy. Surgery is a standard treatment for limited-stage lung cancer, stage I and II, and usually involves a lobectomy, segmentectomy, wedge resection, or pneumonectomy. A lobectomy is the removal of the lobe affected by lung cancer. A segmentectomy is the removal of a piece or
segment of the lobe. Wedge resection is a removal of tissue surrounding the tumor.

Pneumonectomy is removing the entire lung and when it is impossible to remove all lung cancer cells through alternative procedures.

Chemotherapy is a combination of chemical therapies that kill cells in your body. The most common administration route is intravenously, but oral and topical chemotherapies are novel and more popular. It is a systemic treatment that kills cancer cells and healthy cells since it cannot differentiate between the two. This lack of differentiation usually creates severe and debilitating side effects, such as nausea and fatigue, making treatment adherence difficult.

Radiation therapy is a localized treatment and differs slightly from chemotherapy in that it points beams of energy at specific areas, mainly, the tumor and surrounding areas. The radiation damages the DNA of the cells to disrupt the cell division process. Similar to chemotherapy, it cannot differentiate between healthy and cancerous cells, and there are similar side effects but to a lesser degree. Radiation therapy can be used if cancer has not spread and may be used in combination with other treatments and palliative care.

Targeted therapy is a relatively new type of lung cancer treatment that targets specific biomarkers found on cancer cells. Targeted therapies are more tolerable since they do not affect healthy cells. Immunotherapy is more novel than targeted therapy, but there are few options for lung cancer. The guiding principle of immunotherapy is to enhance the immune system to attack abnormal cells further.

The use of these treatments as a standard-of-care in lung cancer is dependent on the type of lung cancer and the stage. Non-small cell lung cancer (NSCLC) is the most common and treatable lung cancer. Surgery is the standard-of-care when it is limited-stage (Stage I or II) or stage IIIA. Chemotherapy is administered as an adjuvant when NSCLC is stage II or
IIIA. When surgery is not possible, chemotherapy and radiation therapy are given concomitantly, mostly for stage III. Chemotherapy is the only treatment when NSCLC is stage IV; however, radiation therapy is a form of palliative care. Targeted therapy and immunotherapy are not standard-of-care; instead, they are adjuvants to treatments for extensive stage lung cancer. Surgery is uncommon for small cell lung cancer (SCLC), and the standard-of-care is chemotherapy, and radiation therapy is an adjuvant to chemotherapy.

2.1.6. Screening and Diagnosis

There are few signs and symptoms in the limited stages of lung cancer, stage I and II. Lung cancer is relatively asymptomatic until malignant cells create tumors dangerous enough to damage normal lung function. Screening is "testing individuals at risk for a disease, but who do not exhibit signs or symptoms of the disease". The goal of screening is to detect cancer early enough to cure it and reduce mortality. It should have little risk, high sensitivity, high specificity, and be easily accessible to the patient. As computers became more accessible, their role in health information technology expanded. Computed Tomography (CT) replaced chest x-rays (CXR) as the most common diagnostic tool. It creates 3-d pictures of the body using a computer to combine several x-rayed images. It became a popular diagnostic imaging tool for lung cancer screening since the National Lung Cancer Screening Trial (NLST) reported CT could reduce mortality.

The largest and the most notable trial assessing the association between CT and reduced mortality was the NLST. The NLST enrolled approximately 54,000 current or former smokers between the ages of 55 and 74 and had the most extended follow-up of any other randomized controlled trial. The NLST changed the outlook of lung cancer screening by demonstrating that annual low-dose CT (LDCT) was superior to CXR and reduced lung cancer mortality among high-risk people by 20%. Through exclusion, it
defined "high-risk" as men and women between the ages of 55 and 74 who are either a
current smoker with 30+ pack-years of smoking or a former smoker having quit smoking in
the past 15 years and are in reasonably good health. When LDCT detects an abnormal lesion
or nodule during screening, the next step for the high-risk patient is to diagnose the nodule.
The most common diagnostic technique for irregular nodules is a biopsy using a needle
(transthoracic biopsy) or a tube (bronchoscopy).

2.2. HIV

2.2.1 Background on viruses

Many organisms live in the human body as part of its normal physiology. Over $10^{14}$
bacteria throughout the entire body make up its normal microbial flora. These organisms
help the body for short periods or co-exist in the body for more extended periods. Despite
the help organisms provide to the human body, many organisms are not part of our normal
microbial flora, and even some that are can cause disease. Organisms that can cause disease
are called pathogens. There are four different types of pathogenic microorganisms. Bacteria
are tiny single-cell organisms with DNA, but no nucleus that can help or hurt the human
body and live and replicate outside of the human body. Protozoa are more considerable in
size than bacteria but have a nucleus and are mostly harmless, but some cause diseases such
as dysentery and malaria. Helminths are multicellular worms that infest in humans, such as
hookworm. Finally, viruses are pathogens that can live outside of the human body for short
periods but cannot replicate.

Viruses have no cellular organelles other than genetic information such as single or
double-stranded DNA or RNA and require other cells to perform normal cell functions. A
protein capsule surrounds the genetic material, and an additional layer of lipids may envelop
this capsid. There are receptor-binding proteins on the surface of the virus particles used to attach to other cells or organisms. Viruses can be spread directly from person-to-person such as airborne droplets or sexual contact, or indirectly through some joint agent such as food and water. Once inside the human body, viruses seek hosts to inject their genetic material. This injection may happen through endocytosis during the body's normal immune response to engulf pathogens. On the other hand, viruses use their viral receptors and their enveloped layer of lipids to either pierce or fuse with another cell. Once the genetic material is inside the host cell, it uses the cellular organelles to reproduce and then kills the host. In exceptional cases, such as in HIV, the genetic material might permanently integrate into the host's genetic material, recoding it and infecting it.

2.2.2 Etiology of HIV

The Human Immunodeficiency Virus (HIV) is a ribonucleic acid (RNA) virus. It is a microscopic disease-causing organism with no cellular organelles. Inside of HIV is the genetic material in a double strand of RNA and protected by a protein shell known as a capsid. HIV is an enveloped protein (Figure 2.2) with a membrane stolen from its host during viral replication. On the surface of the HIV, membrane receptor-binding glycoproteins communicate and bind with other cells. With a membrane and receptors from its host, HIV tricks other cells into thinking it is the same.

HIV is a unique type of virus known as an RNA virus. It is also known as a retrovirus because it is an exception to molecular biology's central dogma. Reverse transcription of HIV occurs as part of a complex viral replication cycle. Viruses cannot perform any cellular functions on their own. They are dependent on other cells for growth, metabolism, and replication. HIV targets T-lymphocytes (T cells) to perform these functions, and this action severely compromises the adaptive immune system.
Figure 2.2: The Human Immunodeficiency Virus (HIV)

White blood cells, known as B-lymphocytes (B cells), attack free-floating pathogens in our body fluids (humors). They are the critical component of antibody-mediated immunity, also known as ‘humoral immunity.’ On the membrane of B cells are antibodies specific to each B cell. The highly differentiated antibodies have approximately $10^{10}$ different types of B cells in the human body. B cells use these membranes bound antibodies to identify the pathogen, join to it, and partially activate it. T cells do not attack pathogens found floating freely in body fluids. T cells recognize viral antigens found in infected cells and non-specific immune cells. T helper cells ($T_h$) are called the “alarm” of the immune system. Their receptors bind to a partially activated B cell and fully activates the B cell. This activation causes protein messengers, also known as cytokines, to be released, stimulating the activation of other immune cells. $T_h$ cells are also known as CD4 cells because they express an additional CD4 protein receptor. Cytotoxic T cells ($T_c$) have one purpose: to kill cells that have been infected by a pathogen, such as HIV, or to kill healthy cells that have become abnormal, such as cancer cells. $T_c$ cells are also known as CD8 cells because they express an additional CD8 protein receptor.
Both CD4 and CD8 cells play an essential role in the HIV replication cycle (Figure 2.3). Glycoproteins on HIV bind with receptors on CD4 cells. This binding triggers a change to the shape of the CD4 cell, allowing additional receptors to draw the HIV molecule closer to it until it cleaves the CD4 cell. Next, the fusion of the HIV viral envelope allows the HIV capsid to penetrate the CD4 cell. The capsid deteriorates, releasing the HIV RNA and HIV enzymes into the CD4 cell. HIV contains three enzymes that play specific roles in the replication process.

Figure 2.3: The HIV Replication Process

Reverse transcriptase is probably the most crucial enzyme in HIV replication since it is the catalyst for creating viral DNA inside the CD4 cell. The reverse transcriptase uses free-floating nucleotides in the cytoplasm of the CD4 cell and combines with HIV RNA to create two single strands of DNA. The two DNA strands combine to create a complementary double strand of DNA. This double strand of DNA contains the HIV genetic information. The reverse transcription of RNA to DNA is highly prone to error, causing numerous mutations that have made HIV challenging to treat and impossible to eradicate. Integrase is the next HIV enzyme used in the replication cycle. It integrates the newly formed HIV
DNA into the nucleus of the CD4 cell. Integrase cleaves the CD4 cell DNA and inserts the HIV DNA into it, permanently infecting it with HIV.

With the HIV genetic information fully integrated into the DNA of the CD4 cell, replication is the next step. Some people experience latent replication, while others experience immediate replication. Whether immediate or latent, the enzyme protease replicates RNA into DNA and then into messenger RNA (mRNA). The mRNA leaves the nucleus of the CD4 cell and moves into the cell cytoplasm, forming the essential proteins for HIV replication. These single “viral polyprotein” are later cleaved in the correct places and code for reverse transcriptase, integrase, and protease. The new HIV RNA and HIV proteins move into the ribosomes of the CD4 cell and begin to assemble as they move towards the membrane of the CD4 cell.

Finally, the newly formed HIV pushes its way outside of the CD4 cell. As it buds from the CD4 cell, it takes part of the CD4 cell membrane to create a new, freely floating, immature HIV. The virus is “immature” because the HIV RNA and proteins are not in the correct order to be “mature.” Once outside of the CD4 cell, the enzyme protease cleaves the HIV viral proteins into smaller enzymes correctly coded for reverse transcriptase, integrase, and protease. The capsid forms around the HIV RNA and proteins maturing HIV and making it readily available to infect other cells.

The HIV replication process produces an estimated 10 Billion\( \left(10^{10}\right) \) new HIV virions per day. Untreated HIV leads to CD4 cell destruction, which ultimately prohibits the full activation of B cells and compromises the immune system. CD4 cell destruction occurs through apoptosis and pyroptosis. Apoptosis only explains the destruction of 5-10% of infected CD4 cells. Pyroptosis of an infected CD4 cell sends cytokines, such as interleukin, into the body fluids and explains the remainder. Cytokines attract other non-infected CD4
cells to the infected CD4 cell and cause an inflammatory response, making them explode. Until treatment is received, healthy CD4 cells deteriorate in large quantities.

2.2.3 Stages of HIV, Signs, and Symptoms

There are three stages of HIV – acute HIV syndrome, clinical latency, and Acquired Immune Deficiency Syndrome (AIDS). The stages of HIV are transient since AIDS is the residual effect of changes in specific clinical characteristics. The signs and symptoms of HIV vary not only by the person but also by the HIV stage. Acute HIV syndrome begins after the primary infection and lasts approximately 2 – 6 weeks. During this time, HIV viral loads (HIV RNA) increase significantly, while CD4 cells decrease drastically. During this phase, most men and women's signs and symptoms are fever, chills, rash, night sweats, swollen lymph nodes, mouth ulcers for days to weeks, muscle aches, sore throat, and fatigue, which can also be asymptomatic for some people.

During the clinical latency phase HIV RNA decreases and then plateau while CD4 cell counts increase but decline over time. This period lasts several years, depending on treatment regimens type and response, and is usually considered asymptomatic. The final stage of HIV is AIDS. People are diagnosed with AIDS when their CD4 cell count is less than 200 cells/mL, or they have a specific “serious and life-threatening disease” that occurs mostly or almost entirely in PWH. These illnesses are called AIDS-defining illnesses (ADIs).

ADIs are diseases, such as an infection or cancer, that have a higher incidence or severity in people with severely compromised immune systems such as HIV. The list of ADIs has evolved and was last updated by the CDC in 2008 to include 27 different illnesses. The list includes variants of candidiasis, cervical cancer, cytomegalovirus, herpes, Kaposi sarcoma, non-Hodgkin Lymphoma, mycobacterium, pneumocystis jirovecii pneumonia, and recurrent pneumonia. PWH, who develops any of these conditions, is diagnosed with AIDS
regardless of CD4 cell count. The signs and symptoms of AIDS in most men and women include rapid weight loss, recurring fever, profuse night sweats, extreme or unexplained tiredness, pneumonia, prolonged swelling of lymph nodes in the armpits, groin, and neck, sores in the mouth, anus, and genitalia, colored blotches (red, pink, or purple) under the skin, and memory loss. Women may have additional signs and symptoms that include more frequent, severe, chronic vaginal yeast infections, pelvic inflammatory disease, and infections of Human Papilloma Virus (HPV).

2.2.4 Risk Factors

For a virus such as HIV, risk factors are behaviors or activities that increase exposure to HIV. According to the CDC, the transmission of HIV is not possible through the air, water, saliva, sweat, tears, closed-mouth kissing, insects, pets, or sharing toilets, food, or drinks. Body fluids from an infected person are the only mechanisms of transmission for HIV. These fluids include blood, semen, vaginal and rectal fluids, and breast milk. Body fluids must enter the body to allow HIV to target CD4 cells in an HIV naïve person. This entrance could happen through any mucous membrane, found on all male and female sexual organs, through damaged tissue, or direct injection into the bloodstream, such as in an intravenous needle.

High risk for HIV is defined by unprotected vaginal or rectal sex or intentional sharing of intravenous needles. Other risk factors for HIV include vertical transmission from a mother to a new-born, breastfeeding, accidental intravenous needle pricks, oral sex, the performance or use of blood procedures and products, and anything else that creates contact between mucous membranes, damaged tissue, and blood between two people. Some consider behaviors that inhibit rational thinking for prevention as risk factors for contracting
HIV. For example, alcohol and oral drug use increase promiscuous and unsafe sex. Likewise, certain socioeconomic factors, such as income and education, might increase HIV exposure.

2.2.5 Diagnosis

The WHO estimates that 21% of PWH are unaware of their HIV status. Testing is a critical step for all people but more important for people engaged in high-risk activities. The mechanisms of HIV testing are dependent on the time between exposure to HIV and its appearance in bodily fluids used by HIV tests ('window period'). No HIV test can detect HIV immediately, but specific tests can detect HIV in as little as ten days after exposure. There are three types of tests for HIV. The first, a nucleic acid test (NAT), detects HIV RNA in the blood. The second, a test that detects both antibodies and antigens, identifies immune response proteins in the blood. Finally, antibody only tests can use blood or oral fluid to detect antibodies 23 to 90 days after exposure.

Tests today have 99.5% sensitivity and specificity. While infrequent, erroneous detections of proteins and misread or misclassified results lead to false-positive outcomes. False negatives are also rare but more likely caused by testing during the window period. Two different prophylaxis medications exist to reduce the risk of transmission and reduce the risk of post-exposure infection. Pre-exposure prophylaxis (PrEP) is a daily medication taken by people at high-risk for exposure to HIV, such as doctors and nurses. Post-exposure prophylaxis (PEP) is a medication that can be taken by people within 72 hours of known exposure to HIV. HIV testing is critical to the care and management of the disease since clinical advances in anti-retroviral therapies (ART) have proven to bring HIV viral loads to undetectable levels and lead to better health.
2.2.6 *Anti-Retroviral Therapy (ART)*

Anti-Retroviral Therapy (ART) has drastically changed the prognosis of HIV. When the epidemic began in the 1980’s the life expectancy for PWH was less than 1 year. HIV was a terminal disease. The first FDA approved treatment was a nucleoside reverse transcriptase inhibitor (NRTI) known as azidothymidine (AZT). It stopped the translation of RNA into DNA during viral replication and reduced the morbidity and mortality from AIDS-related illnesses. Severe adverse effects made the tolerance and adherence to AZT challenging.

The introduction of protease inhibitors led to triple drug combinations known as Highly Active ART (HAART). HAART combined protease inhibitors with two different types of NRTI’s. This led to significant improvements in drug resistance and sustained control of HIV viral loads and CD4 cell counts. HAART regimens were harder to comply with because of its complexity. There were multiple types of drugs which needed to be taken at varying times of the days using varying modalities.

ART has evolved over the last three decades. Today, there are more than 30 types of ART with improved toxicity that has lessened the adverse effects and improved regimen compliance. There have been very few studies to explore the effect of ART on life-expectancy. Some have shown life expectancy without ART is less than 10 years. ART can add as much as 46 years to life expectancy of PWH however life expectancy for PWH is still 30-40% lower than life expectancy of the general population. As a result of less toxicity, less adverse effects, and improved modality, more people, but not all, are using ART. According to the WHO, 23 million people were on ART in 2018 compared to 2 million people in 2005. Unfortunately, this was only 61% of PWH which means 14.6 Million people are not on ART.
2.2.7 Comorbidities, Cancer, and Lung Cancer in PWH

Before ART, PWH did not live long enough to experience illnesses that are prevalent today among PWH. They died from AIDS-defining illnesses. In the pre-ART era, the most widespread diseases were pulmonary lung disease: acute bronchitis, bacterial pneumonia, pulmonary tuberculosis, and pneumocystis pneumonia (PCP). Aside from pulmonary lung disease, PWH suffered from co-infections or three specific cancers before dying: Kaposi's sarcoma, non-Hodgkin's lymphoma, and invasive cervical cancer.

Today PWH live longer, are aging, and are susceptible to many of the same age-related illnesses as those living without HIV. There has been an increase in the number of non-aids-defining diseases, which are comorbidities found in the general population but at a higher frequency in PWH. These include cardiovascular disease, pulmonary lung disease, neurocognitive disease, diabetes, and cancer.

Other than the three AIDS-defining Cancers, all other cancers that occur among PWH are non-AIDS-defining cancer. PWH have a higher risk for cancer because HIV may trigger pro-oncogenes in cells, which disturb the regulation of controlled cell growth and division, and HIV may be an independent risk factor for cancer. ADCs were the primary cause of AIDS-related death before ART; however, NADCs account for the most significant number of deaths today. Some NADCs are associated with co-infections with other viruses such as the Hepatitis C virus (HCV), Human Papilloma Virus (HPV), and Epstein Barr Virus (EBV). Lung cancer is not associated with a viral co-infection and is considered unique cancer in PWH. Other NADCs, such as breast cancer and prostate cancer, occur at similar rates as the general population. Lung cancer is the leading cause of cancer death among PWH occurs at higher rates than the general population. This is in part due to disproportionate smoking incidence in PWH but is also due clinical risk factors of HIV.
2.3. Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) have existed for over four decades. Their importance in health information technology catapulted after a landmark report described human error outcomes in hospitals and the medical system. Since then, mandates from federal agencies have increased their prevalence in mainstream health management due to their benefits. However, fundamental limitations held them back from a broader adoption. While CDSS serve a deeper purpose in the hospital setting, they are used to facilitate decision making in the screening, diagnosis, and treatment of complex diseases such as cancer. They play a large role in many cancers but for lung cancer, their role is limited, and among PWH, they don’t exist.

2.3.1 Overview of CDSS

There are many varieties of definitions for CDSS. The American Health Information Management Association (AHIMA) defines CDSS as a computer system that provides “clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care”47. The most common way to classify them is knowledge-based versus nonknowledge-based. There are three parts to a knowledge-based CDSS (Figure 2.4)48. Knowledge-based CDSS contains a communication mechanism to capture user input and provide user output, a knowledge base, and a reasoning engine or inference mechanism. The knowledge base is a manual or semi-manual collation of information used to derive the inference engine rules. This study used two sets of public use datasets (PDS) from Johns Hopkins University to form the knowledge base. Combining analytical results from both the Multicenter AIDS Cohort Study (MACS) and the Women's Interagency HIV Study (WIHS)
with information obtained from the literature review, the study developed the knowledge base.

Figure 2.4: A general model for a knowledge-based CDSS

![Diagram of a knowledge-based CDSS](image)

Basic mathematical principles such as set, boolean, and probability drive the inference engine's rules. The inference engine logically combines the user input to the system using defined variables, data-driven forward logic, goal-oriented backward logic, and command blocks to transmit information back to the user. Nonknowledge-based CDSS use artificial intelligence to learn from the data and to derive information. One primary difference between the two is that knowledge-based systems require curated data to drive the reasoning.

A brief history of notable CDSS helps understand how CDSS has evolved as information technology advanced. The first notable CDSS was developed in 1959 when Robert Ledley and Lee Lusted published, “Reasoning foundations of medical diagnosis, probability, and value theory aid our understanding of physicians reason” 49,50. Ledley and Lee used set-theory, Bayesian probability, and punch cards to create a knowledge-based
CDSS for an analog computer. This system became the first stand-alone CDSS (computer programs that run separately) but paved the way for other CDSS as technology improved. While the first CDSS to integrate multiple systems was the HELP system, a more notable knowledge-based integrated CDSS was the Regenstrief Medical Record System (RMRS)\(^{49,51}\). RMRS proved physicians followed protocol-driven guidance better under CDSS than without it. MYCIN, a more notable integrated CDSS, is considered one of the first to promote nonknowledge-based reasoning. While it used an existing knowledge base to recommend antibiotic therapy with acceptable sensitivity (75%), it improved with new rules supporting the theory that ‘learning’ from data could improve CDSS\(^{49}\). RMRS and MYCIN are an example of CDSS tools that are specific to single tasks. At the same time, other less notable CDSS were limited to particular diseases. DXplain is an older, but current system which looks across multiple medical domains to provide and explain diagnostic recommendations. The internet paved the way for CDSS by allowing different storage and data access mechanisms and facilitating broader use across multiple healthcare settings. One of the first to do this is SEBASTIAN, a web-based, knowledge-based CDSS that identifies high risk patients, provides alerts to clinicians, and emails reminders to patients.

2.3.2 \textit{CDSS as an Expert System using Exsys Corvid}\(^{\circledR}\)

The primary goal of a CDSS in the healthcare setting is to assist doctors and clinicians in decision making for particular domains such as screening, diagnosis, and treatment of disease. Human decision making is a process of obtaining diverse data from varied sources (e.g., patient input, experience, literature, etc.), cognitively reasoning from data, information, and knowledge, and then communicating the decision. The transformation of data to information to knowledge for clinical reasoning and decision-making is a profoundly focused healthcare process. Doctors and clinicians focus individually
on domains such as screening, diagnosis, and treatment since there are multiple considerations and extensive data to consider before making a decision. Expert systems are knowledge-based CDSS that imitate the human-decision making process for a particular reasoning domain using conditional logic.

Exsys Corvid ("Corvid") is a 28-year old verified tool that supports the design and implementation of expert CDSS. It provides the three essentials of a CDSS: a communication tool, a knowledge base, and a reasoning or inference engine. Corvid is a highly interoperable system since it easily allows access, exchange, and integration to other systems. Another benefit of Corvid is the ease of designing an advanced graphical user interface without a pre-requisite for HTML; its power is with the inference engine's decision-making logic and ability to drive interactive sessions that advise users dynamically.

Corvid's inference engine uses broad rules, known as 'heuristics,' to build probabilistic rules representing decision-making process elements. It does this through logic, action, and command blocks. Logic blocks organize rules in a decision-tree manner, while action blocks organize steps in the decision-making process. Command blocks describe the sequence of procedures and allow for two types of heuristics, forward and backward chaining. Forward chaining is a data-driven process that applies inference rules to available data and acquires more data until a goal is reached in the form of a decision. Backward chaining is a goal-driven process that works backward from a goal or decision and applies inference rules to identify facts that support the goal or decision. Combining the inference engine's vigor with heuristics, probabilistic inference, interoperability, and a visually stimulating interface makes Corvid a useful tool for designing an expert CDSS.
CHAPTER 3

LITERATURE REVIEW

3.1 Lung Cancer Epidemiology

The 5-year overall survival for lung cancer is among the lowest of all other cancers, including liver and pancreatic cancer\(^{25}\). Smoking remains the most significant risk factor for lung cancer; however, other behavioral, clinical, familial, environmental, and social-economic factors increase the risk of lung cancer. Despite considerable improvements in the screening and early detection for lung cancer, gaps exist in current guidelines, and disparities exist in specific demographics targeted for screening.

3.1.1 Incidence and Mortality

Lung cancer is a fatal cancer in both the US and the world and is the leading cause of cancer death\(^{38}\). In 2016, 1.76 million people died globally from lung cancer, while 603,735 died in the US\(^{24,31}\). Some studies forecast 142,760 deaths, which equates to 391 deaths per day for 2019\(^{24}\). Lung cancer incidence increased significantly for more than 100 years, primarily due to increases in smoking prevalence\(^{30}\). The rate of lung cancer has decreased overall since then; however, the number of new cases of lung cancer has been increasing. According to de Groot et al., there were 161,000 new lung cancer cases in 1991, while the estimated number of new cases of lung cancer in 2019 is 228,150, according to Siegel et al. Although it may be true there is a decreasing trend in incidence and mortality overall, it may not be true across specific demographics.
In specific emerging or developing countries, such as Brazil, Russia, India, China, and South Africa, lung cancer incidence has been increasing due to a higher prevalence in smoking\textsuperscript{26,29}. While smoking is the leading risk factor for lung cancer, the primary risk factor across cancers is age. As age increases, so too does the risk for cancer. The median age at diagnosis for lung cancer is 70 years old, and the median age at death from lung cancer is 72\textsuperscript{26}. Differences also exist across gender, and race for both incidence and mortality (Figures 3.1 and 3.2).

Figure 3.1: Number of Cases by Sex and Race/Ethnicity – US – 2017 (CDC)

![Number of Cases by Sex and Race/Ethnicity](image)

Figure 3.2: Number of Deaths by Sex and Race/Ethnicity – US – 2017 (CDC)

![Number of Deaths by Sex and Race/Ethnicity](image)
Incidence and mortality are higher for men than women in the US; however, the gap between them is small, although increasing from 13% in 2017 to 20% in 2018. Some studies forecast that differences will reverse by 2045, with women having higher incidence and mortality than men. The gender difference has to do with “historical, cultural, and regional” differences in smoking. The incidence of smoking in men peaked earlier than for women in developed countries, 1980 and 2000. Culture and societal norms influence smoking in developing countries. As a result, smoking is more prevalent in men. Nonetheless, smoking cessation programs are beginning to affect the high rates of lung cancer, and changes are noticeable.

Race and ethnicity influence the rate of new lung cancer cases and deaths. In developed countries such as the US, white men and women have the highest rates followed by African Americans (“Black”). The demographic differences in annual mortality rates in developing countries are different from those of the 5-year overall survival (OS) rates. Using estimates from 2001-2016, the CDC reported better 5-year OS in women than men (23.1% and 16.5%) and white race compared to the black race (19.7% and 17.1%). These disparities are presumably due to differences in cigarette smoking and access to care. Some studies have also linked differences in incidence and mortality to education and socioeconomic status. Other studies have shown that smoking is more prevalent in those without high school education. Subsequently, lung cancer incidence is higher in the less educated, such as those without a high school diploma than a college graduate, 166.6 per 100,000, and 57.5 per 100,000.

3.1.2 Relevant Findings in Lung Cancer

The most significant and recognized risk factor for lung cancer has been smoking since the Surgeon General reported it as a direct cause of lung cancer. In 1964, the
Surgeon General released a US Public Health Service report drawing attention to the adverse risks of smoking and the "causal relationship" to lung cancer\textsuperscript{26}. Forty years later, the Surgeon General released a new report stating lung cancer is irrefutably the most important risk factor and "major cause" of lung cancer. Smoking prevalence has decreased markedly since then in the US. One study reported a 22\% decrease from 1970 to 2015, 48.1 million compared to 37.5 million\textsuperscript{25}. Another study reported a decline from 20.8\% prevalence in the US in 2005 to 14\% in 2017 \textsuperscript{26, 35}. Men have smoked more than women historically, leading to higher mortality rates; however, men had success in smoking cessation earlier than women. As a result, lung cancer incidence and mortality have decreased in men 2-to-3 times faster than in women\textsuperscript{25}. Although this may be true in developed countries, the same is not valid in developing countries where cultural and societal norms affect smoking rates.

As a solution to reduce smoking, Electronic Nicotine Delivery Systems (ENDS) have been modernized in the last decade and become a popular but dangerous trend among smokers. ENDS are more commonly known as 'e-cigarettes' or 'vaping' and began as healthy alternatives to smoking. Despite this, there is a slow surge in evidence that ENDS is more harmful than helpful in decreasing the smoking incidence and reducing lung cancer. Studies indicate that ENDS use increases the risk of cigarette smoking, even among those who have never smoked, and decreases the chances of smoking cessation\textsuperscript{25, 29}. While it is too soon to evaluate ENDS' effect on the risk of lung cancer, there is some early evidence raising concern. A recent study found that ENDS caused the same molecular and genetic mutations in cells as cigarettes\textsuperscript{29}. As a result, ENDS has neither contributed to a reduction in smoking nor reduced lung cancer incidence. It is more likely contributing to uptake in smoking, specifically among people who never smoke. More than 80\% of lung cancer occurs in
current and former smokers; however, people who have never smoked ("never smokers") also develop lung cancer.

Never smokers are people who have smoked less than 100 cigarettes in their lifetime. Studies have shown that 15% and 53% of global lung cancer occurred in never-smoking men and women, and these cases of lung cancer make between 10-25% of all lung cancer worldwide\textsuperscript{25, 26}. In the last 30 years, the incidence of lung cancer among never-smokers has increased from 8.9% to 17% and would be the 7th leading cause of cancer deaths in the US if considered separately from lung cancer in smokers\textsuperscript{25, 26, 30}. Despite smoking's association with 80-90% of lung cancers, only 20% of smokers will develop lung cancer, and never-smokers have an increased risk for lung cancer. In trying to understand the risk of lung cancer in non-smokers, environmental factors, genetics, and viral co-infections not only explain the risk of non-smokers but, in some cases, explain the heightening of risk in smokers.

Most epidemiologic studies focus on three critical environmental risk factors easily found at home and work: asbestos, radon, and domestic fuel smoke. Among occupational exposures such as arsenic, chromium, and ionizing radiation, asbestos is the most crucial since it increases lung cancer risk five times, and exposure is common in older homes and buildings\textsuperscript{30, 56}. Radon is a vital risk factor among non-smokers since houses built on soil may contain it, making it a household hazard for non-smokers but increasing lung cancer risk 25 times in smokers\textsuperscript{30}. Domestic fuels used in households, such as coal, also significantly increase the risk of lung cancer. The WHO estimates that 3.8 million people globally die from disease attributed to domestic fuel smoke, and 304,000 (8%) are from lung cancer. In an extensive, pooled analysis of 7 studies, domestic fuel smoke from coal increased lung
cancer odds by 2.15 times. Despite the risk of lung cancer from environmental exposures, other factors such genetics play a role as viruses and.

There is a shortage of studies and data on the predisposition of lung cancer due to genetics. One study has shown evidence of an association between a family history of lung cancer and the risk of lung cancer in first-degree relatives. In the study by Matikidou et al., family members of people with lung cancer had an increased risk of 1.84. with lung cancer. Lung cancer is historically not associated with viral coinfections; however, newer evidence incites debate as it suggests there might be an association. Our specific and non-specific immune system recognizes and destroys cell mutations and foreign invaders, such as viruses. A breakdown in our immunity reduces cancerous cells' destruction and allows increases in viral coinfections, explaining some cancers. Three viruses traditionally uncommon to lung cancer have appeared in lung cancer cells. In some small studies, in specific regions, HPV has been found in lung cancer. HPV DNA existed in 20% of Koreans with lung cancer in one study and 39% of Latin Americans with lung cancer in another study. Similarly, EBV was associated with lung cancer in Asian populations. H-Pylori increased the likelihood of lung cancer, with odds ranging from 1.2 to 17.78 in various epidemiological studies. Medications that create immunodeficiency contribute to this phenomenon.

3.1.3 Challenges in Screening and Early Detection

Smoking cessation programs have aimed to reduce the incidence and mortality of lung cancer since the harms of smoking were first reported in 1964 by the Surgeon General. Smoking cessation programs have been successful in mitigating some of the risks of lung cancer. Despite these efforts, lung cancer has been the leading cause of cancer-related deaths for many decades.
The NLST changed lung cancer screening outlook by demonstrating that annual low-dose CT (LDCT) was superior to CXR and reduced lung cancer mortality among high-risk people by 20\%. Despite this significant benefit of LDCT, its false-positive rate is controversial since it could harm patients\(^1\). Twenty-five percent of the scans found by the NLST detected nodules in the lungs, but more than 95\% were benign. Nodules discovered due to a false-positive need invasive biopsy, which may be unnecessary and lead to additional harm. Some harms include surgical complications, anxiety, distress, increased costs, excessive radiation exposure, and cancer discovery that would not have harmed (overdiagnosis)\(^3, 6\). Both Henschke et al. and the Prostate, Lung, Colon, Ovarian Trial contradict the harm of overdiagnosis. Both concluded early-stage lung tumors have a 'malignant natural course,' will be fatal if not treated, and have a significantly higher chance of survival past ten years\(^5, 6\). Likewise, some have refuted the harm of overexposure to radiation. The controversy regarding overexposure to radiation derives from studies of large scale radiation exposure. While this may be true, LDCT produces low-levels of radiation, which the American Association of Physicians in Medicine state is not harmful\(^6, 6\).  

Despite the relative risk associated with the NLST recommendations, many accept that the benefits outweigh the harms\(^3, 6\). First, the high false-positive rate of LDCT was no different than the false-positive rate for the CXR. Second, LDCT resulted in a 20\% reduction in mortality compared to CXR. The USPSTF agreed and adopted the NLST criteria, with small modifications, as the standard-of-care for screening for lung cancer in the US\(^3, 3\). The USPSTF standards are similar to the NLST criteria but extend the upper age limit for inclusion to 80 years of age. Both the NLST and USPSTF criteria have been in use for almost a decade.
There is an ongoing debate over the upper and lower limits of the inclusion criteria for screening and the risk factors which increase the risk of lung cancer, aside from smoking. First, there is debate over the role of LDCT among non-smokers. There is an identified risk for lung cancer among non-smokers. In a more extensive study of 12,176 never-smokers, LDCT discovered nodules in 10% of the patients, and 55 of them were malignant, and almost entirely early-stage\textsuperscript{30,69}. Second, there is debate over strict adherence to the age and smoking criteria recommended by the NLST and USPSTF. In a retrospective, comparative study of LDCT results between 2001 to 2015, a significant increase was not only seen in low-risk, high risk, and never smokers but also people who didn't meet the NLST criteria\textsuperscript{36}. Third, there are considerable differences in the risk of lung cancer based on behavioral risk factors other than smoking. Lung cancer screening should consider significantly different variations of risk factors such as age, socio-economic status, environmental hazards, and smoking intensity and duration\textsuperscript{29}. Finally, discussions continue on the cost-effectiveness of lung cancer screening through LDCT. Some have estimated the cost of screening per-life-year saved to be around $50,000, which is high compared to cervical and breast cancer, which are expected to be $15,600 and $24,000 per life-year saved but low compared to colon cancer, which is estimated to cost $127,000 per life-year saved\textsuperscript{62}. In contrast, others have predicted a much lower cost of lung cancer screening, around $19,000 life-years saved\textsuperscript{62,70,71}. The monthly estimate to treat a 70-year-old with lung cancer with chemotherapy is $9,000 proving lung cancer screening is cost-effective\textsuperscript{72,73}. Despite disagreement on the inclusion and exclusion criteria for lung cancer screening and the benefits and harms of lung cancer screening, there is agreement on the need for shared decision making for lung cancer screening\textsuperscript{61}. 
Shared decision-making (SDM) is "a collaborative process that allows patients and their health care providers to make health care decisions together, taking into account the best scientific evidence available and the patient's values and preferences."\(^{68}\) The USPSTF recommends that the decision to be screened for lung cancer should involve full disclosure about the benefits and harms. Similarly, the Center for Medicare Services (CMS) requires a visit for SDM for the provider to be paid for the screening services. It must cover benefits and harms, false-positive rates, overdiagnosis, radiation exposure, and follow-up testing. A comparative study of pre- and post-SDM visits in 423 patients showed that patients neither understood the age and smoking requirements for screening nor the benefits and harms of screening\(^{61,68}\). The factors which affect a patient's decision-making process are unknown and remain under investigation. In one study, only 50% of patients were screened after engaging in shared decision making while in another study, 95% chose to be screened\(^{38,61,68}\). Shared-decision making is becoming a critical step in decision making in a process that has an unequal value of benefits and harms.

3.1.4 The role of CDSS for Lung Cancer Screening

Lung cancer screening guidelines are relatively new. Results from the NLST were first reported in 2012 and the USPSTF recommendations were published in 2014. The relative newness of lung cancer screening has led to a lack of CDSS\(^{74}\). Among those that do exist very few have been validated through peer review. The primary purpose of CDSS for lung cancer screening is to appropriately identify patients at the greatest risk for lung cancer, who have the greatest potential to benefit from diagnostic tests\(^{74}\). The NLST recommendations and USPSTF guidelines are rigidly dichotomous with strict age and smoking criteria; however, additional risk factors, independent of age and smoking also increase the risk of lung cancer. The National Comprehensive Cancer Network (NCCN)
finds it reasonable to consider these risk factors in people who do not meet NLST and USPSTF criteria in determining lung cancer risk. As a result, risk prediction plays a large role in CDSS for lung cancer screening and has been identified as a critical part of lung cancer screening even among NLST or USPSTF eligible patients.

Risk prediction models predict the probability of condition, such as lung cancer, using known factors. Lung cancer models typically predict the probability of developing lung cancer using logistic regression (Odds Ratios) or Cox regression (Hazards Ratios) modeling independent risk factors such as age and smoking. There are nine models in lung cancer that each use different populations to derive the models and different risk factors making it challenging to generalize and compare. This has resulted in no consensus on an optimal model. Three models have been cited by many as ‘preferred’ models albeit, they used different population and risk factors to derive the models. They include the Bach Model, Liverpool Project Model (LLP), and the PLCO Models. Performance was measured by the model’s ability to predict the probability of an event (Calibration) and the models ability to differentiate between those who will and will not develop disease (Discrimination) as performance measure and listed Bach, LLP, and PLCO as high performing models; however, the measures were dissimilar in each. The Bach Model uses age, gender, cigarettes/day, years smoked, and asbestos exposure as risk factors. The LLP Models uses years smoke, family history, asbestos exposure, pneumonia, and prior malignant tumor as its risk factors. The PLCO model uses smoking status/intensity/duration, years since quitting (if former smoker), education, ethnicity, body mass index, and personal and family history of cancer. Each of the Bach, LLP, and PLCO are available online as electronic CDSS.

After four decades, there are numerous CDSS, each with their own set of benefits and limits. The primary benefit of CDSS is the reduction of human error through
computational human-like reasoning. Two decades ago, the Institute of Medicine released a landmark report stating, “Healthcare in this country is not as safe as it should be, or can be” and said that human error was the root cause of “preventable medical errors.” The authors quantified the deaths attributed to human errors and found they exceeded deaths from car accidents, AIDS, and breast cancer, which were the top three leading causes of death at that time. This became the catalyst for CDSS in the modern-day and a foundational requirement for insurance reimbursement in the US. With the passage of the HITECH Act, insurance providers reimburse clinicians who “meaningfully using” electronic medical records that integrate CDSS into its procedures. The justification for the meaningful-use requirement is empirical evidence. RCTs suggest reductions in human error and improvements in outcomes are achievable through computerized physician orders of treatments and procedures and the diagnosis and identification of patients for screening. For example, when compared to those who do not use a CDSS, there are more outstanding pneumococcal vaccine orders by clinicians (35% and 0.8%), more diagnostic and treatment reminders about osteoporosis fracture (51.5% and 5.9%), and less mortality with the use of beta-blockers during an ECG exam (66% and 74%). While there is surmountable evidence showing that the CDSS better identify patients who can benefit from appropriate treatment and preventive care and facilitate ordering, there is also evidence that refutes these benefits and explains the limits of CDSS.

Several challenges exist today with CDSS. First, there is a lack of research supporting the broad adoption of CDSS. This gap could be due in part to financial costs, the complexity of CDSS design, and lack of standards. Data has grown exponentially and has become highly unstructured, making it large, unpredictable, and challenging to integrate. The variability of ‘big data’ and the lack of standards has made it difficult for providers of
CDSS to develop solutions that can be universally used (“out of the box”) and scalable. There are also reports of user fatigue from systems that are not appropriately designed and the pressure to move faster under ‘fee for service’ cultures within organizations.

Currently, there are no published reports about the use of CDSS for lung cancer screening. Some exist for the treatment of lung cancer. The most applied used for CDSS in lung cancer is through screening alerts and risk calculators programmed into EMR based CDSS in a hospital or clinical settings. Web-based risk-assessment tools, using verified algorithms, have increased due to heightened interest in shared-decision making. This is due mostly to the CMS requirement for shared-decision making but more importantly due to the potentials harms associated with lung cancer screening. Existing web-based risk assessment tools serve as CDSS for both patients and providers. They provide patients with knowledge about the benefits and harms of lung cancer screening, along with person-specific information about their risk of lung cancer that is intelligently filtered and presented in a manner than can enhance their choice to be screened or not to be screened.

3.2 HIV Epidemiology

There have been significant changes in the epidemiology of HIV since it caused the global pandemic of AIDS in 1981. The incidence of HIV and mortality caused by HIV have both decreased while the prevalence has increased. PWH live longer lives with improved quality-of-life due to clinical advances in anti-retroviral therapies (ART). Before ART, the prognosis of HIV was dismal. HIV was mostly prevalent in white men who had sex with men and injection drug users. Without treatment, HIV continued to attack the immune system until an immune deficiency syndrome was acquired, leading to death. The life
expectancy ranged from as little as one year to as most as ten years. Today, HIV is considered a chronic disease\textsuperscript{46}. The life expectancy of PWH is almost the same as the general population of people without HIV. Despite these changes, there has been a shift in demographics most at risk for HIV and an increase in comorbidities historically unseen in PWH.

3.2.1 Incidence and Prevalence of HIV

The World Health Organization (WHO) reported 37.9 million people with HIV globally at the end of 2018, and the global annual incidence of HIV was 1.7 million\textsuperscript{44}. In the United States, the Center for Disease Control and Prevention (CDC) estimated that 1,018,346 people live with HIV at the end of 2017. The CDC estimated 81% of adults and adolescents living with HIV were males, and 19% were females. The annual rate of HIV among males and females at the end of 2017 decreased significantly compared to 2010. In 2018, the annual incidence of HIV was 37,741. The proportion of males and females remained at 81% and 19%, respectively. While the yearly incidence rate has been trending downwards, there have been notable shifts in prevalence and incidence of HIV infections.

One especially interesting shift has been in the age distribution in incidence and prevalence\textsuperscript{44,46}. The population of PWH are living longer and aging due to clinical advances in ART. At the end of 2017, the CDC estimated that 48% of PWH were aged 50 or older (Figure 3.3). Over the last eight years, the annual incidence of HIV increased among 25-34-year-old males and females. The rate decreased among 13-24 and 45-54 by 42% and 30%, respectively, while remaining stable for 35-44 and \( \geq 55 \) years.
African American race is the most frequent race among adult and adolescent males and females living with HIV, 35%, and 58%. The second and third most frequent HIV infections in males are among White and Hispanic men, 33% and 25%, respectively. In contrast to males, the second and third most frequent HIV infections are among Hispanic and White females, 20% and 16%, respectively (Figure 3.4). Over the past seven years, the rate of HIV infection between Hispanics and Whites has reversed. In 2018 Hispanics had a higher incidence than in 2010. Among incident new cases of HIV in adults and adolescents, African American races are the most frequent race in both males and females, 74.8% and 23.1%, respectively. In males, the second and third most frequent are Hispanic and multiple races, 36.4%, and 30.9%. In females, the second and third most common are multiple races and Hispanics, 7.3% and 5.2% (Figure 3.5).

HIV transmission is categorized in five ways: male-to-male sexual contact or men who have sex with men (MSM), heterosexual contact, injection drug use (IDU), both MSM
and IDU, and others. MSM is the most common transmission method followed by heterosexual contact in both males and females (Figure 3.6).

Figure 3.4: Adults and Adolescents Living with Diagnosed HIV Infection, by Sex and Race/Ethnicity, Year-end 2018—United States

Figure 3.5: Rates of Diagnoses of HIV Infection among Adults and Adolescents by Sex and Race/Ethnicity, 2018—United States
3.2.2 Mortality from HIV and Life Expectancy

The World Health Organization (WHO) reported 75 million people died from AIDS-related illnesses globally. At the end of 2018, the global annual mortality was 770,000. In the United States, the Center for Disease Control and Prevention (CDC) reported that 700,000 have died from AIDS-related illnesses. At the end of 2017, the annual mortality was 16,350. Age-adjusted mortality rates in the US linearly increased from 1987 to 1995, where it was at its highest. The increase was primarily due to the lack of treatments. After the introduction of ART in 1995, the age-adjusted mortality rates sharply decreased from 1995 to 1998 with 28%, 46%, and 18% decrease each year. Since then, the annual reduction in mortality rates has ranged between 3 and 13%.

There are similar gender, age group, and racial trends in mortality from HIV as compared to incidence. Males have a higher mortality rate than females; however, the ratio decreased from 10-to-1 in 1987 to 3-to-1 in 1997. While age groups leveled over the past decade compared to the first decade, the groups at the highest risk for death have also
changed. In the first decade, 35-44-year-old males and females died at the highest rate, followed by 25-34 and 45-54 year-olds. The lowest mortality rate was in people older than 55 and younger than 25. While nearly equal rates exist today, those with the highest mortality rates are males and females over the age of 45. There are similar trends in mortality rates in males and females by race, as there were for incidence and prevalence. In both males and females, African Americans have a significantly higher mortality rate than all other races, 11.66% and 5.23%. The group with the second and third higher mortality rates are Hispanics and American India/Alaska Native in males, 3.05% and 2.52%, and 0.73% and 0.96% in females.

The life expectancy for PWH has improved significantly; however, it is not the same as the general population\textsuperscript{5, 37, 86, 87}. The improvement is solely due to clinical advances in ART. Despite this improvement, studies have shown that life expectancy is considerably different across specific clinical, demographics, and socioeconomic groups\textsuperscript{43, 87}. Samji et al. reported differences in life expectancy by transmission group, race, and CD4 cell count at diagnosis of HIV. The differences in race agree with what has been published by the CDC. PWH who are injection drug users or have low CD4 counts had a shorter life expectancy. Similarly, Wandeler et al. reported differences in life expectancy by race, gender, CD4 cell count, transmission group, and drug use in a global comparative study. It also noted that both injection and non-injection drug use increases the risk of death. They reference the Swiss Cohort study and call attention to differences in life expectancy by socioeconomic characteristics such as education and income\textsuperscript{88}. Some believe these variations come from disparities in access to care and socialization and higher incarceration rates\textsuperscript{46}. Today, PWH can live around 70 years, but various clinical, demographic, and socioeconomic characteristics influence this, and there are disparities.
The CDC tracks the incidence of AIDS (Stage 3 Classification), those who are living in AIDS (Stage 3 Prevalence), and deaths of PWH ever classified as having AIDS (Stage 3). The incidence and mortality rates of AIDS is similar to HIV. AIDS incidence and mortality increased, spiked, and then declined between the start of the pandemic and 1995. The prevalence of stage 3 (AIDS) has steadily increased due to clinical advances in treatment and diagnosis (Figure 3.7).

Figure 3.7: Stage 3 (AIDS) Classifications, Deaths, and Persons Living with Diagnosed HIV Infection Ever Classified as Stage 3 (AIDS), 1985–2018—United States

3.2.3 Changes in comorbidities

As the incidence of AIDS diagnoses declined over time, so has the prevalence of AIDS-defining illnesses. There has been an increase in the number of non-aids-defining illnesses, which are comorbidities found in the general population but at a higher frequency in PWH. These include cardiovascular disease, pulmonary lung disease, neurocognitive disease, diabetes, and cancer. Some have argued that the risk factors for these comorbidities are the same as those in the general population. In contrast, some have reported HIV as an independent risk factor that could accelerate the incidence among PWH. HIV leads to chronic inflammation, which could contribute to an increase in non-AIDS-defining illnesses.
Despite the absence of large, prospective studies on pulmonary lung disease, several studies reported pulmonary lung disease as the most common set of comorbidities among PWH. George et al. reported 31.5% of PWH in their cross-sectional study had respiratory comorbidities. Gingo et al. reported 63.5% had respiratory comorbidities in their single-center study. Not surprisingly, the prevalence of smoking in both studies was high, 60% and 76%, respectively. In PWH today, bacterial pneumonia infection, mycobacterium tuberculosis infection, and PCP are common in PWH and ADIs. COPD is a non-infectious pulmonary lung disease with a higher prevalence among PWH despite being a leading cause of death in the general population. There is discordance about the effect of HIV on COPD. Some studies have shown HIV to be an independent risk factor for COPD, while others have not. This disagreement might be due to the recent discovery of HIV infected CD4 cells in lung tissue. Presti et al. reported that pulmonary lung disease associated with HIV is due to the infection itself, increased inflammation, and poor immunodeficiency.

There has been an 80% decrease in the incidence and mortality from AIDS-defining cancers. There has been a noticeable increase in the incidence and mortality from non-AIDS defining cancers (NADCs). The overall incidence of NADCs is projected to increase throughout 2030 as age distributions shift in PWH. In the US, Shiels et al. and Spence et al. project the number of PWH over the age of 65 will triple, and the number of those between the age of 45 and 64 will increase by 10%. Likewise, they project prostate, lung, and liver cancer will be the most common NADCs. Lung cancer will remain the second most frequent cancer and leading cause of cancer-related death among PWH.
3.3 HIV-Associated Lung Cancer

PWH are living longer, and the evolution of HIV has made it a chronic disease. Even on ART, PWH have a higher propensity to develop comorbidities due to the fluctuations in their immune system. Some comorbidities are so rare in the general population that their presence in PWH indicates severe immunodeficiency. These diseases are referred to as AIDS-defining illnesses (ADIs) because of their higher incidence and severity in PWH. Other comorbidities that present in PWH are considered non-AIDS-defining illnesses. These diseases are slightly more common than ADIs in the general population; however, they manifest more severely and with a higher incidence in PWH than the general population. Cancer can be either an AIDS-defining or non-AIDS-defining cancer. Three specific cancers are ADIs: Non-Hodgkin’s Lymphoma (NHL), Kaposi’s Sarcoma (KS), and invasive cervical cancer. These three cancers are known as AIDS-defining cancers (ADC). While any other cancer that occurs in PWH is considered non-AIDS-defining cancer (NADC), the most common have been anal, breast, colon, liver, lung, prostate cancer, and cancer of the oral cavity or pharynx. The epidemiology and pathogenesis of these cancers are different in PWH, and there have been changes in the post-ART era.

3.3.1 Pre-ART vs Post-ART and Epidemiologic Changes Affecting Cancer

PWH in the pre-ART era (1981-1996) had a short life expectancy. At that time HIV rapidly progressed into AIDS and the prognosis was fatal. PWH simply did not live long enough to develop many comorbidities other than ADIs. In the post-ART era (1996 onwards), PWH are living longer lives. Their life expectancy is almost the same as the general population barring some differences in clinical, demographic, and socio-economic characteristics such as CD4 count, age, race, and education. There have been several
significant shifts in the incidence of both ADCs and NADCs in the last ten years which will continue through 2030.

First, there has been a shift in the distribution of age groups among PWH. Using data between 2000 and 2012 from the National Cancer Institute’s HIV/AIDS Cancer Match Study, Shiels et al. reported the overall number of PWH grew between 2010 and 2020 but will decline between 2020 and 2030; however, the 45-64-year-old and 65+ age groups will increase. Second, there has been a shift in the incidence of ADCs and NADCs. With the clinical advances of ART, PWH have become immunocompetent and are living longer. As a result, there has been a decrease in ADCs and an increase in NADCs. The most significant accretions will be seen in PWH 45 years-of-age or older, and the increase will continue through 2030. Finally, there has been a shift in the most common cancers. While in 2010, NHL, KS, and Lung cancer were the three most common cancers, there was a change in 2020. The three most common cancers are prostate, lung cancer, and NHL. In 2030, the three most common cancers will be prostate, lung, and liver cancer.

3.3.2. Differences in Lung Cancer Prevalence among PWH

PWH have an increased risk of developing lung cancer. The estimate of risk is between 2 and 10 times the risk of the general population. The incidence and prevalence of lung cancer among PWH are tracked differently compared to HIV and lung cancer alone. The CDC and the WHO does not follow them annually, if at all. Instead, the reports of incidence come through clinical study reports in primary and secondary sources. There is a shortage of primary reports assessing the prevalence of lung cancer in PWH. Despite this shortcoming, three completed studies had large sample size, a control group, and a prospective design. They showed the incidence of lung cancer significantly higher among PWH compared to the general population (Table 3.1). Three studies in Table
3.1 do not have a comparator but provide similar incidence rates per 100,000 person-years to those studies with a control group.

<table>
<thead>
<tr>
<th>Study</th>
<th>HIV +</th>
<th>HIV -</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sigel et al. 2012</strong></td>
<td>N=37,294</td>
<td>N=75,750</td>
</tr>
<tr>
<td></td>
<td>457 cases</td>
<td>614 cases</td>
</tr>
<tr>
<td></td>
<td>204 cases per 100,000 p-yrs</td>
<td>119 cases per 100,000 p-yrs</td>
</tr>
<tr>
<td></td>
<td>IRR=1.9</td>
<td></td>
</tr>
<tr>
<td><strong>Marcus et al. 2018</strong></td>
<td>N=24,768</td>
<td>N=257,600</td>
</tr>
<tr>
<td></td>
<td>89 cases</td>
<td>506 cases</td>
</tr>
<tr>
<td></td>
<td>66 cases per 100,000 p-yrs</td>
<td>33 cases per 100,000 p-yrs</td>
</tr>
<tr>
<td></td>
<td>IRR = 2.0</td>
<td></td>
</tr>
<tr>
<td><strong>Shiels et al. 2010</strong></td>
<td>N=740</td>
<td>N=1,423</td>
</tr>
<tr>
<td></td>
<td>13 cases</td>
<td>16 cases</td>
</tr>
<tr>
<td></td>
<td>IRR = 2.3*</td>
<td></td>
</tr>
<tr>
<td><strong>Makinson et al. 2020</strong></td>
<td>N=77,819</td>
<td></td>
</tr>
<tr>
<td></td>
<td>285 cases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>56 cases per 100,000 p-yrs</td>
<td></td>
</tr>
<tr>
<td><strong>Shebl et al. 2010</strong></td>
<td>N=322,675</td>
<td></td>
</tr>
<tr>
<td></td>
<td>853 cases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>82.6 cases per 100,000 p-yrs</td>
<td></td>
</tr>
<tr>
<td><strong>Guiguet et al. 2009</strong></td>
<td>N=52,278</td>
<td></td>
</tr>
<tr>
<td></td>
<td>207 cases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>81.7 cases per 100,000 p-yrs</td>
<td></td>
</tr>
</tbody>
</table>

These studies demonstrate a significantly higher prevalence of lung cancer in PWH over the past ten years and the significance of lung cancer in PWH compared to the general population; however, one study contradicts these claims. A small prospective study of 226 HIV+ people in Baltimore, Maryland, found only 1 case of lung cancer over 678 person-years. Hulbert et al. refute the small sample size as a possible cause. Instead, they argue the cohort's age was more likely the cause since the median age was 48-years old. In studies of lung cancer in PWH, there is debate over the age at diagnosis and stage at diagnosis.
The average age at lung cancer diagnosis is 71 in the general population. There is no single source reporting on the average age of lung cancer diagnosis in PWH as there is for the general population. Numerous comparative studies have powerfully shown that lung cancer in PWH occurs at significantly younger ages than the general population. Two cohort studies have shown that more than 60% of HIV+ were younger than 49 years of age\textsuperscript{104,105}. Furthermore, several other studies reported significantly different median ages in HIV+ as low as 50 and as high as 57 compared to median ages in HIV- controls as little as 61 and as high as 71\textsuperscript{7,97,100,106,107}.

On the other hand, there is a small body of evidence from Sigel et al. claiming there is no significant difference in age between PWH and the general population. Using two large cohorts from the Veterans Aging Cohort Study (VACS), Sigel reported two separate but extensive studies. In one study, the mean age was 46 in both the HIV+ and HIV- cohorts, while in the other study, the mean age 60 and 66 in the HIV+ and HIV- groups\textsuperscript{10,108}. In the former, the mean age of 46 correlates with the median ages seen in other studies; however, the latter study does not. It could be possible VACS is not a reflective sample of the population of PWH. Although, the third study from Sigel combining SEER and CMS data also showed an insignificant difference in age\textsuperscript{109}. In this study, 30% and 24% of the HIV+ and HIV- cohorts were younger than 50. Nonetheless, there is a more considerable body of evidence supporting the claim that lung cancer occurs in PWH at a younger age than the general population. The same is not valid for the stage of lung cancer diagnosis, where there is more of a divide between two schools of thought.

Lung cancer is an asymptomatic disease that leads to late-stage diagnosis (Stage III or IV). Some feel that lung cancer is present at a more advanced stage in PWH compared to the general population, while others think there are no differences. There is evidence to support
both but more evidence to support the former. Three large, comparative studies of HIV+ to HIV- have shown significant differences in stage with PWH have either more advanced stage lung cancer or distant cancer (cancer that has spread to other parts of the body). Seven other comparative studies have found the opposite, albeit two studies are from the same cohort. Likewise, all studies except one reported no significant differences in the histology of lung cancer between PWH and the general population. Across nine different comparative studies and six noncomparative studies, the most common histological subtype of lung cancer was adenocarcinoma, followed by squamous cell carcinoma. The significant difference in the one study was likely due to the large sample size since the distribution of stages was similar in both the HIV+ and HIV- groups.

The distributions of gender and race in lung cancer among PWH are similar to the distributions reported in each disease separately. Lung cancer in PWH is more prevalent among men, as HIV and lung cancer are more common in men than in women. Two comparative studies had almost double the number of men in the HIV+ group than the HIV- group insinuating that male PWH are at a higher risk for lung cancer than males in the general population. White and African American race was the most common in almost all of the studies with varying degrees of prevalence. Some studies reported African Americans being the most common race, while others reported White being the most common race. In most, but not all comparative studies, there were significant differences in race. In the study by Marcus et al., prevalence among Whites was 21% greater in the HIV+ group versus the HIV- group (74% and 53%), and African American was only 2% greater in the HIV+ group versus the HIV- group (17% and 15%). In the study by Sigel et al., White and African American race in the HIV+ group was almost identical (47% and 43%); however, there was a 28% difference in the African American race between the two groups.
Finally, in the study by Brock et al., 73% of the HIV+ group was African American, and there was a significant difference of 54.8% to the HIV- group.

3.3.3. Differences in Smoking Prevalence among PWH

The most significant and recognized risk factor for lung cancer is smoking\textsuperscript{25,30}. Smoking prevalence among PWH is particularly crucial since it is twice as high, or greater, as the general population. While the CDC reports the percentage of current smokers in the US has declined over the last thirteen years from 20.9% in 2005 to 13.7% in 2018, evidence has shown the same is not valid for PWH. Some large, comparative studies have shown significant differences as high as four times greater than the general population. In the study by Palacios et al., current smokers accounted for 96.7% of the HIV+ group compared to 57.7% of the HIV- group. Likewise, 90.2% and 40.8% in Brock et al. and 85% and 70% of HIV+ and HIV- groups in D’Jaen et al. were current smokers. Mdodo et al. showed the prevalence of smoking in its 419,945 PWH was twice as high as the general population (42%). Globally, the WHO estimated that 1.1 billion (15%) of the world smoked, and the global prevalence of smoking among PWH is similar in other countries to the US. A nationally representative study in France showed 62% of PWH were current smokers\textsuperscript{23}, and the Swiss HIV Cohort Study showed a significant difference between the HIV+ and HIV- group, 84.6% and 48.9\%. Smoking is a more substantial problem for PWH as it exacerbates the effects of smoking seen in the general population. The two seem to “synergistically work together” to increase the risk of mortality\textsuperscript{110}. An extensive study from Helleberg et al. demonstrated that PWH who smoke lose more years of life from smoking (12 years) than from HIV alone (5.1 years). Smoking is one of the primary risk factors for lung cancer among PWH. At the same time, many have shown several other factors, aside
from smoking, are significantly associated with increased risk of and death from lung cancer even after adjusting for smoking in multivariable models.

3.3.4. \textit{Worse prognosis of lung cancer among PWH}

Lung cancer is one of the most fatal cancers. Its prognosis in the general population of people without HIV is dismal, with one of the most reduced survival rates of all cancers. In the US, the National Cancer Institute (NCI) estimate on 5-year overall survival (OS) for lung cancer using data from 2010 to 2016 is approximately 56%. It is the second leading cause of cancer death among men and women, only preceded by breast cancer in females and prostate cancer in men. The same is not true for PWH, where lung cancer is the leading cause of cancer death and is also significantly more severe\textsuperscript{21, 108, 111}.

PWH and lung cancer do not live long enough to track and compare 5-year OS to the general population leading to a lack of long-term survival data\textsuperscript{23}. Despite this, some comparative studies have been able to contrast and report the significant differences between HIV+ and HIV-. Marcus et al. showed 10\% of HIV+ and 19\% of HIV- survived for five years. Similarly, both Sigel et al. and Smith et al. showed worse 5-year OS in the HIV+ groups than the HIV- groups with 9\% and 23\%, and 9.1\% and 17.9\% survival rates. Finally, Makinson et al. reported 16\% 5-year OS in a group of HIV positive with no control to compare. Despite the lack of 5-year OS data, several other measures have demonstrated disparities in mortality among PWH.

Several other temporal measurements of OS demonstrate the increased risk of death. For example, reports of 1-year, 2-year, and 3-year OS are 31\%, 10\%, and 4\% in some studies\textsuperscript{44, 99, 112}. Some used percentages of the population who died while others used median survival to describe mortality differences. For example, Sigel report 29\% and 13\% all-cause
mortality, and Pakkala reported a median survival of 4 months and 6.1 months in HIV+ and HIV- groups.

Overall survival measures the time from date of cancer diagnosis to death from any cause and is also known as ‘all-cause mortality.’ A more specific measure of mortality is ‘lung cancer-specific mortality,’ which measures the time from date of cancer diagnosis to death from lung cancer. Few studies reported lung cancer-specific mortality, but there were similarities across those that did. Marcus et al., Sigel et al., Suneja et al., and Coghill et al. reported significantly increased adjusted risks of lung cancer mortality in PWH, 1.3, 1.7, 1.25, 1.38, respectively. In contrast, Smith et al. did not describe a significant increase in lung cancer-specific mortality (1.08); however, all-cause mortality was significant (1.44). While D’Jaen et al. did not report on all-cause mortality or lung cancer-specific mortality, they did compare their median survival of 9 months to a registry-based control group of HIV negative and found no differences.

Among PWH, overall death represents a competing risk to lung cancer-specific death. PWH have many risk factors that increase the risk of overall death associated with them. For example, low CD4 cell count significantly increases the risk of mortality among PWH. In one study, the median all-cause mortality of PWH and low CD4 count was eight months compared to 40 months in the group with high CD4 count\textsuperscript{85,113}. It raises the question of whether PWH and low CD4 would have died from lung cancer if they hadn’t died from having low CD4 counts. The factors that increase the risk of developing or dying from lung cancer are different from those of the general population and need to be well understood.
3.3.5. Risk Factors for lung cancer specific to PWH

Several studies have reported immunologic risk factors specific to HIV independently increase the risk of lung cancer. HIV causes systemic inflammation and increases the risk of pulmonary infections and lung cancer. Smoking explains some of the risk of lung cancer in PWH, but these additional risk factors explain more.

3.3.5.1. Smoking

Smoking is the principal risk factor for lung cancer, not only in the general population but also in PWH. It is well established that PWH smoke more, 2–4 times or higher, than the general population. The increased prevalence of smoking in PWH puts them at a higher risk for lung cancer than the general population. A comparative study by Clifford et al. showed HIV+ current smokers had 14.4 times the risk of lung cancer compared to HIV- and an attenuated, but insignificant risk of 3.2 in former smokers. Another study similarly reported significant differences in SIRS comparing current and former smokers to never smokers in HIV+ to HIV-, 6.3, and 3.0, respectively. Even more persuasive is evidence from a study reporting an increased risk of mortality as high as 6.7 in smokers.

Smoking is a difficult factor to measure and report as there are smoking behaviors, specifically duration and severity (e.g., cigarettes per day, packs per day/year), which could affect the risk of lung cancer. Likewise, data sources usually do not collect data on all smoking behaviors, and the sources of data can be vastly different (e.g., Electronic Medical Records, Questionnaires). These differences taken together can impact the assessments of risk and limit or bias its generalizability.

Some have hypothesized that HIV infection and smoking have a “synergistic” effect or interaction that increases the risk of lung cancer. There is a lack of data.
supporting this theory, but the existing data seems to refute the argument. Clifford, Sigel, and Shiels et al. tested for interaction between HIV infection and smoking finding insignificant but attenuated results. The attenuation makes it plausible that interaction could occur, and more research is needed.

Aside from evidence that smoking independently increases the risk of and death from lung cancer in PWH, there is strong evidence that it does not explain all of the risks in PWH. Several studies have controlled the effects of smoking in multivariable analyses and found an increased risk for lung cancer. Increased risk ranges from as low as 1.7 to as high as 3.06 in comparative and non-comparative studies\textsuperscript{20, 103, 111, 113, 116}. Mortality is also increased by 30% and 50% in two separate studies after adjusting for smoking\textsuperscript{111}. These results are a stark comparison to the general population in which age and smoking are the principal risk factors and noted as the primary criteria for screening by the NLST and USPSTF. In PWH, smoking does not explain all of the risks of lung cancer. Instead, factors associated with HIV or its immunosuppression plays a more significant role.

3.3.5.2. HIV and Immunosuppression risk factors

Helper T-cells are among the most critical cells in our adaptive immune system and the principal target for HIV infection. Helper T-cells (CD4) allow B-cells to ingest and kill foreign pathogens, such as HIV, and trigger cytotoxic T-cells (CD8) to detect and kill other cells infected with HIV. There is an inverse relationship between HIV viral load and CD4 counts\textsuperscript{117, 118}. As HIV replicates, more CD4 cells are infected and killed, leading to a lower CD4/CD8 ratio. As a result, CD4 cell count, CD4/CD8 ratio, and HIV viral load are critical markers for immune system deterioration and are under investigation as risk factors for lung cancer.
The threshold for immune system deficiency is 500, and severe immune system deficiency is 200. The CDC set the globally accepted CD4 limit for AIDS as a CD4 cell count less than 200. There have been mixed reviews concerning the association between CD4 cell count, the incidence of lung cancer, and lung cancer mortality. While some have shown associations, others have not. They speculate that historically low sample sizes and lack of detailed patient-level information on CD4 cell count contribute to the lack of concordance\textsuperscript{5, 114}.

Nonetheless, there is debate over the point of estimation. CD4 cell counts fluctuate and are affected by other factors such as treatment compliance and comorbidities. These fluctuations have led to questions about whether the lowest CD4 cell count (nadir), the current CD4 cell count, or longitudinal effects of low CD4 cell count increase the risks of lung cancer. Some have shown the inability to return to a healthy immune system (CD4 > 500, immune reconstitution) is a risk factor of lung cancer\textsuperscript{97, 119}. Others have shown a previous diagnosis of AIDS, regardless of current CD4 count, increases the risk of lung cancer\textsuperscript{97, 101, 105}. Despite all of this, there is no consensus on the effect of CD4 cell count on lung cancer risk.

Most who have proven an association between CD4 cell count and increased risk of lung cancer have longitudinally used 200 as the threshold for comparison\textsuperscript{17, 18, 99, 104, 117, 120}. In studies by Brock et al. and Marcus et al., PWH with CD4 less than 200 had 2.64 and 2.7 times the risk of lung cancer. As CD4 cell counts decrease to 50, the risk of lung cancer increases inversely to 8.5 times in Brock's study. Similarly, Hooker et al. showed that those with CD4 less than 200 had a significantly different survival of 8 months compared to 40 months among those with CD4 greater than 200.
Single point estimations such as nadir and current CD4 cell counts have been disproven, leading to broader acceptance of longitudinal values as predictors of risk for lung cancer\textsuperscript{10, 99, 119}. While it may be true that low CD4 cell count increases the risk of lung cancer, some do not agree. There is a body of evidence, using notable cohorts such as the Swiss HIV Cohort, that show CD4 cell count may be attenuated at lower levels but are not significantly different\textsuperscript{9, 18, 96, 97, 103, 121}. While there may be disagreement over the association between CD4 cell count and lung cancer risk, there is greater concordance on the effect of a more novel marker, CD4/CD8 ratio.

Immunocompetent people generally have higher CD4 cell counts than CD8 and the expected ratio of CD4/CD8 ranges between 1 and 4. In contrast to the general population, PWH have significantly lower CD4/CD8 ratios. There is an established relationship between a low CD4/CD8 rate to an increased risk of lung cancer\textsuperscript{5, 96, 122, 123}. In a comparative study of lung cancer in HIV+ to HIV- controls by Clifford et al., the risk of lung cancer in the group of HIV+ with CD4/CD8 ratio less than 0.25 was increased 2.15 times compared to the HIV- group. Similarly, but in a more extensive comparative study, Sigel et al. prove that both CD4/CD8 ratio < 0.4 and CD4/CD8 ratio between 0.4 to 1 were significantly different in HIV+ lung cancer cases compared to HIV- lung cancer cases, 2.6 and 1.9. Sigel concluded that cumulative exposure to low CD4/CD8 ratio was the "most robust independent immunological predictor of increased risk" of lung cancer independent of CD4 alone.

While CD4 cell count and the CD4/CD8 ratio are significant immunological risk factors for lung cancer, HIV viral load is not. Multiple studies have reported attenuated but insignificant differences in HIV viral load\textsuperscript{96, 103, 104}. One study, Sigel et al., contradicts this claim indicating significantly higher HIV viral loads lagged by 12 months (prior period) are
associated with increased risk for lung cancer. Viral load data is absent in older, pre-ART studies and lacking post-ART studies\textsuperscript{114}. Based on the magnitude of risk, HIV viral load should nonetheless be considered a risk for lung cancer among PWH.

The biological mechanisms for immunosuppression caused by HIV are independent risk factors that increase the risk of lung cancer. CD4 cell count, CD4/CD8 ratio, and HIV viral load independently increase the risk of lung cancer incidence and mortality. Several studies have proven that HIV infection is also an independent risk factor for lung cancer. After adjusting for all known risk factors of lung cancer, including smoking, HIV infection increased the risk of lung cancer, ranging from 1.7 to 3.8 times\textsuperscript{7, 10, 21}. HIV infection is also independently associated with pulmonary inflammatory disease, another set of independent risk factors for lung cancer in PWH.

3.3.5.3. Chronic Inflammation and Pulmonary Infections

HIV is a chronic inflammatory disease. Many believe inflammation causes the comorbidities observed in PWH who have a higher prevalence of cancers and a broad spectrum of conditions, including cardiovascular, neurocognitive, liver, kidney disease, and osteoporosis\textsuperscript{124}. Like cancer, these diseases occur at older ages in the general population due to progressive inflammation (“Inflammaging”)\textsuperscript{125}. These facts have led many to believe PWH are aging at an accelerated rate due to the chronic inflammation caused by HIV infection.

Inflammation is a complex, non-specific immune response to injury or infection. During an inflammatory response, chemicals, such as chemokines and cytokines are released into the bloodstream to direct immune cells to fight the injury or infection. Acute immune responses subside quickly, but chronic immune responses last for a more prolonged duration. These periods exhaust the immune system causing immune dysfunction and eventually end-organ disease\textsuperscript{124}. HIV is a chronic infection that causes persistent
inflammation. Even with undetectable viral load levels, people believe HIV remains in the body and perpetuates inflammation and disease.

HIV causes an immediate inflammatory response at conception. As CD4 cells are infected, they die mostly through pyroptosis, releasing cytokines into the body and perpetuating inflammation\textsuperscript{124}. A collection of evidence is growing to suggest a 'model' of chronic inflammation brought on by HIV\textsuperscript{124}. In this model, the continuous replication of HIV sustains inflammation and triggers permanent T-cell activation. Additionally, persistent immunodeficiency inhibits infection from other pathogens (e.g., HBV, HCV, etc.), and inflammation-induced deterioration of stomach mucosa further supports inflammation\textsuperscript{94, 115, 124}. HIV infection is systemic, and all cells and organs have the potential to be infected and the potential to be harmed by inflammation. HIV found in lung cells, and pulmonary lung inflammation is the root cause of COPD and pneumonia, two of lung cancer's key risk factors\textsuperscript{118, 126}.

COPD is a set of diseases that cause inflammation in the lungs' lining and make it difficult to breathe. It includes emphysema (damaged alveoli in the bronchioles) and chronic bronchitis (inflamed bronchiole tubes); however, other pulmonary diseases that fit the criteria, such as asthma, are sometimes included. In the US, COPD was reported in 15.7 million in 2014 and was the third leading cause of death. It is currently the 5th leading cause of death globally. Smoking and older age are associated with increased incidence in the general population\textsuperscript{94}. Comparative studies have proven HIV an independent risk factor for COPD, independent of smoking, and COPD is significantly associated with increased risk of lung cancer\textsuperscript{89, 92, 93}. For example, in the comparative study by Sigel et al., COPD was significantly associated with a 1.9 increased risk of lung cancer in PWH and lung cancer compared to those with HIV only. Some have theorized that smoking interacts with the
relationship between COPD and lung cancer risk\textsuperscript{54}, but it was insignificant in comparative studies testing for it\textsuperscript{10}.

Pneumonia is an infectious pulmonary disease caused either by bacteria or fungus. The most common form of pneumonia is bacterial pneumonia and pneumocystis pneumonia, which is fungal\textsuperscript{89}. In the pre-ART era, infectious lung diseases such as pneumonia were considered the ‘hallmark’ or universal symptom for AIDS\textsuperscript{89}. Today, they are still quite common, but the prevalence has drastically changed since then. PWH can be diagnosed with pneumonia multiple times a year, and recurrent bacterial and fungal pneumonia is associated with a 1.63 increased risk of lung cancer and an even stronger association, 1.99, in those younger than 50 years of age\textsuperscript{126}. Some have claimed lung cancer causes pneumonia, and the association is causal; however, both Shebl and Marcus et al. refute this. For example, Shebl tested for the association between pneumonia and lung cancer risk using lagged data 5-10 years after pneumonia diagnosis but before lung cancer diagnosis. In this case, the risk for lung cancer is two times greater, 2.41, thereby disproving reverse causation theories\textsuperscript{18,126}.

Bacterial pneumonia has a higher prevalence in PWH than the general population and has been proven a significant risk factor for lung cancer. In one comparison study, Fitzpatrick et al. reported a significant difference in the prevalence of bacterial pneumonia between HIV+ and HIV- women, 8.5 per 100 person-years, and 0.7 per 100 person-years. Likewise, in another comparative study, Sigel et al. reported a history of bacterial pneumonia significantly increased the risk of lung cancer by 1.5 in those with HIV and lung cancer compared to those with HIV but without lung cancer. While Marcus et al. had similar results, their definition of recurrent pneumonia includes bacterial pneumonia and pneumocystis pneumonia\textsuperscript{18}. Bacterial pneumonia as a risk factor for lung cancer is broadly
accepted, despite the disagreement regarding a definition of pneumonia. It is also widely accepted that low CD4 cell count, ART, viral hepatitis infection, and cigarette smoking increase the risks of bacterial pneumonia\textsuperscript{89}.

3.3.6. Disparities in treating PWH for lung cancer

The treatments of lung cancer in PWH are the same as those for the general population, but there are disparities. PWH are less likely to receive treatment and experience worse outcomes. There is neither a set of guidelines nor an optimal treatment strategy for lung cancer in PWH. The reasons for these disparities are not fully known and have led to a lack of data\textsuperscript{97, 99, 108}.

The standard of care for lung cancer treatment in the general population varies by stage. The typical recommendation for early-stage lung malignancies is tumor resection. Other localized treatment includes radiological therapy. The usual advice for late-stage lung cancer that has spread to nearby lymph nodes or has metastasized is systemic treatment using chemotherapy. The general population's guidelines are assumed to apply to PWH with similar survival outcomes\textsuperscript{112, 127}. This assumption is much disputed because HIV infection is considered an exclusion criterion in 20\% of lung cancer clinical trials\textsuperscript{37, 44, 128} but also because empirical evidence has contradicted it. Tumor resection in PWH is more invasive than simple lumpectomy. The more common procedures, lobectomy, and segmentectomy remove more substantial parts of the lungs and tissue\textsuperscript{95}. Combination therapy, including localized and systemic treatment, is the preferred recommendation\textsuperscript{97, 100, 128}.

While the procedures for lung cancer in PWH are the same, there are disparities in their application. PWH are less likely to be treated, less likely to receive the standard of care, and have worse outcomes than the general population. Overall, PWH disproportionately receive treatment compared to the general population. A comparative study by Brock et al.
reported a significant difference in treatment between HIV+ and HIV-, 14.1%, and 27%, while Suneja et al. reported similar results, 13.6% and 32.7%, 2.5 increased odds for no treatment. Marcus et al. also demonstrated a significant difference in treatment between HIV positive and HIV negative albeit the magnitude in proportions was higher than Brock and Suneja, 64% and 76%. Evidence also suggests a reluctance towards surgical resection. Three studies indicated radiation is more common as a local treatment over surgery, and chemotherapy is given adjunctly\textsuperscript{97,100,104}. Even more surprising is the number of PWH receiving no treatment. In a study by D’Jaen et al., 17% of HIV+ received no therapy. In comparison, 24% received no treatment in the study by Brock et al. Additionally, disparities among younger age (45 and 64 years), low CD4 count (less than 200), and race (African Americans) increase the risk of not being treated\textsuperscript{128}. Those who become treated experience worse outcomes compared to the general population. Overall survival is worse for HIV who are treated compared to HIV- people. One study reported a significant difference of 26 and 48 months between HIV+ and HIV- and 8 and 40 months for those with low CD4 cell counts\textsuperscript{113}. Likewise, PWH experience more complications, more treatment-emergent adverse events, and shorter progression-free survival. While many have proven the treatment disparities between PWH and the general population, a few disagree with the most notable from Sigel et al.\textsuperscript{37,108}. Using a sizeable comparative cohort, Sigel et al. reported that PWH were equally likely to be treated, specifically with surgical resection and no differences in postoperative complications or treatment-emergent adverse events.

The reasons for these disparities are not well understood. Some have argued that advanced stage at diagnosis and socioeconomic or demographic differences were principal factors\textsuperscript{95,115}. However, a piece of broader evidence points to problems with the clinicians providing care. Clinicians do not understand or have experience managing PWH and lung
cancer. They do not understand the additional risk factors that drive lung cancer in PWH, nor do they know the differences in lung cancer's pathogenesis. They have failed to screen for lung cancer, assuming infections and inflammation brought on by HIV cause the signs and symptoms. They also expect there will be a lack of tolerance for treatments because of the possible interaction with existing ART. Most importantly, due to the lack of guidelines for treating and screening PWH for lung cancer, clinicians do not feel comfortable or knowledgeable enough to progress screening and treatments appropriately. The disparities in treatment and worse outcomes in PWH have placed an increased emphasis on early detection of lung cancer.

3.3.7. Challenges in Screening, Early Detection, and Prevention

The screening guidelines established in the NLST and adopted by the USPSTF and CMS do not apply to PWH. First, PWH were excluded from the NLST. Second, several studies have demonstrated that the existing guidelines miss inordinate amounts of PWH, 60% to 70%, who develop lung cancer at stages early enough to improve mortality. PWH are diagnosed with lung cancer at younger ages and more advanced stages. Lowering the age and pack-years of smoking thresholds have been proven to include more lung cancer cases in both simulation and registry-based studies; however, there are questions around the feasibility of LDCT in PWH and the harms associated with it.

The data and evidence showing the feasibility of LDCT in PWH is limited, but some have proven it be plausible. The most notable results come from Makinson et al. using a large, prospective study. LDCT detected findings in 21% of the cohort with 10 out of 15 invasively explored nodules being diagnosed with lung cancer among younger ages (less than 55) and less pack-years of smoking (less than 30 pack-years). There were no significant adverse events associated with the invasive procedure.
The most considerable concern with LDCT is the number of false positives and the hypothetical theory that LDCT will lead to more false positives in PWH than in the general population\textsuperscript{2,37,44}. This theory was refuted by Molto et al., citing a comparative VACS study that reported no significant differences in abnormal findings between HIV+ and HIV-. The only evidence to support the theory of increased abnormalities from LDCT stems from simulation studies that lack support and are not empirically correct. A simulation study using an HIV+ cohort with a high CD4 cell count proved that LDCT could be useful in PWH between 45 and 77 years of age and 20 pack-years\textsuperscript{4}. The study showed it possible to identify more screen-eligible cases and avoid more deaths. However, it would require more LDCT examinations and lead to more significant potential for overdiagnosis\textsuperscript{4}. Some feel that additional LDCT examinations, beginning at a younger age, will overexpose PWH to dangerous amounts of radiation, which could further increase cancer risk. This theory, too, has been refuted by several who claim LDCT uses 80% less radiation than other technologies and that PWH has significantly higher mortality and likely won’t live long enough to experience adverse effects from cumulative radiation exposure\textsuperscript{1,2,4}. The lack of evidence supporting these theories and existing, albeit sparse, evidence supporting the benefit of LDCT has led to broad acceptance of LDCT as the recommended screening tool.

The more significant issue is the disparity in the number of eligible people screened. Studies have shown significantly small amounts of PWH being screened for lung cancer, even when they are eligible to be screened according to existing guidelines. In one study, 13% were eligible to be screened with LDCT, but only 3.7% screened, while a similar study showed 14.3% eligible to screen, and only 2.7% screened\textsuperscript{45,130}. This disparity has brought attention to preventive measures until screening guidelines for PWH exist. Smoking cessation is seemingly the most important preventative measure that can begin immediately.
Even though smoking does not explain all of the excess risks of lung cancer, PWH smokes more and lose more life-years to smoking than HIV. As a result, smoking cessation is an essential measure to aide in preventing lung cancer.\textsuperscript{8, 101, 103, 109} Similarly, ART initiative is also considered a preventive measure due to the association between CD4 cell count and increased risk of lung cancer.

It is critically important to enhance the screening guidelines for lung cancer to extend to PWH. Age and pack-years of smoking are different in PWH. It is also clear that additional risk factors for lung cancer exist in PWH that do not exist in the general population. These differences have led to increased interest in identifying high-risk PWH who can benefit the most from lung cancer screening\textsuperscript{6, 8, 14, 44, 108}. Predicting the risk of lung cancer and clinically deciding patients at risk for lung cancer is an urgent, unmet need.

3.3.8. The Absence of Lung Cancer Risk Assessment Tools for PWH

PWH are being diagnosed with lung cancer at younger ages and at advanced stages. It is proven the risks for lung cancer are different in PWH than the general population and that smoking does not explain all of the risk. PWH are less likely to be treated and more likely to experience worse outcomes from lung cancer such as post-operative complications from lung cancer screening and death. There is an absence of CDSS and risk assessment tools to facilitate risk-based identification of PWH but there is an urgent call for them by many.

LDCT is feasible among PWH and can lead to the identification of early stage lung cancer if patients at high risk are identified\textsuperscript{17}. The body of evidence for lung cancer has demonstrated that there is a trajectory towards person-specific risk identification instead of risk based on broad categories and the same is needed for PWH\textsuperscript{2}. It is clear not only that a tool is needed, but also that ‘high-risk’ in PWH is different than the general population. For
PWH, high-risk is defined not just by known risk factors, such as age and smoking, but also by emerging immunologic risk-factors specific to HIV including CD4 cell count, CD4/CD8 ratio, COPD, and pneumonia\textsuperscript{5, 23}.

The purpose of this study will be to determine which risk factors which can be used to create a risk score to identify PWH who are at high risk for lung cancer and use the risk score to develop and validate a CDSS for prevention and earlier diagnosis of lung cancer in PWH. There is sufficient evidence to suggest HIV specific risk factors which describe the excess risk of lung cancer not explained by smoking alone. This study will address the unmet need for a risk score that can aide in the earlier identification of PWH who are at high-risk of lung cancer.
CHAPTER 4

RESEARCH METHODOLOGY

4.1 Introduction

In this study, publicly available data integrated between the Multicenter AIDS Cohort Study (MACS) and the Women’s Interagency HIV Study (WIHS) determined risk factors predictive of lung cancer and created a risk score algorithm to assess the risk of lung cancer in PWH. The CDSS used this risk score to develop a lung cancer risk assessment tool that will prevent lung cancer by identifying existing risks and lead to the earlier diagnosis of lung cancer in PWH. It is the first lung cancer risk assessment tool to consider HIV specific risk factors. It identified more PWH at risk than the existing NLST guidance. PWH live longer lives but are suffering and dying from illnesses unseen in the earlier days of the HIV epidemic. Cancer is the leading cause of death among PWH, and lung cancer is the leading cause of cancer-related death in PWH. Existing guidance for lung cancer screening does not apply to PWH and misses significant amounts of cases when applied to a cohort of PWH. As a result, PWH are diagnosed with and dying from cancer at younger ages and more advanced stages than the general population.

The research in this study draws from both quantitative and quasi-experimental methodology. Data from MACS and WIHS were quantitatively analyzed using bivariable and multivariable analyses. Prospective collection of MACS and WIHS data occurred over a specific period. The natural development of outcomes, such as people with lung cancer,
made this a natural quasi-experiment with no control over results and group assignments. Both the quantitative and quasi-experimental approach to this study supplemented the existing knowledge found in the historical data. Together, it formed the knowledge base of the CDSS making the tool pragmatic and generalizable to the larger population of PWH.

This chapter aims to explain the methodological approach used to achieve the goals of the study. It begins by describing the analysis conducted on curated data from Johns Hopkins University and explaining the creation of the knowledge base and the development of the inference engine in the CDSS. Screenshots of the inference engine's logic and full-text code extensively explain how the logic assesses risk factors and create a risk score for the end-user. This chapter ultimately explains the steps taken to identify significant risk factors for lung cancer in PWH, generate a risk score, and then implement the risk score in a CDSS.

4.2 Research Design

This study's primary goal is to create a risk score to assess lung cancer risk in PWH and use the risk score in a CDSS to identify PWH at risk for lung cancer leading to earlier diagnosis and prevention. This study obtained public-use data sets (PDS) from Johns Hopkins University on a cohort generalizable to the larger population of PWH to develop the risk score algorithm. Analytical results combined with historical data formed the knowledge base. Next, analytic results informed the design of the rules and conditions to be tested by the inference engine. Using Corvid, input and output mechanisms were created and tested against the PDS. Finally, this study used a random sample of subjects reserved from the PDS to validate the CDSS.
4.2.1 Data Analysis

Two separate samples of data from Johns Hopkins University were used to form the knowledge base and create the risk score. Data from the Multicenter AIDS Cohort Study (MACS) and the Women’s Interagency HIV Study (WIHS) were integrated, analyzed, and formed the knowledge base of the CDSS. 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\textsuperscript{131}. 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 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\textsuperscript{131}. 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 used in this study included 12,320 men and women. 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. 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.

A series of predictive models were built using a lung cancer diagnosis, a binary response variable (1 = Lung Cancer Diagnosis, 0 = No Long Cancer Diagnosis), as the primary endpoint. The MACS cohort contained variables to identify cancer cases using the International Classification of Disease for Oncology, 3\textsuperscript{rd} 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 codes for the reporting of cancers. Lung cancer cases were mined over each set of data and combined horizontally to derive one record per subject and identify lung cancer cases. It was impossible 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.

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
- HIV RNA viral load
All analyses utilized the statistical software SAS, version 9.4. This study used logistic regression models to predict the outcome (lung cancer) from a set of predictor variables (risk factors) that built the risk score. Logistic regression is the appropriate model since the primary endpoint, a lung cancer diagnosis, is a binary outcome. The study stratified the analytical data set 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. An existing risk model for the general population created by Spitz et al. stratified predictive models by smoking status with good sensitivity\textsuperscript{132}; however, there are significant differences in smoking behaviors between the general population and PWH. Stratifying predictive models by smoking alone will not provide the same value in PWH as it does in the general population. In the analytical dataset, the incidence of smoking in PWH with lung cancer is more than four times the incidence of smoking historically reported in the general population. Additionally, several studies have reported differences in gender and race distribution in PWH compared to the general population. As a result, this study looked independently at gender (Female, Male), race (White, Black/Other), and smoking status (Smoker, Non-Smoker) and created a risk score from each.

Continuous variables collected longitudinally were analyzed using the last visit on or immediately before lung cancer diagnosis for those with lung cancer or the last visit in the study for those without lung cancer. 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. LOCF is a method for imputing missing data, which takes the last known value of a variable before a missing value and brings it forward until the next known value occurs. Continuous variables were converted into binary indicator variables using the quartiles from those with lung cancer. For example, the first quartile of age representing subjects younger than 32 years of age is reflected in the binary indicator variable 'AGE Q1' where a value of "1" indicates subjects younger than age 32, and a value of "0" indicates subjects not younger than 32. 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. For example, four levels of education (No degree, High School, College, Graduate School) were derived into four binary indicator variables (e.g., NO DEGREE = 1 for subjects with no degree). The use of indicator variables made it easier to interpret in both the predictive models and the CDSS. 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 0.25 alpha using the Wald Chi-Square ($\chi^2$) test in the bivariable models were selected as a candidate for the multivariable models. The p-value cut-off of 0.25 is widely accepted for identifying candidate predictors since the traditional value of 0.05 could erroneously exclude essential predictors.133

For each of the six logistic regression models, the candidate predictor variables significant at the 0.25 alpha level were added to multivariable models using a forward selection method with an entry criterion at the 0.1 alpha level. Forward selection is a variable selection technique in regression analysis used to select the optimal model. The method assesses multiple regression models based on the specified predictors. Forward selection starts by adding the single most influential predictor variable (‘regressor’) and adds the next
best predictor variable. The process continues until the evaluation of all variables is complete, or some criterion is achieved such as the 0.1 alpha criterion for entry. Backward elimination is the opposite of forward selection in that it evaluates the model using all specified predictors and removes the worst predictor from the model. The process is repeated until the optimal model is found. Stepwise selection is a combination of both forward selection and backward elimination. It begins by adding the single best predictor in the model and continues adding the next best predictor; however, each iteration re-evaluates the model as a whole. Predictors that become insignificant in later iterations are removed. 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 a large number of 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. Several iterations of the analysis led to the final six independent models and final variable selection to build a risk score and create the CDSS (Table 4.1).

Table 4.1: Confidence Variables using Odds of Lung Cancer from Multivariable Logistic Regression

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Gender</th>
<th>Race</th>
<th>Smoking Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Black/Other</td>
</tr>
<tr>
<td>Demographics &amp; Socio Demographic Information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Hispanic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE Q4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking Information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GT 1 PD (Pack/Day)</td>
<td>2.2347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk Factor</td>
<td>Gender</td>
<td>Race</td>
<td>Smoking Status</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>--------</td>
<td>------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Black/Other</td>
</tr>
<tr>
<td>Years Smoked Q3</td>
<td>2.4130</td>
<td>1.7144</td>
<td>2.0738</td>
</tr>
<tr>
<td>Years Smoked Q4</td>
<td>3.3921</td>
<td>2.7124</td>
<td>2.0309</td>
</tr>
<tr>
<td><strong>Respiratory Disease</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reoccuring Pneumonia, History Of</td>
<td>1.9384</td>
<td></td>
<td>1.7743  4.3792</td>
</tr>
<tr>
<td>Respiratory Disease, History Of</td>
<td>2.3726</td>
<td>2.3858</td>
<td>2.1263  2.0962  2.2202</td>
</tr>
<tr>
<td><strong>Clinical HIV Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIDS Ever</td>
<td></td>
<td></td>
<td>2.5451</td>
</tr>
<tr>
<td>CD4 Fluctuations, Over Time</td>
<td></td>
<td></td>
<td>2.7450  1.8771</td>
</tr>
<tr>
<td>CD4 Q2</td>
<td>2.2644</td>
<td>1.8980</td>
<td>2.9081  2.3360</td>
</tr>
<tr>
<td>CD4 Q3</td>
<td>2.7000</td>
<td></td>
<td>4.1657  2.5109</td>
</tr>
<tr>
<td>CD4/CD8 Q1</td>
<td>2.2894</td>
<td>2.1489</td>
<td>1.9665  2.4599</td>
</tr>
<tr>
<td>Viral Load GT 500</td>
<td></td>
<td></td>
<td>9.9413</td>
</tr>
</tbody>
</table>

The odds ratios from the six final models served as the probability for lung cancer reflected in the confidence variables in the CDSS. The confidence will be multiplicative for a specific risk factor and account for eight ($2^3$) possible models (e.g., Male, White, Smokers vs. Male, White, Non-Smokers, etc.). A single risk score is the sum of all risk factors and is determined logically by the inference engine using a conditional workflow (Figure 4.1). The CDSS workflow is the sequence of questions used by the inference engine to derive a risk score and assess lung cancer risk. It uses the six multivariable logistic regression models' outcomes, identified in Table 4.1, to conditionally determine the sequence of questions.
Figure 4.1: CDSS Workflow.
The sequence of questions used by the inference engine to derive a risk score and assess lung cancer risk.
4.3 CDSS

The American Health Information Management Association (AHIMA) defines CDSS as a computer system that provides “clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care”\(^3\). A more straightforward definition is that a CDSS is a computer system that facilitates clinical decision-making. A knowledge-based CDSS, such as the one created in this study, contains a communication mechanism to capture user input and provide user output, a knowledge base, and a reasoning mechanism known as the inference engine.

In this study, the development of the CDSS relied on the software development lifecycle (SDLC), including planning, analysis, design, implementation, and testing. The future expectation is that this CDSS will be integrated into electronic medical records (EMRs) and publicly available for PWH as a lung cancer risk assessment. A CDSS introduces complexity to the SDLC due to the knowledge-base and inference engine's addition to the overall software\(^8\). According to Richter et al., the knowledge-base and the inference engine should be treated separately from the software and follow most, if not all, of the SDLC. The software used, Exsys Corvid ("Corvid"), is a validated software for the development of CDSS and did not require the SDLC process. Corvid is an expert system in developing CDSS that quickly and easily creates a knowledge-base, inference engine, and communication tool. Corvid is a highly interoperable system since it allows effortless access, exchange, and integration to other systems. It is also simple to use and requires no programming background to develop the knowledge-base, inference engine, or communication tool. The communication tool for interactions to and from the end-user uses a default mechanism known as the "Corvid Servlet Runtime program." Using this default
mechanism, Corvid will build the CDSS automatically as an HTML page with Java servlets using the knowledge-base and inference engine. This process allows the CDSS to be run by the end-user on the end-user's machine as long as they can access HTML pages. Using the expert CDSS software Corvid, the overall development process for the CDSS in this study is described in three phases (Figure 4.2).

**Figure 4.2: Development for process for the CDSS**

![Diagram of development process for the CDSS]

First, this study identified problems and gaps in the research methodology for lung cancer in PWH. PWH have a higher incidence and mortality from lung cancer compared to the general population. Existing guidelines for the screening and early detection of lung cancer in the general population do not apply to PWH. They exclude significant amounts of lung cancer cases when they are applied. The study's hypotheses addressed these problem areas and curated data to support the study hypothesis. The historical data and the accessibility to the MACS/WIHS data informed the data analysis. The next phase of CDSS development used risk factors for lung cancer specific to PWH. Second, the data was mined, normalized, and analyzed to direct the design and implementation of the knowledge-base and inference engine in the CDSS. Third, historical data and the curated, analyzed data created the knowledge-base. It informed the design and implementation of the boolean
(if/then) rules and conditions that drive the inference engine. This last phase of the CDSS development occurred iteratively until the final CDSS was built and tested.

The formation of the knowledge base prompted the design and implementation of the inference engine. Through bivariable analysis, familiar and novel risk factors were identified as candidate predictors for the multivariable models. The final predictors from the multivariable models formed the knowledge base that instructed the inference engine's design. The first step in designing the inference engine is understanding the workflow for the rules and conditions to be asked by Corvid to the end-user to predict the likelihood of lung cancer. Table 4.3 provided the parameters needed by the strata to predict the likelihood of a lung cancer diagnosis in the form of a risk score and achieve this study's goal.

The purpose of the CDSS in this study is to derive a risk score, though the risk score originates from multiple risk factors stemming from numerous conditions. Suppose the end-user is a black, smoking male with a CD4 count between 78 and 250. The inference engine determines the questions to ask the end-user to calculate the probability of lung cancer and create a risk score. The CDSS will need to know the end-users gender, race, and smoking status to ask the correct questions for CD4. The rules and conditions for each of the risk factors involved in all of the models described in Table 4.3 create the multiplicative probabilities for each risk factor and sums them to derive a single risk score. Corvid allows for this quite easily. Logic blocks using forward chaining will move the system from one question to the next, while logic blocks using backward chaining will capture data as and if it is needed.

The modularly organized workflow fully describes the boolean logic of the inference engine. The modularity of the workflow was the catalyst for the organization of the logic blocks. There is a one-to-one match between workflow modules and the CDSS logic blocks.
The workflow describes the inference engine's data-driven forward chaining process; however, goal-driven backward chaining plays an essential role in the inference engine. Variables describe all logic blocks. For simplicity, the workflow does not represent all variables and backward chaining logic.

4.3.1 *Corvid Variables*

Corvid variables are the key elements of the CDSS. They are used to describe and direct rules and conditions of the logic and command blocks. Variable conventions in Corvid are similar to conventions in other programming languages. There are seven variable types provided by Corvid; however, the CDSS designed in the study only used four: numeric, static lists, confidence, and collection. As an example, quartiles of age are one of the risk factors used in the risk score. Numeric variables capture continuous variables from the end-user, such as age. While the end-user is asked, 'What is your age?,' the inference engine intelligently and automatically derive the quartiles of age, 'AGE_Q,' through backward chaining without asking the end-user. Static list variables are categorical, 'multiple choice' variables. The backward chained variable 'AGE_Q' is a static list with only two possible values required for the risk score: 'Q4,' and 'N/a'.

Corvid's confidence variables reflect the probability that an event will happen and are set by rules in logic blocks. The inference engine never asks the end-user for the value of a confidence variable. They can be combined in many ways, but this study only used multiplication. To derive each risk factor's confidence variable across gender, race, and smoking status, the inference engine multiplied confidence variables while a single, confidence variable summed all multiplicative risk factor confidence variables to create a risk score. Confidence variables in Corvid made it easy to implement the multiplicative probabilities the system needed. For example, females in the 4th quartile of Age have a 1.73
increased risk of a lung cancer diagnosis. The probabilities from table 4.3 are the values for the confidence variables used in the CDSS. A female's value for 'EST_AGE_Q4' is set to 1.73 by the inference engine, asking the end-user what their gender and age are.

Collection variables in Corvid are also known as 'Report' variables because they are used to report information back to the end-user. They are referred to as 'collection' variables because they are a list of text values accumulated throughout the CDSS based on the inference engine's rules. For example, smoking cessation guidance information is provided to smokers when they reach the questions' end. The collection variable, 'Notes,' accumulates values each time the inference engine recognized it should collect information to report, while the collection variable 'Risk Notes' will describe calculated probabilities. In total, the CDSS required 47 variables, which provides the user with a positive experience of the CDSS and accomplishes the goal in as few steps as possible.

The knowledge-base developed from the analysis of the MACS and WIHS informed key variable definitions and thresholds. 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 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.

The quartiles' thresholds are informed by the knowledge-base and use the lung cancer cohort. Not all quartiles were significant predictors for lung cancer, and only those that were significant, by race, gender, or smoking status, were included in the model. For
example, the probability of lung cancer is predicted for years smoked when years smoked is either between 21 and 28 (21 < years smoked <= 28) or greater than 28 (years smoked > 28), the 3rd and 4th quartiles, respectively. Likewise, only the fourth quartile of age (age > 44) is a significant predictor of lung cancer.

The risk score is the sum of all probabilities; however, each probability is multiplicative across gender, race, and smoking status. For example, a female, black, smoker who smokes two or more packs of cigarettes per day and a CD4 cell count in the second quartile (78 < CD4 <= 250) has the following risk predictions:

\[
\begin{align*}
\text{EST\_PACKS\_DAY\_GT1} &= 2.23 \times 1.79 = 3.99 \\
\text{EST\_CD4\_Q2} &= 2.26 \times 2.91 \times 2.34 = 15.39 \\
\text{RISK} &= 19.38
\end{align*}
\]

The inference engine calculates the risk based on the gender, race, and smoking status of the end-user and the rules and conditions assigned to each risk factor. The risk threshold is calculated using the quartiles of the risk score informed from the knowledge-base. The quartiles of risk are 14.8, 25.2, and 39.1, with a minimum score of 0 and a maximum score of 89.66. A risk score in the first quartile is considered low risk. A risk score within the second and third quartiles is regarded as a medium risk, and a risk score in the fourth quartile is viewed as high risk. These thresholds of risk provide a sensitivity of 77% and a specificity of 60%. This is comparable to prominent and well-known CDSS such as MYCIN.

4.3.2 Forward and Backward Chaining Logic Blocks

The rules and conditions that make up the inference engine in Corvid are defined in logically related blocks using boolean logic (True or False) consisting of "IF/THEN" rules. Humans use boolean logic in everyday decision making; however, in a different context. The inference engine of a clinical decision support system, such as this study, uses similar rules to
aid clinical decision-making. For example, "If age is greater than 44, then the probability of lung cancer is 1.73". In this example, when the "IF" condition is true, and the person is greater than 44, "THEN" set the probability. It is important to note if the rule is false, the action taken ("THEN") is not always false. In that case, different rules test for other conditions and take further steps.

The boolean, IF/THEN rules of Corvid can be several to many individual rules and logically grouped. Logically grouped rules are referred to as "tree diagrams," as seen in Figure 4.3 below. The 'tree' is simply a way to logically view the code, a benefit to using Corvid.

Logic is defined textually as "full-text" with the step-by-step resolution of rules for each action taken. Regardless, the complete solution of each 'IF' and 'THEN' rule of Corvid's logic is referred to as a 'node.'

Each 'IF' component's boolean test is a node, while each action taken by the 'THEN' part is a node. For example, the full-text rule below is an excerpt from the screening logic block in Figure 4.3 and excludes subjects that should not be using the CDSS. In this full-text example, two nodes are created, one from the boolean "IF" condition and another from the "THEN" part of the rule setting the variable's value.
IF: HIV_POSITIVE (Are you HIV Positive?) = No
THEN: SCREENING_CONFIDENCE (Screening Criteria Met?): Confidence = 0

Not all rules are as simple as a single "IF" and a single "THEN" rule taking up two nodes. Much of the logic in this study's CDSS required more complex rules and used more nodes. For example, the full text below is the complete rule to test for meeting the screening criteria. Four nodes were used for the "IF" part of the logic and one used for the "THEN" part of the logic. In total, this one particular rule required five nodes to complete in the inference engine.

IF: HOME_PAGE ('Would you like to Continue?') = Yes AND
    HIV_POSITIVE ('Are you HIV Positive?') = Yes AND
    LUNG_CANCER_NOW ('Do you currently have lung cancer?') = No AND
    LUNG_CANCER_EVER ('Have you ever had lung cancer?') = No
THEN: Screening Criteria Met: Confidence = 1

Variables are critical to the design and implementation of rules in Corvid. Corvid supports data-driven and goal-driven logic known as forward and backward chaining, respectively. Thirteen logic blocks and 207 nodes were used, including forward-chaining logic blocks to drive the inference engine from each module of logic to another and backward chained logic blocks to derive variables. They are organized modularly and intuitively to match the CDSS flow diagram (Figure 4.3). It is important to understand the full-text resolution of rules and conditions applied to derive the risk score.

4.3.2.1 Screening Logic Block

The screening logic block is the first logic block the inference engine uses and consists of eight rules (Figure 4.4). The purpose of the screening logic block is to present the home page to the end-user with instructions and determine if the end-user meets the tool's three inclusion criteria. For the end-user to qualify for the tool, they must be HIV positive and have no current or historical lung cancer diagnosis. The outcome of the screening logic
block is a confidence variable used as a condition for advancing to the remaining logic blocks.

The screening logic block uses four static list variables with binary outcomes (Yes or No): HOME_PAGE, HIV_POSITIVE, LUNG_CANCER_NOW, and LUNG_CANCER_EVER (Table 4.2). As the boolean rules are resolved to TRUE or FALSE, the screening logic block creates a confidence variable, SCREENING_CONFIDENCE, with a numeric value of 1 or 0, where 1 indicates the screening criteria are met, and 0 indicates it is not met. The confidence variable becomes a condition in all subsequent forward chained logic blocks. Table 4.2 also reports the Corvid Prompts the end-user will see upon running the CDSS.

Table 4.2: Variables used in the Screening Logic Block

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOME_PAGE</td>
<td>Would you like to continue?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>HIV_POSITIVE</td>
<td>Are you HIV Positive?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>LUNG_CANCER_NOW</td>
<td>Do you currently have lung cancer?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>LUNG_CANCER_EVER</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The screening logic block rules are data-driven, forward chained, and accomplished through boolean rules. The rules are conditional upon each other (nested) and progress iteratively. The screening logic block contains eight boolean ‘IF/THEN’ rules that return a value of ‘TRUE’ or ‘FALSE’. The tests for inclusion criteria in the screening logic block are determined through eight forward chained rules and five variables through 17 nodes.

Rule (I), in full text below, asks the end-user if they would like to continue (Yes or No). If they choose to continue, then the rule is true, and the inference engine advances to the next question to begin testing the screening criteria. If the rule is false, the inference engine moves to rule (VIII).

(I)  IF: HOME_PAGE (Would you like to continue?) = ‘Yes’,
     THEN: Rule (I) = TRUE. Continue to rule (II) HIV_POSITIVE (Are you HIV Positive?). If it is FALSE, move to rule (VIII).

For end-users who chose to continue, rule (II) tests if the end-user is HIV positive (Yes or No). If they are HIV positive, then the rule is true, and the inference engine will move to the next rule to test if they currently have lung cancer rule (III). If the rule is false, the inference engine moves to rule (VII).

(II) IF: HOME_PAGE (Would you like to continue?) = ‘Yes’, AND
     IF: HIV_POSITIVE (Are you HIV Positive?) = ‘Yes’,
     THEN: Rule (II) = TRUE. Continue to Rule (III) LUNG_CANCER_NOW (Do you currently have lung cancer?). If it is false, then move to rule (VII).

HIV positive end-users that chose to continue move to rule (III) to test if they currently have lung cancer (Yes or No), the second screening criteria. Rule (III) tests if they currently have lung cancer. If the rule is true, the inference engine sets the screening
confidence variable to 0, indicating the inclusion criteria has not been met, ultimately stopping the system and displaying a message through the collection variable NOTE. If this rule is false, the inference engine moves to rule (IV).

(I) IF: HOME_PAGE (Would you like to continue?) = ‘Yes’, AND
(II) IF: HIV_POSITIVE (Are you HIV Positive?) = ‘Yes’, AND
(III) IF: LUNG_CANCER_NOW (Do you currently have lung cancer?) = ‘Yes’, THEN: Rule (III) = TRUE. Set SCREENING_CONFIDENCE (‘Screening Criteria Met?’) to 0. Add to NOTES “You indicated you currently have lung cancer. You do not qualify for the tool.” The inference engine will stop and display the note. If false, move to rule (IV).

Rule (IV) is identical to rule (III) but tests the end-user does not have lung cancer now. If they do not have lung cancer, then the rule is true. The inference engine will move to test the third and final screening criteria in rule (V).

(I) IF: HOME_PAGE (Would you like to continue?) = ‘Yes’, AND
(II) IF: HIV_POSITIVE (Are you HIV Positive?) = ‘Yes’, AND
(IV) IF: LUNG_CANCER_NOW (Do you currently have lung cancer?) = ‘No’, THEN: Rule (IV) is TRUE. Move to rule (V) LUNG_CANCER_EVER (Did you ever have lung cancer?).

HIV positive end-users that chose to continue and do not currently have lung cancer move to test the final screening criteria have you ever had lung cancer (Yes or No). Rule (V) tests if they have previously had lung cancer. If they have, the rule is true. The inference engine sets the screening confidence variable to 0, indicating the inclusion criteria has not been met, ultimately stopping the system and displaying a message through the collection variable NOTE. If this rule is false, the inference engine moves to rule (VI).

(I) IF: HOME_PAGE (Would you like to continue?) = ‘Yes’, AND
(II) IF: HIV_POSITIVE (Are you HIV Positive?) = ‘Yes’, AND
(IV) IF: LUNG_CANCER_NOW (Do you currently have lung cancer?) = ‘No’, AND
(V) IF: LUNG_CANCER_EVER (Have you ever had lung cancer?) = ‘Yes’, THEN: Rule (V) = TRUE. Set the SCREENING_CONFIDENCE = 0. Add to NOTES “You indicated you have previously had lung cancer. You do not qualify for the tool.” The inference engine will stop and display the note.

Rule (VI) is identical to rule (V) but tests that the end-user has not previously had lung cancer. If they have not yet had lung cancer, then the rule is true. The screening
confidence variable resolves to 1, indicating all inclusion criteria have been met, and the inference engine progress to the next logic block, Smoking Status.

(I) IF: HOME_PAGE (Would you like to continue?) = ‘Yes’, AND
(II) IF: HIV_POSITIVE (Are you HIV Positive?) = ‘Yes’, AND
(IV) IF: LUNG_CANCER_NOW (Do you currently have lung cancer?) = ‘No’, AND
(VI) IF: LUNG_CANCER_EVER (Have you ever had lung cancer?) = ‘No’,
      THEN: Rule (VI) = TRUE. Set the SCREENING_CONFIDENCE = 1. Inference engine will continue to next forward-chained logic blocks.

The inference engine moves to rule (VII) for those who continued from the home page, but rule (II) was false. Rule (VII) tests if they are not HIV Positive. If they are not HIV Positive, rule (VII) is true. The inference engine sets the screening confidence variable to 0, indicating the inclusion criteria has not been met, ultimately stopping the system and displaying a message through the collection variable NOTE.

(I) IF: HOME_PAGE (Would you like to continue?) = ‘Yes’, AND
(VII) IF: HIV_POSITIVE (Are you HIV Positive?) = ‘No’,
      THEN: Rule (VII) = TRUE. Set the SCREENING_CONFIDENCE = 0. Add to NOTES “You indicated you do not have HIV. You do not qualify for the tool.” The inference engine will stop and display the note.

If the inference engine determined rule (I) to be false, it would move to the final rule of the screening logic block. Rule (VIII) tests if the end-user does not want to continue. If they do not, then the rule is true. The inference engine sets the screening confidence variable to 0, indicating the inclusion criteria has not been met, ultimately stopping the system and displaying a message through the collection variable NOTE.

(VIII) IF: HOME_PAGE (Would you like to continue?) = ‘No’
       THEN: Rule (I) = FALSE. Set the SCREENING_CONFIDENCE = 0. Add to NOTES “Thank you for your understanding.”. The inference engine will stop and display the note.

The rationale for using a confidence variable instead of a static list variable to report the screening criteria' outcome is easier facilitation of the command block. Using a static list variable would have required extra commands to report results to the end-user and explain their exclusion from the tool. The outcome of the logic block is a value of 0 or 1 for the
screening confidence variable. End-users with a value of 0 were excluded either because they chose not to continue or did not meet all screening criteria. Text was added to the collection variable, NOTES, to explain why they were excluded. Those with a value of 1 met all of the screening criteria. The screening confidence variable is used as the first test in each of the logic blocks that follow.

4.3.2.2 Smoking Status Logic Block

The smoking status logic block is the second logic block the inference engine uses and consists of seven forward chained and two backward chained rules (Figures 4.5 and 4.6). The inference engine will only test these rules when it determines that all inclusion criteria are met. The purpose of the smoking status logic block is to determine if the end-user is a smoker or a non-smoker. The smoking status logic block outcome is a static list variable SMOKING_STATUS with a value of ‘Smoker’ or ‘Non-Smoker.’ Smoking status is a critical variable in determining the risk score since smoking status is one of the strataums used to derive the final risk score. A separate logic block (Figure 4.7) is used to create a message about smoking for the results and is not part of the logic to derive smoking status. This logic block is executed through the command block.

Figure 4.5: Smoking Status Logic Block

- [screening_confidence] = 1
- smoker_current = Yes
  - smoking_status = Smoker
- smoker_current = No
  - smoker_ever = Yes
    - quit_24_months = Yes
      - smoking_status = Smoker
    - quit_24_months = No
      - smoking_status = NonSmoker
- smoker_ever = No
  - smoking_status = NonSmoker
Figure 4.6: Months Quit Backward Chaining Logic Block

Figure 4.7: Smoking Message Logic Block

The smoking status logic block is a data-driven forward-chained logic block using three static list variables with binary outcomes (Yes or No): SMOKER_CURRENT, SMOKER_EVER, and QUIT_24_MONTHS (Table 4.3). The static list variable, QUIT_24_MONTHS, is a backward chained variable managed separately in a different logic block and derived from the numeric variable QUIT_MONTHS.

Backward chaining is a powerful feature of Corvid, allowing the inference engine to obtain information from other rules. In this case, the forward-chained rules use the binary outcome of the static list variable, QUIT_24_MONTHS, but need the numeric variable QUIT_MONTHS to derive it, which is in a different rule. The inference engine processes both rules but asks the end-user the question it needs to satisfy both rules. In this case, the inference engine asks the end-user QUIT_MONTHS, “How many months ago did you quit smoking?” instead of QUIT_24_MONTHS, “Quit smoking in the past 24 months?”
Table 4.3: Variables used in the Smoking Status Logic Block

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOKER_CURRENT</td>
<td>Do you currently smoke cigarettes?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>SMOKER_EVER</td>
<td>Have you ever smoked cigarettes?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>QUIT_24_MONTHS</td>
<td>Quit smoking in past 24 months?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>QUIT_MONTHS</td>
<td>How many months ago did you quit smoking?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>SMOKING_STATUS</td>
<td>Smoking Status (Derived)</td>
<td>Static List</td>
<td>Smoker, NonSmoker</td>
</tr>
</tbody>
</table>

The outcome of the smoking status logic block is a static list variable SMOKING_STATUS with a value of ‘Smoker’ or ‘Non-Smoker’. If the end-user currently smokes, they are categorized as a ‘Smoker’; however, if the end-user is not a current smoker, the inference engine will require the end-users smoking history. Former smokers who quit smoking within the last 24 months are categorized as a ‘Smoker’ while those who quit smoking more than 24 months prior are classified as a ‘Non-Smoker.’ Including the backward chained logic block and the screening condition, smoking status is derived using 10 rules and five variables through 15 nodes explained below.

Rule (I) of the smoking status logic block begins by checking that all inclusion criteria were met using the screening confidence variable derived in the screening logic block. The inference engine will continue with rules (II) through (VII) only if rule (I) is true. All remaining forward-chained logic blocks begin with the same condition, excluding people from the CDSS if they did not meet the inclusion criteria. As a result, there is no action for a false value as the system does not allow for a false value.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1
    THEN: Rule (I) = TRUE. Continue to rule (II).
For those meeting the screening criteria, the inference engine moves to rule (II) to test if the end-user currently smokes cigarettes (Yes or No). Rule (II) tests if they do presently smoke. If the end-user is a current smoker, then rule (II) is true. The inference engine classifies them as a ‘Smoker’ and does not require any further information from the logic block. If rule (II) is false, the inference engine moves on to rule (III) to test if they are not current smokers.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1, AND
(II) IF: SMOKER_CURRENT (Do you currently smoke cigarettes?) = ‘Yes’, THEN: Rule (II) =TRUE. Set SMOKING_STATUS to ‘Smoker’.

Rule (III) is identical to rule (II) but tests if the end-user does not currently smoke. If they do not currently smoke, rule (III) is true, and the inference engine moves on to rule (IV) to obtain more information about their smoking history.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1, AND
(III) IF: SMOKER_CURRENT (Do you currently smoke cigarettes?) = ‘No’, THEN: Rule (III) = TRUE. The inference engine needs more information about smoking history to determine the smoking status and continues to rule (IV).

The inference engine tests those who met the screening criteria and are not current smokers if they ever smoked (Yes or No). Rule (IV) tests if the end-user has smoked previously. If they have, rule (IV) is true, and the inference engine moves on to rule (V). If they have not, rule (IV) is false, and the inference engine moves on to rule (VII).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1, AND
(III) IF: SMOKER_CURRENT (Do you currently smoke cigarettes?) = ‘No’, AND
(IV) IF: SMOKER_EVER (Did you ever smoke cigarettes?) = ‘Yes’, THEN: Rule (IV) =TRUE. The end-user is a previous smoker; however, the inference engine needs to know many months it has been since they stopped smoking to determine to the smoking status. Continue to rule (V). If false, continue to rule (VII).

The inference engine moves to rule (V) for those who are not current smokers but have smoked in their past and met the inclusion criteria. Rule (V) will test if they quit smoking in the past 24 months. The inference engine checks to see if other rules will
determine its value, which is in another rule managed separately (Figure 4.6). As a result, the inference engine pauses testing rule (V). It does not ask the end-user the static list variable. Instead, it moves to evaluate the backward chained rule.

The backward chained rule below has two rules. Each of the rules asks the end-user, “How many months ago did you quit smoking?” to determine a ‘Yes’ or ‘No’ value for the static list variable. Rule (A) tests if the end-user quit smoking more than 24 months ago. If it is true, it sets the value of QUIT_24_MONTHS to ‘No.’ Rule (B) tests if the end-user quit smoking less than 24 months ago. If it is true, it sets the value of QUIT_24_MONTHS to ‘Yes.’ After the inference engine tests these rules in this logic block, and it returns to rule (V) to use the values of QUIT_24_MONTHS to determine if rule (V) is true.

(A) IF: QUIT_MONTHS (How many months ago did you quit smoking?) > 24
    THEN: QUIT_24_MONTHS (Quit smoking in past 24 months?) = ‘No’
(B) IF: QUIT_MONTHS (How many months ago did you quit smoking?) ≤ 24
    THEN: QUIT_24_MONTHS (Quit smoking in past 24 months?) = ‘Yes’

With the value of QUIT_24_MONTHS determined, the inference engine can test if the end-user quit smoking in the past 24 months (Yes or No). Rule (V) tests if they have quit smoking in the past 24 months. If they have, rule (V) is true, and smoking status is set to ‘Smoker’. The inference engine has gathered all of the information it needs. If it is false, it will move to rule (VI).

(V) IF: QUIT_24_MONTHS (Quit smoking past 24 months?) = ‘Yes’, AND
    THEN: Rule (V) = TRUE. The end-user is not a current smoker but is a former smoker who quit in the past 24 months, the inference engine categorizes them as a smoker and set SMOKING_STATUS = ‘Smoker’. Rule (VI) is identical to rule (V) but tests if the end-user has not quit smoking in the past 24 months. If they haven’t, then rule (VI) is true, and the smoking status is set to ‘Non-Smoker.’
IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1, AND

(III) IF: SMOKER_CURRENT (Do you currently smoke cigarettes?) = ‘No’, AND

(IV) IF: SMOKER_EVER (Did you ever smoke cigarettes?) = ‘Yes’, AND

(VI) IF: QUIT_24_MONTHS (Quit smoking past 24 months?) = ‘No’, THEN: Rule (VI) = TRUE. The end-user is not a current smoker. Rather, they are a former smoker who quit more than 24 months ago. The inference engine categorizes them as a non-smoker and set SMOKING_STATUS = ‘NonSmoker’.

Rule (VII) is the final rule in the smoking status logic block and is the next step if rule (IV) was false. If the end-user was not a current smoker, it tests if they never smoked previously. If they have never previously smoked, rule (VII) is true, and the smoking status is set to ‘Non-Smoker.’

IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1, AND

(III) IF: SMOKER_CURRENT (Do you currently smoke cigarettes?) = ‘No’, AND

(VII) IF: SMOKER_EVER (Did you ever smoke cigarettes?) = ‘No’, THEN: Rule (VII) = TRUE. Since the end-user is neither a current smoker nor a former smoker, SMOKING_STATUS = ‘NonSmoker’.

After each of these tests has been completed, the inference engine has set the value of SMOKING_STATUS, a required variable to predict lung cancer’s probabilities using the stratum in table 4.3. For all non-smokers, the inference engine will move to the demographics logic block. The inference engine will move to the next logic block for all smokers to determine risk factors of lung cancer based on specific smoking information.

4.3.2.3 Smoking Information Logic Block

The smoking information logic block is the first logic block in the CDSS to begin setting probabilities for individual risk factors and consists of 14 forward chained rules and three backward chained rules (Figures 4.8 and 4.9). The purpose of the smoking information logic block is to use information about end-user smoking characteristics (e.g., packs per day, years smoked) and determine the probability for each risk factor based on their gender, race, and derived smoking status. Three smoking risk factors are significant determinants for lung cancer: the number of packs smoked per day and the third and fourth quartiles of years
smoked; however, each is significant within a specific stratum and with different probabilities. For example, smoking more than one pack per day is a significant predictor for women and black or other races. The probability of lung cancer risk for females smoking more than a pack of cigarettes per day is 2.23, while it is 1.80 for people of black or other races.

Figure 4.8: Smoking Information Logic Block

Figure 4.9: Years Smoked Backward Chained Logic Block
The smoking information logic block uses a combination of nine different static lists, numeric, and confidence variables (Table 4.4). The first three static list variables, GENDER, RACE, and SMOKING_STATUS, conditionally determine the presence of risk factors that will establish a probability for that risk factor. For the first time in the CDSS, the inference engine will evaluate gender (Female or Male) and race (White, Black, or Other). Smoking status was derived by the smoking status logic block (Smoker or Non-Smoker).

The logic block uses two numeric variables: PACKS_DAY and YEARS_SMOKED; however, YEARS_SMOKED is asked to the end-user through backward chaining in a separate rule to derive the static list variable YEARSSMK_Q, which has three possible values: ‘Na’ (Not Applicable), Q3 (Quartile 3), and Q4 (Quartile).

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>What is your sex?</td>
<td>Static List</td>
<td>Male, Female</td>
</tr>
<tr>
<td>RACE</td>
<td>What is your race?</td>
<td>Static List</td>
<td>White, Black, Other</td>
</tr>
<tr>
<td>SMOKING_STATUS</td>
<td>Smoking Status (Derived)</td>
<td>Static List</td>
<td>Smoker, NonSmoker</td>
</tr>
<tr>
<td>YEARSSMK</td>
<td>How many years have you smoked?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>YEARSSMK_Q</td>
<td>Years Smoked by Quartiles</td>
<td>Static List</td>
<td>Q3, Q4, Na</td>
</tr>
<tr>
<td>EST_YEARSSMK_Q3</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>EST_YEARSSMK_Q4</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>PACKS_DAY</td>
<td>How many packs per day do you smoke?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>EST_PACKS_DAY_GT1</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
</tbody>
</table>
The smoking information logic block has three outcomes that calculate lung cancer's probability for three different risk factors. The first is the probability of lung cancer if more than a pack of cigarettes is smoked per day (EST_PACKS_DAY_GT1). The second is the probability of lung cancer if the end-user has smoked between 21 and 28 years (EST_YEARSSMK_Q3), while the third is based on smoking more than 28 years (EST_YEARSSMK_Q4). All confidence variables for the probability are multiplicative. As the inference engine tests the rules sequentially, it multiplies the confidence variable with the rule's value. For example, suppose the end-user is a black female smoker who smokes more than a pack per day. The probability of lung cancer for smoking more than one pack per day (EST_PACKS_DAY_GT1) is first set to 2.23 when rule (V) resolves because "IF GENDER = 'Female'" is TRUE. The confidence variable is then multiplied by 1.80 when rule (XII) resolves because "IF RACE = "Black" or "Other" is TRUE. As a result, the probability of lung cancer is based on smoking more than one pack per day for the black female smoker is 2.23 X 1.80 = 4.014. The inference engine uses six variables, including the screening criteria condition, and 17 rules, including the backward chained logic block, to create the three confidence variables as explained below.

As is true for all of the logic blocks, the first rule, rule (I), begins by testing that all inclusion criteria were met using the screening confidence variable derived in the screening logic block. The inference engine will continue with rules (II) through (XIV) only if rule (I) is true. All remaining forward-chained logic blocks begin with the same condition, excluding people from the CDSS if they did not meet the inclusion criteria.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1
THEN: Rule (I) = TRUE. Continue to rule (II).

The rules of the smoking information logic block only apply to smokers, so the second rule, rule (II), tests for smoking status. For those who met the screening criteria, if
the inference engine determines the end-user to be a smoker, it will continue to test rules (III) to (XIV). If the end-user is not a smoker, it will move to the next logic block, demographics.

(I) \[ \text{IF: SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1, AND} \]
(II) \[ \text{IF: SMOKING\_STATUS (Smoking Status (Derived)) = ‘Smoker’} \]
THEN: Rule (II) = TRUE. Continue to rule (III).

The inference engine moves to rule (III) for all Smokers who met the inclusion criteria. Smoking for more than 21 and less than 28 years increases the risk of lung cancer. In rule (III), the inference engine tests whether the end-user has been smoking for 21 to 28 years. The rule requires the static list variable YEARSSMK\_Q. The inference engine checks to see if other rules will determine its value, which is in another rule managed separately (Figure 4.9). As a result, the inference engine pauses testing rule (III). It does not ask the end-user which quartile applies to them. Instead, it moves to evaluate the backward chained rule.

The backward chained rule uses three rules. Each of the rules asks the end-user, “How many years have you smoked?” to determine one of the three values for the static list variable “Years Smoked by Quartiles.” Rule (A) tests if the end-user smoked less than 21 years. If it is true, it sets the quartile value to ‘Not-Applicable.’ Rule (B) tests if the end-user smoked between 21 and 28 years. If it is true, it sets the quartile value to ‘Q3’. Finally, Rule (C) tests if the end-user smoked more than 28 years and sets the quartile value to ‘Q4’ if it is true. After the inference engine tests the three rules in this logic block, it returns to rule (III) to use the values of YEARSSMK\_Q to determine if rule (III) is true.

(A) \[ \text{IF: YEARSSMK (How many years have you smoked?) \leq 21,} \]
THEN: YEARSSMK\_Q = ‘Na’
(B) \[ \text{IF: YEARSSMK (How many years have you smoked?) > 21 AND YEARSSMK (How many years have you smoked?) \leq 28,} \]
THEN: YEARSSMK\_Q = ‘Q3’
(C) \[ \text{IF: YEARSSMK (How many years have you smoked?) > 28,} \]
THEN: YEARSSMK_Q = ‘Q4’

Returning to rule (III), If the years smoked is between 21 and 28 (Q3), the test is true, and the inference engine will set the confidence variable, EST_YEARSSMK_Q3 = 2.07 and move to rule (IV). This reflects the end-users probability of lung cancer due to smoking between 21 and 28 years.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(III) IF: YEARSSMK_Q (Years Smoked by Quartiles) = ‘Q3’
    THEN: Rule (III) = TRUE. Set the confidence variable EST_YEARSSMK_Q3
         (Probability of lung cancer for Years Smoked Q3) = 2.07.

Rule (IV) is identical to rule (III) except it uses the value ‘Q4’ from the years smoked by quartiles variable. If it is true, the inference engine will set the confidence variable,

EST_YEARSSMK_Q4 = 2.03. This reflects the end-users probability of lung cancer due to smoking more than 28 years.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(IV) IF: YEARSSMK_Q (Years Smoked by Quartiles) = ‘Q4’,
    THEN: Rule (IV) = TRUE. Set the confidence variable EST_YEARSSMK_Q4
         (Probability of lung cancer for Years Smoked Q4) = 2.03.

The inference engine moves to rule (V) for all Smokers who met the inclusion criteria. There are different risk factors and probabilities of female versus male, rule (V) tests if the end-user is a female. If it is true, it will move to rule (VI), which is a test to determine whether they have smoked more than a pack of cigarettes per day. If it is false, it will move to rule (VII).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(V) IF: GENDER = ‘Female’,
    THEN: Rule (V) = TRUE. Continue to rule (VI).

For female smokers who met all screening criteria, rule (VI) will test if they smoke more than a pack of cigarettes per day. If they do, rule (VI) is true, and the confidence
variable is set to 2.23 and reflects the end-users probability of lung cancer due to smoking more than a pack of cigarettes per day.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(V) IF: GENDER = ‘Female’ AND
(VI) IF: PACKS_DAY greater than 1,

THEN: Rule (VI) = TRUE. Set the confidence variable EST_PACKS_DAY_GT1 (Probability of lung cancer for smoking more than a pack per day) = 2.23.

The inference engine moves to rule (VII) for all Smokers who met the inclusion criteria. There are different risk factors and probabilities of male versus female. Rule (VII) tests if the end-user is a male. If it is true, it will move to rule (VIII), which is a test to determine how years they smoked.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(VII) IF: GENDER = ‘Male’,

THEN: Rule (VII) = TRUE. Continue to rule (VIII).

The inference engine moves to rule (VIII) for all male smokers who met the inclusion criteria. Smoking between 21 and 28 years increases lung cancer risk. The inference engine follows a similar backward chaining process to rule (III), where it will pause, find a rule to determine the value of YEARSSMK_Q, examine that rule, and then return to test rule (VIII). If the end-user smoked between 21 and 28 years (Q3), then the rule is true, and the confidence variable is set to 2.23, reflecting the end-users probability of lung cancer due to smoking between 21 and 28 years.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(VII) IF: GENDER = ‘Male’ AND
(VIII) IF: YEARSSMK_Q = ‘Q3’,

THEN: Rule (VIII) = TRUE. Set the confidence variable EST_YEARSSMK_Q3 (Probability of lung cancer for Years Smoke Q3) = 2.41.

Rule (IX) is identical to rule (VIII), except it uses the value ‘Q4’ from the years smoked by quartiles variable YEARSSMK_Q. If it is true, the inference engine will set the confidence variable, EST_YEARSSMK_Q4 to 3.39.
IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
IF: GENDER = ‘Male’ AND
IF: YEARSSMK_Q = ‘Q4’,
THEN: Rule (IX) = TRUE. Set the confidence variable EST_YEARSSMK_Q4
(Probability of lung cancer for Years Smoke Q4) = 3.39.

The inference engine moves to rule (X) for all Smokers who met the inclusion criteria. There are different risk factors and probabilities for people whose race is white versus black or other races. Rule (X) tests if the end-user is white. If it is true, it will move to rule (XI), which is a test to determine how many years they smoked. If it is false, it will move to rule (XII).

IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
IF: RACE = ‘White’,
THEN: Rule (X) = TRUE. Continue to rule (XI).

The inference engine moves to rule (XI) for all white smokers who met the inclusion criteria. Smoking for more than 28 years increases the risk of lung cancer. The inference engine follows a similar backward chaining process to rule (III), where it will pause, find a rule to determine the value of YEARSSMK_Q, process that rule, and then return to test rule (XI). If the end-user smoked more than 28 years (Q4), then the rule is true, and the confidence variable is set to 2.71, reflecting the end-user's probability of lung cancer due to smoking more than 28 years.

IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
IF: RACE = ‘White’ AND
IF: YEARSSMK_Q = ‘Q4’,
THEN: Rule (XI) = TRUE. Set the confidence variable EST_YEARSSMK_Q4
(Probability of lung cancer for Years Smoke Q4) = 2.71.

The inference engine moves to rule (XII) for all Smokers who met the inclusion criteria. There are different risk factors and probabilities for people whose race is black or other versus white races. Rule (XII) tests if the end-user is black or of another race. If it is
true, it will move to rules (XIII) and (XIV), which is a test to determine whether they have specific risk factors.

(I)  \[ \text{IF: SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1 AND} \]

(II) \[ \text{IF: SMOKING\_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND} \]

(XII) \[ \text{IF: RACE = ‘Black’ or ‘Other’,} \]

THEN: Rule (XII) = TRUE. Continue to Rule 13.

The inference engine moves to rule (XIII) for all smokers of black or other races who met the inclusion criteria. Smoking for more than 21 years and less than 28 years increases the risk of lung cancer 1.71 times. The inference engine follows a similar backward chaining process to rule (III), where it will pause, find a rule to determine the value of YEARSSMK\_Q, process that rule, and then return to test rule (XIII). If the end-user smoked between 21 and 28 years (Q3), then the rule is true, and the confidence variable is set to 1.71, reflecting the end-users probability of lung cancer due to smoking between 21 and 28 years.

(I)  \[ \text{IF: SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1 AND} \]

(II) \[ \text{IF: SMOKING\_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND} \]

(XII) \[ \text{IF: RACE = ‘Black’ or ‘Other’ AND} \]

(XIII) \[ \text{IF: YEARSSMK\_Q = ‘Q3’,} \]

THEN: Rule (XIII) = TRUE. Set the confidence variable EST\_YEARSSMK\_Q3 (Probability of lung cancer for Years Smoke Q3) = 1.71.

For smokers of black or other races who met all screening criteria, rule (XIV) will test if they smoke more than a pack of cigarettes per day. If they do, rule (XIV) is true, and the confidence variable is set to 1.80, reflecting the end-users probability of lung cancer due to smoking more than a pack of cigarettes per day.

(I)  \[ \text{IF: SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1 AND} \]

(II) \[ \text{IF: SMOKING\_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND} \]

(XII) \[ \text{IF: RACE = ‘Black’ or ‘Other’ AND} \]

(XIV) \[ \text{IF: PACKS\_DAY greater than 1,} \]

THEN: Rule (XIV) = TRUE. Set the confidence variable EST\_PACKS\_DAY\_GT1 (Probability of lung cancer for smoking more than a pack per day) = 1.80.
4.3.2.4 Demographics Logic Block

The demographics logic block is the second logic block in the CDSS to set probabilities for individual risk factors and contains 11 forward chained rules (Figure 4.10). The purpose of the demographic logic block is similar to the smoking information logic block. It uses demographic information to evaluate the probability of lung cancer for each risk factor based on gender, race, and smoking status. By this point in the CDSS, the inference engine has ascertained the values of gender, race, and smoking status and uses them conditionally to test rules and collect demographic information.

Figure 4.10: Demographics Logic Block

Three demographic risk factors are significant predictors of lung cancer: non-Hispanic ethnicity, having only high school education and being older than 44 years of age. As in all other logic blocks that set probabilistic confidence variables, the probabilities are significant to that stratum. For example, having only a high-schooled education is a significant predictor for white men. The probability of lung cancer for males with only a high schooled education is 3.09, while it is 2.37 for white people. The multiplicative probability of lung cancer for a white male with only a high school education is 3.09 X 2.37 = 7.323.
The demographics logic block uses a combination of nine different static lists, numeric, and confidence variables (Table 4.5). The numeric variable AGE will use the end-users age in years to test rules, while the static lists variables, HISPANIC (Yes or No) and EDUCATION (No Degree, High School, College, or Graduate School) will use fixed values to test rules.

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>What is your sex?</td>
<td>Static List</td>
<td>Male, Female</td>
</tr>
<tr>
<td>RACE</td>
<td>What is your race?</td>
<td>Static List</td>
<td>White, Black, Other</td>
</tr>
<tr>
<td>SMOKING_STATUS</td>
<td>Smoking Status (Derived)</td>
<td>Static List</td>
<td>Smoker, NonSmoker</td>
</tr>
<tr>
<td>AGE</td>
<td>What is your age (in years)?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>EST_AGE_Q4</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>Are you Hispanic or Latino?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>EST_NOTHISPANIC</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>What is your highest level of education (degree received)?</td>
<td>Static List</td>
<td>No Degree, High School, College, Graduate School</td>
</tr>
<tr>
<td>EST_HIGHSCHOOL</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
</tbody>
</table>

The demographics logic block has three outcomes that calculate lung cancer's probability for three different risk factors. The first is the probability of lung cancer if age is greater than 44 years, referred to as the fourth quartile (EST_AGE_Q4). The second is the probability of lung cancer if the end-user is not Hispanic (EST_NOTHISPANIC). The final probability is for end-users with only a high-school education (EST_HIGHSCHOOL). The
The inference engine uses six variables, including the screening criteria condition, and 11 rules to create the three confidence variables as explained below.

Once again, the first rule, rule (I), begins by testing that all inclusion criteria were met using the screening confidence variable derived in the screening logic block. The inference engine will continue with rules (II) through (XI) only if rule (I) is true.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1
THEN: Rule (I) = TRUE. Continue to rule (II).

The inference engine moves to rule (II) for all end-users who met the screening criteria. There are demographic risk factors for smokers, so the inference engine tests if the end-user is a smoker. If they are, the rule is true, and the inference engine moves to rule (III). If they were non-smoker, the inference engine would move to rule (IV) since there are no demographic risk-factors for non-smokers.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived))= ‘Smoker’,
THEN: Rule (II) = TRUE. Continue to rule (III).

For smokers who met all screening criteria, rule (III) will test if they are not-Hispanic (Yes or No). If they are not Hispanic, then the rule is true, and the confidence variable is set to 1.81 and reflects the end-users probability of lung cancer due to being non-Hispanic.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: SMOKING_STATUS (Smoking Status (Derived))= ‘Smoker’ AND
(III) IF: HISPANIC (Are you Hispanic or Latino?)= ‘No’,
THEN: Rule (III) = TRUE. Set the confidence variable EST_NOTHISPANIC (Probability of lung cancer for non-Hispanic Ethnicity ) = 1.81.

The inference engine moves to rule (IV) for those who met the inclusion criteria. There are different risk factors and probabilities of black or other race versus white. Rule (IV) tests if the end-user is black or of another race. If it is true, it will move to rule (V), which is a test to determine whether they are not Hispanic, a significant risk factor for lung cancer in black or other races.
(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

(IV) IF: RACE (What is your race?) = ‘Black’ or ‘Other’,

THEN: Rule (IV) = TRUE. Continue to rule (V).

For smokers who are black or of another race who met all screening criteria, rule (V) will test if they are not Hispanic. If they are not, rule (V) is true, and the confidence variable is set to 2.02 and reflects the end-users probability of lung cancer due to not being Hispanic.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

(IV) IF: RACE (What is your race?) = ‘Black’ or ‘Other’ AND

(V) IF: HISPANIC (Are you Hispanic or Latino?) = ‘No’,

THEN: Rule (V) = TRUE. Set the confidence variable EST_NOTHISPANIC (Probability of lung cancer for non-Hispanic Ethnicity) = 2.02.

The inference engine moves to rule (VI) for those who met the inclusion criteria.

There are different risk factors and probabilities of white versus black or other race. Rule (VI) tests if the end-user is white. If it is true, it will move to rule (VII), which is a test to determine whether they have only a high school education, a significant risk factor for lung cancer in white people.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

(VI) IF: RACE (What is your race?) = ‘White’,

THEN: Rule (VI) = TRUE. Continue to rule (VII).

For white smokers who met all screening criteria, rule (VII) will test if they have only a high-school education. If they do, rule (VII) is true, and the confidence variable is set to 2.37 and reflects the end-users probability of lung cancer due to having a high school education.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

(VI) IF: RACE (What is your race?) = ‘White’ AND

(VII) IF: EDUCATION (What is your highest level of education (degree received)?) = ‘High School’,

THEN: Rule (VII) = TRUE. Set the confidence variable EST_HIGHSCHOOL (Probability of lung cancer for High School Education) = 2.37.

The inference engine moves to rule (VIII) for those who met the inclusion criteria.
There are different risk factors and probabilities of males versus female. Rule (VIII) tests if the end-user is male. If it is true, it will move to rule (IX), which is a test to determine whether they have only a high school education, a significant risk factor lung cancer in men.

(VIII) IF: GENDER (What is your sex?) = ‘Male’,
THEN: Rule (VIII) = TRUE. Continue to rule (IX).

For men who met all screening criteria, rule (IX) will test if they have only a high-school education. If they do, rule (IX) is true, and the confidence variable is set to 3.09 and reflects the end-users probability of lung cancer due to having a high school education.

(IX) IF: EDUCATION (What is your highest level of education (degree received)?) = ‘High School’,
THEN: Rule (IX) = TRUE. Set the confidence variable EST_HIGHSCHOOL (Probability of lung cancer for High School Education) = 3.09.

Like rule (VIII), the inference engine moves to rule (X) for those who met the inclusion criteria. There are different risk factors and probabilities of females versus male.

Rule (X) tests if the end-user is female. If it is true, it will move to rule (XI), which is a test to determine whether they are greater than 44 years of age, a significant risk factor for females.

(X) IF: GENDER (What is your sex?) = ‘Female’,
THEN: Rule (X) = TRUE. Continue to rule (XI).

For women who met all screening criteria, rule (XI) will test if they are greater than 44 years of age. If they are, rule (XI) is true, and the confidence variable is set to 1.73 and reflects the end-users probability of lung cancer due to age greater than 44 (Q4).

(XI) IF: AGE (What is your age (in years)?) > 44,
THEN: Rule (XI) = TRUE. Set the confidence variable EST_AGE_Q4 (Probability of lung cancer for Q4 Age) = 1.73.
4.3.2.5 Respiratory Disease Logic Block

The respiratory disease logic block is the third logic block in the CDSS to set probabilities for respiratory disease risk factors and has 15 forward chained rules (Figure 4.11). Like the smoking information and demographics logic block, the purpose of the respiratory disease logic block is to assess respiratory risk factors specific to gender, race, and smoking status.

Two risk factors, determined from the knowledge base, are significant predictors of lung cancer, a history of reoccurring pneumonia (pneumonia occurring two or more times) and a history of respiratory disease. In the knowledge base, respiratory disease included Acute Respiratory Infections, Chronic Obstructive Pulmonary Disease (COPD), Tuberculosis, MAI/MAC Infection, Pneumoconiosis and Other Lung Disease due to
external agents, Candida in esophagus or lungs, Herpes simplex in lungs, Cytomegalovirus in lungs, and any other diseases of the upper or lower respiratory tract. These are used in the prompt from the inference engine to the end-user when asked about respiratory disease. The two outcomes of the respiratory disease logic block are probabilistic confidence variables, one for a history of respiratory disease, and another for a history of reoccurring pneumonia. The confidence variables are multiplicative across gender, race, and smoking status. Suppose a non-smoking female has a history of reoccurring pneumonia. The probability of lung cancer for women with reoccurring pneumonia is 1.94, while it is 4.38 for non-smokers. The multiplicative probability of lung cancer for a female non-smoker with a history of reoccurring pneumonia is 1.94 X 4.38 = 8.497.

The respiratory disease logic block uses five static lists and two confidence variables (Table 4.6). Once again, GENDER, RACE, and SMOKING_STATUS, conditionally determine the presence of risk factors that will establish a probability for lung cancer based on the knowledge base. Both PNEUMONIA_REOCCURRING, “Have you had pneumonia two or more times in the past?” and RESPIRATORY_DISEASE, “Have you ever been diagnosed with a respiratory disease?”, are static list variables with binary outcomes (Yes or No). They each have a corresponding confidence variable for the probability of lung cancer tested by the inference engine in 15 forward chained rules.

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>What is your sex?</td>
<td>Static List</td>
<td>Male, Female</td>
</tr>
<tr>
<td>RACE</td>
<td>What is your race?</td>
<td>Static List</td>
<td>White, Black, Other</td>
</tr>
</tbody>
</table>
Once again, the first rule, rule (I), begins by testing that all inclusion criteria were met using the screening confidence variable derived in the screening logic block. The inference engine will continue with rules (II) through (XV) only if rule (I) is true.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1
    THEN: Rule (I) = TRUE. Continue to rule (II).

The inference engine moves to rule (II) for those who met the inclusion criteria.

There are different risk factors and probabilities of male versus female. Rule (II) tests if the end-user is male. If it is true, it will move to rule (III), which is a test to determine whether they have a history of respiratory disease, a significant risk factor for lung cancer in men. If it is false, it will move to rule (IV) to test if they are a female.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: GENDER (What is your sex?) = ‘Male’
    THEN: Rule (II) = TRUE. Continue to rule (III).

For men who met all screening criteria, rule (III) will test if they have a history of respiratory disease diagnosis. If they do, rule (III) is true, and the confidence variable is set to 2.39 and reflects the end-users probability of lung cancer due to having a history of respiratory disease diagnosis.
IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

IF: GENDER (What is your sex?) = ‘Male’ AND

IF: RESPIRATORY_DISEASE (Have you ever been diagnosed with a respiratory disease?) = ‘Yes’

THEN: Rule (III) = TRUE. Set the confidence variable

EST_RESPIRATORY_DISEASE (Probability of lung cancer for a history of respiratory disease) = 2.39.

The inference engine moves to rule (IV) for those who met the inclusion criteria.

There are different risk factors and probabilities of female versus male. Rule (IV) tests if the end-user is female. If it is true, it will move to rule (V) and (VI), tests to determine whether they have a history of respiratory disease or a history of reoccurring pneumonia, two significant risk factors for lung cancer in women.

IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

IF: GENDER (What is your sex?) = ‘Female’

THEN: Rule (IV) = TRUE. Continue to rule (V).

For women who met all screening criteria, rule (V) will test if they have a history of reoccurring pneumonia. If they do, rule (V) is true, and the confidence variable is set to 1.94 and reflects the end-users probability of lung cancer due to having a history of reoccurring pneumonia.

IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

IF: GENDER (What is your sex?) = ‘Female’

IF: PNEUMONIA_REOCCURRING (Have you had pneumonia 2 or more times in the past?) = ‘Yes’

THEN: Rule (V) = TRUE. Set the confidence variable EST_PNEUMONIA_REOC (Probability of lung cancer for a history of reoccurring pneumonia) = 1.94.

Rule (VI) is identical to rule (V) for women who met all screening criteria. Instead, rule (VI) will test if they have a history of respiratory disease diagnosis. If they do, rule (VI) is true, and the confidence variable is set to 2.37 and reflects the end-users probability of lung cancer due to having a history of respiratory disease diagnosis.

IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND

IF: GENDER (What is your sex?) = ‘Female’

IF: RESPIRATORY_DISEASE (Have you ever been diagnosed with a respiratory disease?) = ‘Yes’
THEN: Rule (VI) = TRUE. Set the confidence variable
EST_RESPIRATORY_DISEASE (Probability of lung cancer for a history of respiratory disease) = 2.37.

The inference engine moves to rule (VII) for those who met the inclusion criteria.

There are different risk factors and probabilities of the white race versus black or other race.

Rule (VII) tests if the end-user is white. If it is true, it will move to rule (VIII), tests to determine whether they have a history of respiratory disease, a significant risk factor for lung cancer in people of white race. If it is false, then the inference engine will move to rule (IX).

(I)  IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(VII) IF: RACE (What is your race?) = ‘White’
THEN: Rule (VII) = TRUE. Continue to rule (VIII)

For people of white race who met all screening criteria, rule (VIII) will test if they have a history of respiratory disease. If they do, rule (VIII) is true, and the confidence variable is set to 2.13 and reflects the end-users probability of lung cancer due to having a history of respiratory disease.

(I)  IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(VII) IF: RACE (What is your race?) = ‘White’ AND
(VIII) IF: RESPIRATORY_DISEASE (Have you ever been diagnosed with a respiratory disease?) = ‘Yes’
THEN: Rule (VIII) = TRUE. Set the confidence variable
EST_RESPIRATORY_DISEASE (Probability of lung cancer for a history of respiratory disease) = 2.13.

Rule (IX) is identical to rule (VII) but tests if the end-user is of black or other race. If it is true, it will move to rule (X), to test if they have a history of respiratory disease, two significant risk factors for lung cancer in those of black or other race.

(I)  IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(IX) IF: RACE (What is your race?) = ‘Black’ or ‘Other’
THEN: Rule (IX) = TRUE. Continue to rule (X).

Rule (X) is identical to rule (VIII) and will test if the end-user has a history of respiratory disease. If they do, rule (X) is true, and the confidence variable is set to 2.10 and
reflects the end-users probability of lung cancer due to having a history of respiratory disease.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(IX) IF: RACE (What is your race?) = ‘Black’ or ‘Other’ AND
(X) IF: RESPIRATORY_DISEASE (Have you ever been diagnosed with a respiratory disease?) = ‘Yes’

THEN: Rule (X) = TRUE. Set the confidence variable
EST_RESPIRATORY_DISEASE (Probability of lung cancer for a history of respiratory disease) = 2.10.

The inference engine moves to rule (XI) for those who met the inclusion criteria.

There are different risk factors and probabilities of smokers versus non-smoker. Rule (XI) tests if the end-user is a smoker. If it is true, it will move to rules (XII) and (XIII), tests to determine whether they have a history of respiratory disease, a significant risk factor for lung cancer in people of white race. If it is false, then the inference engine will move to rule (XIV).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’

THEN: Rule (XI) = TRUE. Continue to rule (XII).

For smokers who met all screening criteria, rule (XII) will test if they have a history of reoccurring pneumonia. If they do, rule (XII) is true, and the confidence variable is set to 1.77 and reflects the end-users probability of lung cancer due to having a history of reoccurring pneumonia.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(XII) IF: PNEUMONIA_REOCCURRING (Have you had pneumonia 2 or more times in the past?) = ‘Yes’

THEN: Rule (XII) = TRUE. Set the confidence variable
EST_PNEUMONIA_REOCC (Probability of lung cancer for a history of reoccurring pneumonia) = 1.77.
Rule (XIII) is identical to rule (XII) and will test if the end-user has a history of respiratory disease. If they do, rule (XII) is true, and the confidence variable is set to 2.22 and reflects the end-users probability of lung cancer due to having a history of respiratory disease.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(XIII) IF: RESPIRATORY_DISEASE (Have you ever been diagnosed with a respiratory disease?) = ‘Yes’
THEN: Rule (XIII) = TRUE. Set the confidence variable EST_RESPIRATORY_DISEASE (Probability of lung cancer for a history of respiratory disease) = 2.22.

Rule (XIV) is identical to rule (XI) but tests if the end-user is a non-smoker. If it is true, it will move to rule (XV), to test if they have a history of reoccurring pneumonia, a significant risk factor for lung cancer in non-smokers.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XIV) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘NonSmoker’
THEN: Rule (XIV) = TRUE. Continue to rule (XII).

For non-smokers who met all screening criteria, rule (XV) will test if they have a history of reoccurring pneumonia. If they do, rule (XV) is true, and the confidence variable is set to 4.38 and reflects the end-users probability of lung cancer due to having a history of reoccurring pneumonia.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XIV) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘NonSmoker’ AND
(XV) IF: PNEUMONIA_REOCCURRING (Have you had pneumonia 2 or more times in the past?) = ‘Yes’
THEN: Rule (XV) = TRUE. Set the confidence variable EST_PNEUMONIA_REOC (Probability of lung cancer for a history of reoccurring pneumonia) = 4.38.
4.3.2.6 *Clinical HIV Characteristics Logic Block*

The Clinical HIV Characteristics Logic Block is the fourth and final logic block in the CDSS to set probabilities for individual clinical HIV risk factors. It is the most extensive logic block in the CDSS with 22 forward chained rules (Figure 4.12) and six additional rules from backward chaining. The knowledge base informs us that six clinical HIV risk factors predict lung cancer across gender, race, and smoking status. The numeric variable CD4 cell count is used to derive predictors in every stratum (Male, Female, White, Black or Other race, and Smoker) except for non-smokers. CD4 cell counts between 78 and 250 (Q2) and CD4 cell counts between 250 and 446 (Q3) are backward chained in a separate rule (Figure 4.13) and used to test many rules. Likewise, a CD4 to CD8 ratio less than 0.12 (Q1) is also backward chained in a separate rule (Figure 4.14) and is used to test many rules. The fourth predictor is the stability of CD4 values. The knowledge base indicates that drastic changes (“fluctuations”) in the CD4 cell count increase lung cancer risk. Analytical results from the knowledge-base suggest that viral loads greater than 500 are significant predictors for lung cancer. Finally, having a diagnosis of AIDS is also a significant predictor of lung cancer.

Acquired Immunodeficiency Syndrome (AIDS) occurs when either the CD4 cell counts fall below 200 cubic millimeters of blood (200 cells/mm³) or an opportunistic disease develops. The inference engine presents this definition to the end-user when asking the question.

There are six outcomes of this logic block a confidence variable to predict the probability for lung cancer for each of the six risk factors: Q2 of CD4, Q3 of CD4, CD4/CD8 Ratio, CD4 Fluctuations, Viral load greater than 500, and a history of AIDS diagnosis.
Figure 4.12: Clinical HIV Characteristics Logic Block

- screening_confidence = 1
  - gender = Male
    - cd4_q = Q2
      - [est_cd4_q2] = 1.90
    - cd4cd8ratio_q = Q1
      - [est_cd4cd8ratio_q1] = 2.29
  - gender = Female
    - cd4_q = Q2
      - [est_cd4_q2] = 2.26
    - cd4_q = Q3
      - [est_cd4_q3] = 2.70
  - race = White
    - cd4cd8ratio_q = Q1
      - [est_cd4cd8ratio_q1] = 2.15
  - race = Black, Other
    - aidsEver = Yes
      - [est_aidsEver] = 2.55
    - cd4_fluctuations = Never_the_same_Fluctuates
      - [est_cd4_fluct] = 2.75
    - cd4_q = Q2
      - [est_cd4_q2] = 2.91
    - cd4_q = Q3
      - [est_cd4_q3] = 4.17
    - cd4cd8ratio_q = Q1
      - [est_cd4cd8ratio_q1] = 1.97
  - smoking_status = Smoker
    - cd4_fluctuations = Never_the_same_Fluctuates
      - [est_cd4_fluct] = 1.88
    - cd4_q = Q2
      - [est_cd4_q2] = 2.34
    - cd4_q = Q3
      - [est_cd4_q3] = 2.51
    - cd4cd8ratio_q = Q1
      - [est_cd4cd8ratio_q1] = 2.46
  - smoking_status = NonSmoker
    - [viral_load] > 500
      - [est_vloadgt500] = 9.94
The clinical HIV characteristics logic block uses a combination of 17 static lists, numeric, and confidence variables (Table 4.7). Once again, GENDER, RACE, and SMOKING_STATUS, conditionally determine the presence of risk factors that will establish a probability for lung cancer based on the knowledge base. The variable AIDSEVER, "Have you ever been diagnosed with AIDS?" is a static list variable with a binary outcome (Yes or No). To test for fluctuations in CD4 cell count over time, the CD4_FLUCTUATIONS variable, "Over the history of your HIV diagnosis, how would you classify your CD4 fluctuations?" asks the end-user if their CD4 values have always been less than 200, always between 200 and 500, always greater than 500, or never the same (Fluctuates). Quartiles of CD4 cell count, CD4_q are backward chained. The inference engine prompts the end-user to enter their CD4 value, "What is/was your most recent CD4 value?" and determines if it is in the second or third quartiles, Q2 or Q3, else it is not applicable. Similarly, the CD4/CD8 ratio's first quartile is also backward chained and derives CD4CD8RAT_Q with a binary outcome (Q1 or Not-Applicable). The inference engine
determined the CD4 value for other risk factors, so the inference engine prompts the end-user to specify their CD8 value, "What is/was your most recent CD8 value?" Finally, the numeric variable VIRAL_LOAD, "What is your most recent viral load?" captures the end-user's most recent viral load count.

Table 4.7: Variables used in the Clinical HIV Characteristics Logic Block

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>What is your sex?</td>
<td>Static List</td>
<td>Male, Female</td>
</tr>
<tr>
<td>RACE</td>
<td>What is your race?</td>
<td>Static List</td>
<td>White, Black, Other</td>
</tr>
<tr>
<td>SMOKING_STATUS</td>
<td>Smoking Status (Derived)</td>
<td>Static List</td>
<td>Smoker, NonSmoker</td>
</tr>
<tr>
<td>AIDSEVER</td>
<td>Have you ever been diagnosed with AIDS?</td>
<td>Static List</td>
<td>Yes, No</td>
</tr>
<tr>
<td>EST_AIDSEVER</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>CD4_FLUCTUATIONS</td>
<td>Over the history of your HIV diagnosis, how would you classify your CD4 fluctuations?</td>
<td>Static List</td>
<td>Always less than 200, Always between 200 and 500, Always greater than 500, Never the same (Fluctuates).</td>
</tr>
<tr>
<td>EST_CD4_FLUCT</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>CD4</td>
<td>What is/was your most recent CD4 value?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>CD4_Q</td>
<td>CD4 in Quartiles</td>
<td>Static List</td>
<td>Q2, Q3, N/a</td>
</tr>
<tr>
<td>EST_CD4_Q2</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>EST_CD4_Q3</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>CD8</td>
<td>What is/was your most recent CD8 value?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>CD4CD8RATIO</td>
<td>CD4/CD8 Ratio</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>CD4CD8RATIO_Q</td>
<td>Quartiles of CD4/CD8 Ratio</td>
<td>Static List</td>
<td>Q1, N/a</td>
</tr>
<tr>
<td>EST_CD4CD8RATIO_Q1</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
<tr>
<td>VIRAL_LOAD</td>
<td>What is your most recent viral load?</td>
<td>Numeric</td>
<td>Continuous</td>
</tr>
<tr>
<td>EST_VLOADGT500</td>
<td>Probability</td>
<td>Confidence</td>
<td>Table 4.1</td>
</tr>
</tbody>
</table>
The clinical HIV characteristic logic block has six outcomes, one for each of the six risk factors. The confidence variables EST_AIDSEVER, EST_CD4_Q2, EST_CD4_Q3, EST_CD4_FLUCT, EST_CD4CD8RATIO_Q1, and EST_VLOADGT500 are the probabilities of lung cancer based on a history of AIDS diagnosis, CD4 values falling within the Q2 or Q3 ranges (78 to 250 or 250 to 446), fluctuating CD4 values over time (‘Never the same (Fluctuates)’), a ratio of CD4 to CD8 falling within the Q1 range (lower than 0.12), and viral load greater than 500. The inference engine uses 28 rules to accomplish these outcomes, including backward chaining as explained below.

Once again, the first rule, rule (I), begins by testing that all inclusion criteria were met using the screening confidence variable derived in the screening logic block. The inference engine will continue with rules (II) through (XXII) only if rule (I) is true.

(I) \[ \text{IF: SCREENING\_CONFIDENCE (Screening Criteria Met?)} = 1 \] \[ \text{THEN: Rule (I) = TRUE. Continue to rule (II).} \]

As seen in previous logic blocks, the inference engine moves to rule (II) for those who met the inclusion criteria. There are different risk factors and probabilities of male versus female. Rule (II) tests if the end-user is male. If it is true, it will move to rule (III), which is a test to determine what quartile range the end-user CD4 values are. If rule (II) is false, and the end-user is a female, the inference engine will move to rule (V).

(II) \[ \text{IF: GENDER (What is your sex?) = ‘Male’} \] \[ \text{THEN: Rule (II) = TRUE. Continue to rule (III).} \]

For men who met all screening criteria, rule (III) will test if they have a CD4 value between 78 and 250. The rule requires the static list variable CD4_Q; however, the inference engine will check to see if other rules will determine its value, a process that happens with every rule. The inference engine finds a separate rule (Figure 4.13) that has more information to derive CD4_Q. It pauses testing rule (III) and does not ask the end-user which quartile
applies to them. Instead, it goes to the logic block “Backward Chaining – CD4 Quartiles” and tests its rules.

The backward chained rule, “Backward Chaining – CD4 Quartiles,” has three rules. Each rule uses the numeric CD4 variable, asking the end-user, “What is/was your most recent CD4 value?” If the specified value is between 78 and 250, then the inference engine set the value of CD4_Q, “CD4 in Quartiles”, to ‘Q2’. Likewise, if the specified value is between 250 and 446, then the inference engine set the value of CD4_Q, “CD4 in Quartiles”, to ‘Q3’. Since only Q2 and Q3 are significant predictors of lung cancer, the inference engine will set any other value of CD4, CD4 lower than 78, or greater than 446 to ‘Not Applicable.’ Each of the tests (A), (B), (C) will be tested and resolved; although, only one will return true. Once the inference engine completes these tests, the inference engine will return to rule (III) to use the information to complete the test of the rule.

(A) IF: CD4 (What is/was your most recent CD4 value?) > 78 AND CD4 (What is/was your most recent CD4 value?) ≤ 250, THEN: CD4_Q (CD4 in Quartiles) = ‘Q2’.
(B) IF: CD4 (What is/was your most recent CD4 value?) > 250 AND CD4 (What is/was your most recent CD4 value?) ≤ 446, THEN: CD4_Q (CD4 in Quartiles) = ‘Q3’.
(C) IF: CD4 (What is/was your most recent CD4 value?) ≤ 78 OR CD4 (What is/was your most recent CD4 value?) > 446, THEN: CD4_Q (CD4 in Quartiles) = ‘Not Applicable’.

With CD4_Q determined through backward chaining, the inference completes the rule (III) test, where it tests if the CD4 value is in Q2. If it is, rule (III) is true, and the confidence variable is set to 1.90 and reflects the end-users probability of lung cancer due to a CD4 value between 78 and 250 (Q2).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(II) IF: GENDER (What is your sex?) = ‘Male’ AND
(III) IF: CD4_Q (CD4 in Quartiles) = ‘Q2’ THEN: Rule (III) = TRUE. Set the confidence variable EST_CD4_Q2 (Probability of lung cancer for CD4 values in Q2) = 1.90.
A very similar process occurs in rule (IV) for men who met all screening criteria. Rule (IV) will test if the end-user has a CD4/CD8 ratio that is lower than 0.12 (Q1). The rule requires the static list variable CD4CD8RAT_Q; however, the inference engine will check to see if other rules will determine its value, a process that happens with every rule. The inference engine finds a separate rule (Figure 4.14) that has more information to derive CD4CD8_Q. It pauses testing rule (IV) and does not ask the end-user which quartile applies to them. Instead, it goes to the logic block “Backward Chaining – CD4CD8 Ratio Quartiles” and tests its rules.

The backward chained rule, “Backward Chaining – CD4CD8 Ratio Quartiles,” has three rules. Each rule uses the numeric CD4 variable, asking the end-user, “What is/was your most recent CD4 value?” and the numeric CD8 variable, asking the end-user, “What is/was your most recent CD8 value?” If the inference engine already acquired the CD4 value through another test, it will use that value and not ask it again. For the ratio to be valid, a minim value of 1 is required in both CD4 and CD8. The logic block first tests that CD4 and CD8 values are greater than 0. If both CD4 and CD8 are greater than zero, then rule (A) resolves to true and derives the CD4CD8RATIO as CD4 divided by CD8 and moves to rule (B) to test if the ratio is lower than 0.12. This value, 0.12, is informed by the knowledge base as a threshold that significantly predicts lung cancer. If the ratio is less than 0.12, the rule is true, and the value of CD4CD8RATIO_Q, “Quartiles of CD4/CD8 Ratio,” is set to ‘Q1’. Rule (C) tests if the ratio is greater than 0.12. If it is, then rule (C) is true instead of rule (B), and the value of CD4CD8RATIO_Q, “Quartiles of CD4/CD8 Ratio” is set to ‘Not-Applicable’. Once the inference engine completes these tests, the inference engine will return to rule (IV) to use the information to complete the examination of the rule.

(A) IF: CD4 (What is/was your most recent CD4 value?) > 0 AND CD8 > 0 (What is/was your most recent CD4 value?) >0,
THEN: \( CD4\text{CD}8\text{RATIO} (CD4/CD8 Ratio) = CD4/CD8. \)

(B) IF: \( CD4\text{CD}8\text{RATIO} (CD4/CD8 Ratio) < 0.12, \)
THEN: \( CD4\text{CD}8\text{RATIO}_Q (\text{Quartiles of CD4/CD8 Ratio}) = 'Q1'. \)

(C) IF: \( CD4\text{CD}8\text{RATIO} (CD4/CD8 Ratio) \geq 0.12, \)
THEN: \( CD4\text{CD}8\text{RATIO}_Q (\text{Quartiles of CD4/CD8 Ratio}) = 'Not Applicable'. \)

With \( CD4\text{CD}8\text{RATIO}_Q \) determined through backward chaining, the inference completes the rule (IV) test, where it tests if the \( CD4\text{CD}8\text{RATIO} \) value is in Q1. If it is, rule (IV) is true, and the confidence variable is set to 2.29 and reflects the end-users probability of lung cancer due to a CD4/CD8 ratio lower than 0.12 (Q1).

(I) IF: \( \text{SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1 AND} \)

(II) IF: \( \text{GENDER (What is your sex?) = 'Male'} \)

(IV) IF: \( CD4\text{CD}8\text{RATIO}_Q (\text{Quartiles of CD4/CD8 Ratio}) = 'Q1' \)
THEN: Rule (IV) = TRUE. Set the confidence variable \( \text{EST\_CD4\text{CD}8\text{RATIO}_Q1} \) (Probability of lung cancer for CD4/CD8 ration in Q1) = 2.29

As in rule (II), there are different risk factors and probabilities of female versus male.

Rule (V) tests if the end-user is female. If it is true, it will move to rule (VI), which is a test to determine what quartile range the end-user CD4 values are.

(I) IF: \( \text{SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1 AND} \)

(V) IF: \( \text{GENDER (What is your sex?) = 'Female'} \)
THEN: Rule (V) = TRUE. Continue to rule (VI).

For women who met all screening criteria, rule (VI) will test if they have a CD4 value between 78 and 250. The rule requires the static list variable \( CD4\text{\_Q} \) and will follow the same backward chaining process as in rule (III) where it will pause, find a rule to determine the value of \( CD4\text{\_Q} \), conduct a test of that rule, and then return to test rule (VI) to use the information. If the end-users CD4 value is in Q2, the rule is true, and the confidence variable is set to 2.26 and reflects the end-users probability of lung cancer due to a CD4 value between 78 and 250 (Q2).

(I) IF: \( \text{SCREENING\_CONFIDENCE (Screening Criteria Met?) = 1 AND} \)

(V) IF: \( \text{GENDER (What is your sex?) = 'Female AND} \)

(VI) IF: \( CD4\text{\_Q (CD4 in Quartiles)} = 'Q2' \)
THEN: Rule (VI) = TRUE. Set the confidence variable \( \text{EST\_CD4\text{\_Q2} (Probability of lung cancer for CD4 values in Q2)} = 2.26. \)
The test of rule (VII) is identical to rule (VI). It, too, uses backward chaining to obtain information for CD4_Q and tests if the value is in Q3. The rule is true if CD4 is in Q3, and the confidence variable is set to 2.70 and reflects the end-users probability of lung cancer due to a CD4 value between 250 and 446 (Q3).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(V) IF: GENDER (What is your sex?) = ‘Female’ AND
(VII) IF: CD4_Q (CD4 in Quartiles) = ‘Q3’
THEN: Rule (VII) = TRUE. Set the confidence variable EST_CD4_Q3 (Probability of lung cancer for CD4 values in Q3) = 2.70.

The inference engine moves to rule (VIII) for those who met the inclusion criteria.

There are different risk factors and probabilities of white versus black or other race. Rule (VIII) tests if the end-user is white. If it is true, it will move to rule (IX), which is a test to determine what quartile range their CD4/CD8 Ratio is. If rule (VIII) is false, the inference engine will move to rule (X).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(VIII) IF: RACE (What is your race?) = ‘White’
THEN: Rule (VIII) = TRUE. Continue to rule (IX).

For people of the white race who met all screening criteria, rule (IX) will test if they have a CD4/CD8 Ratio value less than 0.12. It will follow the same backward chaining process as in rule (IV) where it will pause, find a rule to determine the value of CD4CD8RATIO_Q, conduct a test of that rule, and then return to test rule (IX) to use the information. Suppose the end-user's CD4/CD8 ratio value is in Q1. In that case, the rule is true, and the confidence variable is set to 2.15 and reflects the end-users probability of lung cancer due to a CD4/CD8 ratio value less than 0.12 (Q1).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(VIII) IF: RACE (What is your race?) = ‘White’ AND
(IX) IF: CD4CD8RATIO_Q (Quartiles of CD4/CD8 Ratio) = ‘Q1’
THEN: Rule (IX) = TRUE. Set the confidence variable EST_CD4CD8RATIO_Q1 (Probability of lung cancer for CD4/CD8 ration in Q1) = 2.15.
The inference engine moves to rule (X) for those who met the inclusion criteria.

There are different risk factors and probabilities of black or other race versus white. Rule (X) tests if the end-user is black or of another race. If it is true, it will move to rule (XI), which is a test to determine if the end-user has ever had an AIDS diagnosis. If rule (X) is false, and the end-user is not black or from another race, the inference engine will move to rule (XVI).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(X) IF: RACE (What is your race?) = ‘Black’ or ‘Other’
THEN: Rule (X) = TRUE. Continue to rule (XI).

People of the black or other races would have an increased risk of lung cancer if they had a previous diagnosis of AIDS. Acquired Immunodeficiency Syndrome (AIDS) occurs when either the CD4 cell counts fall below 200 cubic millimeters of blood (200 cells/mm3) or an opportunistic disease develops. Rule (XI) tests for this among blacks or other races who met all screening criteria. If the end-user has been diagnosed with AIDS in their past, the rule is true. The confidence variable is set to 2.55 and reflects the end-users probability of lung cancer due to a history of an AIDS diagnosis.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(X) IF: RACE (What is your race?) = ‘Black’ or ‘Other’ AND
(XI) IF: AIDSEVER (Have you ever been diagnosed with AIDS?) = ‘Yes’
THEN: Rule (XI) = TRUE. Set the confidence variable EST_AIDSEVER (Probability of lung cancer for a history of AIDS diagnosis) = 2.55.

The knowledge base informed the inference engine that over time, people of black or other races with fluctuating CD4 values have an increased risk of lung cancer. Rule (XII) tests if the end-user’s CD4 values have fluctuated over time. If so, rule (XII) is true, and the confidence variable is set to 2.75 and reflects the end-users probability of lung cancer due to fluctuating CD4 values.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(X) IF: RACE (What is your race?) = ‘Black’ or ‘Other’ AND
(XII) IF: CD4_FLUCTUATIONS (Over the history of your HIV diagnosis, how would you classify your CD4 fluctuations?) = ‘Never the same (Fluctuates)’
THEN: Rule (XII) = TRUE. Set the confidence variable EST_CD4_FLUCT (Probability of lung cancer for CD4 fluctuations over time) = 2.75.

For people of the black or other race who met all screening criteria, rule (XIII) will test if they have a CD4 value between 78 and 250. The rule requires the static list variable CD4_Q and will follow the same backward chaining process as in rule (III) where it will pause, find a rule to determine the value of CD4_Q, conduct a test of that rule, and then return to test rule (XIII) to use the information. If the end-users CD4 value is in Q2, the rule is true, and the confidence variable is set to 2.91 and reflects the end-users probability of lung cancer due to a CD4 value between 78 and 250 (Q2).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(X) IF: RACE (What is your race?) = ‘Black’ or ‘Other’ AND
(XIII) IF: CD4_Q (CD4 in Quartiles) = ‘Q2’
    THEN: Rule (XIII) = TRUE. Set the confidence variable EST_CD4_Q2 (Probability of lung cancer for CD4 values in Q2) = 2.91.

The test of rule (XIV) is identical to rule (XIII). It, too, uses backward chaining to obtain information for CD4_Q and tests if the value is in Q3. The rule is true if it is, and the confidence variable is set to 4.17 and reflects the end-users probability of lung cancer due to a CD4 value between 250 and 446 (Q3).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(X) IF: RACE (What is your race?) = ‘Black’ or ‘Other’ AND
(XIV) IF: CD4_Q (CD4 in Quartiles) = ‘Q3’
    THEN: Rule (XIV) = TRUE. Set the confidence variable EST_CD4_Q3 (Probability of lung cancer for CD4 values in Q3) = 4.17.

For people of the black or other race who met all screening criteria, rule (XV) will test if they have a CD4/CD8 Ratio value less than 0.12. It will follow the same backward chaining process as in rule (IV) where it will pause, find a rule to determine the value of CD4CD8RATIO_Q, conduct a test of that rule, and then return to test rule (XV) to use the information. If the end-users CD4/CD8 ratio value is in Q1, then the rule is true, and the
confidence variable is set to 1.97. This confidence value reflects the end-users probability of lung cancer due to a CD4/CD8 ratio value less than 0.12 (Q1).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND (X) IF: RACE (What is your race?) = 'Black' or 'Other' AND (XV) IF: CD4CD8RATIO_Q (Quartiles of CD4/CD8 Ratio) = 'Q1' THEN: Rule (XV) = TRUE. Set the confidence variable EST_CD4CD8RATIO_Q1 (Probability of lung cancer for CD4/CD8 ration in Q1) = 1.97

The inference engine moves to rule (XVI) for those who met the inclusion criteria.

There are different risk factors and probabilities of smokers versus non-smoker. Rule (XVI) tests if the end-user is a smoker. If it is true, it will move to rule (XVII), which is a test to determine if the end-user’s CD4 values fluctuate over time. If rule (XVI) is false, and the end-user is a non-smoker, the inference engine will move to rule (XXI).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND (XVI) IF: SMOKING_STATUS (Smoking Status (Derived)) = 'Smoker' THEN: Rule (XVI) = TRUE. Continue to rule (XVII).

Rule (XVII) is identical to rule (XII) and tests if the end-user’s CD4 values have fluctuated over time. If so, rule (XVII) is true, and the confidence variable is set to 1.88 and reflects the end-users probability of lung cancer due to fluctuating CD4 values.

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND (XVI) IF: SMOKING_STATUS (Smoking Status (Derived)) = 'Smoker' AND (XVII) IF: CD4_FLUCTUATIONS (Over the history of your HIV diagnosis, how would you classify your CD4 fluctuations?) = 'Yes' THEN: Rule (XVII) = TRUE. Set the confidence variable EST_CD4_FLUCT (Probability of lung cancer for CD4 fluctuations over time) = 1.88.

Rule (XVIII) is identical to rule (III) and will test if the end-user has a CD4 value between 78 and 250. For smokers who met the screening criteria, the rule requires the static list variable CD4_Q. It will follow the same backward chaining process as in rule (III) where it will pause, find a rule to determine the value of CD4_Q, conduct a test of that rule, and then return to test rule (XVIII) to use the information. If the end-users CD4 value is in Q2,
the rule is true, and the confidence variable is set to 2.34 and reflects the end-users probability of lung cancer due to a CD4 value between 78 and 250 (Q2).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XVI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(XVIII) IF: CD4_Q (CD4 in Quartiles) = ‘Q2’
THEN: Rule (XVIII) = TRUE. Set the confidence variable EST_CD4_Q2 (Probability of lung cancer for CD4 values in Q2) = 2.34.

The test of rule (XIX) is identical to rule (XVIII). It, too, uses backward chaining to obtain information for CD4_Q and tests if the value is in Q3. The rule is true if it is, and the confidence variable is set to 2.51 and reflects the end-users probability of lung cancer due to a CD4 value between 250 and 446 (Q3).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XVI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(XIX) IF: CD4_Q (CD4 in Quartiles) = ‘Q3’
THEN: Rule (XIX) = TRUE. Set the confidence variable EST_CD4_Q3 (Probability of lung cancer for CD4 values in Q3) = 2.51.

Rule (XX) is the same as rule (IV) and will test if they have a CD4/CD8 Ratio value less than 0.12. It will follow the same backward chaining process as in rule (IV) where it will pause, find a rule to determine the value of CD4CD8RATIO_Q, conduct a test of that rule, and then return to test rule (XX) to use the information. If the end-users CD4/CD8 ratio value is in Q1, then the rule is true, and the confidence variable is set to 2.46. This confidence value reflects the end-users probability of lung cancer due to a CD4/CD8 ratio value less than 0.12 (Q1).

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XVI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘Smoker’ AND
(XX) IF: CD4CD8RATIO_Q (Quartiles of CD4/CD8 Ratio) = ‘Q1’
THEN: Rule (XX) = TRUE. Set the confidence variable EST_CD4CD8RATIO_Q1 (Probability of lung cancer for CD4/CD8 ratio in Q1) = 2.46.

The inference engine moves to rule (XXI) for those who met the inclusion criteria. There are different risk factors and probabilities of non-smoker versus smoker. Rule (XXI) tests if the end-user is a non-smoker. If it is true, it will move to rule (XXII), which is a test
to determine if the end-user’s viral load count is greater than 500. If rule (XXII) is false, and the end-user is a non-smoker, the inference engine will move to the final logic block, the risk score logic block

(I) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XXI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘NonSmoker’
THEN: Rule (XXI) = TRUE. Continue to rule (XXII).

Non-smokers who met the screening criteria have an increased risk for lung cancer if their viral loads are greater than 500. This threshold, 500, is a standard threshold and a value deemed significant in the knowledge base. Rule (XXII), the final rule in the clinical HIV characteristic logic block, tests if viral loads are greater than 500. If they are, rule (XXII) is true, and the confidence variable is set to 9.94, the highest confidence value in the CDSS. It reflects the end-users probability of lung cancer due to a viral load greater than 500.

(II) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1 AND
(XXI) IF: SMOKING_STATUS (Smoking Status (Derived)) = ‘NonSmoker’ AND
(XXII) IF: VIRAL_LOAD (What is your most recent viral load?) > 500
THEN: Rule (XXII) = TRUE. Set the confidence variable EST_VLOADGT500
(Probability of lung cancer for viral load greater than 500) = 9.94.

4.3.2.7 Risk Score Logic Block

The risk score logic block is the culmination of all the probabilistic logic blocks. Its primary purpose is to sum each of the 14 probabilities of lung cancer from each risk factor. It does this through 14 rules (Figure 4.15). Once the inference engine calculates the risk score, a separate logic block uses backward chaining to stratify risk into three levels, low, medium, and high. The secondary purpose of the risk score logic block is to set notes displayed in the final end-user screen. The notes describe the risk factor probabilities used to derive the risk score to the end-user. There are three outcomes of the risk score logic block. First is the risk score itself, a confidence variable, which is the sum of all 14 risk factor specific confidence variables used to predict lung cancer. The second outcome is the
collection variable that will collate a list of messages to report back to the end-user in the results screen. The numeric risk score is backward chained in the final logic block to stratify risk into three levels, low, medium, and high, the final outcome of the logic block.

Figure 4.15: Risk Score Logic Block

The risk score logic block uses 15 confidence variables and one collection variable (Table 4.8). The risk score sums the multiplicative confidence variables derived from smoking information, demographics, respiratory disease, and clinical HIV characteristics logic blocks.
Table 4.8: Variables used in the Risk Score Logic Block

<table>
<thead>
<tr>
<th>Name</th>
<th>Prompt or Label</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST_AGE_Q4</td>
<td>Probability: Age &gt; 44</td>
<td>Confidence</td>
<td>Multiplicative</td>
</tr>
<tr>
<td>EST_AIDSEVER</td>
<td>Probability: AIDS Diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_CD4_FLUCT</td>
<td>Probability: Fluctuating CD4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_CD4_Q2</td>
<td>Probability: 78 &lt; CD4 &lt;=250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_CD4_Q3</td>
<td>Probability: 250 &lt; CD4 &lt;= 446</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_CD4CD8RATIO_Q1</td>
<td>Probability: CD4/CD8 &lt; 0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_HIGHSCHOOL</td>
<td>Probability: High School Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_NOTHISPANIC</td>
<td>Probability: Non-Hispanic Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_PACKS_DAY_GT1</td>
<td>Probability: Smoking &gt; 1 Pack/day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_PNEUMONIA_REOC</td>
<td>Probability: Reoccurring Pneumonia (2 or more times)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_RESPIRATORY_DISEASE</td>
<td>Probability: History of Respiratory Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_VLOADGT500</td>
<td>Probability: Viral Load &gt; 500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_YEARSSMK_Q3</td>
<td>Probability: 21 &lt; Smoking Years &lt;= 28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST_YEARSSMK_Q4</td>
<td>Probability: Smoking &gt; 28 Years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RISK_SCORE</td>
<td>Risk Score (Calculated)</td>
<td>Summed</td>
<td></td>
</tr>
<tr>
<td>RISK_NOTE</td>
<td>Risk Score Note</td>
<td>Collection</td>
<td>List</td>
</tr>
<tr>
<td>RISK_SCORE_LEVEL</td>
<td>Risk Score Level</td>
<td>Static List</td>
<td>Low, Medium, High</td>
</tr>
</tbody>
</table>
The format for each of the 14 rules in the logic is exactly the same. It can be understood as:

IF: \( \text{ESTIMATE} > 1 \)
THEN: \( \text{RISK\_SCORE} = \text{ESTIMATE} \)
\( \text{RISK\_NOTE} = \text{Text to be displayed} \)

Each test begins by testing the specific confidence variable is greater than 1. By default, the inference engine sets a null response for the multiplicative confidence variable to a value of 1. Without this test, the risk score would be 14 for someone who did not have any risk factor for lung cancer since each confidence variable would have defaulted to 1.

For example, rule (IV) of the risk score logic block tests if the confidence variable, \( \text{EST\_CD4\_Q2} \), is greater than 1. A value greater than one indicates that a rule in the clinical HIV characteristics logic block, testing if CD4 was in Q2, was true and set a probability for \( \text{EST\_CD4\_Q2} \). Suppose the end-user is a white male smoker with a CD4 value of 300. In this case, rule (III) and (XVIII) are used from the clinical HIV characteristics logic block and resolve to be false, and the confidence variable, \( \text{EST\_CD4\_Q2} \), is set to 1. In contrast, another white male smoker with a CD4 value of 200 would have a value of \( 1.90 \times 2.34 = 4.446 \) when both rules (III) and (XVIII) are true. If rule (IV) in the risk score logic block is true, then the confidence variable \( \text{RISK\_SCORE} \) is set to the confidence variable, \( \text{EST\_CD4\_Q2} \) and a message is added to the collection variable \( \text{RISK\_NOTES} \).

(IV) IF: \( \text{EST\_CD4\_Q2} \) (Probability for CD4 Q2) \( >1 \)
THEN: \( \text{RISK\_SCORE} \) (Risk Score (Calculated)) = \( \text{EST\_CD4\_Q2} \) AND \( \text{RISK\_NOTES} = \text{Your CD4 value ([CD4]) is between 78 and 250. This increased your risk of lung cancer by [[EST\_CD4\_Q2.FORMAT ###.###]]} \)

In Corvid, the variable values can be reported in the collection variable's text output using double brackets. For example, the text \([CD4]\) would return the value of the numeric variable CD4. The output for variable values can be formatted using the '.FORMAT' method in Corvid. The probabilities for each risk factor are formatted to limit the values to
three values before and after the decimal point (e.g., EST_CD4_Q2.FORMAT ###.###).

The risk note message for the white male smoker with a CD4 value of 300 would read as
follows using rule (IV): "Your CD4 value (200) is between 78 and 250. This increased your risk of
lung cancer by 4.446".

Each of the 14 rules in the risk score logic block is tested the same way as rule (IV).
The confidence variable tested in each rule will change, as will the risk note's message. The
inference engine recursively sums the risk score confidence variable each time one of the 14
tests resolves to true. Finally, the inference engine uses backward chaining to stratify the risk
score into low, medium, and high using five rules (Figure 4.16). The threshold for low,
medium, and high were informed from the quartiles of the risk score analyzed in the
knowledge base. Quartile 1 is low risk, Quartiles 2 and 3 are medium risk, and Quartile 4 is
high risk. With these thresholds the CDSS is estimated to be 77% sensitive.

Figure 4.16: Risk Level Backward Chained Logic Block

As seen in other rules, rule (A) begins by testing that all inclusion criteria are met
using the screening confidence variable derived in the screening logic block. The inference
engine will continue with rules (B), (C), and (D) only if rule (A) is true. If rule (A) is false,
then the inference engine will move to rule (E).

(A) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 1
    THEN: Rule (A) = TRUE. Continue to rule (B).
Rule (B) tests if the risk score is between 0 and 14.8. If it is true, then the risk level is set to ‘low.’ Rule (C) tests if the risk score is between 14.8 and 39.1 and sets the risk level to ‘medium’ if it is true. Likewise, rule (D) tests if the risk score greater than 39.1 and sets the risk level to ‘high’ if it is true.

(B) IF: RISK_SCORE (Risk Score (Calculated)) >= 0 AND RISK_SCORE (Risk Score (Calculated)) <= 14.8
    THEN: Rule (B) = TRUE. Set RISK_LEVEL (Risk Score Level) to ‘Low’.

(C) IF: RISK_SCORE (Risk Score (Calculated)) > 14.8 AND RISK_SCORE (Risk Score (Calculated)) <= 39.1
    THEN: Rule (C) = TRUE. Set RISK_LEVEL (Risk Score Level) to ‘Medium’.

(D) IF: RISK_SCORE (Risk Score (Calculated)) > 39.1
    THEN: Rule (D) = TRUE. Set RISK_LEVEL (Risk Score Level) to ‘High’.

Finally, rule (E) tests if the screening confidence variable is 0. This occurs when some chose not to continue from the main screen or did not meet the screening criteria. The reason for this was so to provide a clearer message to the end-user when they stop using the CDSS or don’t meet the criteria and allow for a simpler command block.

(E) IF: SCREENING_CONFIDENCE (Screening Criteria Met?) = 0
    THEN: Rule (E) = TRUE. Set RISK_LEVEL (Risk Score Level) to ‘Not Applicable’.

4.3.3 Command Block

All Corvid systems require at least one command block. Command blocks tell the system what to do and can be as simple or complex as designed by the CDSS developer. The logic blocks instruct the system how to do everything, so the CDSS has a straightforward command block with four simple commands (Figure 4.17).
The first command, 'DERIV CONF,' instructs the inference engine to derive all confidence variables' value using backward chaining. The inference engine has been designed with all of the conditional rules to drive the probabilities. This first command instructs the inference engine to automatically and intelligently test the rules required to derive each risk factors probability. The second command, 'DERIVIVE [risk_score_level]' instructs the inference engine to determine the value of the static list variable 'risk score level.' This variable will tell the end-user if they are at low, medium, or high risk. This is a critical variable to the end-user and needed a separate command since no rules instructed the inference engine to derive it. The third command, ‘FORWARD BLOCK = Smoking message’ tells the inference engine to run the ‘Smoking Message’ logic block which derives a message for smokers. Finally, the results command reports all of the guidance and specific variables to the end-user.
4.3.4 Running the Corvid CDSS

Corvid provides two mechanisms for interacting with the end-user to gather data, test rules, and draw conclusions. While both use Java, a powerful, object-oriented programming language often used in web page design, one runs on the end-users computer, and the other runs on a server. The former is an Applet and the latter is a Servlet. This study utilized the Corvid Servlet runtime process because of numerous benefits such as easy HTML design, a dynamic interface based on end-user input, and the capacity to build and test in the Corvid development environment. Additionally, Corvid's Servlet will be easily transferrable to URL for integration into a website. It will be accessible via smartphone technology, such as iPhones and iPads, because it runs on a server instead of the computer's local users.

This study required three components to run the Corvid Servlet (Figure 4.18). The end-user utilizes their personal computing device (e.g., a computer or laptop) to connect to a servlet engine such as Apache Tomcat ("Tomcat"). Tomcat interprets the Java language necessary to communicate to the inference engine. The inference engine presents HTML pages to the end-user to collect information it needs (input), stores the data, and report results (output). This study did not need to communicate with any external databases or programs.
4.4 Testing and Validation

The knowledge-base and inference-engine development went through several iterations of the SDLC, including planning, designing, analysis, and implementation. The knowledge-based required planning to identify variables to normalize, designing the normalization process for the data, selecting the appropriate analysis method, implementing data to derive the risk score, and then testing the algorithms to ensure they function accurately. Similarly, the inference engine went through the steps of the SDLC. Planning involved assessing Corvid for its capabilities, designing the variables and logic blocks' architecture, implementing the rules and conditions, and testing the results to ensure the rules execute correctly and as expected. There were several iterations of the SDLC in developing the two components independently. Several iterations of the SDLC were performed comparing the inference engine to the knowledge base. For example, if the
knowledge-base manual assessment informed a risk-score of 1.97 and the inference engine informed of a different value, then the knowledge-base and inference-engine were re-evaluated and adjusted as appropriate, leading to more rounds of the SDLC.

The final step before integration is validation. It was challenging to obtain case-reports or a second, independent validation dataset to validate the risk score algorithm. Since these challenges were known before beginning the second phase of the CDSS development, 20 subjects were excluded from the analysis and served as use-cases for testing the CDSS. The survey select procedure in SAS with simple random samples (“SRS”) as the method and 499812 as a randomly selected seed identified the 20 subjects. The analytical sample used to form the knowledge-base and inform the inference engine used 7,587 subjects instead of 7,607. The algorithm developed for the risk score yields a sensitivity of 77%. It is expected that the validation results will produce a true positive at a similar rate.

4.5 Limitations

There were several limitations to this study. The most significant burden was in obtaining public use data reporting lung cancer in PWH. HIV status is considered protected health information (PHI) and is not readily available. The MACS and WIHS data from Johns Hopkins University was the only organization that did not require IRB approval for public use data sets (PDS). The PDS contained all of the necessary variables planned for in this study; however, the PDS structure and organization differed significantly between MACS and WIHS.

In most cases, the data was normalized smoothly, but in other cases, data was more challenging to normalize, or data did not exist in both cohorts. For example, the use of e-
cigarettes was a risk factor of interest that was collected in MACS but not WIHS. Likewise, for annual income, a risk factor that ended up being used in the final risk models was collected continuously in MACS but categorically in WIHS, which led to a loss of information when the MACS data was categorized. There was also very little support for understanding data issues that arose from poor documentation or dirty data.

It was also difficult to identify primary versus secondary data in the WIHS. The MACS did have an indicator flag for metastasized cancer and concluded that approximately 90% of MACS lung cancer cases were primary. While it was not possible to identify primary lung cancer in the WIHS, very few secondary cancer cases were found. It is believed that the majority of the WIHS is primary lung cancer, similar to the MACS.

Corvid also posed limitations since the version of the software being used is an academic version, which restricted the software to 250 nodes. The original design of the CDSS inference engine exceeded 250 nodes and required refinement. The restrictions forced the creative style of questioning that wouldn't have been required in non-academic settings. Another limitation of Corvid is that it only works in Windows environments, which limits the use of Macintosh computers, such as the one used in this study. Corvid also had particular requirements for Macintosh users' virtual computing software, which rendered some of the more straightforward, inexpensive virtual machines impossible to use.
CHAPTER 5

RESULTS

5.1 Overview

This chapter aims to summarize the knowledge-base results, the inference engine, and the CDSS and demonstrate their roles in supporting this study's hypothesis. This study's primary goal was to develop a risk score for lung cancer in PWH and use that risk score in a CDSS as a tool to assess lung cancer risk for prediction and earlier diagnosis in PWH. The analysis data consisted of public use data sets (PDS) managed by Johns Hopkins University. The data was analyzed using the MACS and WIHS cohorts and formed the knowledge-base of the CDSS.

The analysis used multivariable logistic regression, which led to the design of the inference engine's rules and conditions explained in the research methodology. This chapter focuses on the development and results of the CDSS, a tool that has been branded for integration and use by PWH and their clinicians. Screenshots from the CDSS provide the end-users experience and demonstrate the workflow of questions asked by the inference engine to determine a risk score. Finally, it provides results from patient narratives for CDSS testing, using the 20 subjects excluded from the data analysis.
5.2 Data Analyzed

The purpose of the data analysis was to identify significant risk factors for lung cancer and form predictive models for lung cancer. The models are combined multiplicatively and sum to create a single risk score. Integrated data analysis of two longitudinal cohorts of men and women led to creating the knowledge base for the CDSS. Johns Hopkins University governs the Multicenter AIDS Cohort Study (MACS) and the Women's Interagency HIV Study (WIHS) (Figure 5.1). Both cohorts became available through public-use data sets (PDS) with different cut-offs, September 30, 2017, and 2015, respectively. The MACS contained 7,338 men, and the WIHS included 4,982 women leading to an integrated cohort of 12,320 men and women with 7,607 participants having HIV. A total of 100 lung cancer cases were identified over 73,401 person-years in PWH using ICD-0-3 topography code C34 in the MACS and a combination of text fields and study-specific outcome codes in the WIHS (IR = 123 per 100,000 person-years). A total of 26 lung cancer cases were identified over 63,163 person-years among the HIV negative (IR=41 per 100,000 person-years). The incidence rate ratio for the entire cohort, 3.31, is significantly greater than other studies. As a result, this analytical data set is generalizable to PWH and other studies previously reported.

A lung cancer diagnosis was derived as a binary response variable (1= Lung Cancer Diagnosis, 0=No Long Cancer Diagnosis) and was the study's primary endpoint. The analytical sample for the analysis and the knowledge base only included HIV positive men and women. A total of 20 HIV positive subjects, 10 with lung cancer and 10 without lung cancer, were reserved and excluded from the analytical sample for testing the CDSS. The analytical cohort became 7,587 HIV positive men and women. The analysis's explanation is
divided into the primary analysis, which led to the final predictive models and CDSS, and exploratory analysis to explain the effects of separating combined race categories.

Figure 5.1: MACS/WIHS Analytical Sample

5.2.1 Primary Analysis

The literature review highlighted traditional and novel risk factors for lung cancer-specific to PWH. Risk factors available in the MACS and WIHS cohorts are described in Table 5.1 below by lung cancer diagnosis to show the differences and similarities between those with lung cancer and those without lung cancer. Binary and categorical variables are summarized using frequency and percent and use Chi-Square ($\chi^2$) tests for differences. Continuous variables are described using the mean and standard deviation and use t-tests for differences. First, second, and third quartiles describe continuous variables since indicator variables used in the predictive models were derived from quartiles.
Table 5.1: Descriptive Statistics, MACS/WIHS risk factors by lung cancer diagnosis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lung Cancer</th>
<th>No Lung Cancer</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics Information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>38.63 (8.93)</td>
<td>35.89 (8.57)</td>
<td>0.0015</td>
</tr>
<tr>
<td>Gender (Female)</td>
<td>48 (48.00)</td>
<td>3856 (51.37)</td>
<td>0.5036</td>
</tr>
<tr>
<td>Gender (Male)</td>
<td>52 (52.00)</td>
<td>3561 (48.63)</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td>0.0369</td>
</tr>
<tr>
<td>White</td>
<td>46 (46.00)</td>
<td>3292 (43.85)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>44 (44.00)</td>
<td>2723 (36.27)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>10 (10.00)</td>
<td>1492 (19.87)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity (Not Hispanic)</td>
<td></td>
<td></td>
<td>0.0090</td>
</tr>
<tr>
<td>N</td>
<td>12 (12.00)</td>
<td>1730 (23.05)</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>88 (88.00)</td>
<td>5777 (76.95)</td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td>23.92 (5.74)</td>
<td>25.67 (6.94)</td>
<td>0.0038</td>
</tr>
<tr>
<td><strong>Sociodemographic Information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Income</td>
<td>20285 (19458)</td>
<td>23233 (20542)</td>
<td>0.2331</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Degree</td>
<td>51 (51.00)</td>
<td>3755 (50.02)</td>
<td>0.0763</td>
</tr>
<tr>
<td>Completed High School</td>
<td>29 (29.00)</td>
<td>1535 (20.45)</td>
<td></td>
</tr>
<tr>
<td>Completed College</td>
<td>14 (14.00)</td>
<td>1396 (18.60)</td>
<td></td>
</tr>
<tr>
<td>Attended/Completed Graduate School</td>
<td>6 (6.00)</td>
<td>821 (10.94)</td>
<td></td>
</tr>
<tr>
<td>Alcohol use</td>
<td></td>
<td></td>
<td>0.0335</td>
</tr>
<tr>
<td>N</td>
<td>58 (58.00)</td>
<td>3551 (47.32)</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>42 (42.00)</td>
<td>3954 (52.68)</td>
<td></td>
</tr>
<tr>
<td>Cocaine use</td>
<td></td>
<td></td>
<td>0.0010</td>
</tr>
<tr>
<td>N</td>
<td>26 (26.00)</td>
<td>3178 (42.35)</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>74 (74.00)</td>
<td>4327 (57.65)</td>
<td></td>
</tr>
<tr>
<td>Marijuana use</td>
<td></td>
<td></td>
<td>0.0307</td>
</tr>
<tr>
<td>N</td>
<td>41 (41.00)</td>
<td>3894 (51.87)</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>59 (59.00)</td>
<td>3613 (48.13)</td>
<td></td>
</tr>
<tr>
<td>Smoking Information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking Status (Smoker)</td>
<td>18 (18.00)</td>
<td>3590 (47.82)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Y</td>
<td>82 (82.00)</td>
<td>3917 (52.18)</td>
<td></td>
</tr>
<tr>
<td>Years smoked</td>
<td>21.17 (9.17)</td>
<td>17.44 (9.40)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Packs smoked per day</td>
<td></td>
<td></td>
<td>0.7133</td>
</tr>
<tr>
<td>&lt; 1 per day</td>
<td>42 (51.85)</td>
<td>2264 (56.39)</td>
<td></td>
</tr>
<tr>
<td>&gt;= 1 per day</td>
<td>28 (34.57)</td>
<td>1244 (30.98)</td>
<td></td>
</tr>
<tr>
<td>2 or more per day</td>
<td>11 (13.58)</td>
<td>507 (12.63)</td>
<td></td>
</tr>
<tr>
<td>Months since quit</td>
<td>52.13 (60.74)</td>
<td>156.9 (1403)</td>
<td>0.0252</td>
</tr>
<tr>
<td></td>
<td>[12.00, 27.00, 79.50]</td>
<td>[24.00, 60.00, 99.00]</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Lung Cancer</td>
<td>No Lung Cancer</td>
<td>P-Value</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>---------</td>
</tr>
<tr>
<td><strong>Respiratory Disease</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reoccurring Pneumonia, History of</td>
<td>N</td>
<td>74 (74.00)</td>
<td>6795 (90.52)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>26 (26.00)</td>
<td>712 (9.48)</td>
</tr>
<tr>
<td>Respiratory Disease, History of</td>
<td>N</td>
<td>70 (70.00)</td>
<td>6500 (86.59)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>30 (30.00)</td>
<td>1007 (13.41)</td>
</tr>
<tr>
<td><strong>Clinical HIV Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIDS Diagnosis ever</td>
<td>N</td>
<td>50 (50.00)</td>
<td>4739 (63.13)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>50 (50.00)</td>
<td>2768 (36.87)</td>
</tr>
<tr>
<td>CD4 cell counts</td>
<td></td>
<td>294.9 (246.6)</td>
<td>451.3 (355.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[78.00, 250.5, 446.5]</td>
<td>[138.0, 416.0, 674.0]</td>
</tr>
<tr>
<td>CD4, fluctuations over time</td>
<td>N</td>
<td>15 (15.00)</td>
<td>1959 (26.10)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>85 (85.00)</td>
<td>5548 (73.90)</td>
</tr>
<tr>
<td>CD4, Nadir</td>
<td></td>
<td>180.5 (172.6)</td>
<td>245.9 (225.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[42.50, 142.5, 262.0]</td>
<td>[59.00, 201.0, 360.0]</td>
</tr>
<tr>
<td>CD8 cell counts</td>
<td></td>
<td>733.8 (517.1)</td>
<td>783.7 (492.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[385.0, 600.5, 970.0]</td>
<td>[460.0, 701.0, 1002]</td>
</tr>
<tr>
<td>CD4/CD8 Ratio</td>
<td></td>
<td>0.45 (0.42)</td>
<td>0.64 (0.56)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.12, 0.35, 0.60]</td>
<td>[0.20, 0.51, 0.92]</td>
</tr>
<tr>
<td>Viral load greater than 500</td>
<td>N</td>
<td>36 (37.11)</td>
<td>3723 (51.74)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>61 (62.89)</td>
<td>3472 (48.26)</td>
</tr>
</tbody>
</table>

Race was condensed into three levels, White, Black, or Other, due to extremely low counts among those with lung cancer (Table 5.2). 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. The final predictive models used a race variable collapsed into a binary variable, White versus Black/Other, due to the extremely low counts of “other” races among those with lung cancer.
Table 5.2: Explanation of categorized versus collected Race

<table>
<thead>
<tr>
<th>Race (Categories)</th>
<th>Race (Collected)</th>
<th>Lung Cancer</th>
<th>No Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>White</td>
<td>46</td>
<td>3292</td>
</tr>
<tr>
<td>Black</td>
<td>Black or African American</td>
<td>44</td>
<td>2723</td>
</tr>
<tr>
<td>Other</td>
<td>American Indian or Alaska Native</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Asian or Pacific Islander</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Other (Specify)</td>
<td>5</td>
<td>669</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>5</td>
<td>745</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
<td><strong>100</strong></td>
<td><strong>7,507</strong></td>
</tr>
</tbody>
</table>

PWH with a diagnosis of lung cancer are significantly different on many risk factors. Specifically, they are 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.

The literature highlighted discernable differences in lung cancer risk by gender, race, and smoking status, evidenced by the outcomes in Table 5.1. As a result, this study looked independently at gender (Female, Male), race (White, Black/Other), and smoking status (Smoker, Non-Smoker) and created a composite risk score from each. The first step in developing the risk score was assessing each risk factor's association to 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. Table 5.3 below reports the p-values for each risk factor in each of the 306 bivariable logistic regressions highlighting significant predictors of lung cancer. Predictors with a p-value less than 0.25 were considered candidate risk factors for inclusion in multivariable models.
Table 5.3: Bivariable logistic regression results, Candidate risk factors for multivariable models

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Female</th>
<th>Male</th>
<th>White</th>
<th>Black/Other</th>
<th>Smokers</th>
<th>Non-Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics Information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE Q1</td>
<td>0.028</td>
<td>0.312</td>
<td>0.762</td>
<td>0.003</td>
<td>0.005</td>
<td>0.328</td>
</tr>
<tr>
<td>AGE Q2</td>
<td>0.804</td>
<td>0.467</td>
<td>0.456</td>
<td>0.796</td>
<td>0.757</td>
<td>0.853</td>
</tr>
<tr>
<td>AGE Q3</td>
<td>0.146</td>
<td>0.588</td>
<td>0.662</td>
<td>0.022</td>
<td>0.102</td>
<td>0.388</td>
</tr>
<tr>
<td>AGE Q4</td>
<td>0.132</td>
<td>0.657</td>
<td>0.613</td>
<td>0.096</td>
<td>0.083</td>
<td>0.880</td>
</tr>
<tr>
<td>Female</td>
<td>0.467</td>
<td></td>
<td>0.467</td>
<td>0.038</td>
<td>0.092</td>
<td>0.111</td>
</tr>
<tr>
<td>White</td>
<td>0.320</td>
<td>0.082</td>
<td></td>
<td>0.241</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Black/Other</td>
<td>0.320</td>
<td>0.082</td>
<td></td>
<td>0.241</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>0.018</td>
<td>0.298</td>
<td>0.470</td>
<td>0.024</td>
<td>0.090</td>
<td>0.475</td>
</tr>
<tr>
<td>BMI Q1</td>
<td>0.057</td>
<td>0.245</td>
<td>0.362</td>
<td>0.106</td>
<td>0.721</td>
<td>0.464</td>
</tr>
<tr>
<td>BMI Q2</td>
<td>0.955</td>
<td>0.236</td>
<td>0.263</td>
<td>0.749</td>
<td>0.556</td>
<td>0.007</td>
</tr>
<tr>
<td>BMI Q3</td>
<td>0.396</td>
<td>0.845</td>
<td>0.584</td>
<td>0.772</td>
<td>0.744</td>
<td>0.363</td>
</tr>
<tr>
<td>BMI Q4</td>
<td>0.030</td>
<td>0.878</td>
<td>0.434</td>
<td>0.262</td>
<td>0.868</td>
<td>0.071</td>
</tr>
<tr>
<td><strong>Sociodemographic Information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Income Q1</td>
<td>0.028</td>
<td>0.851</td>
<td>0.868</td>
<td>0.085</td>
<td>0.496</td>
<td>0.998</td>
</tr>
<tr>
<td>Annual Income Q2</td>
<td>0.083</td>
<td>0.430</td>
<td>0.223</td>
<td>0.101</td>
<td>0.058</td>
<td>0.681</td>
</tr>
<tr>
<td>Annual Income Q3</td>
<td>0.285</td>
<td>0.264</td>
<td>0.524</td>
<td>0.519</td>
<td>0.855</td>
<td>0.372</td>
</tr>
<tr>
<td>Annual Income Q4</td>
<td>0.314</td>
<td>0.362</td>
<td>0.425</td>
<td>0.907</td>
<td>0.996</td>
<td>0.807</td>
</tr>
<tr>
<td>No Degree</td>
<td>0.882</td>
<td>0.487</td>
<td>0.835</td>
<td>0.889</td>
<td>0.984</td>
<td>0.118</td>
</tr>
<tr>
<td>High School</td>
<td>0.403</td>
<td>&lt;.001</td>
<td>0.010</td>
<td>0.100</td>
<td>0.063</td>
<td>0.144</td>
</tr>
<tr>
<td>College</td>
<td>0.284</td>
<td>0.756</td>
<td>0.695</td>
<td>0.174</td>
<td>0.257</td>
<td>0.360</td>
</tr>
<tr>
<td>Grad School</td>
<td>0.582</td>
<td>0.161</td>
<td>0.219</td>
<td>0.557</td>
<td>0.294</td>
<td>0.719</td>
</tr>
<tr>
<td>Alcohol Use</td>
<td>0.399</td>
<td>0.074</td>
<td>0.167</td>
<td>0.361</td>
<td>0.262</td>
<td>0.032</td>
</tr>
<tr>
<td>Cocaine Use</td>
<td>0.001</td>
<td>0.332</td>
<td>0.797</td>
<td>&lt;.001</td>
<td>0.691</td>
<td>0.023</td>
</tr>
<tr>
<td>Marijuana Use</td>
<td>0.075</td>
<td>0.320</td>
<td>0.731</td>
<td>0.007</td>
<td>0.702</td>
<td>0.830</td>
</tr>
<tr>
<td><strong>Smoking Information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smokers</td>
<td>&lt;.001</td>
<td>0.003</td>
<td>0.020</td>
<td>&lt;.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years Smoked Q1</td>
<td>0.689</td>
<td>0.804</td>
<td>0.938</td>
<td>0.926</td>
<td>0.050</td>
<td>0.856</td>
</tr>
<tr>
<td>Years Smoked Q2</td>
<td>0.151</td>
<td>0.905</td>
<td>0.892</td>
<td>0.151</td>
<td>0.269</td>
<td>0.058</td>
</tr>
<tr>
<td>Years Smoked Q3</td>
<td>&lt;.001</td>
<td>0.012</td>
<td>0.151</td>
<td>&lt;.001</td>
<td>0.006</td>
<td>0.955</td>
</tr>
<tr>
<td>Years Smoked Q4</td>
<td>0.043</td>
<td>0.010</td>
<td>0.012</td>
<td>0.025</td>
<td>0.089</td>
<td>0.158</td>
</tr>
<tr>
<td>Risk Factor</td>
<td>Female</td>
<td>Male</td>
<td>White</td>
<td>Black/Other</td>
<td>Smokers</td>
<td>Non-Smokers</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>-------------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>LT 1 PD (pack/day)</td>
<td>0.007</td>
<td>0.147</td>
<td>0.250</td>
<td>0.005</td>
<td>0.570</td>
<td>0.571</td>
</tr>
<tr>
<td>GT 1 PD (pack/day)</td>
<td>&lt;.001</td>
<td>0.284</td>
<td>0.315</td>
<td>&lt;.001</td>
<td>0.190</td>
<td>0.322</td>
</tr>
<tr>
<td>Quit Months Q1</td>
<td>0.999</td>
<td>0.130</td>
<td>0.203</td>
<td>0.863</td>
<td>0.796</td>
<td></td>
</tr>
<tr>
<td>Quit Months Q2</td>
<td>0.730</td>
<td>0.523</td>
<td>0.507</td>
<td>0.703</td>
<td>0.457</td>
<td>0.040</td>
</tr>
<tr>
<td>Quit Months Q3</td>
<td>0.699</td>
<td>0.790</td>
<td>0.588</td>
<td>0.297</td>
<td></td>
<td>0.255</td>
</tr>
<tr>
<td>Quit Months Q4</td>
<td>0.284</td>
<td>0.333</td>
<td>0.288</td>
<td>0.297</td>
<td></td>
<td>0.820</td>
</tr>
<tr>
<td><strong>Respiratory Disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reoccuring Pneumonia</td>
<td>&lt;.001</td>
<td>0.041</td>
<td>0.111</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>0.081</td>
</tr>
<tr>
<td>Respiratory Disease</td>
<td>&lt;.001</td>
<td>0.013</td>
<td>0.031</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>0.449</td>
</tr>
<tr>
<td><strong>Clinical HIV Factors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aids Ever</td>
<td>&lt;.001</td>
<td>0.776</td>
<td>0.401</td>
<td>&lt;.001</td>
<td>0.005</td>
<td>0.972</td>
</tr>
<tr>
<td>CD4 Q1</td>
<td>0.898</td>
<td>0.085</td>
<td>0.113</td>
<td>0.916</td>
<td>0.558</td>
<td>0.068</td>
</tr>
<tr>
<td>CD4 Q2</td>
<td>0.019</td>
<td>0.046</td>
<td>0.261</td>
<td>0.002</td>
<td>0.009</td>
<td>0.267</td>
</tr>
<tr>
<td>CD4 Q3</td>
<td>0.010</td>
<td>0.766</td>
<td>0.319</td>
<td>0.002</td>
<td>0.064</td>
<td>0.998</td>
</tr>
<tr>
<td>CD4 Q4</td>
<td>&lt;.001</td>
<td>0.018</td>
<td>0.244</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>0.093</td>
</tr>
<tr>
<td>CD4 Fluctuations</td>
<td>0.056</td>
<td>0.300</td>
<td>0.965</td>
<td>0.010</td>
<td>0.015</td>
<td>0.977</td>
</tr>
<tr>
<td>Nadir Q1</td>
<td>0.748</td>
<td>0.518</td>
<td>0.953</td>
<td>0.379</td>
<td>0.619</td>
<td>0.364</td>
</tr>
<tr>
<td>Nadir Q2</td>
<td>0.311</td>
<td>0.099</td>
<td>0.155</td>
<td>0.202</td>
<td>0.144</td>
<td>0.215</td>
</tr>
<tr>
<td>Nadir Q3</td>
<td>0.496</td>
<td>0.383</td>
<td>0.612</td>
<td>0.281</td>
<td>0.096</td>
<td>0.383</td>
</tr>
<tr>
<td>Nadir Q4</td>
<td>0.165</td>
<td>0.034</td>
<td>0.202</td>
<td>0.023</td>
<td>0.008</td>
<td>0.617</td>
</tr>
<tr>
<td>CD4/C8 Q1</td>
<td>0.232</td>
<td>0.013</td>
<td>0.034</td>
<td>0.149</td>
<td>0.011</td>
<td>0.479</td>
</tr>
<tr>
<td>CD4/C8 Q2</td>
<td>0.208</td>
<td>0.680</td>
<td>0.705</td>
<td>0.053</td>
<td>0.809</td>
<td>0.066</td>
</tr>
<tr>
<td>CD4/C8 Q3</td>
<td>0.276</td>
<td>0.533</td>
<td>0.337</td>
<td>0.429</td>
<td>0.304</td>
<td>0.958</td>
</tr>
<tr>
<td>CD4/C8 Q4</td>
<td>0.024</td>
<td>0.011</td>
<td>0.056</td>
<td>0.007</td>
<td>0.008</td>
<td>0.134</td>
</tr>
<tr>
<td>Viral Load GT 500</td>
<td>0.161</td>
<td>0.013</td>
<td>0.017</td>
<td>0.172</td>
<td>0.307</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Multicollinearity was detected between viral load and the CD4/CD8 ratio to CD4 Q1 values. To address multicollinearity, viral load and CD4 Q1 were removed as applicable. The chi-square goodness of fit statistic, the C-statistic, quantifies the model's performance. Receiver Operating Curves (ROC) are diagnostic tool measuring the accuracy of the model by plotting sensitivity to specificity (1-Specificity). As measured by the 'C-Statistic,' the area under the curve quantifies the performance reflecting high sensitivity and specificity. The closer to 1 the C-statistic is, the greater the accuracy. The final predictors and accuracy of each of the final six models are summarized in Table 5.4 below and through Figures 5.2 through 5.7. Each model performed well with accuracy ranging from approximately 70% to 90%.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors of Lung Cancer</th>
<th>C-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>Age, Q4, Smoking &gt; 1 Pack / day, History of reoccurring pneumonia, History of respiratory disease, CD4 Q2, CD4 Q3</td>
<td>82% (Figure 5.2)</td>
</tr>
<tr>
<td>Males</td>
<td>High school education, Years smoked Q3, Years smoked Q4, History of Respiratory Disease, CD4 Q2, CD4/CD8 Q1</td>
<td>77% (Figure 5.3)</td>
</tr>
<tr>
<td>White</td>
<td>High school education, Years smoked Q4, History of Respiratory Disease, CD4/CD8 Q1</td>
<td>68% (Figure 5.4)</td>
</tr>
<tr>
<td>Black/Other</td>
<td>Non-Hispanic, Smoking &gt; 1 Pack /day, Years smoked Q3, History of Respiratory Disease, AIDS Diagnosis, CD4 Fluctuation, CD4 Q2, CD4 Q3, CD4/CD8 Q1</td>
<td>88% (Figure 5.5)</td>
</tr>
<tr>
<td>Smokers</td>
<td>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</td>
<td>73% (Figure 5.6)</td>
</tr>
<tr>
<td>Non-Smokers</td>
<td>History of Reoccurring pneumonia, Viral load &gt; 500</td>
<td>83% (Figure 5.7)</td>
</tr>
</tbody>
</table>
Figure 5.2: ROC for Final Logistic Regression Modeling Females

![ROC Curve for Selected Model](image)

Area Under the Curve = 0.8193

Figure 5.3: ROC for Final Logistic Regression Modeling Males

![ROC Curve for Selected Model](image)

Area Under the Curve = 0.7720
Figure 5.4: ROC for Final Logistic Regression Modeling White Race

Figure 5.5: ROC for Final Logistic Regression Modeling Black or Other Race
Figure 5.6: ROC for Final Logistic Regression Modeling Smokers

Figure 5.7: ROC for Final Logistic Regression Modeling Non-Smokers
Using the odds ratios determined by the final models (Table 4.1), each risk factor's risk score is multiplied across gender, race, and smoking status. 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. A female black smoker has an increased likelihood of 15.38 (Table 5.5). Likewise, a black male smoker has an increased likelihood of 12.94 (Table 5.6). The maximum summed risk scores listed in Table 5.5 and Table 5.6 reflect the highest risk scores, per model, for someone having each risk factor. For example, a female, black, smoker with each of the 14 risk factors would have a maximum risk score of 85.7, while a male, black smoker with each of the 10 risk factors would have a maximum risk score of 79.14.

Table 5.5: Multiplicative Risk Scores for Each Risk Factor by Model for Females

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>White, Smoker</th>
<th>White, Non-Smoker</th>
<th>Black, Smoker</th>
<th>Black, Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics &amp; Socio Demographic Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>1.82</td>
<td>3.66</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>AGE Q4</td>
<td>1.74</td>
<td>1.74</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>High School</td>
<td>2.38</td>
<td>2.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GT 1 PD (Pack/Day)</td>
<td>2.24</td>
<td>4.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years Smoked Q3</td>
<td>2.08</td>
<td>3.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years Smoked Q4</td>
<td>5.5</td>
<td>2.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory Disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reoccurring Pneumonia, History</td>
<td>3.44</td>
<td>8.5</td>
<td>3.44</td>
<td>8.5</td>
</tr>
<tr>
<td>Respiratory Disease, History</td>
<td>11.2</td>
<td>5.04</td>
<td>11.04</td>
<td>4.98</td>
</tr>
<tr>
<td>Clinical HIV Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIDS Ever</td>
<td>2.56</td>
<td>2.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD4 Fluctuations, Over Time</td>
<td>1.88</td>
<td>5.18</td>
<td>2.76</td>
<td></td>
</tr>
<tr>
<td>CD4 Q2</td>
<td>5.28</td>
<td>2.26</td>
<td>15.38</td>
<td>6.58</td>
</tr>
<tr>
<td>CD4 Q3</td>
<td>6.78</td>
<td>2.7</td>
<td>28.26</td>
<td>11.26</td>
</tr>
<tr>
<td>CD4/CD8 Q1</td>
<td>5.28</td>
<td>2.16</td>
<td>4.84</td>
<td>1.98</td>
</tr>
<tr>
<td>Viral Load GT 500</td>
<td>9.94</td>
<td></td>
<td>9.94</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td><strong>49.62</strong></td>
<td><strong>34.72</strong></td>
<td><strong>85.7</strong></td>
<td><strong>52.32</strong></td>
</tr>
</tbody>
</table>
The risk score's quartiles from the analytical sample of 7,587 HIV positive men and women (Figure 5.8) 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.

To better understand the sample of false positives, table 5.7 below summarizes those in which the risk score predicted lung cancer but who did not have a lung cancer diagnosis. The cohort, on average is age 37, predominantly Black (46%) and White (35%), Not Hispanic (81%), a high school degree or less (76%), Cocaine Users (64%), Smokers (62%), 20 Years of Smoking, less than one pack per day (55%), no history of reoccurring pneumonia (82%), no history of respiratory disease (75%), CD4 fluctuations over time (82%), and viral loads greater than 500 (69%). Of the 582 ‘Other’ race, 28 were American Indian or Alaska Natives, 21 were Asian or Pacific Islander, 307 were indeed ‘Other,’ and the remainder were unknown.

Table 5.7: Descriptive Statistics for False Positives.

<table>
<thead>
<tr>
<th>Variable</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics Information</strong></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>37.02 (8.51)</td>
</tr>
<tr>
<td></td>
<td>[31.00, 36.00, 42.00]</td>
</tr>
<tr>
<td>Gender (Female)</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>1467 (48.19)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>1577 (51.81)</td>
</tr>
<tr>
<td>Race</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td>1064 (34.95)</td>
</tr>
<tr>
<td></td>
<td>Black</td>
</tr>
<tr>
<td></td>
<td>1398 (45.93)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>582 (19.12)</td>
</tr>
<tr>
<td>Ethnicity (Not Hispanic)</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>592 (19.45)</td>
</tr>
<tr>
<td></td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>2452 (80.55)</td>
</tr>
<tr>
<td>BMI</td>
<td>25.15 (6.87)</td>
</tr>
<tr>
<td></td>
<td>[21.56, 23.82, 27.62]</td>
</tr>
<tr>
<td><strong>Sociodemographic Information</strong></td>
<td></td>
</tr>
<tr>
<td>Annual Income</td>
<td>19448 (18621)</td>
</tr>
<tr>
<td></td>
<td>[4998, 14998, 24998]</td>
</tr>
<tr>
<td>Education</td>
<td>No Degree</td>
</tr>
<tr>
<td></td>
<td>1534 (50.39)</td>
</tr>
<tr>
<td></td>
<td>Completed High School</td>
</tr>
<tr>
<td></td>
<td>774 (25.43)</td>
</tr>
<tr>
<td>Variable</td>
<td>False Positives</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Completed College</td>
<td>471 (15.47)</td>
</tr>
<tr>
<td>Attended/Completed Graduate School</td>
<td>265 ( 8.71)</td>
</tr>
<tr>
<td>Alcohol use</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1561 (51.30)</td>
</tr>
<tr>
<td>Y</td>
<td>1482 (48.70)</td>
</tr>
<tr>
<td>Cocaine use</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1105 (36.31)</td>
</tr>
<tr>
<td>Y</td>
<td>1938 (63.69)</td>
</tr>
<tr>
<td>Marijuana use</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1520 (49.93)</td>
</tr>
<tr>
<td>Y</td>
<td>1524 (50.07)</td>
</tr>
<tr>
<td>Smoking Information</td>
<td></td>
</tr>
<tr>
<td>Smoking Status (Smoker)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1171 (38.47)</td>
</tr>
<tr>
<td>Y</td>
<td>1873 (61.53)</td>
</tr>
<tr>
<td>Years smoked</td>
<td>19.30 ( 9.82)</td>
</tr>
<tr>
<td>[12.00, 19.00, 26.00]</td>
<td></td>
</tr>
<tr>
<td>Packs smoked per day</td>
<td></td>
</tr>
<tr>
<td>&lt; 1 per day</td>
<td>1059 (55.04)</td>
</tr>
<tr>
<td>&gt;= 1 but &lt; 2 per day</td>
<td>616 (32.02)</td>
</tr>
<tr>
<td>2 or more per day</td>
<td>249 (12.94)</td>
</tr>
<tr>
<td>Months since quit</td>
<td>114.3 (1109)</td>
</tr>
<tr>
<td>[12.00, 37.50, 99.00]</td>
<td></td>
</tr>
<tr>
<td>Respiratory Disease</td>
<td></td>
</tr>
<tr>
<td>Reoccurring Pneumonia, History of</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>2502 (82.19)</td>
</tr>
<tr>
<td>Y</td>
<td>542 (17.81)</td>
</tr>
<tr>
<td>Respiratory Disease, History of</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>2279 (74.87)</td>
</tr>
<tr>
<td>Y</td>
<td>765 (25.13)</td>
</tr>
<tr>
<td>Clinical HIV Characteristics</td>
<td></td>
</tr>
<tr>
<td>AIDS Diagnosis ever</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1468 (48.23)</td>
</tr>
<tr>
<td>Y</td>
<td>1576 (51.77)</td>
</tr>
<tr>
<td>CD4 cell counts</td>
<td>260.4 (273.2)</td>
</tr>
<tr>
<td>[37.00, 196.0, 388.0]</td>
<td></td>
</tr>
<tr>
<td>CD4, fluctuations over time</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>562 (18.46)</td>
</tr>
<tr>
<td>Y</td>
<td>2482 (81.54)</td>
</tr>
<tr>
<td>CD4, Nadir</td>
<td>141.3 (149.9)</td>
</tr>
<tr>
<td>[21.00, 92.00, 224.0]</td>
<td></td>
</tr>
<tr>
<td>CD8 cell counts</td>
<td>732.4 (509.8)</td>
</tr>
<tr>
<td>[381.0, 634.5, 956.0]</td>
<td></td>
</tr>
<tr>
<td>CD4/CD8 Ratio</td>
<td>0.38 ( 0.43)</td>
</tr>
<tr>
<td>[0.08, 0.25, 0.54]</td>
<td></td>
</tr>
<tr>
<td>Viral load greater than 500</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>943 (31.37)</td>
</tr>
<tr>
<td>Y</td>
<td>2063 (68.63)</td>
</tr>
</tbody>
</table>
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.

![Distribution of Predicted Lung Cancer Risk by Lung Cancer Diagnosis – Box Plots](image)

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.
5.2.2 Exploratory Analysis

A separate analysis explored the effects of dividing the combined Black/Other race category into two distinct race categories creating a three-level race variable (White, Black, and Other). Table 5.8 compares the bivariable logistic regression results identifying risk factors for lung cancer across the black race and other races to the combined Black/Other race used in the primary analysis. The validity of the bivariable logistic regression models for 'Other Race' was unreliable since the maximum likelihood estimate did not exist due to "Quasi-complete separation of data points," a situation when a predictor perfectly or almost entirely predicts the outcome. Quasi-separation is common when the response variable has a low count, which it does with the 'Other Race'. Similar to linear regression, which uses least-squares estimation (LSE) to produce the best estimate of a regressor's effect on its outcome, logistic regression uses maximum likelihood estimation (MLE). In these models, an MLE could not be determined. To address the validity issue, a penalized likelihood estimation method, Firth Regression, was used. Using firth regression, all models were valid and produced strong models to show which race category, black or other, contributed to the primary analysis's combined variable.

<table>
<thead>
<tr>
<th>Demographics Information</th>
<th>Black/Other</th>
<th>Black</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE Q1</td>
<td>0.003</td>
<td>0.020</td>
<td>0.213</td>
</tr>
<tr>
<td>AGE Q2</td>
<td>0.796</td>
<td>0.643</td>
<td>0.389</td>
</tr>
<tr>
<td>AGE Q3</td>
<td>0.022</td>
<td>0.065</td>
<td>0.175</td>
</tr>
<tr>
<td>AGE Q4</td>
<td>0.096</td>
<td>0.594</td>
<td>0.026</td>
</tr>
<tr>
<td>Female</td>
<td>0.038</td>
<td>0.028</td>
<td>0.410</td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>0.024</td>
<td>0.479</td>
<td>0.813</td>
</tr>
<tr>
<td>BMI Q1</td>
<td>0.106</td>
<td>0.284</td>
<td>0.166</td>
</tr>
<tr>
<td>Risk Factor</td>
<td>Black/Other</td>
<td>Black</td>
<td>Other</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>BMI Q2</td>
<td>0.749</td>
<td>0.810</td>
<td>0.984</td>
</tr>
<tr>
<td>BMI Q3</td>
<td>0.772</td>
<td>0.701</td>
<td>0.813</td>
</tr>
<tr>
<td>BMI Q4</td>
<td>0.262</td>
<td>0.322</td>
<td>0.705</td>
</tr>
</tbody>
</table>

**Sociodemographic Information**

| Annual Income Q1     | 0.085       | 0.105 | 0.994  |
| Annual Income Q2     | 0.101       | 0.091 | 0.763  |
| Annual Income Q3     | 0.519       | 0.524 | 0.953  |
| Annual Income Q4     | 0.907       | 0.651 | 0.107  |
| No Degree            | 0.889       | 0.592 | 0.155  |
| High School          | 0.100       | 0.040 | 0.280  |
| College              | 0.174       | 0.179 | 0.924  |
| Grad School          | 0.557       | 0.632 | 0.855  |
| Alcohol Use          | 0.361       | 0.303 | 0.872  |
| Cocaine Use          | <.001       | <.001 | 0.284  |
| Marijuana Use        | 0.007       | 0.017 | 0.138  |

**Smoking Information**

| Smokers              | <.001       | <.001 | 0.023  |
| Years Smoked Q1      | 0.926       | 0.653 | 0.674  |
| Years Smoked Q2      | 0.151       | 0.458 | 0.079  |
| Years Smoked Q3      | <.001       | 0.001 | 0.003  |
| Years Smoked Q4      | 0.025       | 0.104 | 0.156  |
| LT 1 PD (pack/day)   | 0.005       | 0.005 | 0.590  |
| GT 1 PD (pack/day)   | <.001       | 0.002 | 0.002  |
| Quit Months Q1       | 0.863       | 0.659 | 0.691  |
| Quit Months Q2       | 0.703       | 0.823 | 0.499  |
| Quit Months Q3       | 0.297       | 0.427 | 0.939  |
| Quit Months Q4       | 0.297       | 0.398 | 0.995  |

**Respiratory Disease**

| Reoccuring Pneumonia | <.001       | <.001 | <.001  |
| Respiratory Disease  | <.001       | 0.010 | <.001  |

**Clinical HIV Factors**

| Aids Ever            | <.001       | <.001 | 0.025  |
| CD4 Q1               | 0.916       | 0.416 | 0.049  |
| CD4 Q2               | 0.002       | 0.001 | 0.416  |
| CD4 Q3               | 0.002       | 0.001 | 0.374  |
| CD4 Q4               | <.001       | <.001 | 0.058  |
| CD4 Fluctuations     | 0.010       | 0.008 | 0.502  |
| Nadir Q1             | 0.379       | 0.698 | 0.222  |
| Nadir Q2             | 0.202       | 0.119 | 0.871  |
| Nadir Q3             | 0.281       | 0.132 | 0.905  |
A multivariable model for the other race category could not be built using any stepwise selection techniques for the same validity issues seen in the bivariable logistic regression models. Firths regression was used to remediate the validity issue and build the six independent models leading to final variable selection and risk score derivation (Table 5.9). Table 5.9 is identical to Table 4.1 on gender, smoking status, and white race. The only difference is Black/Other has been split into two race categories, black or other. The risk score was built the same as in the primary analysis; multiplicative for a specific risk factor across twelve models, instead of eight, and then summed for a single risk score.

Table 5.9: Exploratory Race Analysis Odds of Lung Cancer from Multivariable Logistic Regression

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Black/Other</th>
<th>Black</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nadir Q4</td>
<td>0.023</td>
<td>0.018</td>
<td>0.707</td>
</tr>
<tr>
<td>CD4/CD8 Q1</td>
<td>0.149</td>
<td>0.563</td>
<td>0.038</td>
</tr>
<tr>
<td>CD4/CD8 Q2</td>
<td>0.053</td>
<td>0.126</td>
<td>0.180</td>
</tr>
<tr>
<td>CD4/CD8 Q3</td>
<td>0.429</td>
<td>0.608</td>
<td>0.338</td>
</tr>
<tr>
<td>CD4/CD8 Q4</td>
<td>0.007</td>
<td>0.101</td>
<td>0.027</td>
</tr>
<tr>
<td>Viral Load GT 500</td>
<td>0.172</td>
<td>0.789</td>
<td>0.037</td>
</tr>
</tbody>
</table>

Demographics & Socio Demographic Information

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Gender</th>
<th>Race</th>
<th>Smoking Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Smoker</td>
</tr>
<tr>
<td>Not Hispanic</td>
<td></td>
<td></td>
<td>1.8107</td>
</tr>
<tr>
<td>AGE Q4</td>
<td>1.7316</td>
<td></td>
<td>4.5790</td>
</tr>
<tr>
<td>High School</td>
<td>3.0883</td>
<td>2.3661</td>
<td>1.7563</td>
</tr>
<tr>
<td>Cocaine Use</td>
<td></td>
<td></td>
<td>2.1211</td>
</tr>
</tbody>
</table>

Smoking Information

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>GT 1 PD (Pack/Day)</th>
<th>Years Smoked Q3</th>
<th>Years Smoked Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.2347</td>
<td>2.4130</td>
<td>3.3921</td>
</tr>
<tr>
<td></td>
<td>1.7767</td>
<td>2.1677</td>
<td>2.1767</td>
</tr>
<tr>
<td></td>
<td>1.7767</td>
<td>2.1677</td>
<td>4.3709</td>
</tr>
<tr>
<td></td>
<td>4.0977</td>
<td>2.0738</td>
<td>2.0309</td>
</tr>
<tr>
<td></td>
<td>4.0977</td>
<td>2.0738</td>
<td>2.0309</td>
</tr>
</tbody>
</table>
A few notable changes can be seen by cross-comparison of Table 5.9 to Table 4.1. Ethnicity (‘Not Hispanic) is no longer a risk factor. New risk factors increase the probability of lung cancer in the black race, such as cocaine use, history of reoccurring pneumonia, and high school education. Older age, such as the 4th quartile of age, increases the probability of lung cancer for other races. The remaining changes are shifts into either the black race or other race. Despite these changes, there was no significant improvement in the accuracy of the model. Sensitivity decreased slightly to 76%, and specificity increased slightly to 64%. Again, the lack of improvement is due to the low counts of subjects in the other race category.
5.3 CDSS Screen Shots

The final CDSS used 17 confidence variables, two collection variables, eight numeric variables, and 20 static list variables to fully describe eight forward chained, data-driven logic blocks and five backward chained, goal-driven logic blocks. The formation of the knowledge base and the inference engine followed all of the software development life cycle (SDLC) except integration. While the integration of the CDSS is beyond the scope of this study, Corvid's use will allow for easy integration into the IT environment of both inpatient and outpatient clinical organizations such as hospitals. It will be easy to publish the tool to a publicly available website. The tool was given a name and a brand to make it easily identifiable in preparation for its future integration. The tool is called "LUCENT," an acronym for "Lung Cancer Risk Assessment" (Figure 5.9). All interaction between the end-user and the tool were designed to provide artificially intelligent prompts and responses between LUCENT, the CDSS, and end-user as if they were interacting with a clinical decision-maker.

Figure 5.9: CDSS Name and Brand for Integration

The 47 variables and 13 logic blocks have been explained in detail in Chapter 4, Research Methodology, and explained how the CDSS, LUCENT, was developed. The design of the CDSS mirrors the workflow described in Figure 4.3. While much of the CDSS is forward chained, backward chaining is used collaboratively to derive the risk score. The inference engine will obtain the appropriate information and take the end-user down a path to derive the risk score based on their responses to the inference engine's questions. As a result, the order and experience of the CDSS will not be entirely sequential, as explained in
the workflow, and will differ for each end-user. The screenshots that follow describe each possible question the inference engine will ask, by module, to derive the risk score and provide the probability of lung cancer.

5.3.1 Screening

The primary purpose of the screening logic is to determine if the end-user meets the screening criteria and follows a simple workflow (Figure 5.10). The CDSS is intended to predict lung cancer risk, probability, for those who are living with HIV, and do not have a current or a previous diagnosis of lung cancer. First, the end-user is presented with the home page (Figure 5.11). This is the first screen they see when the CDSS is published to a website or when the tools is opened at an inpatient or outpatient medical facility. The CDSS forecasts how long it will take to complete the questions and advises them to have a copy of their most recent lab report or specifically know their CD4, CD8, and viral load values from their last medical exam. If they do not wish to continue, they are brought to a results screen (Figure 5.12) with a simple “Thank you for using LUCENT” message. If they do continue, their next question is “Are you HIV Positive?” (Figure 5.13). If they are not, they are brought to the results screen (Figure 5.14) with a simple explanation “You indicated you do not have HIV. You do not qualify for this tool”. If they do have HIV, they move to the next screen (Figure 5.15) to determine the second screening criteria and are asked “Do you currently have cancer?”. If they do not, they are brought to the results screen (Figure 5.16) with a simple message “You indicated you currently have lung cancer. You do not qualify for this tool”. If they don’t have lung cancer, they progress to the final screening criteria (Figure 5.17) and are asked “Have you ever had lung cancer?”. If they have, they are brought to the results screen (Figure 5.18) with a simple message “You indicated you have previously had lung cancer. You do not qualify for this tool”. If they have not had a previous diagnosis of
lung cancer, they have met all of the screening criteria and will advance to answer basic demographic information in the next block of logic. All end-users who either choose not to use the tool or do not meet the screening criteria have a risk score of 0 and are flagged as ‘Not Assessable’ in the results screen (Figures 5.12, 5.14, 5.16, and 5.18).

**Figure 5.10: Workflow - Screening**

![Workflow - Screening](image1)

**Figure 5.11: CDSS Screenshot: Main Screen**

This survey will take approximately 5 minutes to complete.

You should have a copy of your most recent laboratory reports. Specifically, you should know your CD4, CD8, and Viral Load counts.

**Would you like to Continue?**

- Yes
- No

**Figure 5.12: CDSS Screenshot: Screening Criteria Not Met – Does not want to continue**

**Results:**

Your risk score is 0.

**Risk Score Level: Not Assessable**

LUCENT used your responses and scored your risk for lung cancer based on ....

*Based on your answers here are general recommendations:*

Thank you for using LUCENT!
Figure 5.13: CDSS Screenshot: HIV Positive

*LUCENT first needs to determine if you qualify for the tool...*

**Screening Information:**

Are you HIV Positive?

- Yes
- No

Figure 5.14: CDSS Screenshot: Screening Criteria Not Met – HIV Negative

**Results:**

Your risk score is 0.

**Risk Score Level: Not Assessable**

LUCENT used your responses and scored your risk for lung cancer based on ....

*Based on your answers here are general recommendations:*

You indicated you do not have HIV. You do not qualify for this tool.

Figure 5.15: CDSS Screenshot: Currently have Lung Cancer

**Screening Information:**

Do you currently have lung cancer?

- Yes
- No

Figure 5.16: CDSS Screenshot: Screening Criteria Not Met – Current Lung Cancer

**Results:**

Your risk score is 0.

**Risk Score Level: Not Assessable**

LUCENT used your responses and scored your risk for lung cancer based on ....

*Based on your answers here are general recommendations:*

You indicated you currently have lung cancer. You do not qualify for this tool.

Figure 5.17: CDSS Screenshot: Ever had lung cancer

**Screening Information:**

Have you ever had lung cancer?

- Yes
- No
5.2.3 Demographics

Gender (Female or Male) and Race (White, or Black/Other) are two critical variables in the CDSS. They are crucial primarily because the risk score comprises fourteen risk factors multiplicative across six strata of gender, race, and smoking status. The workflow describes them separately to explain the flow of questions (Figure 5.19) and highlight their importance. The colors in the workflow identify the independent paths used to describe risk factors specific to females (purple), males (blue), white race (orange), and black or other race (green).

There are different risk factors for lung cancer in PWH across female, male, white, and black or other races. The inference engine asks additional demographic questions in the workflow to determine the demographic risk factor's existence and establish the probability of lung cancer; however, there is only one demographic logic block in the CDSS. The inference engine recursively tests gender and race through many inference engine rules; however, the inference engine collects their values once. Gender is asked first (Figure 5.20) as "What is your gender?", followed by race (Figure 5.21) as "What is your race?". With these two questions answered, the CDSS moves to the next block of logic to determine smoking status. The inference engine must determine the end-users' smoking status before asking the additional demographic question. This is because demographic risk factors will differ by smoking status.
5.2.4 Smoking Status

This study defined smoking status as 'Smoker' or 'Non-Smoker.' This is due to the high prevalence of smoking in the knowledge base; however, this also aligns with the prevalence of smoking the literature review identified and NIH HIV registry studies reported. Studies have reported rates between 80 and 100% prevalence of smoking in PWH. The inference engine seeks to derive the end-users smoking status using three questions as seen in the workflow (Figure 5.22). If the end-user currently smokes, their status is 'Smoker'; however, if the end-user is not a current smoker, the inference engine will require the end-users smoking history. Former smokers who quit smoking within the last 24 months are categorized as a 'Smoker' while those who quit smoking more than 24 months prior are classified as a 'Non-Smoker.' Smoking status is crucial because the risk score comprises
fourteen risk factors multiplicative across six strata of gender, race, and smoking status. The workflow describes them separately to explain the flow of questions and highlight their importance. The colors in the workflow identify the independent paths used to describe risk factors specific to smokers (red) and non-smokers (yellow).

The inference engine assesses their current smoking status (Figure 5.23), asking, "Do you currently smoke cigarettes?". If they do, they are a 'Smoker,' and the inference engine does not ask the remaining smoking status questions. If they do not, then the inference engine assesses their smoking history (Figure 5.24), asking, "Have you ever smoked cigarettes?". If they have not, then the inference engine does not request any further smoking status questions, and they are a 'Non-Smoker.' If they have smoked previously, then the inference engine assesses how many months it is has been since they quit (Figure 5.25). If it has been less than 24 months, they are a 'Smoker' but, if it has been more than 24 months, they are a 'Non-Smoker.' The end-user will not know their classification until the results are displayed. The inference engine will use the derived categorization of smoking to determine the existence of smoking specific risk factors and predicts lung cancer.
5.2.5 Smoking Information

With gender, race, and smoking status collected by the inference engine, the CDSS will identify additional smoking risk factors. Specifically, the inference engine asks about the number of packs smoked per day and the years of smoking (Figure 5.26). Using the colored lines and arrows in the workflow makes it easy to see which data the inference engine will collect and ask the end-user. For example, the inference engine asks a white, smoking female both questions while it only asks one to a white, smoking male. Packs per day is only a risk factor for females, so they are asked, "How many packs per day do you smoke?" (Figure 5.27). Years smoked is a risk factor for males, white races, black or other races, and smokers but not for females and smokers, so they are asked, "How many years have you smoked?" (Figure 5.28). With answers to each question, the inference engine intelligently calculates each risk factor's probability of lung cancer. They will be revealed to the end-user in the results screen when all data has been retrieved.
5.2.6 Demographics Continued

Age, Education, and Ethnicity are significant risk factors for lung cancer across gender, race, and ethnicity for PWH (Figure 5.29). The workflows show these questions separate from gender and race; however, there is one logic block used by the inference engine to collect demographic information. The inference engine gathers information as it is needed to test logic rules. As a result, the inference engine will ask gender and race earlier than age, education, and ethnicity.

The inference engine will ascertain the values for age, education, and ethnicity depending on the end-users gender, race, and smoking status. Females are asked, "What is your age (in years)?" (Figure 5.30). The inference engine asks males and white races, "What is your highest level of education (diploma received)?" (Figure 5.31) to determine education. Likewise, it asks smokers and black or other races, "Are you Hispanic or Latino?" (Figure 5.32) to determine ethnicity. Once again, the inference engine intelligently calculates each
risk factor's probability of lung cancer. They will be revealed to the end-user in the results screen when all data has been retrieved.

5.2.7 Respiratory Disease

The respiratory disease logic gathers information about significant respiratory diseases that increase lung cancer probability for PWH. There are two risk factors, a history of lung disease and reoccurring pneumonia (Figure 5.33). The inference engine asks all six strata (females, males, white, black or other races, smokers and non-smokers) the respiratory disease question, "Have you ever been diagnosed with a respiratory disease?" (Figure 5.34).
The inference engine prefaces the question with examples for the end-user to consider as respiratory disease. These examples are informed by the knowledge-base using the MACS and WIHS. The second respiratory question asks women, smokers, and non-smokers, "Have you had pneumonia 2 or more times in the past?" (Figure 5.35). As in other logic, the inference engine intelligently calculates each risk factor’s probability of lung cancer. They will be revealed to the end-user in the results screen when all data has is retrieved.

![Figure 5.33: Workflow - Respiratory Disease](image)

![Figure 5.34: CDSS Screenshot: History of Respiratory Disease](image)

**History of Respiratory Disease:**

Examples of respiratory disease include, but are not limited to:

1. Acute Respiratory Infections
2. Chronic Obstructive Pulmonary Disease (COPD)
3. Tuberculosis
4. MAI/MAC Infection
5. Pneumococci and Other Lung Diseases due to external agents
6. Candida in esophagus or lungs
7. Herpes simplex in lungs
8. Cytomegalovirus in lungs
9. Other diseases of upper or lower respiratory tract

Have you ever been diagnosed with a respiratory disease?

- Yes
- No

![Figure 5.35: CDSS Screenshot: History of Reoccurring Pneumonia](image)

**History of Respiratory Disease:**

Have you had pneumonia 2 or more times in the past?

- Yes
- No
5.2.8 Clinical HIV Characteristics

The inference engine's final sets of questions to determine lung cancer risk factors resides in the logic for clinical HIV characteristics. Depending on the end-users gender, race, and smoking status, the inference engine tests the presence of five different risk factors (Figure 5.36). The only clinical HIV characteristic to increase the probability of lung cancer for non-smokers is their viral load. Non-smokers are asked, "What is/was you're your most recent viral load?" (Figure 5.37). Women, men, smokers, white, black, and other races are all asked about their most recent CD4 value, "What is/was your most recent CD4 value?" (Figure 5.38). All but females will progress to answer, "What is/was your most recent CD8 value?" (Figure 5.39). The inference engine uses CD8 to calculate the CD4/CD8 ratio, a significant lung cancer predictor. Smokers, black and other races, answer a question, "Over the history of your HIV diagnosis, how would you classify your CD4 fluctuations?" (Figure 5.40). Those with fluctuations have a higher risk of lung cancer compared to those who do not. Finally, the inference engine will ask blacks and other races, "Have you ever been diagnosed with AIDS?" (Figure 5.41). The inference engine reminds the end-user what the definition of AIDS is in the prompt. The inference engine intelligently calculates each risk factor's probability of lung cancer and is revealed next in the results screen.

Figure 5.36: Workflow - Clinical HIV Characteristics
Figure 5.37: CDSS Screenshot: Viral Load

**Clinical HIV Characteristics:**

What is/was your most recent viral load?

---

Figure 5.38: CDSS Screenshot: CD4 value

**Clinical HIV Characteristics:**

What is/was your most recent CD4 value?

---

Figure 5.39: CDSS Screenshot: CD8 value

**Clinical HIV Characteristics:**

What is/was your most recent CD8 value?

---

Figure 5.40: CDSS Screenshot: CD4 Fluctuations

**Clinical HIV Characteristics:**

Over the history of your HIV diagnosis, how would you classify your CD4 fluctuations?

- Always less than 200
- Always between 200 and 500
- Always greater than 500
- Never the same - Fluctuates

---

Figure 5.41: CDSS Screenshot: History of AIDS Diagnosis

**Clinical HIV Characteristics:**

An AIDS definition, per the CDC, is defined as:

1.) CD4 cells fall below 200 cells per cubic millimeter of blood (200 cells/mm³) OR
2.) You developed one of the below opportunistic infections regardless of your CD4 cell count.

List of opportunistic infections:
- Candidiasis
- Invasive cervical cancer
- Coccidioidomycosis
- Cryptococcosis
- Cryptosporidiosis (Crypto)
- Cytomegalovirus
- Encephalopathy, HIV-related
- Herpes Simplex Virus (HSV)
- Histoplasmosis
- Kaposi's sarcoma (KS)
- Lymphoma
- Tuberculosis (TB)
- Mycobacterium avium complex (MAC)
- Pneumocystis pneumonia (PCP)
- Pneumonia
- Progressive multifocal leukoencephalopathy
- Salmonella septicemia
- Toxoplasmosis
- Wasting syndrome due to HIV

Have you ever been diagnosed with AIDS?

- Yes
- No
5.2.9  Results Screen

The results screen is only presented to the end-user when the system stops processing, as it does in the screening logic, or when the inference engine has asked the questions required to create the risk score based on gender, race, and smoking status (Figure 5.42). The results screen provides detailed information to the end-user to help them understand their risk score. It will enable a deeper understanding of their risk factors, but more importantly, it is a communication tool for them and their clinicians.

The results screen provides five levels of information (Figure 5.43). First, it provides a quantitative risk score. Second, and more importantly, it gives their risk strata a qualitative classification as low, medium, or high risk. Third, it explains each of the risk factors contributing to their risk scores with the probabilities for each risk factor. These explanations will be different for every end-user, but the inference engine is intelligent enough to modify the screen using each end-users input. Key immunocompetency risks will trigger guidance about improving their CD4 values. The results point them to an instrument to help improve their CD4 values and encourages them to start a conversation with their clinician. Finally, the fifth piece of the results is guidance on smoking cessation for smokers. The results suggest visiting the CDC guidance on smoking cessation and encourage them to have a conversation with their clinician. The results is an HTML page and are easily printable or downloadable depending on their computing settings. It contains all of the information required to assess the risk of lung cancer with a clinician.
5.4 Patient Narratives

Twenty subjects were reserved from the analytical sample used to develop the knowledge-base and inform the inference engine's rules. For each subject, risk factors experienced in the MACS or WIHS are summarized and compared to the results from the CDSS developed by this study.
Subject M-1017 is a black, male, non-smoker. He is 31 years old, not Hispanic, and has a college education. He is a former smoker but quit 216 months ago. As a result, he is classified as a non-smoker. He has no history of pneumonia and does have a history of respiratory disease. Subject M-1017 does not have a history of AIDS. His CD4 value is 165 and is always less than 200. His CD8 value is 839 making the CD4/CD8 ratio 0.2. He reported his viral load at 6,457.

The CDSS calculated its probability based on him being a black male non-smoker (Figure 5.44). Based on risk factors and his responses, CD4 values in the second quartile, non-Hispanic ethnicity, history of respiratory disease, and viral load increased his lung cancer risk. They determined him to have a medium risk of developing cancer. In the MACS database, subject M-1017 does not have a lung cancer diagnosis. While this was not the expected outcome, if the patient doesn’t address these risk factors, he has a medium risk for developing lung cancer in the future. To assist in mitigating risk, the CDSS also suggested he speak with his clinician to manage his low CD4 count.

Figure 5.44: Results Screen from Patient Narrative 1 – Subject M-1017

**Results:**
Your risk score is 22.51.

**Risk Score Level: Medium**
LUCENT used your responses and scored your risk for lung cancer based on ....

Your CD4 value (165.0) is between 78 and 250. This increased your risk of lung cancer by 5.529 times. Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 2.02 times. A history of respiratory disease increased your risk of lung cancer 5.019 times. Your viral load (6457.0) is greater than 500 and increased your risk of lung cancer 9.94 times.

**Based on your answers here are general recommendations:**
Clinical HIV Guidance:
LUCENT as identified you as someone with low CD4 cell counts (between 78 and 250). This is a significant risk factor for a lung cancer diagnosis. CD4 cell counts should be at or above 500. Please speak with your general practitioner.
Subject W-2547 is a white, female, non-smoker. She is 31 years old, Hispanic, and has not completed any education. She has never smoked. As a result, she is classified as a non-smoker. She has no history of pneumonia and no history of respiratory disease. Subject W-2547 does not have a history of AIDS. Her CD4 value is 9 and is always less than 200. Her CD8 value is 133 making the CD4/CD8 ratio 0.1. She reported her viral load at 941,000.

The CDSS calculated its probability based on her being a white, female, non-smoker (Figure 5.45). Based on risk factors and her responses, a low CD4/CD8 ratio and high viral load increased her lung cancer risk. They determined her to have a low risk of developing cancer. In the WIHS database, subject W-2547 does not have a lung cancer diagnosis, the expected outcome for low risk. The CDSS also suggested she speak with her clinician to manage her CD4/CD8 ratio.

Figure 5.45: Results Screen from Patient Narrative 2 – Subject W-2547

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 12.09.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ...</td>
</tr>
</tbody>
</table>
Based on your CD4 (9.0) and CD8 (133.0), your CD4/CD8 ratio (0.07) is significantly low. This increased your risk of lung cancer 2.15 times.
Your viral load (941000.0) is greater than 500 and increased your risk of lung cancer 9.94 times.

Based on your answers here are general recommendations:

LUCENT Clinical HIV Guidance:

LUCENT has calculated your ratio of CD4 and CD8 ratio based on your CD4 of 9.0 and your CD8 of 133.0. Your ratio 0.068 is significantly lower than the 1.0 threshold.
Please speak with your general practitioner about how to raise your CD4 values.

For more educational material, please see the following website:
https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count
Subject M-1040 is a black male smoker. He is 24 years old, not Hispanic, and has a graduate degree. He is a current smoker. He smokes one pack per day and has smoked for three years. As a result, he is classified as a smoker. He has no history of pneumonia but does have an account of respiratory disease. Subject M-1040 does not have a history of AIDS. His CD4 value is 1182 and is never the same. It fluctuates. His CD8 value is 1100 making the CD4/CD8 ratio 1.1. He reported his viral load at 10.

The CDSS calculated its probability based on him being a black male smoker (Figure 5.46). Based on risk factors and his responses, fluctuating CD4 values, non-Hispanic ethnicity, and respiratory disease history increased his lung cancer risk. They determined him to have a medium risk of developing cancer. In the MACS database, subject M-1040 does not have a lung cancer diagnosis. While this was not the expected outcome, if the patient doesn’t address these risk factors, he has a medium risk for developing lung cancer in the future. To assist in mitigating risk, the CDSS also guided smoking cessation.

![Figure 5.46: Results Screen from Patient Narrative 3 – Subject M-1040](image_url)

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 19.97.</td>
</tr>
</tbody>
</table>

**Risk Score Level: Medium**

LUCENT used your responses and scored your risk for lung cancer based on ....

- Fluctuating CD4 values increased your risk of lung cancer by 5.17 times.
- Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 3.656 times.
- A history of respiratory disease increased your risk of lung cancer 11.142 times.

**Based on your answers here are general recommendations:**

LUCENT Smoking Cessation Guidance:

- You have been identified as a smoker.
- Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.

For help with smoking cessation, please visit the CDC website. [https://www.cdc.gov/tobacco/campaign/lips/quit-smoking/index.html](https://www.cdc.gov/tobacco/campaign/lips/quit-smoking/index.html)
Subject W-2825 is a black female smoker. She is 40 years old, not Hispanic, and has high school education. She is a current smoker. She smokes 1 pack per day and has smoked for 22 years. As a result, she is classified as a smoker. She has no history of pneumonia and no history of respiratory disease. Subject W-2825 does not have an account of AIDS. Her CD4 value is 1114 and is never the same. It fluctuates. Her CD8 value is 1200 making the CD4/CD8 ratio 0.9. She reported her viral load at 112.

The CDSS calculated its probability based on her being a black female smoker (Figure 5.47). Based on risk factors and her responses, fluctuating CD4 values, non-Hispanic ethnicity, and years smoked in the third quartile increased her lung cancer risk. They determined her to have a low risk of developing cancer. In the WIHS database, subject W-2825 does not have a lung cancer diagnosis, the expected outcome for low risk. The CDSS also guided smoking cessation.

**Figure 5.47: Results Screen from Patient Narrative 4 – Subject W-2825**

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 12.37.</td>
</tr>
<tr>
<td>Risk Score Level: Low</td>
</tr>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Fluctuating CD4 values increased your risk of lung cancer by 5.17 times.</td>
</tr>
<tr>
<td>Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 3.656 times.</td>
</tr>
<tr>
<td>You have smoked for 22.0 years. Since this is between 21 and 28, your risk of lung cancer increased 3.54 times.</td>
</tr>
</tbody>
</table>

**Based on your answers here are general recommendations:**

<table>
<thead>
<tr>
<th>LUCENT Smoking Cessation Guidance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>You have been identified as a smoker.</td>
</tr>
<tr>
<td>Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.</td>
</tr>
</tbody>
</table>

For help with smoking cessation, please visit the CDC website.  
https://www.cdc.gov/tobacco/campaign/lips/quit-smoking/index.html
Subject M-1154 is a white, male, non-smoker. He is 27 years old, not Hispanic, and has not completed any education. He has never smoked. As a result, he is classified as a non-smoker. He has a history of pneumonia two or more times and no history of respiratory disease. Subject M-1154 does not have a history of AIDS. His CD4 value is 21 and is never the same. It fluctuates. His CD8 value is 180 making the CD4/CD8 ratio 0.1. He reported his viral load at 988,097.

The CDSS calculated its probability based on him being a white, male, non-smoker (Figure 5.48). Based on risk factors and his responses, a low CD4/CD8 ratio, history of reoccurring pneumonia, and high viral load increased his lung cancer risk. They determined him to have a medium risk for developing cancer. In the MACS database, subject M-1154 has a lung cancer diagnosis, the expected outcome for medium risk. The CDSS also suggests he speak with his clinician about his low CD4/CD8 ratio.

Figure 5.48: Results Screen from Patient Narrative 5 – Subject M-1154

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 19.24.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on.....</td>
</tr>
</tbody>
</table>

Based on your CD4 (21.0) and CD8 (180.0), your CD4/CD8 ratio (0.12) is significantly low. This increased your risk of lung cancer 4.923 times. A history of reoccurring pneumonia increased your risk of lung cancer 4.38 times. Your viral load (988097.0) is greater than 500 and increased your risk of lung cancer 9.94 times.

**Based on your answers here are general recommendations:**

LUCENT Clinical HIV Guidance:

LUCENT has calculated your ratio of CD4 and CD8 ratio based on your CD4 of 21.0 and your CD8 of 180.0. Your ratio 0.117 is significantly lower than the 1.0 threshold. Please speak with your general practitioner about how to raise your CD4 values.

For more educational material, please see the following website:
https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count
Subject W-3274 is a black, female, non-smoker. She is 35 years old, not Hispanic, and has high school education. She has never smoked. As a result, she is classified as a non-smoker. She has no history of pneumonia and no history of respiratory disease. Subject W-3274 does not have an account of AIDS. Her CD4 value is 852 and is never the same. It fluctuates. Her CD8 value is 598 making the CD4/CD8 ratio 1.4. She reported her viral load at 20.

The CDSS calculated its probability based on her being a black, female, non-smoker (Figure 5.49). Based on risk factors and her responses, fluctuating CD4 values and non-Hispanic ethnicity increased her risk of lung cancer but determined her to have a low risk for developing lung cancer. In the WIHS database, subject W-3274 does not have lung cancer, the expected outcome for low risk.

Figure 5.49: Results Screen from Patient Narrative 6 – Subject W-3274

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 4.77.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Fluctuating CD4 values increased your risk of lung cancer by 2.75 times.</td>
</tr>
<tr>
<td>Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 2.02 times.</td>
</tr>
</tbody>
</table>
Subject M-1797 is a white male smoker. He is 33 years old, not Hispanic, and has not completed any education. He is a current smoker. He smokes 4 packs per day and has smoked for 22 years. As a result, he is classified as a smoker. He has no history of pneumonia and does have a history of respiratory disease. Subject M-1797 does not have a history of AIDS. His CD4 value is 12 and is never the same. It fluctuates. His CD8 value is 140 making the CD4/CD8 ratio 0.1. He reported his viral load at 112,692.

The CDSS calculated its probability based on him being a white male smoker (Figure 5.50). Based on risk factors and his responses, a fluctuating CD4 count, low CD4/CD8 ratio, non-Hispanic ethnicity, history of respiratory disease, and years smoked in the third quartile increased his lung cancer risk. They determined him to have a medium risk for developing lung cancer. In the MACS database, subject M-1797 has a lung cancer diagnosis, the expected outcome for medium risk. The CDSS also guided smoking cessation and suggested he speak with his clinician about managing his low CD4/CD8 ratio.

Figure 5.50: Results Screen from Patient Narrative 7 – Subject M-1797

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 32.09</td>
</tr>
<tr>
<td>Risk Score Level: Medium</td>
</tr>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on....</td>
</tr>
<tr>
<td>Fluctuating CD4 values increased your risk of lung cancer by 1.88 times.</td>
</tr>
<tr>
<td>Based on your CD4 (12.0) and CD8 (140.0), your CD4/CD8 ratio (0.09) is significantly low. This increased your risk of lung cancer 12.112 times.</td>
</tr>
<tr>
<td>Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 1.81 times.</td>
</tr>
<tr>
<td>A history of respiratory disease increased your risk of lung cancer 11.301 times.</td>
</tr>
<tr>
<td>You have smoked for 22.0 years. Since this is between 21 and 28, your risk of lung cancer increased 4.989 times.</td>
</tr>
<tr>
<td>Based on your answers here are general recommendations:</td>
</tr>
<tr>
<td>LUCENT Clinical HIV Guidance:</td>
</tr>
<tr>
<td>LUCENT has calculated your ratio of CD4 and CD8 ratio based on your CD4 of 12.0 and your CD8 of 140.0. Your ratio 0.065 is significantly lower than the 1.0 threshold. Please speak with your general practitioner about how to raise your CD4 values.</td>
</tr>
<tr>
<td>For more educational material, please see the following website:</td>
</tr>
<tr>
<td><a href="https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count">https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count</a></td>
</tr>
<tr>
<td>LUCENT Smoking Cessation Guidance:</td>
</tr>
<tr>
<td>You have been identified as a smoker.</td>
</tr>
<tr>
<td>Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.</td>
</tr>
<tr>
<td>For help with smoking cessation, please visit the CDC website.</td>
</tr>
<tr>
<td><a href="https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html">https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html</a></td>
</tr>
</tbody>
</table>
Subject W-3472 is a black female smoker. She is 38 years old, not Hispanic, and has not completed any education. She currently smokes 2 packs per day and has smoked for 15 years. She has a history of pneumonia 2 or more times and has a history of respiratory disease. She does not have an account of AIDS. Her CD4 value is 398 and is never the same. Her CD8 value is 810 making the CD4/CD8 ratio 0.5. She reported her viral load at 20. The CDSS calculated its probability based on her being a black female smoker (Figure 5.51). Based on risk factors and her responses, fluctuating CD4 values, CD4 values in the third quartile, non-Hispanic ethnicity, packs smoked per day, a history of respiratory disease and a history of reoccurring pneumonia increased her lung cancer risk. They determined her to have a high risk of developing cancer. In the WIHS database, subject W-3472 does not have a lung cancer diagnosis. While this was not the expected outcome, if the patient doesn’t address these significant risk factors, she has a high risk for developing lung cancer in the future. To assist in mitigating risk, the CDSS also guided smoking cessation and suggested she speak with her clinician about her low CD4 count.

Figure 5.51: Results Screen from Patient Narrative 8 – Subject W-3472

Results:
Your risk score is 55.58.

Risk Score Level: High
LUCENT used your responses and scored your risk for lung cancer based on ....
Fluctuating CD4 values increased your risk of lung cancer by 5.17 times.
Your CD4 value (398.0) is between 250 and 446. This increased your risk of lung cancer 28.26 times.
Having an ethnicity that is not hispanic or latino increased your risk of lung cancer 3.656 times.
You said you smoked 2.0 packs per day.
Smoking more than 1 pack a day increased your risk of lung cancer 4.014 times.
A history of reoccurring pneumonia increased your risk of lung cancer 3.434 times.
A history of respiratory disease increased your risk of lung cancer 11.049 times.

Based on your answers here are general recommendations:

Clinical HIV Guidance:
LUCENT as identified you as someone with low CD4 cell counts (between 250 and 446).
This is a significant risk factor for a lung cancer diagnosis.
CD4 cell counts should be at or above 500.
Please speak with your general practitioner.

LUCENT Smoking Cessation Guidance:
You have been identified as a smoker.
Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.
For help with smoking cessation, please visit the CDC website.
https://www.cdc.gov/tobacco/campaign/tips/quilt-smoking/index.html
Subject M-3682 is a white, male, non-smoker. He is 29 years old, not Hispanic, and has a graduate degree. He has never smoked. As a result, he is classified as a non-smoker. He has no history of pneumonia and no history of respiratory disease. Subject M-3682 does have a history of AIDS. His CD4 value is 23 and is never the same. It fluctuates. His CD8 value is 108 making the CD4/CD8 ratio 0.2. He reported his viral load at 280,987.

The CDSS calculated its probability based on him being a white male non-smoker (Figure 5.52). Based on risk factors and his responses, a high viral load increased his lung cancer risk. It determined him to have a low risk of developing cancer. In the MACS database, subject M-3682 has a lung cancer diagnosis. While this was not the expected outcome, if the patient doesn’t address these risk factors, he has a low risk of developing lung cancer in the future.

Figure 5.52: Results Screen from Patient Narrative 9 – Subject M-3682

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 9.94.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Your viral load (280987.0) is greater than 500 and increased your risk of lung cancer 9.94 times.</td>
</tr>
</tbody>
</table>
Subject W-4017 is a black female smoker. She is 36 years old, not Hispanic, and has not completed any education. She is a current smoker. She smokes 3 packs per day and has smoked for 17 years. As a result, she is classified as a smoker. She has a history of pneumonia 2 or more times and no history of respiratory disease. Subject W-4017 does have an account of AIDS. Her CD4 value is 338 and is never the same. It fluctuates. Her CD8 value is 776 making the CD4/CD8 ratio 0.4. She reported her viral load at 80.

The CDSS calculated its probability based on her being a black female smoker (Figure 5.53). Based on risk factors and her responses, a previous history of AIDS, fluctuating CD4 values, CD4 values in the third quartile, non-Hispanic ethnicity, packs smoked per day, and history of reoccurring pneumonia increased her lung cancer risk. They determined her to have a high risk of developing cancer. In the WIHS database, subject W-4017 has a lung cancer diagnosis, the expected outcome for high risk. The CDSS also guided smoking cessation and suggested she speak with her clinician about her low CD4 count.

Figure 5.53: Results Screen from Patient Narrative 10 – Subject W-4017

Results:

Your risk score is 47.08.

Risk Score Level: High

LUCENT used your responses and scored your risk for lung cancer based on ....

Based on your answers here are general recommendations:

Clinical HIV Guidance:

LUCENT as identified you as someone with low CD4 cell counts (between 250 and 446). This is a significant risk factor for a lung cancer diagnosis. CD4 cell counts should be at or above 500. Please speak with your general practitioner.

LUCENT Smoking Cessation Guidance:

You have been identified as a smoker. Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.

For help with smoking cessation, please visit the CDC website.
https://www.cdc.gov/tobacco/campaign/tips/quitsmoking/index.html
Subject M-6107 is a white, male, non-smoker. He is 30 years old, not Hispanic, and has a college education. He has never smoked. As a result, he is classified as a non-smoker. He has no history of pneumonia and no history of respiratory disease. Subject M-6107 does not have a history of AIDS. His CD4 value is 274 and is never the same. It fluctuates. His CD8 value is 616 making the CD4/CD8 ratio 0.4. He reported his viral load at 66,019.

The CDSS calculated its probability based on him being a white, male, non-smoker (Figure 5.54). Based on risk factors and his responses, a high viral load increased his lung cancer risk. It determined him to have a low risk of developing cancer. In the MACS database, subject M-6107 has a lung cancer diagnosis.

Figure 5.54: Results Screen from Patient Narrative 11 – Subject M-6107

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 9.94.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Your viral load (66019.0) is greater than 500 and increased your risk of lung cancer 9.94 times.</td>
</tr>
</tbody>
</table>

**Based on your answers here are general recommendations:**

**Clinical HIV Guidance:**

LUCENT as identified you as someone with low CD4 cell counts (between 250 and 446). This is a significant risk factor for a lung cancer diagnosis. CD4 cell counts should be at or above 500. Please speak with your general practitioner.
Subject W-4745 is a black female smoker. She is 52 years old, not Hispanic, and has not completed any education. She is a current smoker. She smokes 2 packs per day and has smoked for 31 years. As a result, she is classified as a smoker. She has a history of pneumonia 2 or more times and does have an account of respiratory disease. Subject W-4745 does not have a history of AIDS. Her CD4 value is 464 and is never the same. It fluctuates. Her CD8 value is 696 making the CD4/CD8 ratio 0.7. She reported her viral load at 20.

The CDSS calculated its probability based on her being a black female smoker (Figure 5.55). Based on risk factors and her responses, age greater than 44, fluctuating CD4 values, non-Hispanic ethnicity, packs smoked per day, history of both respiratory disease and reoccurring pneumonia, and years smoked in the fourth quartile increased her lung cancer risk. They determined her to have a medium risk for developing cancer. In the WIHS database, subject W-4745 has a lung cancer diagnosis, the expected outcome for medium risk. The CDSS also guided smoking cessation.

Figure 5.55: Results Screen from Patient Narrative 12 – Subject W-4745

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 31.08.</td>
</tr>
</tbody>
</table>

**Risk Score Level: Medium**

LUCENT used your responses and scored your risk for lung cancer based on ....

- Your age (52.0), is greater than 44 and increased your risk of lung cancer by 1.73 times.
- Fluctuating CD4 values increased your risk of lung cancer by 5.17 times.
- Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 3.656 times.
- You said you smoked 2.0 packs per day.
- Smoking more than 1 pack a day increased your risk of lung cancer 4.014 times.
- A history of reoccurring pneumonia increased your risk of lung cancer 3.434 times.
- A history of respiratory disease increased your risk of lung cancer 11.049 times.
- You have smoked for 31.0 years. Since this is greater than 28, your risk of lung cancer increased 2.03 times.

**Based on your answers here are general recommendations:**

LUCENT Smoking Cessation Guidance:

- You have been identified as a smoker.
- Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.

For help with smoking cessation, please visit the CDC website. [https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html](https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html)
Subject M-6732 is a white, male, non-smoker. He is 30 years old, not Hispanic, and has high school education. He has never smoked. As a result, he is classified as a non-smoker. He has no history of pneumonia and no history of respiratory disease. Subject M-6732 does not have a history of AIDS. His CD4 value is 406 and is never the same. It fluctuates. His CD8 value is 959 making the CD4/CD8 ratio 0.4. He reported his viral load at 65.

The CDSS calculated its probability based on him being a white, male, non-smoker (Figure 5.56). Based on risk factors and his responses, having only a high-school education increased his risk of lung cancer but determined him to have a low risk for developing lung cancer. In the MACS database, subject M-6732 does not have lung cancer, the expected outcome for low risk. The CDSS suggests he speak with his clinician to manage his CD4 count even though it wasn’t a significant risk factor.

Figure 5.56: Results Screen from Patient Narrative 13 – Subject M-6732

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 7.32.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Having only a high school education increased your risk of lung cancer 7.323 times.</td>
</tr>
</tbody>
</table>

**Based on your answers here are general recommendations:**

**Clinical HIV Guidance:**

LUCENT as identified you as someone with low CD4 cell counts (between 250 and 446). This is a significant risk factor for a lung cancer diagnosis. CD4 cell counts should be at or above 500. Please speak with your general practitioner.
Subject W-9445 is a black female smoker. She is 38 years old, not Hispanic, and has not completed any education. She is a current smoker. She smokes 2 packs per day and has smoked for 5 years. As a result, she is classified as a smoker. She has no history of pneumonia and no history of respiratory disease. Subject W-9445 does not have an account of AIDS. Her CD4 value is 159 and is never the same. It fluctuates. Her CD8 value is 300 making the CD4/CD8 ratio 0.5. She reported her viral load at 48.

The CDSS calculated its probability based on her being a white female smoker (Figure 5.57). Based on risk factors and her responses, fluctuating CD4 values, CD4 counts in the second quartile, non-Hispanic ethnicity, and packs smoked per day increased her lung cancer risk. They determined her to have a medium risk for developing cancer. In the WIHS database, subject W-9445 has a lung cancer diagnosis, the expected outcome for medium risk. The CDSS also guided smoking cessation and suggested she speak with her clinician about her low CD4 count.

Figure 5.57: Results Screen from Patient Narrative 14 – Subject W-9445

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 28.23.</td>
</tr>
</tbody>
</table>

**Risk Score Level: Medium**

- Fluctuating CD4 values increased your risk of lung cancer by 5.17 times.
- Your CD4 value (159.0) is between 78 and 250. This increased your risk of lung cancer by 15.389 times.
- Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 3.656 times.
- You said you smoked 2.0 packs per day.
- Smoking more than 1 pack a day increased your risk of lung cancer 4.014 times.

**Based on your answers here are general recommendations:**

**Clinical HIV Guidance:**

- LUCENT as identified you as someone with low CD4 cell counts (between 78 and 250). This is a significant risk factor for a lung cancer diagnosis.
- CD4 cell counts should be at or above 500.
- Please speak with your general practitioner.

**LUCENT Smoking Cessation Guidance:**

- You have been identified as a smoker.
- Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.

For help with smoking cessation, please visit the CDC website.

https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html
Subject M-7025 is a white male smoker. He is 40 years old, not Hispanic, and has high school education. He is a current smoker. He smokes 4 packs per day and has smoked for 35 years. As a result, he is classified as a smoker. He has no history of pneumonia and no history of respiratory disease. Subject M-7025 does not have an account of AIDS. His CD4 value is 150 and is never the same. It fluctuates. His CD8 value is 459 making the CD4/CD8 ratio 0.3. He reported his viral load at 166,323. The CDSS calculated its probability based on him being a white male smoker (Figure 5.58). Based on risk factors and his responses, fluctuating CD4 values, CD4 count in the second quartile, having only a high-school education, non-Hispanic ethnicity, and years smoked in the fourth quartile increased his lung cancer risk. They determined him to have a medium risk of developing cancer. In the MACS database, subject M-7025 does not have a lung cancer diagnosis. While this was not the expected outcome, if the patient doesn’t address these risk factors, he has a medium risk for developing lung cancer in the future. To assist in mitigating risk, the CDSS also guided smoking cessation and suggested he speak with his clinician about his low CD4 count.

Figure 5.58: Results Screen from Patient Narrative 15 – Subject M-7025

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 34.11.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Fluctuating CD4 values increased your risk of lung cancer by 1.88 times.</td>
</tr>
<tr>
<td>Your CD4 value (150.0) is between 78 and 250. This increased your risk of lung cancer by 4.446 times.</td>
</tr>
<tr>
<td>Having only a high school education increased your risk of lung cancer 7.323 times.</td>
</tr>
<tr>
<td>Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 1.81 times.</td>
</tr>
<tr>
<td>You have smoked for 35.0 years. Since this is greater than 28, your risk of lung cancer increased 18.649 times.</td>
</tr>
</tbody>
</table>

**Based on your answers here are general recommendations:**

**Clinical HIV Guidance:**

- LUCENT as identified you as someone with low CD4 cell counts (between 78 and 250). This is a significant risk factor for a lung cancer diagnosis.
- CD4 cell counts should be at or above 500.
- Please speak with your general practitioner.
- LUCENT Smoking Cessation Guidance:

You have been identified as a smoker. Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.

For help with smoking cessation, please visit the CDC website.

https://www.cdc.gov/tobacco/campaign/tips/quilt-smoking/index.html
Subject W-9896 is a black, female, non-smoker. She is 37 years old, Hispanic, and has not completed any education. She has never smoked. As a result, she is classified as a non-smoker. She has no history of pneumonia and no history of respiratory disease. Subject W-9896 does have a history of AIDS. Her CD4 value is 1 and is always less than 200. Her CD8 value is 241 making the CD4/CD8 ratio 0.004. She reported her viral load at 3,700,000.

The CDSS calculated its probability based on her being a black, female, non-smoker (Figure 5.59). Based on risk factors and her responses, a previous history of AIDS, low CD4/CD8 ratio, and a high viral load increased her risk of lung cancer but determined her to have a low risk for developing lung cancer. In the WIHS database, subject W-9896 does not have lung cancer, the expected outcome for low risk.

Figure 5.59: Results Screen from Patient Narrative 16 – Subject W-9896

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 14.46.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>A previous history of AIDS, your risk of lung cancer is increased 2.55 times.</td>
</tr>
<tr>
<td>Based on your CD4 (1.0) and CD8 (241.0), your CD4/CD8 ratio (0) is significantly low. This increased your risk of lung cancer 1.97 times.</td>
</tr>
<tr>
<td>Your viral load (3700000.0) is greater than 500 and increased your risk of lung cancer 9.94 times.</td>
</tr>
</tbody>
</table>

Based on your answers here are general recommendations:

LUCENT Clinical HIV Guidance:

LUCENT has calculated your ratio of CD4 and CD8 ratio based on your CD4 of 1.0 and your CD8 of 241.0. Your ratio 0.004 is significantly lower than the 1.0 threshold. Please speak with your general practitioner about how to raise your CD4 values.

For more educational material, please see the following website:
https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count!
Subject M-7170 is a white, male, non-smoker. He is 45 years old, not Hispanic, and has a college education. He is a former smoker but quit 99 months ago. As a result, he is classified as a non-smoker. He has no history of pneumonia and does have a history of respiratory disease. Subject M-7170 does not have a history of AIDS. His CD4 value is 451 and is never the same. It fluctuates. His CD8 value is 234 making the CD4/CD8 ratio 1.9. He reported his viral load at 477,626.

The CDSS calculated its probability based on him being a white male non-smoker (Figure 5.60). Based on risk factors and his responses, a history of respiratory disease and a high viral load increased his lung cancer risk. They determined him to have a medium risk for developing lung cancer. In the MACS database, subject M-7170 has a lung cancer diagnosis, the expected outcome for medium risk.

Figure 5.60: Results Screen from Patient Narrative 17 – Subject M-7170

Results:

Your risk score is 15.03.

Risk Score Level: Medium

LUCENT used your responses and scored your risk for lung cancer based on ....

A history of respiratory disease increased your risk of lung cancer 5.091 times.
Your viral load (477626.0) is greater than 500 and increased your risk of lung cancer 9.94 times.
Subject W-9929 is a black female smoker. She is 46 years old, not Hispanic, and has not completed any education. She smokes 3 packs per day and has smoked for 30 years. As a result, she is classified as a smoker. She has a history of pneumonia 2 or more times and no history of respiratory disease. Subject W-9929 does have an account of AIDS. Her CD4 value is 57 and is always less than 200. Her CD8 value is 503 making the CD4/CD8 ratio 0.1. She reported her viral load at 560,000. The CDSS calculated its probability based on her being a black female smoker (Figure 5.61). Based on risk factors and her responses, age greater than 44, a previous history of AIDS, low CD4/CD8 ratio, non-Hispanic ethnicity, a history of reoccurring pneumonia, packs smoked per day, and years smoked in the fourth quartile increased her lung cancer risk. They determined her to have a medium risk of developing cancer. In the WIHS database, subject W-9929 does not have a lung cancer diagnosis. While this was not the expected outcome, if the patient doesn’t address these risk factors, she has a medium risk for developing lung cancer in the future. The CDSS suggests smoking cessation and suggested she speak with her clinician about her low CD4/CD8 ratio.

Figure 5.61: Results Screen from Patient Narrative 18 – Subject W-9929

Results:
Your risk score is 22.26.

Risk Score Level: Medium

LUCENT used your responses and scored your risk for lung cancer based on ...

- Your age (46.0), is greater than 44 and increased your risk of lung cancer by 1.73 times.
- A previous history of AIDS, your risk of lung cancer is increased 2.55 times.
- Based on your CD4 (57.0) and CD8 (503.0), your CD4/CD8 ratio (0.11) is significantly low. This increased your risk of lung cancer 4.846 times.
- Having an ethnicity that is not-Hispanic or latino increased your risk of lung cancer 3.656 times.
- You said you smoked 3.0 packs per day.
- Smoking more than 1 pack a day increased your risk of lung cancer 4.014 times.
- A history of reoccurring pneumonia increased your risk of lung cancer 3.434 times.
- You have smoked for 30.0 years. Since this is greater than 28, your risk of lung cancer increased 2.03 times.

Based on your answers here are general recommendations:

LUCENT Clinical HIV Guidance:
LUCENT has calculated your ratio of CD4 and CD8 ratio based on your CD4 of 57.0 and your CD8 of 503.0. Your ratio 0.113 is significantly lower than the 1.0 threshold. Please speak with your general practitioner about how to raise your CD4 values.

For more educational material, please see the following website:
https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count

LUCENT Smoking Cessation Guidance:
You have been identified as a smoker. Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.

For help with smoking cessation, please visit the CDC website:
https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html
Subject M-7867 is a white, male, non-smoker. He is 32 years old, Hispanic, and has not completed any education. He has never smoked. As a result, he is classified as a non-smoker. He has no history of pneumonia and no history of respiratory disease. Subject M-7867 does not have a history of AIDS. His CD4 value is 32 and is never the same. It fluctuates. His CD8 value is 801 making the CD4/CD8 ratio 0.04. He reported his viral load at 440,189.

The CDSS calculated its probability based on him being a white, male, non-smoker (Figure 5.62). Based on risk factors and his responses, a low CD4/CD8 ratio and high viral load increased his lung cancer risk. They determined him to have a medium risk for developing cancer. In the MACS database, subject M-7867 has a lung cancer diagnosis, the expected outcome for medium risk. The CDSS also guided smoking cessation and suggested he speak with his clinician about his low CD4 count.

Figure 5.62: Results Screen from Patient Narrative 19 – Subject M-7867

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 14.86.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
<tr>
<td>Based on your CD4 (32.0) and CD8 (891.0), your CD4/CD8 ratio (0.04) is significantly low. This increased your risk of lung cancer 4.923 times. Your viral load (440189.0) is greater than 500 and increased your risk of lung cancer 9.94 times.</td>
</tr>
</tbody>
</table>

Based on your answers here are general recommendations:

<table>
<thead>
<tr>
<th>LUCENT Clinical HIV Guidance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT has calculated your ratio of CD4 and CD8 ratio based on your CD4 of 32.0 and your CD8 of 801.0. Your ratio 0.04 is significantly lower than the 1.0 threshold. Please speak with your general practitioner about how to raise your CD4 values.</td>
</tr>
</tbody>
</table>

For more educational material, please see the following website: https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count
Subject M-9272 is a white male smoker. He is 44 years old, not Hispanic, and has not completed any education. He is a current smoker. He smokes 3 packs per day and has smoked for 30 years. As a result, he is classified as a smoker. He has no history of pneumonia and no history of respiratory disease. Subject M-9272 does not have an account of AIDS. His CD4 value is 693 and is always greater than 500. His CD8 value is 484 making the CD4/CD8 ratio 1.4. He reported his viral load at 35,861.

The CDSS calculated its probability based on him being a white, male, smoker (Figure 5.63). Based on risk factors and his responses, non-Hispanic ethnicity, and years smoked in the fourth quartile increased his lung cancer risk. They determined him to have a medium risk for developing cancer. In the MACS database, subject M-9272 has a lung cancer diagnosis, the expected outcome for medium risk. The CDSS also guided smoking cessation.

Figure 5.63: Results Screen from Patient Narrative 20 – Subject M-9272

<table>
<thead>
<tr>
<th>Results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your risk score is 20.46.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk Score Level: Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUCENT used your responses and scored your risk for lung cancer based on ....</td>
</tr>
</tbody>
</table>

| Having an ethnicity that is not-hispanic or latino increased your risk of lung cancer 1.81 times. |
| You have smoked for 30.0 years. Since this is greater than 28, your risk of lung cancer increased 18.649 times. |

Based on your answers here are general recommendations:

<table>
<thead>
<tr>
<th>LUCENT Smoking Cessation Guidance:</th>
</tr>
</thead>
<tbody>
<tr>
<td>You have been identified as a smoker.</td>
</tr>
<tr>
<td>Many risk factors contribute to your overall risk of lung cancer; however, smoking is a significant risk factor.</td>
</tr>
</tbody>
</table>

For help with smoking cessation, please visit the CDC website.  
https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html
5.5 Results Conclusion

Figure 5.64 summarizes the results of the 20 randomly selected subjects. The risk score on this sample outperformed the expected sensitivity of 77%. The CDSS accurately predicted a medium or high risk for a total of 8 subjects diagnosed with lung cancer. In this small sample, the risk score slightly underperformed the expected specificity of 60%. Four subjects with medium risk and one subject with high risk did not have a lung cancer diagnosis. Despite this, the accuracy remained at 70%, which is expected since sensitivity can only increase while sacrificing specificity. The moderate increase in sensitivity and decrease in specificity is likely due to the small subset of the 7,587 subjects used from the knowledge base. It is crucial to point out that the NLST criteria applied to the small sample of 20 subjects would not have recommended screening for any of them and demonstrate significant improvement over the existing guidance.

Figure 5.64: Summary of Validation Results

These results demonstrate the successful completion of the study's three primary goals. First, it proves that CDSS will accurately predict the risk of lung cancer in PWH using
risk factors specific to PWH instead of the general population. It also demonstrates the study's second primary goal to stratify PWH into three categories of risk for lung cancer. More than 300 logistic regression models led to creation of six stratified models using a total of 14 patient-specific variables to identify PWH at low, medium, and high risk for cancer. It has proven to be a stark contrast to the recommendation for screening in the general population. These results have clearly shown that this study's risk score and CDSS significantly outperforms the screening criteria recommended for the general population. Finally, the development of this CDSS, branded for integration as LUCENT, demonstrates the third and final goal of using the new risk score. Neither a risk score nor a CDS tool exists for lung cancer in PWH. This CDSS will allow PWH to assess their risk of lung cancer accurately and take the necessary steps for prevention and treatment at a sufficiently early stage. This tool is the first of its kind to address this urgent and unmet need.
6.1 Summary

People with HIV (PWH) have a significantly greater incidence and mortality from lung cancer than the general population of people without HIV. While PWH's age and stage at lung cancer diagnosis are controversial, several studies have shown that PWH are diagnosed with lung cancer at younger ages and later cancer stages. This study's results support these findings and have demonstrated significant differences, specifically in age.

The two most significant risk factors for lung cancer in the general population have historically been age and smoking pack-years. The NLST and USPSTF are the only two recommended guidelines for lung cancer screening in the general population. They both consider men and women at high risk for lung cancer if they are between 55 and 74 years of age and current smokers with 30-pack years of smoking or former smokers who quit smoking in the past 15 years. While these guidelines have demonstrated acceptable sensitivity and specificity in the general population, they have proved to be inadequate guidelines for PWH. When applied to the analytical sample of 7,587 PWH in this study, these guidelines overwhelmingly missed 97% of lung cancer cases. The reason for these differences is due to risk factors that clinically manifest only in PWH.

Several HIV specific risk factors explain the risk of lung cancer greater than age and smoking do. The most notable risk factors are those which describe immunocompetency in
PWH. HIV is a retrovirus that proliferates using CD4 cells. Uncontrolled HIV will increase viral loads and destroy CD4 cells leading to opportunistic illnesses in PWH. Leading opportunistic infections in PWH include respiratory diseases, which directly correlate to lung cancer. CD4 cell count, the ratio of CD4/CD8 cell counts, HIV viral load, history of reoccurring pneumonia, and the respiratory disease history proved significant risk factors for lung cancer in this study. Smoking and certain demographic risk factors such as age and ethnicity also play a significant role in lung cancer in PWH. Several other demographic risk factors and novel risk factors, such as alcohol, cocaine, and marijuana use, were not significant in multivariable models.

This study is the first study to use HIV specific risk factors to develop a CDSS that can assess lung cancer risk for prevention and earlier diagnosis in PWH. Data was curated from public-use data sets (PDS) managed by Johns Hopkins University to build the CDSS. The data contained 12,320 subjects, and the analytical sample included 7,607 HIV-positive men and women, of which 100 were diagnosed with lung cancer during the study. It was difficult to identify primary lung cancer cases; however, available data shows that more than 90% of the cases are primary. Twenty subjects were reserved for validating the CDSS, making the analytical sample 7,587 HIV positive men and women. The MACS and WIHS data were normalized for integration, and missing data in longitudinal data sets, such as laboratory data, were imputed using last observation carried forward. One value was selected from longitudinal data either at the visit preceding lung cancer diagnosis for those diagnosed with cancer or the last visit in the study for those without lung cancer. As a result, the analytical sample was one record per subject for 7,587 records and was used to develop the knowledge base for the CDSS.
To form the knowledge-base of the CDSS, fifty-one independent risk factors from the analytical sample were conditioned on gender, race, and smoking status were tested in 306 bivariable logistic regression models to identify candidate risk factors for lung cancer in PWH. Risk factors significant in the bivariable models formed the final predictive models creating eight risk score algorithms. Quartiles of risk were used as the risk-stratification thresholds.

The knowledge-based risk score algorithms formed the workflow and rules tested by the inference engine in the CDSS. The CDSS required 47 variables, seven data-driven forward chained rules, and six goal-oriented backward chained rules. Two hundred seven nodes produce a web-based lung cancer risk assessment tool that can be accessed individually by PWH or their clinicians. The sensitivity in the analytical sample that formed the knowledge base was 77%. The sensitivity from a small validation of 20 patients, excluded from the analytical sample and reserved for validation, demonstrated 80% sensitivity using the CDSS. The study accomplished its goals to produce a risk score specific to PWH that can assess lung cancer risk, stratify the risk into categories, and build a CDSS that PWH and their clinicians can use. This tool significantly outperformed the existing guidance from the NLST and USPSTF by detecting approximately 96% more lung cancer cases.

6.2 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 tool created by this study has not only addressed
the need but has demonstrated significantly better results than existing criteria.\textsuperscript{2,8,115} 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.

An analytic sample of 7,607 HIV-positive men and women was created to build the knowledge-base for the CDSS using publicly available datasets of two studies managed by Johns Hopkins University. This dataset represents both men and women with HIV and those with lung cancer in the United States. It was used to develop the knowledge-base of the CDSS, inform the rules for the inference engine in the CDSS, and validate the CDSS using a small, reserved sample. HIV status is considered protected health information (PHI), making the curation of an additional dataset impracticable. Traditional risk factors used by risk assessment tools in the general population were identified, as were HIV-specific risk factors proven associated with lung cancer in existing studies.

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 contrariety is likely due to the small number of non-smokers in the study.

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, as seen in Table 5.2. This distribution is consistent with the number of lung cancer cases by sex and race/ethnicity seen in Figures 3.1 and 3.2. This shortcoming led to the combination of black and other races into a single category. The exploratory analysis explains the effect of separating the black race from the ‘Other’ races. The models predicting lung cancer using ‘other’ race are questionable due to the small number of patients with lung cancer who were an ‘other’ race. When logistic regression models perfectly predict an outcome or almost perfectly predict an outcome, they are referred to as ‘quasi-separating.’ Instead, the exploratory analysis provides better insight into the actual risk of lung cancer for black PWH. The twelve overall models seen in Table 5.9 are very similar to the primary results and do not improve the sensitivity or specificity. Of interest is the addition of cocaine use into the final model for the black race. 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 CDSS used the subjects race, gender, and smoking status to conditionally determine which risk factors the end-user has. The CDSS
dynamically builds a risk score for each risk factor using multiplication and then sums the probabilities of all risk factors. 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 and used in the CDSS 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 analytical sample 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 sample 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\(^{18}\).

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.

Finally, the academic version of Corvid posed limitations on the realism of the tool for real-world use. There were node limitations that required rule modifications due to the version used. Despite this, this study successfully built a CDSS to conditionally capture the correct information to assess lung cancer risk based on gender, race, and smoking status.
6.3 Future Direction

This study is the first study to create a risk score for lung cancer specific to PWH, which can be risk-stratified. It is also the first study to create a web-based risk assessment tool that can be accessed via the internet by PWH and their clinicians. It will allow PWH to assess and modify factors putting them at risk for lung cancer, thereby preventing disease. It will also allow for shared decision making between PWH and their clinicians to pursue lung cancer screening if deemed appropriate.

The CDSS has been branded as the "Lung Cancer Risk Assessment Tool (LUCENT) for HIV" and is ready for integration into a URL for web-access; however, it needs 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 were challenging to normalize, integrate and were 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. 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.

In addition to investigating other traditional and novel risk factors, the potential to incorporate longitudinal data needs further exploration. A future CDSS should have the capacity to connect to electronic health records and directly pull longitudinal data instead of summarizing longitudinal data into a single record. 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 will be challenging to do but will provide more insight into PWH risk factors that do not smoke.

Future work should also explore utilizing the full version of Exsys Corvid or utilizing other programming languages to build the CDSS. The version of Corvid used for this study was an academic version that restricted the number of nodes used. These limitations led to modifications to variables and logic. As a result, some questions could have been asked differently to provide a better user experience for the end-user; however, Corvid's limitations affected this.

6.4 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


