Pre-Analysis:

This study uses a cross-sectional, correlational design based on secondary data from the Korean National Health Insurance Service (NHIS), an open-access administrative dataset covering adults aged 20–85 years from the Korean population. No primary data collection was conducted. The analytic sample comprises 991,299 individuals with complete information on smoking status, age, anthropometrics, lipid parameters, and relevant covariates.

The pre-analysis focused on data preparation, descriptive inspection, and systematic testing of model assumptions prior to hypothesis testing. Smoking status was recoded into three ordered categories (never smoked, former smoker, current smoker). Body mass index (BMI) was computed from measured height and weight. Distributions of all variables were examined visually (histograms, Q–Q plots), and no missing values were detected in the analytic dataset. Extremely implausible lipid values (LDL > 1000 mg/dL; HDL > 500 mg/dL) were excluded for analyses involving lipid metabolism, as these values are physiologically unlikely and would unduly influence regression estimates.

For H1 (age and smoking status), ordinal regression models were evaluated, including tests of linearity using spline terms and likelihood-ratio comparisons. A spline specification for age provided superior fit, while the proportional-odds assumption was met. For H2 (smoking status and body weight), linear regression assumptions were assessed via residual diagnostics, Q–Q plots, variance inflation factors, and homoscedasticity checks using hexbin residual plots; all assumptions were adequately satisfied. For H3 (smoking status and lipid profiles), linearity, residual normality, and multicollinearity were examined separately for LDL and HDL cholesterol, adjusting for age, sex, and alcohol consumption. Given the very large sample size, formal normality tests were intentionally avoided, as they are known to be overly sensitive; instead, visual diagnostics guided model adequacy.

Overall, the pre-analysis confirms that the data structure and statistical assumptions are appropriate for testing the proposed hypotheses: (H1) decreasing likelihood of current smoking with increasing age, (H2) lower body weight among current smokers compared to non-smokers and former smokers, and (H3) higher LDL and lower HDL cholesterol levels among smokers. The analyses are explicitly observational and do not permit causal inference, but they are well-suited to identifying robust population-level associations between smoking behavior and key physiological indicators in a Korean adult population.