Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.38253

Predicting Self-Reported Lung Cancer Diagnosis Using Statistical and Machine Learning Approaches: A 2024 NHIS Study

Tianhao Xiang

School of Arts and Sciences, Massachusetts College of Pharmacy and Health Sciences, Boston, USA
13601187093@163.com

Abstract. Lung cancer remains an important public health concern, and cigarette smoking is one of its most established associated factors. This study examined demographic, behavioral, socioeconomic, and health-related factors associated with self-reported lung cancer diagnosis among U.S. adults using the 2024 National Health Interview Survey (NHIS). Survey-weighted descriptive analyses and logistic regression were used to estimate population-level prevalence and adjusted associations. Logistic regression and random forest were also compared for respondent-level prediction using a stratified 80/20 training and test split. Among 32,573 adults with a valid lung cancer outcome, 145 reported a previous diagnosis. The weighted prevalence was 0.312%. Older age and smoking history showed the strongest adjusted associations. Each 10-year increase in age was associated with 2.13 times the odds of reported lung cancer, while former and current smokers had odds ratios of 4.05 and 5.63 compared with never smokers. In the test sample, logistic regression achieved an ROC-AUC of 0.899 and identified 24 of 28 cases, while random forest identified only one. Greater model complexity did not improve prediction in this rare-outcome survey dataset.

Keywords: lung cancer, NHIS, logistic regression, random forest, machine learning, epidemiology

1. Introduction

Lung cancer remains an important public health issue, with cigarette smoking representing a major behavioral exposure related to the disease. Although cigarette smoking among U.S. adults has declined over time, tobacco use remains an important preventable health concern [1]. Smoking patterns are also not equally distributed across the population. Previous research using NHIS data has shown persistent differences in smoking initiation, cessation, and prevalence across levels of educational attainment [2]. Socioeconomic differences have also been observed in lung cancer incidence across education, occupation, and income [3].

Traditional statistical models, particularly logistic regression, are widely used to examine binary health outcomes and develop prediction models. Lung cancer models have combined demographic and smoking-related information to estimate individual risk and identify screening populations [4].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.38253

More recently, machine-learning methods have been increasingly applied to clinical prediction because they can model complex relationships without requiring all patterns to be specified in advance. However, current evidence does not show that greater model complexity automatically improves prediction performance [5].

This study therefore examined which demographic, behavioral, and socioeconomic factors were associated with self-reported lung cancer diagnosis in the 2024 NHIS and whether random forest improved prediction compared with logistic regression.

2. Background and literature review

2.1. Smoking and socioeconomic factors

Smoking remains central to research on lung cancer because tobacco exposure is strongly related to the disease. Recent national surveillance based on NHIS data showed that exclusive cigarette smoking decreased substantially between 2017 and 2023, although overall tobacco product use did not show the same decline [1]. This changing pattern makes recent population-based data useful for examining smoking-related health outcomes.

Smoking behavior is also socially patterned. Cao et al. analyzed NHIS data from 1966 to 2018 and found that differences in smoking behavior across educational groups widened across more recent birth cohorts. In particular, smoking initiation and cessation differed according to educational attainment. These findings suggest that education may be related to lung cancer outcomes indirectly through differences in smoking behavior and other social conditions.

Socioeconomic differences have also been documented more directly for lung cancer. A systematic review and meta-analysis by Sidorchuk et al. [3] included 64 eligible studies and reported higher lung cancer incidence among populations with lower educational, occupational, and income-based socioeconomic position. Importantly, these associations were observed even in many analyses that adjusted for smoking. Therefore, examining demographic and socioeconomic characteristics alongside smoking may provide a broader understanding of differences in self-reported lung cancer diagnosis.

2.2. Statistical and machine-learning approaches

Logistic regression is commonly used when the outcome has two categories because its coefficients can be transformed into odds ratios that are relatively easy to interpret. Existing lung cancer prediction research also shows why model evaluation must extend beyond simply fitting a model. Katki et al. [4] compared nine lung cancer risk models and found substantial differences in the screening populations selected by different models. Better-performing models showed stronger calibration and discrimination when evaluated in independent cohorts.

Machine learning provides a different approach. Algorithms such as random forest can automatically identify nonlinear patterns and interactions. In a nationwide Korean cohort Lee et al. [6] compared logistic regression with several machine-learning approaches for cancer prediction using demographic, laboratory, anthropometric, and medical-history information, and they also performed external validation. Their study demonstrates the potential value of machine learning when richer clinical data are available.

However, machine learning is not automatically superior. Christodoulou et al. [5] reviewed 71 studies comparing logistic regression with machine-learning methods. Among 145 comparisons judged to have a low risk of bias, the average difference in logit-transformed AUC between logistic

DOI: 10.54254/2753-8818/2026.38253

regression and machine learning was 0.00, providing no evidence of superior discrimination for machine learning. The authors also found possible problems with validation procedures in 68% of the included studies and noted that calibration was frequently not reported [5]. These findings emphasize that fair model comparison and appropriate validation are more important than model complexity alone.

Evaluation is particularly challenging when the positive outcome is rare. Saito and Rehmsmeier [7] showed that ROC curves may give an overly favorable impression in highly imbalanced classification settings and argued that precision-recall measures provide important additional information about positive predictions. This issue is directly relevant to the present study because self-reported lung cancer was very uncommon in the NHIS sample.

3. Data and methodology

3.1. Data source and study population

This study used the 2024 National Health Interview Survey Sample Adult public-use file. NHIS is conducted by the National Center for Health Statistics and uses a complex sampling design to collect health information from the U.S. civilian noninstitutionalized population [8].

The original Sample Adult file contained 32,629 respondents. After coding the lung cancer outcome, 32,573 respondents had a valid outcome, including 145 respondents who reported a previous lung cancer diagnosis. Fifty-six respondents had missing, refused, or unknown outcome information.

The primary survey-weighted regression model included 31,649 respondents and 140 lung cancer cases with complete information for the primary predictors. The expanded complete-case sample contained 31,115 respondents and 139 lung cancer cases and was used for Table 1 and the prediction comparison.

3.2. Variables

The outcome was a self-reported history of lung cancer diagnosis, constructed from the NHIS variables CANEV_A and LUNGCAN_A. The outcome therefore represents whether respondents reported previously being told that they had lung cancer. It does not represent new or future lung cancer incidence.

The primary predictors were age, cigarette smoking status, sex, race and ethnicity, and educational attainment. Smoking was classified as never, former, or current smoking. Race and ethnicity were categorized as non-Hispanic White, Hispanic, non-Hispanic Black, non-Hispanic Asian, and other or multiple race. Education was grouped into bachelor's degree or higher, some college or associate degree, high school or GED, and less than high school. BMI category and self-reported general health were included in an expanded analysis.

3.3. Statistical analysis

All analyses were performed in R. Survey-weighted descriptive estimates incorporated the final annual sample weight (WTFA_A), strata (PSTRAT), and primary sampling unit (PPSU). Because lung cancer was rare, survey-weighted prevalence estimates used logit-based 95% confidence intervals.

DOI: 10.54254/2753-8818/2026.38253

Survey-weighted logistic regression was used to estimate adjusted odds ratios. Age was modeled per 10-year increase. The primary model adjusted for age, smoking status, sex, race and ethnicity, and education. A secondary expanded model additionally included BMI and general health.

Prediction was evaluated separately at the respondent level. The expanded complete-case sample was divided using a stratified 80/20 split. The training set contained 24,891 respondents and 111 cases, and the held-out test set contained 6,224 respondents and 28 cases. Logistic regression and random forest used the same predictor set. The random forest consisted of 1,000 trees and used class weighting. Classification thresholds were selected from the training data using Youden's index and then applied to the test sample. Prediction metrics were not survey-weighted.

4. Results and discussion

4.1. Study population and weighted prevalence

As shown in Table 1, respondents reporting lung cancer differed most clearly from respondents without reported lung cancer in age, smoking history, race and ethnicity, and general health. The survey-weighted mean age was 71.1 years among respondents reporting lung cancer compared with 48.1 years among those without lung cancer. Former smokers represented 58.8% of the weighted lung cancer group, compared with 22.3% of respondents without lung cancer. Current smokers represented 20.0% and 9.9%, respectively.

Table 1. Characteristics of the study population by self-reported lung cancer diagnosis

Variable	Level	Overall (N=31,115)	No lung cancer (N=30,976)	Lung cancer (N=139)	P-value
Age, years	Weighted mean (SE)	48.2 (0.2)	48.1 (0.2)	71.1 (1.0)	<0.001
Sex	Female	16,722 (51.1%)	16,653 (51.1%)	69 (44.5%)	0.207
	Male	14,393 (48.9%)	14,323 (48.9%)	70 (55.5%)	
Smoking status	Never	19,832 (67.6%)	19,805 (67.7%)	27 (21.2%)	<0.001
	Former	7,987 (22.4%)	7,905 (22.3%)	82 (58.8%)	
	Current	3,296 (10.0%)	3,266 (9.9%)	30 (20.0%)	
Race and ethnicity	Non-Hispanic White	20,892 (61.6%)	20,779 (61.6%)	113 (80.3%)	<0.001
	Hispanic	4,471 (17.7%)	4,468 (17.8%)	3 (1.8%)	
	Non-Hispanic Black	3,078 (11.6%)	3,060 (11.6%)	18 (14.3%)	
	Non-Hispanic Asian	1,781 (6.2%)	1,778 (6.2%)	3 (2.8%)	
	Other/Multiple	893 (2.8%)	891 (2.8%)	2 (0.7%)	
Educational attainment	Bachelor degree or higher	11,920 (33.9%)	11,886 (33.9%)	34 (25.6%)	0.321
	Some college/Associate	8,694 (29.1%)	8,653 (29.1%)	41 (30.5%)	
	High school/GED	8,002 (27.0%)	7,958 (27.0%)	44 (31.8%)	
	Less than high school	2,499 (10.1%)	2,479 (10.1%)	20 (12.1%)	
BMI category	Underweight/Healthy weight	10,234 (33.0%)	10,175 (33.0%)	59 (35.9%)	0.454
	Overweight	10,482 (33.6%)	10,437 (33.6%)	45 (36.5%)	
	Obesity	10,399 (33.5%)	10,364 (33.5%)	35 (27.6%)	
General health status	Excellent/Very good	15,541 (50.1%)	15,522 (50.1%)	19 (15.6%)	<0.001
	Good	9,503 (30.2%)	9,459 (30.2%)	44 (30.8%)	
	Fair/Poor	5,071 (14.7%)	4,995 (14.6%)	76 (53.7%)	

DOI: 10.54254/2753-8818/2026.38253

The overall survey-weighted prevalence of self-reported lung cancer was 0.312% (95% CI: 0.254%–0.383%).

As shown in Figure 1, prevalence increased sharply with age and was also higher among respondents with a smoking history. Prevalence increased from 0.006% among adults aged 18–44 to 0.202% among those aged 45–64 and 1.076% among those aged 65 years or older. By smoking status, prevalence was 0.100% among never smokers, 0.786% among former smokers, and 0.604% among current smokers.

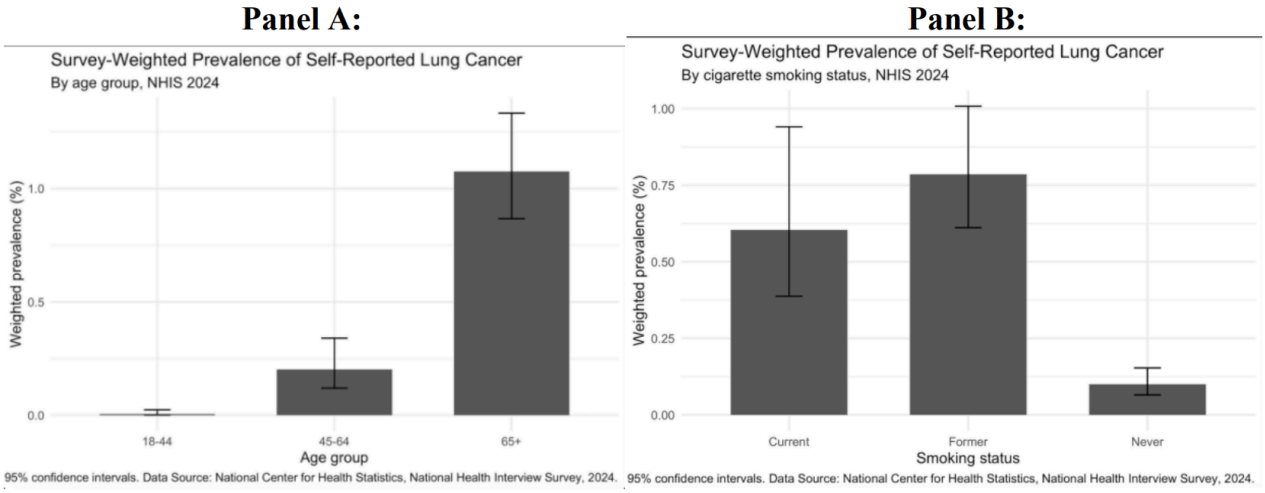


Figure 1. Survey-weighted prevalence of self-reported lung cancer by age group and cigarette smoking status, NHIS 2024

The higher descriptive prevalence among former smokers than among current smokers should not be interpreted as evidence that former smoking is more harmful. Former smokers may be older or may have stopped smoking after receiving a diagnosis. Adjustment for age and other characteristics is therefore important.

4.2. Survey-weighted logistic regression

As shown in Table 2, age and smoking status remained strongly associated with self-reported lung cancer after adjustment. Each 10-year increase in age was associated with more than twice the odds of reported lung cancer (OR = 2.13, 95% CI: 1.88–2.43, $p < .001$).

Compared with never smokers, former smokers had approximately four times the adjusted odds (OR = 4.05, 95% CI: 2.57–6.38, $p < .001$), while current smokers had more than five times the odds (OR = 5.63, 95% CI: 2.92–10.86, $p < .001$). Sex and educational attainment were not statistically significant after adjustment.

Table 2. Survey-weighted logistic regression estimates for self-reported lung cancer diagnosis

model	term	log_odds	standard_err	statistic	p_value	odds_ratio	ci_lower	ci_upper
Primary survey-weighted model	(Intercept)	-11.373462	0.58494047	-19.443794	1.19E-65	1.15E-05	3.64E-06	3.63E-05

DOI: 10.54254/2753-8818/2026.38253

Table 2. (continued)

Primary survey-weighted model	age 10	0.75813537	0.06540577	11.5912612	3.60E-28	2.13429285	1.87701523	2.426834851
Primary survey-weighted model	smoking Former	1.3988987	0.23135253	6.04661074	2.60E-09	4.05073645	2.571625029	6.380582552
Primary survey-weighted model	smoking Current	1.72850049	0.33424913	5.17129385	3.17E-07	5.63220202	2.921385062	10.85844534
Primary survey-weighted model	sex Male	0.17525432	0.20285366	0.86394458	0.38796415	1.19154921	0.80005625	1.774724415
Primary survey-weighted model	race ethnicity Hispanic	-1.8257283	0.71593719	-2.5501236	0.01101627	0.16110027	0.039487131	0.657259607
Primary survey-weighted model	race ethnicity Non-Hispanic Black	0.24712721	0.30543555	0.80909772	0.41878019	1.28034198	0.70276856	2.32596308
Primary survey-weighted model	race ethnicity Non-Hispanic Asian	-0.3433465	0.68261121	-0.5029898	0.61515631	0.70939236	0.185639496	2.710832211
Primary survey-weighted model	race ethnicity Other/Multiple	-0.9954112	0.74960101	-1.3279213	0.18470989	0.36957146	0.084790076	1.610837834
Primary survey-weighted model	education Some college/ Associate degree	0.10471813	0.29210504	0.35849477	0.72009939	1.11039758	0.625654874	1.970707547
Primary survey-weighted model	education High school/GED	0.16375181	0.27514918	0.5951383	0.55197579	1.17792193	0.686175036	2.022078931
Primary survey-weighted model	education Less than high school	0.31477716	0.36483207	0.86280014	0.38859267	1.369954	0.669162525	2.804660901

Although the estimate for Hispanic respondents was statistically significant, only three lung cancer cases occurred in this category in the complete-case sample. This estimate should therefore be interpreted cautiously.

The expanded model produced similar associations for age and smoking. Fair or poor general health was strongly associated with reported lung cancer compared with excellent or very good health (OR = 6.48, 95% CI: 3.38-12.42). However, because general health and previous lung cancer diagnosis were measured in the same cross-sectional survey, poor health may partly reflect the

DOI: 10.54254/2753-8818/2026.38253

consequences of illness rather than a factor that existed before diagnosis. General health was therefore treated as a secondary finding.

4.3. Logistic regression versus random forest

The prediction comparison showed a clearer advantage for logistic regression. As shown in Figure 2A, logistic regression achieved a ROC-AUC of 0.899, compared with 0.852 for random forest. The precision-recall results in Figure 2B showed the same overall pattern, with a PR-AUC of 0.0276 for logistic regression and 0.0174 for random forest.

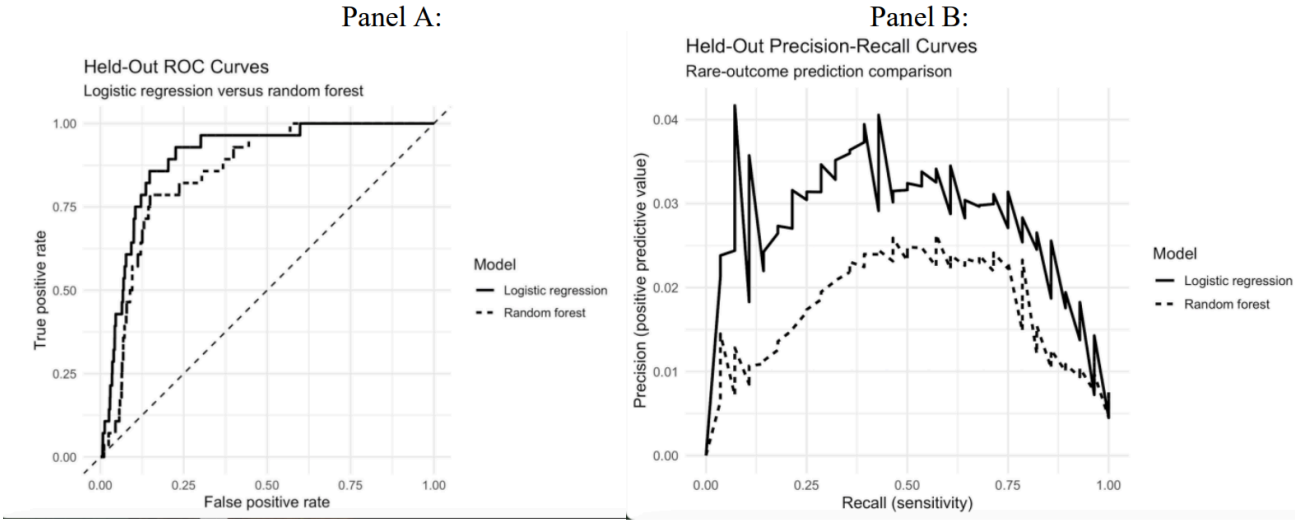


Figure 2. Held-out prediction performance of logistic regression and random forest

The largest practical difference was sensitivity. In the held-out test sample, logistic regression correctly identified 24 of 28 lung cancer cases, corresponding to 85.7% sensitivity. Random forest identified only 1 of 28 cases, giving a sensitivity of 3.6%. Logistic regression had a specificity of 79.9% and PPV of 1.89%, while random forest had a specificity of 98.1% and PPV of 0.84%.

Random forest achieved an overall accuracy of 97.7%, compared with 79.9% for logistic regression. However, this high accuracy was misleading because almost all respondents in the test set were negative cases. Random forest produced only one true positive while missing 27 positive cases. This illustrates why accuracy alone is inadequate for analyzing highly imbalanced outcome data and why precision-recall measures provide useful supplementary information [7].

The results are also consistent with the broader findings of Christodoulou et al., who found no evidence that machine-learning approaches yield better discriminative performance than logistic regression in analyses with a low risk of bias [5]. In the present study, the structured dataset contained relatively few predictors and only 139 complete-case lung cancer cases. Under these conditions, the additional flexibility of random forest did not translate into improved prediction.

The model comparison should still be considered exploratory. Christodoulou et al [5]. emphasized that validation procedures can strongly affect apparent differences between statistical and machine-learning models. The present study selected thresholds using training data and evaluated performance on a separate held-out test sample, but only one random split was used and the test set contained only 28 positive cases. Stronger conclusions would require repeated validation or evaluation in an independent population.

DOI: 10.54254/2753-8818/2026.38253

5. Limitations

This study has several limitations. NHIS is cross-sectional, so temporal or causal relationships cannot be established. Lung cancer diagnosis was self-reported, and only 139 cases were available in the complete-case sample. The test set contained only 28 cases, and prediction was evaluated using one train-test split without external validation. Detailed information, such as pack-years, smoking duration, time since quitting, radon exposure, occupational exposure, secondhand smoke, and family history was not included.

6. Conclusion

This study used 2024 NHIS data to examine demographic, behavioral, socioeconomic, and health-related factors associated with self-reported lung cancer diagnosis and to compare logistic regression with random forest for prediction. Older age and smoking history showed the strongest independent associations. Each 10-year increase in age was associated with approximately twice the odds of reported lung cancer, while both former and current smokers had substantially higher adjusted odds than never smokers.

The prediction analysis also showed that logistic regression performed better than random forest in the held-out test sample. Logistic regression had higher ROC-AUC and PR-AUC and identified 24 of the 28 positive cases, whereas random forest identified only one. The high overall accuracy of random forest was mainly a consequence of the very low prevalence of the positive outcome rather than effective case identification.

These findings show that a more complex model does not automatically provide better prediction, particularly when the outcome is rare and the number of available predictors is limited. Logistic regression offered both stronger performance and easier interpretation in this analysis. However, neither model had sufficient positive predictive value for clinical screening. Future work should use larger longitudinal datasets, more detailed smoking and environmental exposure information, repeated or external validation, and a greater number of lung cancer cases before similar models are considered for clinical or public health applications.

References

[1] Arrazola, R. A., Husten, C. G., Cornelius, M. E., & Armour, B. S. (2025). Notes from the field: Tobacco product use among adults—United States, 2017–2023. *Morbidity and Mortality Weekly Report, 74*(7), 118–121. https://doi.org/10.15585/mmwr.mm7407a3

[2] Cao, P., Jeon, J., Tam, J., Fleischer, N. L., Levy, D. T., Holford, T. R., & Meza, R. (2023). Smoking disparities by level of educational attainment and birth cohort in the U.S. *American Journal of Preventive Medicine, 64*(4 Suppl. 1), S22–S31. https://doi.org/10.1016/j.amepre.2022.06.021

[3] Sidorchuk, A., Agardh, E. E., Aremu, O., Hallqvist, J., Allebeck, P., & Moradi, T. (2009). Socioeconomic differences in lung cancer incidence: A systematic review and meta-analysis. *Cancer Causes & Control, 20*, 459–471.

[4] Katki, H. A., Kovalchik, S. A., Petito, L. C., Cheung, L. C., Jacobs, E., Jemal, A., Berg, C. D., & Chaturvedi, A. K. (2018). Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. *Annals of Internal Medicine, 169*(1), 10–19. https://doi.org/10.7326/M17-2701

[5] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[6] Lee, E., Jung, S. Y., Hwang, H. J., & Jung, J. (2021). Patient-level cancer prediction models from a nationwide patient cohort: Model development and validation. *JMIR Medical Informatics, 9*(8), e29807. https://doi.org/10.2196/29807

DOI: 10.54254/2753-8818/2026.38253

[7] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
[8] National Center for Health Statistics. (2025). National Health Interview Survey, 2024 survey description. Centers for Disease Control and Prevention.