Smoking and Mortality Risk (NHANES 2013–2014)

December 3, 2025

1 Data

This section describes the data sources, how the files are linked, and the basic sample structure. DEMO and SMQ are joined on the NHANES respondent identifier `SEQN` within the 2013–2014 cycle. Mortality is then linked from the public use National Center for Health Statistics follow up file.

Table 1: Raw NHANES and Mortality Assets

file	wave	years	n	k
DEMO_H.csv	H	2013–2014	10 175.00	47.00
SMQ_H.csv	H	2013–2014	7168.00	32.00
NHANES_2013_2014_MORT_2019_PUBLIC.csv	NA	Mortality 2019 linkage	10 175.00	8.00

Table 1 documents the raw input files, their row counts, and key variables. It shows that the DEMO and SMQ files have similar numbers of observations and that the mortality linkage covers nearly all respondents in the cycle. This confirms that merge keys and file coverage are consistent before any cleaning.

Table 2: Analytic Working Sample After Merges

Data source: Cleaned matrices (label = dead) from clean-data/
NHANES cycles: H (2013–2014)
Observations (n): 6098
Variables (k): 8
Total missing cells: 0 (0.00%)

Table 2 summarizes the analytic dataset after joining DEMO and SMQ on `SEQN` and merging to the mortality file. It shows the final number of unique individuals and the frequency of nonmissing mortality labels. This table makes explicit that the sample is restricted only by data availability on the outcome, not by predictor missingness.

2 Cleaned Data

We construct a feature matrix X and a binary outcome y from the merged dataset with the following design:

- The label y_i is an indicator for mortality by the follow up time. Rows are dropped only if this label is missing.
- Core predictors in X include age, sex, income to poverty ratio (PIR), smoking status, and basic demographic controls.
- Numeric variables are imputed using the median and accompanied by missingness flags. Categorical variables are imputed to an explicit `Unknown` level.

- Age and PIR are standardized; we add squared age and a log income transformation for flexibility.
- Smoking and sex remain categorical factors that enter the models through dummy variables.

Table 3: Clean Data Overview and Feature Construction

Label variable: dead
Features used: age, sex, smoker, smoker_code, age_sq, log_income
Initial n: 6111; **Dropped on label only:** 13; **Final n:** 6098
Columns with missing (pre-impute counts): income, dead
Modeled numeric features: age (standardized), age_sq, log_income (from PIR; raw PIR is standardized for preprocessing but not used directly in modeling) **Categoricals in X:** smoker and sex as factors. Race and education are excluded from X and kept only for descriptive/meta. **Smoker categories:** Never/Former/Current coded as 1/2/3 in smoker_code (counts shown in summary stats).

Table 3 lists the main variables in the clean matrix and the choices for coding and imputation. The design preserves the full sample of labeled observations while making the underlying numeric variables comparable in scale and explicitly tracking missingness.

Illustrative Heads of the Clean Matrices

The next tables show the top rows of the key objects. They are mainly diagnostic and help verify that the pipeline produced the intended structure.

Table 4: Head of Feature Matrix X

age	sex	smoker	smoker_code	age_sq	log_income
1.17	male	Former	2.00	1.37	0.61
0.36	male	Current	3.00	0.13	1.02
1.33	male	Former	2.00	1.78	1.71
1.39	female	Never	1.00	1.92	1.79
0.47	male	Former	2.00	0.22	1.76
0.74	female	Never	1.00	0.54	1.79

Table 4 confirms that the feature matrix has the expected standardized numeric columns, dummy variables, and missingness flags. Each row corresponds to an individual respondent.

Table 5: Head of Label Vector y

dead
FALSE
TRUE
FALSE
TRUE
FALSE
FALSE

Table 5 shows the corresponding outcome vector, which is binary with values equal to zero or one. This confirms that the label extraction is consistent with the mortality indicator.

Table 6: Head of Cleaned Full Data Frame

id	age	sex	income	smoker	smoker_code	dead	wave	age_missing	income_missing	smoker_code_missing	age_sq	log_income
73 557.00	1.17	male	−0.99	Former	2.00	FALSE	H	0.00	0.00	0.00	1.37	0.61
73 558.00	0.36	male	−0.40	Current	3.00	TRUE	H	0.00	0.00	0.00	0.13	1.02
73 559.00	1.33	male	1.32	Former	2.00	FALSE	H	0.00	0.00	0.00	1.78	1.71
73 561.00	1.39	female	1.62	Never	1.00	TRUE	H	0.00	0.00	0.00	1.92	1.79
73 562.00	0.47	male	1.49	Former	2.00	FALSE	H	0.00	0.00	0.00	0.22	1.76
73 564.00	0.74	female	1.62	Never	1.00	FALSE	H	0.00	0.00	0.00	0.54	1.79

Table 6 presents the head of the full cleaned frame that includes all predictors, the label, and any auxiliary variables. It is useful for checking that variable names, units, and coding of categorical levels align with the intended design.

Table 7: Head of Meta Data (IDs and Missingness Flags)

id	wave	age_missing	income_missing	smoker_code_missing
73 557.00	H	0.00	0.00	0.00
73 558.00	H	0.00	0.00	0.00
73 559.00	H	0.00	0.00	0.00
73 561.00	H	0.00	0.00	0.00
73 562.00	H	0.00	0.00	0.00
73 564.00	H	0.00	0.00	0.00

Table 7 focuses on identifier columns and missingness indicators. It confirms that the data are de duplicated on `SEQN` and that missingness flags behave as expected.

Table 8: Imputation Counts by Variable

variable	n_imputed
age	0.00
income	480.00
sex	0.00
smoker	0.00
smoker_code	0.00

Table 8 reports the number of imputed values by variable. It shows that core predictors like age and smoking status have little or no imputation, while income (PIR) may require more. This pattern justifies a simple median+flag strategy and alerts the reader that inference on heavily imputed variables should be interpreted with care.

3 Feature and Label Dictionaries and Metrics

We next document the variables in X and y and summarize their distributions. This provides context for the modeling results.

Table 9: Feature Dictionary for X

column	description
age	Age (years), standardized
sex	Sex factor: female/male
smoker	Smoking status factor: Never/Former/Current from SMQ020/040
smoker_code	Ordinal smoking code: Never=1, Former=2, Current=3; 0 if imputed
age_sq	Age squared, from standardized age
log_income	log1p(PIR) from imputed income

Table 9 defines each feature, its source, its transformation, and units. This makes clear, for example, which variables are standardized, which are indicators, and which are derived from original NHANES items.

Table 10: Label Dictionary for y

label	description
dead	Mortality status from public-use linked mortality file (MORTSTAT==1)

Table 10 spells out the coding of the mortality outcome and any auxiliary risk index that mirrors the binary label. The outcome is a true binary indicator, which is important when interpreting later residual diagnostics.

Table 11: Numeric Metrics for Features in X

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
age	63	0	0.0	1.0	−1.6	0.0	1.8
smoker_code	3	0	1.6	0.8	1.0	1.0	3.0
age_sq	63	0	1.0	0.9	0.0	0.7	3.1
log_income	425	0	1.1	0.5	0.0	1.1	1.8

Table 11 summarizes means, standard deviations, and ranges for numeric predictors. Standardization implies that many transformed variables have mean close to zero and variance close to one. This scaling helps the LASSO penalty treat coefficients on a comparable footing and stabilizes neural network training.

Table 12: Categorical Metrics for Features in X

		N	%
sex	female	3187	52.3
	male	2911	47.7
	Unknown	0	0.0
smoker	Never	3523	57.8
	Former	1344	22.0
	Current	1231	20.2
	Unknown	0	0.0

Table 12 reports category frequencies and shares for categorical variables such as smoking status and sex.

These proportions indicate that the sample is reasonably balanced across key groups and that each smoking category has enough observations to estimate separate effects.

Table 13: Outcome Distribution for y

y	count
FALSE	5632.00
TRUE	466.00

Table
reftab:y-metrics shows the number and share of deaths versus survivors. The mortality indicator is relatively rare, so the outcome distribution is imbalanced. We therefore use inverse-prevalence class weights in our logistic models and report classification metrics (accuracy, log loss, Brier score, ROC AUC).

Table 14: Class Imbalance and Suggested Weights

quantity	value
n_total	6098.00
n_positive (y=1)	466.00
n_negative (y=0)	5632.00
prevalence (y=1)	0.08
w1 = 1/prevalence	13.09
w0 = 1/(1-prevalence)	1.08
w1_normalized	6.54
w0_normalized	0.54

4 Transformations and Smoking Codes

The transformations in the data pipeline are motivated by both statistical considerations and alignment with standard epidemiologic coding of NHANES smoking variables.

Key NHANES smoking items include:

- **SMQ020**: indicator for whether the respondent has smoked at least 100 cigarettes in their life. This defines ever versus never smokers.
- **SMQ040**: indicator for whether the respondent currently smokes cigarettes. This splits ever smokers into current and former smokers.
- Related items such as SMQ050Q or SMQ060 capture intensity or age at smoking, but these are not central in the current specification.

Using these variables, we define a three level smoking factor: Never, Former, and Current. Never smokers have SMQ020 equal to zero. Former smokers have SMQ020 equal to one and SMQ040 equal to zero. Current smokers have SMQ020 equal to one and SMQ040 equal to one.

Age and PIR are transformed for both modeling stability and interpretability. Age is standardized and used both linearly and through a squared term. PIR is standardized and a log transformation is used when appropriate to compress the upper tail.

Table 15: Processing Notes and Transformation Summary

Processing Notes

Joins and Scope: Person-level joins are performed on SEQN within each NHANES cycle (DEMO + SMQ), then rows are stacked across cycles.
Smoker Derivation (SMQ):

- If SMQ020 == 2 (*has not smoked 100 cigarettes*): label as Never.
- If SMQ020 == 1 and SMQ040 == 3 (*currently do not smoke*): label as Former.
- If SMQ020 == 1 and SMQ040 equals 1 or 2 (*every day or some days*): label as Current.
- Otherwise: smoker is set to NA (insufficient or inconsistent information).

Smoker Coding: Factor levels Never/Former/Current and an ordinal code smoker_code = 1/2/3 are created for modeling.
Other Harmonization:

- Sex normalized to female/male factor.
- Income set to NHANES PIR (INDFMPIR) and standardized jointly with age.
- Rows with missing among id, age, sex, income, smoker, smoker_code, and the label (dead) are dropped for the clean matrix.

Standardization: Numeric predictors (age, income) are centered and scaled to unit variance.
Label: The modeling label is dead.

Table 15 summarizes these choices in a compact format. It emphasizes that label only dropping preserves the analytic sample size and that we do not impute the outcome.

Table 16: Smoking Categories: Counts and Shares

smoker	count
Never	3523.00
Former	1344.00
Current	1231.00
Unknown	0.00

Table 16 reports the final counts and shares of Never, Former, and Current smokers in the clean sample. The distribution shows substantial representation in each category, so the estimated differences in mortality risk across smoking groups are based on nontrivial sample sizes rather than a few outliers.

5 Models

We estimate two logistic models on the cleaned data: a weighted logit with HC1 robust standard errors, and a weighted regularized logit (LASSO) with cross validation.

5.1 Model 1: Logit (Weighted, HC1 robust)

We model the probability of death via the log-odds:

$$\log \frac{p_i}{1-p_i} = \beta_0+\beta_1 \mathbf{1}\{\text{smoker}_i = \text{Former}\}+\beta_2 \mathbf{1}\{\text{smoker}_i = \text{Current}\}+\beta_3\, \text{age}_i+\beta_4 \mathbf{1}\{\text{sex}_i = \text{male}\}+\beta_5\, \log(1+\text{PIR}_i). \tag{1}$$

Weights are inverse-prevalence and normalized to have mean one.

Table 17: Logit Coefficients (HC1 robust)

	title
(Intercept)	0.267
	(0.193)
smokerFormer	−0.540***
	(0.184)
smokerNever	−1.142***
	(0.188)
age	1.303***
	(0.095)
age_sq	0.451***
	(0.069)
sexmale	0.174
	(0.128)
log_income	−0.921***
	(0.138)
Num.Obs.	6098
RMSE	0.39
Std.Errors	Custom

* p <0.1, ** p <0.05, *** p <0.01

Table 18: Logit Classification Metrics (Training Split)

Metric	Value
Accuracy	0.78
LogLoss	0.46
Brier	0.15
ROC_AUC	0.87
PR_AUC	0.38

Table 19: Logit Classification Metrics (Validation Split)

Metric	Value
Accuracy	0.79
LogLoss	0.46
Brier	0.15
ROC_AUC	0.88
PR_AUC	0.39

5.2 Model 2: Regularized Logit (LASSO, CV)

The LASSO model penalizes the absolute size of coefficients to encourage sparsity. In vector notation,

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{2}$$

where $\lambda \geq 0$ controls the strength of the penalty. We choose λ by cross validation, using both the minimum error choice and the one standard error rule.

Table 20: LASSO Coefficients at Cross Validated λ

lambda	term	estimate
lambda.min	(Intercept)	0.24
lambda.min	smokerFormer	−0.50
lambda.min	smokerNever	−1.11
lambda.min	age	1.29
lambda.min	age_sq	0.44
lambda.min	sexmale	0.16
lambda.min	log_income	−0.91
lambda.1se	(Intercept)	−0.13
lambda.1se	smokerNever	−0.54
lambda.1se	age	1.09
lambda.1se	age_sq	0.32
lambda.1se	log_income	−0.60

Table 20 lists the nonzero coefficients at the selected penalty. The variables that survive penalization mirror the OLS specification: age and PIR remain central, along with sex and smoking indicators. Many weaker covariates are set exactly to zero, which simplifies the model without materially changing its main substantive conclusions.

Table 21: Regularized Logit Classification Metrics (Training Split)

Metric	Value
Accuracy	0.78
LogLoss	0.46
Brier	0.15
ROC_AUC	0.87
PR_AUC	0.38

Table 22: Regularized Logit Classification Metrics (Validation Split)

Metric	Value
Accuracy	0.79
LogLoss	0.46
Brier	0.15
ROC_AUC	0.88
PR_AUC	0.39

Tables 21 and 22 show that the LASSO achieves training and validation performance very close to the OLS model. The gains in out of sample error are small. The main value of the LASSO here is interpretive: it confirms that only a small subset of the available predictors is truly useful for prediction and that age and smoking status dominate.

6 Coefficient Comparison: OLS versus LASSO

Table 23: Coefficient Comparison: OLS and LASSO at λ_{min} and λ_{1se}

term	base_estimate	lasso_min	lasso_1se
(Intercept)	0.27	0.24	−0.13
age	1.30	1.29	1.09
age_sq	0.45	0.44	0.32
log_income	−0.92	−0.91	−0.60
sexmale	0.17	0.16	
smokerFormer	−0.54	−0.50	
smokerNever	−1.14	−1.11	−0.54

Table 23 compares OLS coefficients with LASSO solutions at two penalty levels. The LASSO estimates for age, smoking dummies, sex, and PIR closely track the OLS values at λ_{min}, which indicates that the OLS coefficients are not driven by a few high leverage observations. At the more conservative λ_{1se}, many smaller coefficients are shrunk to zero, but the sign and ordering of the surviving effects remain consistent. This reinforces the conclusion that age and smoking status are the most stable and important predictors in this modeling framework.

7 Monte Carlo Evidence for LASSO Stability

Finally, we use bootstrap resampling of the cleaned dataset to examine how often LASSO selects each predictor and how large the typical active set is.

Table 24: Bootstrap LASSO Selection Frequencies

Table 25: Bootstrap Distribution of LASSO Nonzero Coefficients

Table 24 shows how often each variable is selected with a nonzero coefficient across bootstrap samples. Core predictors such as age and the smoking indicators are chosen in a large fraction of resamples, while many auxiliary variables are selected rarely. This pattern indicates that the importance of age and smoking for mortality risk is robust to sampling variability.

Table 25 reports the distribution of the number of active coefficients in each bootstrap draw. The distribution is relatively tight, which means that the effective model size under the chosen penalty does not fluctuate wildly across resamples. The combination of stable selection frequencies for key variables and a narrow distribution of active set sizes suggests that the LASSO model is well calibrated and that its conclusions about which predictors matter are not an artifact of a particular random split.