EDA and Regression analysis_Indhira

2025-12-06

Methods

Descriptive Analysis

We analyzed 493,282 participants from the NLMS Tobacco Use Cohort using a big-data workflow centered on Apache Arrow to enable efficient, memory-safe processing of the large dataset. The raw CSV file was converted to Parquet and loaded as an Arrow Dataset, allowing fast, columnar I/O, lazy evaluation, and on-disk querying without importing the entire file into memory—an approach far more efficient than base R, which fully materializes data frames and incurs substantially higher memory and compute costs. We used Arrow and dplyr to subset variables, perform grouped operations, and collect only the necessary columns into R for recoding. Smokeless tobacco variables (CURRUSE, EVERUSE) were recoded into binary indicators based on digit patterns, and marital status was collapsed into two categories (married, not married/separated). Labeled factor variables were created for readability. Descriptive statistics were computed stratified by sex, with categorical variables summarized using counts and percentages and continuous variables using means and standard deviations. A formatted Table 1 was generated using kableExtra to present all variables with clear labels and hierarchical indentation.

Regression Analysis

Variable Selection

Covariates were selected a priori based on established epidemiological evidence demonstrating their associations with both smoking behavior and mortality. Age is a strong determinant of baseline mortality risk (Arias & Xu, 2022), while sex differences in survival are well-documented (Kruger & Nesse, 2006). Race/ethnicity influences mortality through structural and socioeconomic inequities (Williams & Mohammed, 2009). Socioeconomic status, particularly income, is closely linked to life expectancy (Chetty et al., 2016). Health insurance status affects access to healthcare and subsequently mortality risk (Wilper et al., 2009). Marital status is included due to its protective association with lower mortality (Rendall et al., 2011). Smoking duration, measured by age at initiation, has also been shown to independently influence long-term mortality (USDHHS, 2014).

Logistic Regression

We evaluated the association between cigarette smoking behavior and mortality using logistic regression models applied to the NLMS Tobacco Use dataset. Mortality (inddea) was modeled as a binary outcome. Two main exposures were examined: (1) smoking status (never, everyday, some-days, former) and (2) smoking status plus smoking duration, operationalized using age at smoking initiation (agesmk). Covariates selected a priori based on demographic and socioeconomic relevance included age, race, sex, health insurance status, income, and marital status.

Model 1 : Smoking Status

$$\text{logit}\big[P(\text{Death}_i = 1)\big] = \beta_0 + \beta_1\,\text{Smokestat}_i + \beta_2\,\text{Age}_i + \beta_3\,\text{Race}_i + \beta_4\,\text{Sex}_i + \beta_5\,\text{Insurance}_i + \beta_6\,\text{Income}_i + \beta_7\,\text{MaritalStatus}_i$$

Model 2: Smoking Status+Snoking Duration

$$\text{logit}\big[P(\text{Death}_i = 1)\big] = \gamma_0 + \gamma_1\,\text{Smokestat}_i + \gamma_2\,\text{AgeStartedSmoking}_i + \gamma_3\,\text{Age}_i + \gamma_4\,\text{Race}_i$$

$$+\gamma_5\,\text{Sex}_i + \gamma_6\,\text{Insurance}_i + \gamma_7\,\text{Income}_i + \gamma_8\,\text{MaritalStatus}_i$$

All models were fit using generalized linear models with a binomial distribution and logit link in R. A univariate model estimated the crude association between smoking status and mortality, followed by multivariable models adjusting for confounders. Model performance was compared and selected using AIC and residual deviance. Odds ratios and 95% confidence intervals were computed by exponentiating model coefficients. Missing data were handled using complete-case analysis to ensure comparability across models.

Results

Descriptive Analysis

The analytic sample included 493,282 participants, with 230,967 males and 262,315 females. The sex-stratified descriptive statistics showed broadly similar age distributions between groups, while race, income, and health insurance patterns exhibited modest variation. Marital status differed by sex, with a higher proportion of males classified as married and females more frequently not married or separated. Tobacco-related behaviors demonstrated clear sex patterns: males reported higher rates of current and ever use of smokeless tobacco and were more likely to have smoked at least 100 cigarettes. The mean age of smoking initiation was slightly lower among males. Mortality proportions also varied slightly by sex. Overall, Table 1 summarizes these demographic, socioeconomic, and tobacco-use differences, providing a foundational characterization of the NLMS tobacco cohort prior to inferential analysis.

Table 1. Descriptive Statistics Stratified by Sex

Male (N = 230,967) & Female (N = 262,315)

Variable	Level	Male	Female
Age	NA	43 (17.8)	44.6 (18.7)
Age_Started_Smoking	NA	17.2 (4.2)	18.7 (5.4)
Cigarrette_smoking_Status	Everyday Smoker	59618 (25.8%)	52732 (20.1%)
	Former_smoker	46057 (19.9%)	37244 (14.2%)
	NA	313 (0.1%)	274 (0.1%)
	Never Smoker	117535 (50.9%)	165040 (62.9%)
	Someday Smoker	7444 (3.2%)	7025 (2.7%)
Current_Smokeless_Use	NA	28639 (12.4%)	31410 (12%)
	No	186022 (80.5%)	229538 (87.5%)
	Yes	16306 (7.1%)	1367 (0.5%)
Ever_Smokeless_Use	Ever used	41722 (18.1%)	5591 (2.1%)
	NA	28639 (12.4%)	31410 (12%)
	Never used	160606 (69.5%)	225314 (85.9%)
Health_Insurance_Status	Insured	157023 (68%)	182716 (69.7%)
	NA	47929 (20.8%)	54107 (20.6%)
	Uninsured	26015 (11.3%)	25492 (9.7%)
Income	NA	9 (3.7)	8.3 (3.9)
Marital_Status	Married	143262 (62%)	145472 (55.5%)
	NA	1337 (0.6%)	1393 (0.5%)
	Not married/Separated	86368 (37.4%)	115450 (44%)
Mortality	Alive	220312 (95.4%)	252815 (96.4%)
	Dead	10655 (4.6%)	9500 (3.6%)
Race	American Indian/Alaskan Native	2417 (1%)	2805 (1.1%)
	Asian/Pacific Islander	7226 (3.1%)	8329 (3.2%)
	Black	19022 (8.2%)	26539 (10.1%)
	NA	7 (0%)	7 (0%)
	Other	1131 (0.5%)	1327 (0.5%)
	White	201164 (87.1%)	223308 (85.1%)
Smoked_100_Cigarettes	No	117535 (50.9%)	165040 (62.9%)
	Yes	113432 (49.1%)	97275 (37.1%)

1 n (%); Mean (SD)

Logistic Regression

Predictor	Interpretation
Everyday smoker (vs never)	119% higher mortality (OR=2.19)
Some-days smoker (vs never)	78% higher mortality (OR=1.78)
Former smoker (vs never)	37% higher mortality (OR=1.37)
Age (per year)	9% increase in mortality per year (OR=1.09)
Race:Black (vs White)	24% higher mortality (OR=1.24)
Race:American Indian/Alaskan Native (vs White)	38% higher mortality (OR=1.38)
Race:Asian/Pacific Islander(vs White)	26% lower mortality (OR=0.74)
Race: Other (vs White)	Not significant
Female (vs Male)	43% lower mortality (OR=0.57)
Insured (vs Uninsured)	17% higher mortality (OR=1.17)
Income (per level increase)	5% lower mortality per income level (OR=0.95)
Married (vs Separated/Not married)	27% lower mortality (OR=0.73)

Predictor	Interpretation
Someday smoker (vs Everyday Smoker)	13.7% lower mortality (OR = 0.86)
Former smoker (vs Everyday Smoker)	38.8% lower mortality (OR = 0.61)
Age started smoking (per year)	2.1% lower mortality per year started later (OR = 0.98)
Age (per year)	9.4% higher mortality per year of age (OR = 1.09)
Race:Black (vs White)	14.7% higher mortality (OR = 1.15)
Race:American Indian/Alaskan Native (vs White)	27.9% higher mortality (OR = 1.28)
Race:Asian/Pacific Islander(vs White)	30.0% lower mortality (OR = 0.70)
Race: Other (vs White)	Not significant (OR = 0.97)
Female (vs Male)	35.7% lower mortality for females (OR = 0.64)
Health insurance: Yes (vs No)	19.4% higher mortality (OR = 1.19)
Income (per level)	5.9% lower mortality per income level (OR = 0.94)
Married (vs Seperated/Not married)	22.4% lower mortality for married individuals (OR = 0.78)

In the model adjusting for demographic and socioeconomic covariates, cigarette smoking status showed a strong graded association with mortality. Compared with never smokers, everyday smokers had more than double the odds of death (OR = 2.19), some-day smokers had 78% higher odds (OR = 1.78), and former smokers had 37% higher odds (OR = 1.37), demonstrating a clear dose–response pattern (Table 2).

When age at smoking initiation was added to the model, the associations shifted. After adjustment, some-day smokers showed approximately 14% lower odds of mortality (OR = 0.86) and former smokers showed approximately 39% lower odds (OR = 0.61) relative to everyday smokers, likely reflecting differences in cumulative smoking duration. Age at smoking initiation was independently associated with mortality, with each additional year of later initiation reducing mortality risk by about 2% (Table 3).

Across both models, age, sex, income, race, and marital status remained strong independent predictors. Age showed the largest effect, with each additional year associated with roughly 9% higher mortality risk. Females had substantially lower mortality (about 36–43% lower), while higher income and being married were consistently protective. Race differences persisted, with several groups showing higher or lower mortality risks relative to the reference category.

Overall, smoking status was a robust predictor of mortality, and incorporating smoking duration through age at initiation refined these associations. These findings suggest that both current smoking behavior and lifetime smoking history contribute meaningfully to mortality risk.

References

Arias, E., & Xu, J. Q. (2022). United States life tables, 2019 (National Vital Statistics Reports, Vol. 70, No. 19). National Center for Health Statistics. https://doi.org/10.15620/cdc:113096

Kruger, D. J., & Nesse, R. M. (2006). An evolutionary life-history framework for understanding sex differences in human mortality rates. Evolutionary Psychology, 4(1), 66–85. https://doi.org/10.1177/147470490600400114

Williams, D. R., & Mohammed, S. A. (2009). Discrimination and racial disparities in health: Evidence and needed research. Journal of Behavioral Medicine, 32(1), 20–47. https://doi.org/10.1007/s10865-008-9185-0

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., & Cutler, D. (2016). The association between income and life expectancy in the United States, 2001–2014. JAMA, 315(16), 1750–1766. https://doi.org/10.1001/jama.2016.4226

Wilper, A. P., Woolhandler, S., Lasser, K. E., McCormick, D., Bor, D. H., & Himmelstein, D. U. (2009). Health insurance and mortality in US adults. American Journal of Public Health, 99(12), 2289–2295. https://doi.org/10.2105/AJPH.2008.157685

Rendall, M. S., Weden, M. M., Favreault, M. M., & Waldron, H. (2011). The protective effect of marriage for survival: A review and update. Demography, 48(2), 481–506. https://doi.org/10.1007/s13524-011-0032-5

U.S. Department of Health and Human Services. (2014). The health consequences of smoking—50 years of progress: A report of the Surgeon General. https://www.ncbi.nlm.nih.gov/books/NBK179276/