Smoking, Cholesterol, and Heart Disease:
An Analysis of Framingham Heart Study Data

Introduction

Heart disease is a debilitating group of cardiovascular conditions that consistently constitute the leading cause of death in the United States (NCHS, 2023) and produce significant decreases in quality of life for those afflicted.

Smoking may cause heart disease due to the increased risk of plaque buildup along arterial walls (thrombosis). Moreover, smoking is also associated with arterial inflammation, producing additional risk of clotting (Ambrose and Barua, 2003).

Serum cholesterol, particularly increased low-density lipoprotein cholesterol (LDL-C), may also increase risk of heart disease (Wadhera et al., 2016). The relationships between high-density lipoprotein cholesterol (HDL-C) and triglycerides to heart disease appear more complex (e.g., Després et al., 2000; Fernandez et al., 2008).

The current report analyzes Framingham Heart Study data to confirm previous findings regarding the effects of smoking and cholesterol components on heart disease.

Data Description

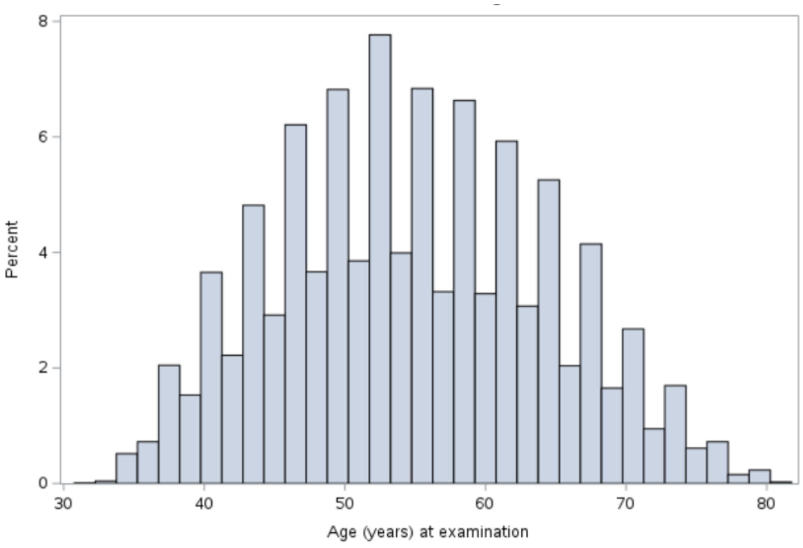
There are a total of 4,434 subjects included in this study. Since participants are followed over time with repeated measurements, there are a total of 11,627 observations. Data was collected at three examination periods approximately 6 years apart during the time period from 1956 to 1968. There are a total of 38 variables in the dataset, including risk factors (age, sex, period, blood pressure, cholesterol, BMI, etc.) and event data (stroke, heart disease, death, etc.), which includes the time to event. During the first examination period, all 4,434 participants were present. During the second examination period, 3,930 subjects remained. And during the final examination period, 3,263 subjects were left. The first five observations from the dataset with select variables of interest are presented below.

Obs	RANDID	SEX	AGE	TOTCHOL	CURSMOKE	CIGPDAY	BMI	PERIOD	PREVCHD	CVD	TIMECVD
1	2448	1	39	195	0	0	26.97	1	0	1	6438
2	2448	1	52	209	0	0	.	3	0	1	6438
3	6238	2	46	250	0	0	28.73	1	0	0	8766
4	6238	2	52	260	0	0	29.43	2	0	0	8766
5	6238	2	58	237	0	0	28.50	3	0	0	8766

Sex is a binary variable, where the value "1" denotes male and the value "2" denotes female. Across all time periods, approximately 56.8% (6,605 / 11,627) of the observations were female. During the first examination period, 56.2% (2,490 / 4,434) of the study population identified as female. During the second examination period, 57.0% (2,239 / 3,930) of the study population were women. And during the final examination period, 57.5% (1,876 / 3,263) of study participants identified as female.

Age is a continuous variable, which represents the age of a study subject at a given examination. The distribution of age is generally normally distributed; however, there are many peaks and valleys (Figure 1). Across all time periods, the mean age is approximately 54.8 years (SD = 9.56 years). During the first examination period, the mean age of study subjects is 49.9 years (SD = 8.68 years). During the second examination period, the mean age of study participants is 55.4 years (SD = 8.54 years). And during the final examination period, the mean age of the study population is 60.7 years (SD = 8.30 years). The interesting shape of the distribution with repeated peaks and valleys is likely due to subjects more often appearing during years when their age was an even number compared to years when their age was an odd number.

Figure 1: Distribution of Age of Study Participants



Framingham Heart Study

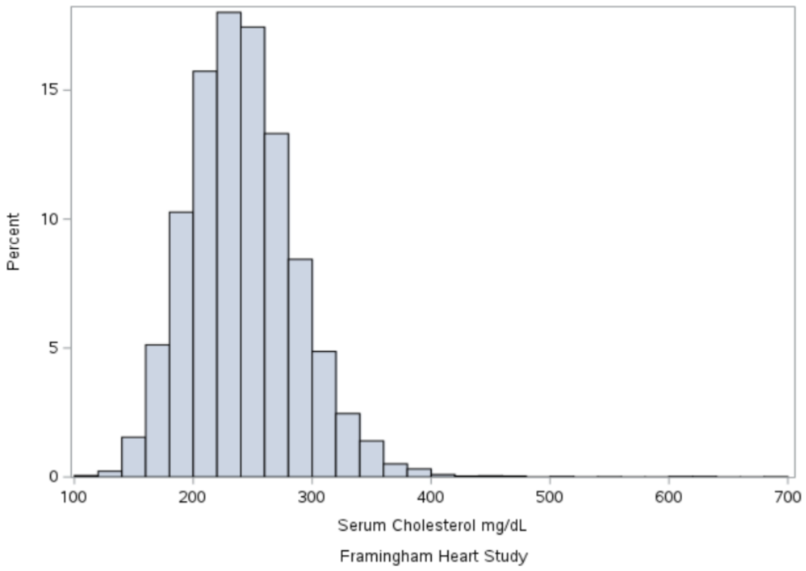
Current smoking status (CURSMOKE) is a binary variable, where the value "1" denotes current smoker and the value "0" denotes not a current smoker. Across all time periods, approximately 43% (5,029 / 11,627) of the study observations were current smokers. During the first examination period, 49% (2,181 / 4,434) of the study population was smoking. During the second examination period, 44% (1,727 / 3,930) of the study population smoked. And during the final examination period, just 34% (1,121 / 3,263) of study participants were smoking. One potential explanation for the reduction in smoking prevalence over time is that smokers die younger, so fewer remain by the final examination period.

Prevalent coronary heart disease status (PREVCHD) is a binary variable, where the value "1" denotes the presence of coronary heart disease at a given examination cycle and the value "0" denotes the absence of coronary heart disease at a given examination cycle. This variable is measured at each examination cycle for each subject. Across all time periods, approximately 7.2% (842 / 11,627) of the study observations had prevalent coronary heart disease, which is defined by: pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina). During the first examination period, 4.4% (194 / 4,434) of the study population had prevalent coronary heart disease. During the second examination period, 7.3% (288 / 3,930) of the study population had prevalent coronary heart disease. And during the final examination period, 11.0% (360 / 3,263) of study participants had prevalent coronary heart disease. This variable will be used in log linear models and as the response in logistic regression models.

Cardiovascular disease (CVD) is a binary variable, where the value "1" denotes the occurrence of a cardiovascular disease event during follow-up while the value "0" denotes the absence of the event during follow-up. This variable is measured just *once* for each subject (in contrast to prevalent coronary heart disease, which is recorded at each visit). This variable is defined by the occurrence of: Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease. During follow-up, 26.1% (1,157 / 4,434) of study participants experienced at least one cardiovascular disease event while 73.9% (3,277 / 4,434) of study subjects did not experience a cardiovascular disease event. The corresponding variable time to cardiovascular disease (TIMECVD) is a time-to-event variable that measures the time from baseline to the *first* cardiovascular disease event (subjects can have multiple cardiovascular disease events) or the time from baseline to loss to follow-up. These variables will be used in survival analysis models.

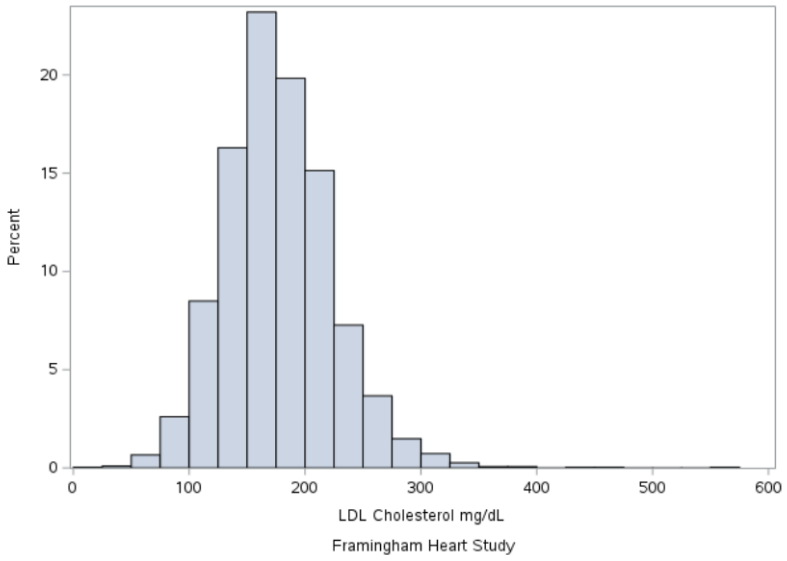
Total cholesterol (TOTCHOL) is a continuous variable, which is measured in mg/dL. The distribution of total cholesterol is approximately normal, with perhaps a slight skew to the right (due to some outliers in the right tail) (Figure 2). Across all time periods, mean total cholesterol is approximately 241 mg/dL (SD = 45.37 mg/dL). During the first examination period, mean total cholesterol among study subjects is 237 mg/dL (SD = 44.65 mg/dL). During the second examination period, mean total cholesterol among study participants is 250 mg/dL (SD = 45.75 mg/dL). And during the final examination period, mean total cholesterol in the study population is 237 mg/dL (SD = 44.45 mg/dL).

Figure 2: Distribution of Total Cholesterol of Study Participants



Low-density lipoprotein (LDL) cholesterol is a continuous variable, which is measured in mg/dL. The distribution of total cholesterol is approximately normal, with perhaps a slight skew to the right (due to some outliers in the right tail) (Figure 3). LDL cholesterol is only measured during the third examination period. During this final period, mean LDL cholesterol in the study population is 176 mg/dL (SD = 46.86 mg/dL).

Figure 3: Distribution of LDL Cholesterol of Study Participants



Methods

Log Linear Models

The primary variables of interest are current smoking status (CURSMOKE), prevalent coronary heart disease status (PREVCHD), and cholesterol. Recall that current smoking status and prevalent coronary heart disease status are binary variables. Prevalent coronary heart disease status is the primary outcome of interest for this analysis. Total cholesterol is a continuous variable, which is measured in mg/dL, and equals the sum of low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, and one-fifth of triglycerides levels. For examination cycles 1 and 2, only total cholesterol is reported.

For log-linear models, all variables are treated equally (regardless of whether they are predictors or the response in future models). In order to include cholesterol as a main effect, it is turned into a categorical variable. The cut points are chosen from the medical literature ("Lipid Panel"). For total cholesterol, levels below 200 mg/dL are considered "normal"; levels between 200 and 239 mg/dL are considered "borderline high"; levels at 240 mg/dL and above are considered "high" (Table 1).

Table 1: Total Cholesterol Variable Cutoffs

Total Cholesterol (Continuous)	**Total Cholesterol (Categorical)**	**Label**
< 200 mg/dL	0	Normal
200 – 239 mg/dL	1	Borderline High
≥ 240 mg/dL	2	High

For LDL cholesterol, levels below 100 mg/dL are considered "optimal"; levels between 100 and 129 mg/dL are considered "near optimal"; levels between 130 and 159 mg/dL are considered "borderline high"; levels between 160 and 189 mg/dL are considered "high"; levels at 190 mg/dL and above are considered "very high" (Table 2). The model can be described as a generalized linear model with the following components: the "response" variable here is the count of individuals with the i-th cholesterol level, j-th smoking status, and k-th prevalent coronary heart disease status. The response variable follows a Poisson distribution. The link function is the natural logarithm function. The variance function is the identity.

Table 2: LDL Cholesterol Variable Cutoffs

LDL Cholesterol (Continuous)	LDL Cholesterol (Categorical)	Label
< 100 mg/dL	0	Optimal
100 – 129 mg/dL	1	Near Optimal
130 – 159 mg/dL	2	Borderline High
160 – 189 mg/dL	3	High
≥ 190 mg/dL	4	Very High

The association between current smoking status, prevalent coronary heart disease status, and total cholesterol is separately studied for each examination cycle. For the final examination cycle, the association between current smoking status, prevalent coronary heart disease status, and LDL cholesterol is also studied. Models treating cholesterol as a continuous variable were also considered. When total cholesterol is considered a continuous variable, the categorical values 0–2 are simply viewed as continuous (2nd column of Table 1). When LDL cholesterol is treated as a continuous variable, the categorical values 0–4 are regarded as continuous (2nd column of Table 2). Nested models were compared using the likelihood ratio test, which follows a chi-square distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models. Non-nested models were compared using AIC. Model diagnostics, such as the distribution of deviance residuals, were also used in model selection.

Logistic Regression

First, the effect of current smoking status (CURSMOKE) and LDL cholesterol on the odds of prevalent coronary heart disease (PREVCHD) is studied for the final examination cycle. The response variable prevalent coronary heart disease status (measured at each examination cycle) is binary, so logistic regression was used. As a generalized linear model, logistic regression has the following components: the response variable follows a Binomial distribution, with the probability of prevalent heart disease p. The link function is the logit function. The variance function is quadratic and equals the probability of prevalent coronary heart disease multiplied by
one minus the probability of prevalent coronary heart disease [that is, $Var(p) = p * (1 - p)$].

Since LDL cholesterol is only measured during the third examination cycle, earlier cycles are ignored. The cut points for LDL cholesterol are the same as before (Table 2 in the “Log Linear Models” Subsection of the “Methods” Section). Models treating cholesterol as a continuous variable were also considered. When LDL cholesterol is treated as a continuous variable, the categorical values 0–4 are regarded as continuous (2nd column of Table 2). Nested models were compared using the likelihood ratio test (difference in the -2 Log Likelihoods), which follows a chi-square distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models. Non-nested models were compared using AIC. Model diagnostics, such as the c-statistic (AUC), Somers' D statistic, and percent of correctly classified responses were also considered in model selection.

The results of model selection are displayed below (Table 3). A model assuming the independence of current smoking status and LDL cholesterol (treated as continuous) was selected (Model 4). This model is as good as the full model (Model 1) according to the likelihood ratio test (LR = 2,073.862 – 2,071.341 = 2.521, df = 7, p-value = 0.9255).

Table 3: Logistic Regression Models for Examination Cycle 3 (Ignoring Age)

	Logistic Regression Model Predictors	**-2 Log Likelihood**	**DF**	**P Value**	**AIC**
1	Smoking | LDL Cholesterol (Categorical)	2,071.341	–	–	2,091.341
2	Smoking LDL Cholesterol (Categorical)	2,072.672	4	0.8561	2,084.672
3	Smoking | LDL Cholesterol (Continuous)	2,073.780	6	0.8752	2,081.780
4	**Smoking LDL Cholesterol (Continuous)**	**2,073.862**	**7**	**0.9255**	**2,079.862**

Since age may confound the relationship between the predictors current smoking status and LDL cholesterol and the response prevalent coronary heart disease status, models including the variable age were considered next. The same model selection techniques described above were used here. Note that Models 1–9 treat LDL Cholesterol as categorical while Models 10–18 treat LDL Cholesterol as continuous. The results of model selection are displayed below (Table 4). A model assuming the mutual independence of current smoking status, LDL cholesterol (treated as continuous), and age was selected (Model 18). This model is as good as the full model (Model 1) according to the likelihood ratio test (LR = 1,981.248 – 1,969.512 = 11.736, df = 16, p-value = 0.7619).

Table 4: Logistic Regression Models for Examination Cycle 3 (Including Age)

Logistic Regression Model Predictors		-2 Log Likelihood	DF	P Value	AIC
LDL Cholesterol is treated as categorical in Models 1–9					
1	Smoking | LDL Cholesterol | Age	1,969.512	–	–	2,009.512
2	All Two Way Interactions	1,972.335	4	0.5879	2,004.335
3	Conditional Independence of Smoking and LDL Cholesterol	1,974.256	8	0.7846	1,998.256
4	Conditional Independence of Smoking and Age	1,975.911	5	0.2693	2,005.911
5	Conditional Independence of LDL Cholesterol and Age	1,974.813	8	0.7250	1,998.813
6	Joint Independence of Smoking and LDL Cholesterol from Age	1,978.428	9	0.4451	2,000.428
7	Joint Independence of Smoking and Age from LDL Cholesterol	1,976.916	12	0.8298	1,992.916
8	Joint Independence of LDL Cholesterol and Age from Smoking	1,977.599	9	0.5254	1,999.599
9	Mutual Independence of Smoking, LDL Cholesterol, and Age	1,980.322	13	0.6267	1,994.322
LDL Cholesterol (cLDL) is treated as continuous in Models 10–18					
10	Smoking | cLDL Cholesterol | Age	1,975.817	12	0.8999	1,991.817
11	All Two Way Interactions	1,976.010	13	0.9261	1,990.010
12	Conditional Independence of Smoking and cLDL Cholesterol	1,976.488	14	0.9356	1,988.488
13	Conditional Independence of Smoking and Age	1,979.447	14	0.7669	1,991.447
14	Conditional Independence of cLDL Cholesterol and Age	1,977.515	14	0.8892	1,989.515

15	Joint Independence of Smoking and cLDL Cholesterol from Age	1,981.061	15	0.7128	1,991.061
16	Joint Independence of Smoking and Age from cLDL Cholesterol	1,977.760	15	0.9135	1,987.760
17	Joint Independence of cLDL Cholesterol and Age from Smoking	1,979.979	15	0.7894	1,989.979
18	**Mutual Independence of Smoking, cLDL Cholesterol, and Age**	**1,981.248**	**16**	**0.7619**	**1,989.248**

Results

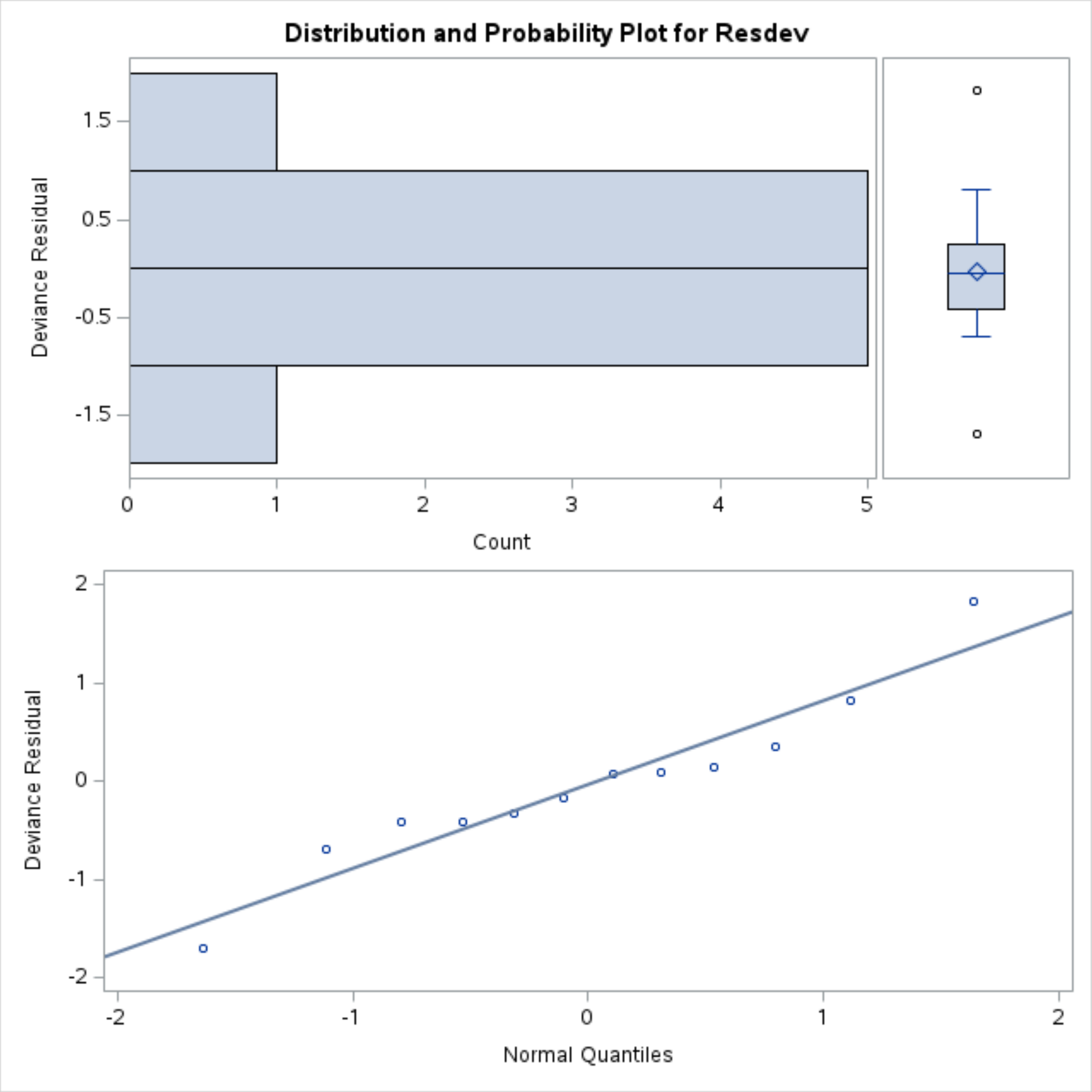
Log Linear Models

For the first examination cycle, a model assuming the joint independence of total cholesterol category and current smoking status from prevalent coronary heart disease status was selected (Table 5, Model 6). This model is as good as the saturated model (Model 1) according to the likelihood ratio test (Deviance = 8.0237, df = 5, p-value = 0.1549). The distribution of the deviance residuals is symmetric, with the mean equal to the median. From the box plot of deviance residuals, note the presence of two “outliers,” one in the lower tail and one in the upper tail (Figure 10). In the quantile-quantile plot, the deviance residuals closely follow the Q-Q line (Figure 10). Despite the pair of “outliers” in the box plot, all deviance residuals are smaller in magnitude than 2 (which is the usual cutoff for outliers); the largest deviance residual in magnitude is 1.83.

Table 5: Log Linear Models for Examination Cycle 1

Log Linear Model		Deviance	DF	P Value	AIC
1	Smoking | Heart Disease | Total Cholesterol	0	0	–	105.30
2	All Two Way Interactions	3.3052	2	0.1916	104.60
3	Conditional Independence of Total Cholesterol and Smoking	12.9123	4	0.0117	110.21
4	Conditional Independence of Total Cholesterol and Heart Disease	6.5537	4	0.1614	103.85
5	Conditional Independence of Smoking and Heart Disease	4.6905	3	0.1959	103.99
6	**Joint Independence of Total Cholesterol and Smoking from Heart Disease**	**8.0237**	**5**	**0.1549**	**103.32**
7	Joint Independence of Total Cholesterol and Heart Disease from Smoking	14.3823	5	0.0134	109.68
8	Joint Independence of Smoking and Heart Disease from Total Cholesterol	16.2455	6	0.0125	109.54
9	Mutual Independence of Smoking, Heart Disease, and Total Cholesterol	17.7155	7	0.0133	109.01

Figure 10: Distribution and Q-Q Plot of Deviance Residuals for Final Model for Examination Cycle 1 (Joint Independence of Total Cholesterol and Smoking from Heart Disease)

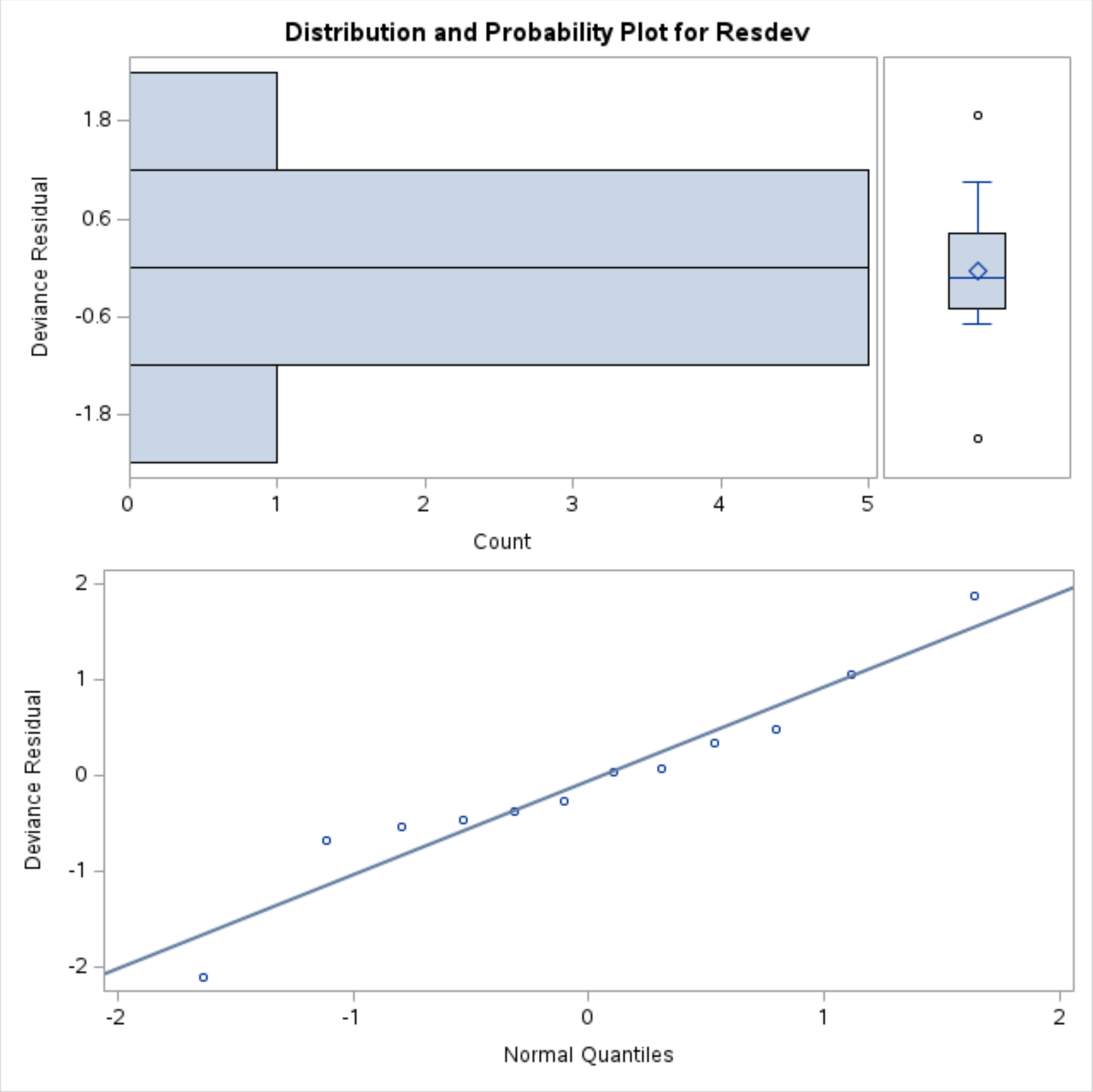


For the second examination cycle, a mutual independence model of current smoking status, total cholesterol category, and prevalent coronary heart disease status was selected (Table 6, Model 9). This model is as good as the saturated model (Model 1) according to the likelihood ratio test (Deviance = 10.6077, df = 7, p-value = 0.1567). The distribution of the deviance residuals appears symmetric in the histogram; however, the box plot of deviance residuals shows that the mean is larger than median (Figure 11). In addition, the box plot indicates the presence of two "outliers," one in the lower tail and one in the upper tail. In the quantile-quantile plot, the deviance residuals generally follow the Q-Q line (Figure 11). There is one deviance residual larger in magnitude than 2 (–2.11), which indicates the presence of an outlier. This outlier corresponds to the count of subjects who are current smokers, have prevalent coronary heart disease, and belong to the borderline high total cholesterol category; the observed count is 26 while the model's predicted count is approximately 38.

Table 6: Log Linear Models for Examination Cycle 2

	Log Linear Model	**Deviance**	**DF**	**P Value**	**AIC**
1	Smoking | Heart Disease | Total Cholesterol	0	0	–	104.83
2	All Two Way Interactions	0.9077	2	0.6352	101.74
3	Conditional Independence of Total Cholesterol and Smoking	2.4885	4	0.6467	99.32
4	Conditional Independence of Total Cholesterol and Heart Disease	4.5467	4	0.3371	101.38
5	Conditional Independence of Smoking and Heart Disease	5.1749	3	0.1594	104.01
6	Joint Independence of Total Cholesterol and Smoking from Heart Disease	8.9204	5	0.1123	103.75
7	Joint Independence of Total Cholesterol and Heart Disease from Smoking	6.8622	5	0.2311	101.70
8	Joint Independence of Smoking and Heart Disease from Total Cholesterol	6.2340	6	0.3975	99.07
9	**Mutual Independence of Smoking, Heart Disease, and Total Cholesterol**	**10.6077**	**7**	**0.1567**	**101.44**

Figure 11: Distribution and Q-Q Plot of Deviance Residuals for Final Model for Examination Cycle 2 (Mutual Independence of Smoking, Heart Disease, and Total Cholesterol)

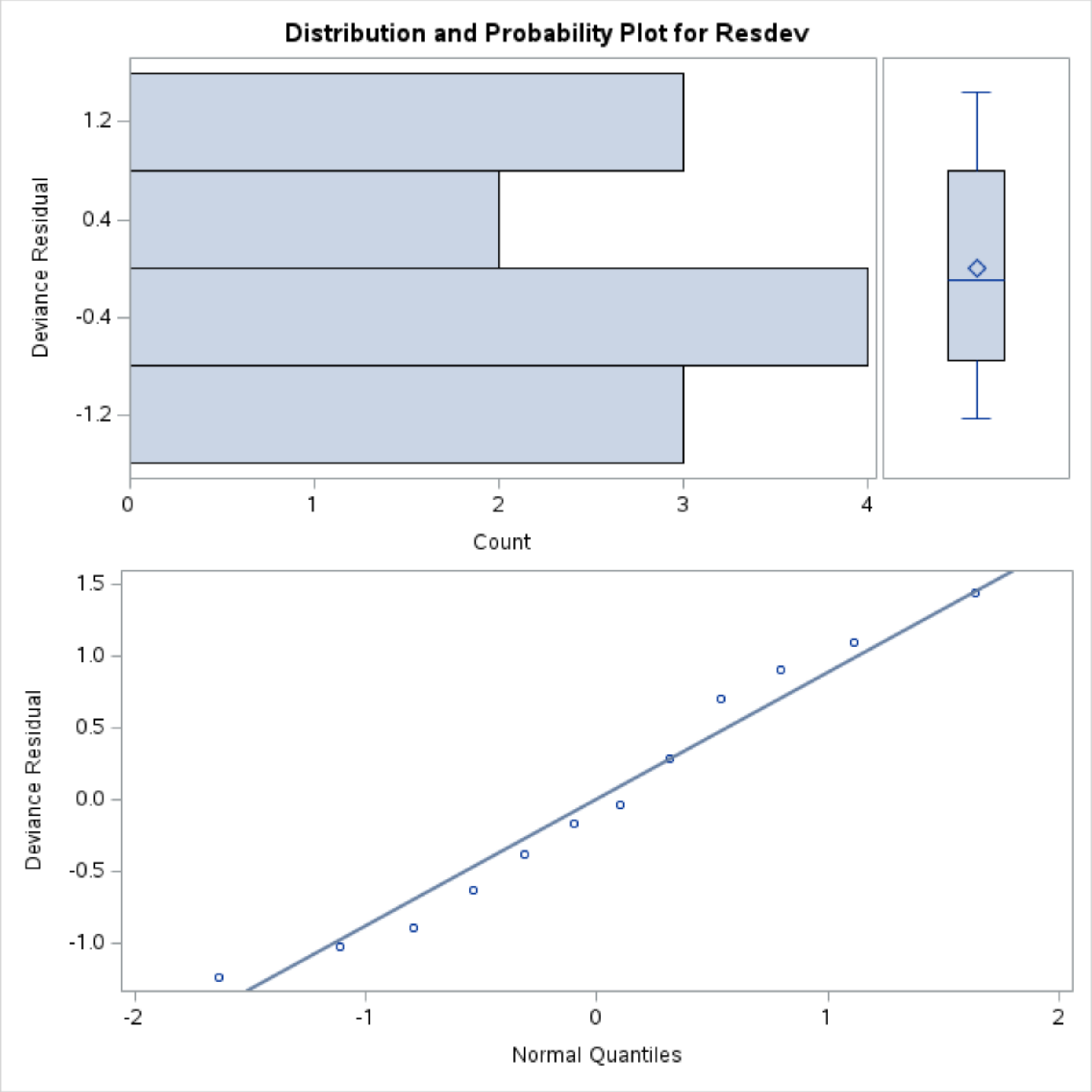


For the third examination cycle, a model assuming the joint independence of current smoking status and prevalent coronary heart disease status from total cholesterol category was selected (Table 7, Model 8). This model is as good as the saturated model (Model 1) according to the likelihood ratio test (Deviance = 8.5930, df = 6, p-value = 0.1978). The distribution of the deviance residuals is bimodal according to the histogram, and the box plot of deviance residuals shows that the mean is larger than median (Figure 12). In the quantile-quantile plot, the deviance residuals generally follow the Q-Q line (Figure 12). All deviance residuals are smaller in magnitude than 2; the largest deviance residual in magnitude is 1.44.

Table 7: Log Linear Models for Examination Cycle 3

Log Linear Model		**Deviance**	**DF**	**P Value**	**AIC**
1	Smoking | Heart Disease | Total Cholesterol	0	0	–	105.27
2	All Two Way Interactions	0.8669	2	0.6483	102.14
3	Conditional Independence of Total Cholesterol and Smoking	8.1459	4	0.0864	105.42
4	Conditional Independence of Total Cholesterol and Heart Disease	1.2069	4	0.8770	98.48
5	Conditional Independence of Smoking and Heart Disease	8.6618	3	0.0341	107.94
6	Joint Independence of Total Cholesterol and Smoking from Heart Disease	9.1088	5	0.1048	104.38
7	Joint Independence of Total Cholesterol and Heart Disease from Smoking	16.0479	5	0.0067	111.32
8	**Joint Independence of Smoking and Heart Disease from Total Cholesterol**	**8.5930**	**6**	**0.1978**	**101.87**
9	Mutual Independence of Smoking, Heart Disease, and Total Cholesterol	16.4949	7	0.0210	107.77

Figure 12: Distribution and Q-Q Plot of Deviance Residuals for Final Model for Examination Cycle 3 (Joint Independence of Smoking and Heart Disease from Total Cholesterol)



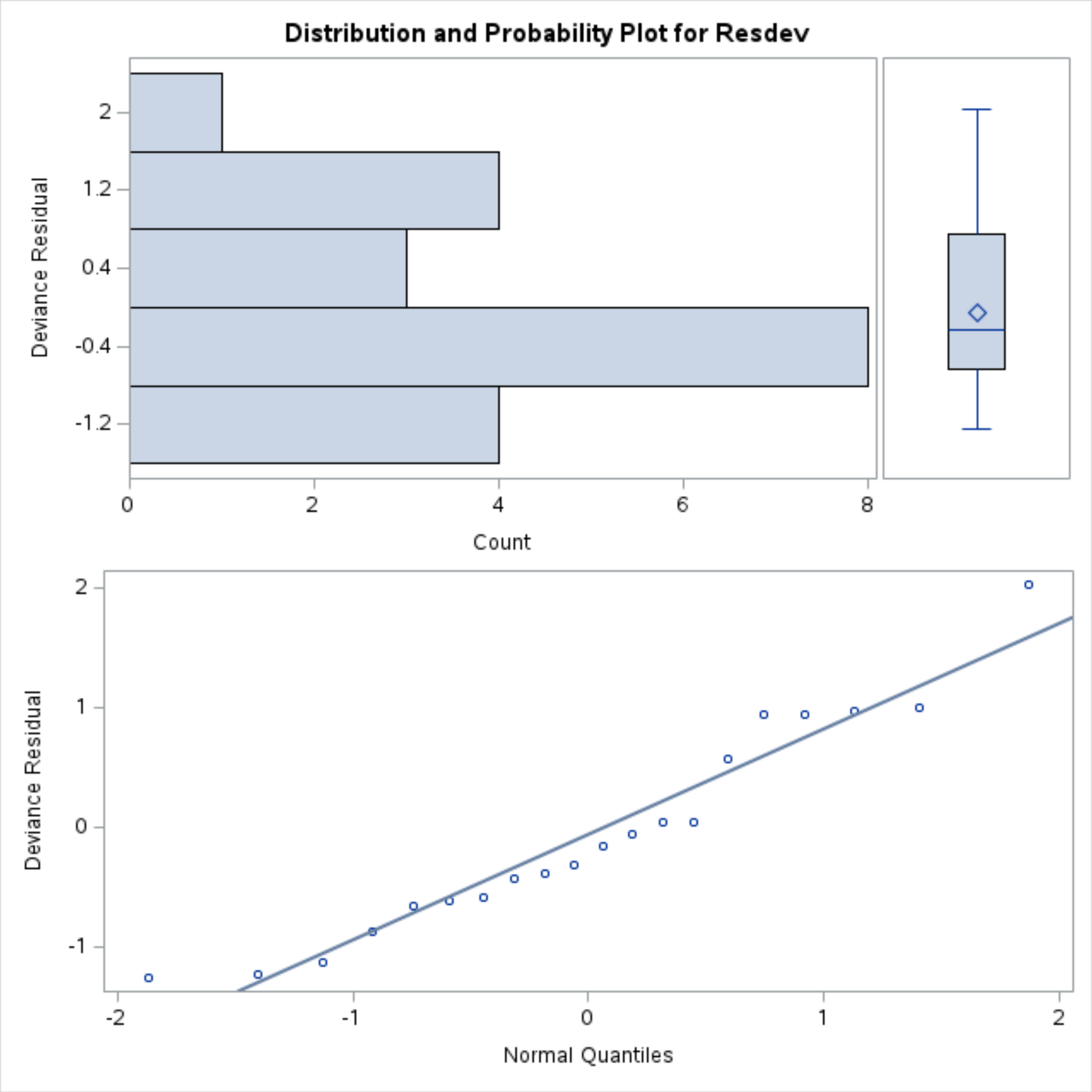
When examining the association between current smoking status, prevalent coronary heart disease status, and LDL cholesterol category during the final examination cycle, a model assuming the joint independence model of current smoking status and prevalent coronary heart disease status from LDL cholesterol category was selected (Table 8, Model 8). This model is as good as the saturated model (Model 1) according to the likelihood ratio test (Deviance = 14.8013, df = 12, p-value = 0.2525). The distribution of the deviance residuals is bimodal according to the histogram, and the box plot of deviance residuals shows that the mean is larger than median (Figure 13). In the quantile-quantile plot, the deviance residuals generally follow the Q-Q line (Figure 13). There is one deviance residual larger in magnitude than 2 (2.03), which indicates the presence of an outlier. This outlier corresponds to the count of subjects who are current smokers, do NOT have prevalent coronary heart disease, and belong to the near optimal LDL cholesterol category; the observed count is 125 while the model's predicted count is approximately 104.

Table 8: Log Linear Models with LDL Cholesterol for Examination Cycle 3

Log Linear Model		Deviance	DF	P Value	AIC
1	Smoking | Heart Disease | LDL Cholesterol	0	0	–	160.09
2	All Two Way Interactions	1.3305	4	0.8562	153.42
3	Conditional Independence of LDL Cholesterol and Smoking	10.5369	8	0.2293	154.63
4	Conditional Independence of LDL Cholesterol and Heart Disease	5.6112	8	0.6907	149.70
5	Conditional Independence of Smoking and Heart Disease	9.5184	5	0.0901	159.61
6	Joint Independence of LDL Cholesterol and Smoking from Heart Disease	13.7827	9	0.1303	155.87
7	Joint Independence of LDL Cholesterol and Heart Disease from Smoking	18.7085	9	0.0278	160.80
8	**Joint Independence of Smoking and Heart Disease from LDL Cholesterol**	**14.8013**	**12**	**0.2525**	**150.89**
9	Mutual Independence of Smoking, Heart Disease, and LDL Cholesterol	22.9728	13	0.0420	157.06
	LDL Cholesterol (cLDL) is treated as continuous for interaction terms in Models 10–25 (Denotes cLDL [Continuous LDL] interaction terms with Smoking and/or CHD [Heart Disease])				
10	All Two Way Interactions (cLDL*Smoking, cLDL^2*Smoking, cLDL*CHD, cLDL^2*CHD)	11.4424	8	0.1779	155.53
11	Conditional Independence of Smoking and Heart Disease (cLDL*Smoking, cLDL^2*Smoking, cLDL*CHD, cLDL^2*CHD)	19.5489	9	0.0209	161.64
12	All Two Way Interactions (cLDL*Smoking, cLDL^2*Smoking, cLDL*CHD)	11.5489	9	0.2400	153.64

13	Conditional Independence of Smoking and Heart Disease (cLDL*Smoking, cLDL^2*Smoking, cLDL*CHD)	19.6592	10	0.0327	159.75
14	All Two Way Interactions (cLDL*Smoking, cLDL*CHD, cLDL^2*CHD)	11.4544	9	0.2458	153.54
15	Conditional Independence of Smoking and Heart Disease (cLDL*Smoking, cLDL*CHD, cLDL^2*CHD)	19.5657	10	0.0336	159.66
16	All Two Way Interactions (cLDL*Smoking, cLDL*CHD)	11.5646	10	0.3153	151.65
17	Conditional Independence of Smoking and Heart Disease (cLDL*Smoking, cLDL*CHD)	19.6759	11	0.0500	157.77
18	Conditional Independence of LDL Cholesterol and Heart Disease (cLDL*Smoking, cLDL^2*Smoking)	14.6389	10	0.1458	154.73
19	Joint Independence of LDL Cholesterol and Smoking from Heart Disease (cLDL*Smoking, cLDL^2*Smoking)	22.8104	11	0.0188	160.90
20	Conditional Independence of LDL Cholesterol and Smoking (cLDL*CHD, cLDL^2*CHD)	11.5398	10	0.3170	151.63
21	Joint Independence of (LDL Cholesterol and Heart Disease) from Smoking (cLDL*CHD, cLDL^2*CHD)	19.7113	11	0.0495	157.80
22	Conditional Independence of LDL Cholesterol and Heart Disease (cLDL*Smoking)	14.6556	11	0.1988	152.75
23	Joint Independence of (LDL Cholesterol and Smoking) from Heart Disease (cLDL*Smoking)	22.8272	12	0.0292	158.92
24	Conditional Independence of LDL Cholesterol and Smoking (cLDL*CHD)	11.6500	11	0.3905	149.74
25	Joint Independence of (LDL Cholesterol and Heart Disease) from Smoking (cLDL*CHD)	19.8215	12	0.0705	155.91

Figure 13: Distribution and Q-Q Plot of Deviance Residuals for Final Model for Examination Cycle 3 (Joint Independence of Smoking and Heart Disease from LDL Cholesterol)



Logistic Regression Models

When ignoring age, the final model for the effect of smoking status and LDL cholesterol on the odds of prevalent coronary heart disease is the independence model, with the parameters displayed in Table 14. The model diagnostics are quite poor. The c-Statistic (AUC) is 0.554, which is very close to the lower bound of 0.50. Note that the c-Statistic for the full model is not much better (0.560). The percentage of correctly classified responses is just 47.7%, which is worse than the 50-50 probability of flipping a coin (compared with an essentially identical 47.7% in the full model). Therefore, the predictive capability of this model is inadequate.

Table 9: Final Model for Logistic Regression (Ignoring Age)

Effect	Estimate	Standard Error	DF	Wald Chi Square	P Value
Intercept	-2.2497	0.1679	1	179.4877	< 0.0001
Smoking Status (1 vs. 0)	-0.3626	0.1301	1	7.7723	0.0053
LDL Cholesterol Category	0.0917	0.0527	1	3.0277	0.0819
Percent Correctly Classified = 47.7%		Somers' D = 0.109		c-Statistic (AUC) = 0.554	

Table 10: Odds Ratios for Final Model for Logistic Regression (Ignoring Age)

Effect	Odds Ratio	95% Confidence Interval
Smoking Status (1 vs. 0)	0.696	(0.539, 0.898)
LDL Cholesterol Category	1.096	(0.988, 1.215)

Nevertheless, the parameters from this model can still be interpreted (with the understanding that this model is not very good). The parameter estimate for the smoking status variable is –0.3626. This indicates that the log odds of prevalent coronary heart disease are 0.3626 *lower* in smokers compared with nonsmokers when adjusting for LDL cholesterol. In a more understandable way, the odds of prevalent coronary heart disease are 30.4% *lower* in smokers compared with nonsmokers when LDL cholesterol category is fixed (OR = 0.696, 95% CI: (0.539, 0.898)) (Table 15). Observe that the associated p-value is 0.0053, so smoking is statistically significant at the 5% significance level but in the *opposite direction* of what would be expected, which is surprising. One would assume that smoking status would be associated with *higher* odds of prevalent coronary heart disease instead of lower odds. The parameter estimate for the LDL cholesterol category variable is 0.0917. This indicates that the log odds of prevalent coronary heart disease increase by 0.0917 for each unit increase in LDL cholesterol category when adjusting for smoking status. In a more interpretable manner, the odds of heart disease increase by 9.6% for each unit increase in LDL cholesterol category when smoking status is fixed (OR = 1.096, 95% CI: (0.988, 1.215)). Observe that the associated p-value is 0.0819, which indicates that LDL cholesterol is NOT statistically significant at the 5% significance level, which is surprising.

When including the confounding variable age, the final model for the effect of smoking status, LDL cholesterol, and age on the odds of prevalent coronary heart disease is the mutual independence model, with the parameters displayed in Table 16. The model diagnostics are much better than before. The c-Statistic (AUC) is 0.673, which is considerably higher than the 0.554 value from the previous model. The percentage of correctly classified responses is now 67.0%, which is a solid improvement from the 47.7% value before. Therefore, adjusting for the variable age significantly improves the predictive capability of the model.

Table 11: Final Model for Logistic Regression (Including Age)

Effect	Estimate	Standard Error	DF	Wald Chi Square	P Value
Intercept	-6.7405	0.5231	1	166.0138	< 0.0001
Smoking Status (1 vs. 0)	-0.0412	0.1366	1	0.0911	0.7628
LDL Cholesterol Category	0.0930	0.0533	1	3.0375	0.0814
Age	0.0702	0.00748	1	88.0033	< 0.0001
Percent Correctly Classified = 67.0%		Somers' D = 0.345		c-Statistic (AUC) = 0.673	

Table 12: Odds Ratios for Final Model for Logistic Regression (Including Age)

Effect	Odds Ratio	95% Confidence Interval
Smoking Status (1 vs. 0)	0.960	(0.734, 1.254)
LDL Cholesterol Category	1.097	(0.988, 1.218)
Age	1.073	(1.057, 1.089)

The parameter estimate for the smoking status variable is now –0.0412, which is a considerable change from the –0.3626 value in the prior model. This indicates that the log odds of prevalent coronary heart disease are 0.0412 *lower* in smokers compared with nonsmokers when adjusting for LDL cholesterol and age. In a more understandable way, the odds of prevalent coronary heart disease are 4.0% *lower* in smokers compared with nonsmokers when LDL cholesterol category and age are held constant (OR = 0.960, 95% CI: (0.734, 1.254)) (Table 17). Despite the fact that the direction of association is opposite what one would expect, observe that the p-value is 0.7628, so smoking is no longer statistically significant. Clearly, age confounds the relationship between smoking status and the odds of prevalent coronary heart disease. The parameter estimate for the LDL cholesterol category variable is 0.0930, which is nearly unchanged from the 0.0917 value in the previous model. This indicates that the log odds of prevalent coronary heart disease increase by 0.0930 for each unit increase in LDL cholesterol category when adjusting for smoking status and age. In a more interpretable manner, the odds of prevalent coronary heart disease increase by 9.7% for each unit increase in LDL cholesterol category when smoking status and age are held fixed (OR = 1.097, 95% CI: (0.988, 1.218)). The associated p-value is 0.0814, which is nearly the same as before, and indicates that LDL cholesterol remains NOT statistically significant at the 5% significance level. The parameter estimate for the age variable is 0.0702. This indicates that the log odds of prevalent coronary heart disease increase by 0.0702 for each one year increase in age when adjusting for smoking status and LDL cholesterol category. In a more interpretable manner, the odds of prevalent coronary heart disease increase by 7.3% for each one year increase in age when smoking status and LDL cholesterol category are held constant (OR = 1.073, 95% CI: (1.057, 1.089)). The associated p-value is < 0.0001, so age is highly statistically significant, which is expected because older age is a known risk factor for coronary heart disease.

Discussion

We ran separate log linear models for each examination cycle using the collected data and treating cholesterol as both categorical and continuous. Across all log linear models, prevalent coronary heart disease and cholesterol category (total or LDL) were found to be independent. The independence of coronary heart disease and cholesterol is surprising and may be due to failing to control for confounding variables, such as age. For logistic regression models during the final examination cycle, being in a higher LDL cholesterol category was associated with higher odds of prevalent coronary heart disease, but this association did not rise to the level of statistical significance at the 5% significance level regardless of adjustment for age. Smoking was surprisingly associated with *lower* odds of prevalent coronary heart disease compared with not smoking. This association was statistically significant when failing to adjust for the confounding variable age. Once adjusting for age, this association was no longer significant at the 5% significance level. Explanations for this unexpected finding include not adjusting for other confounders and survivorship bias. That is, only the healthiest smokers are still alive at the final examination cycle and thus have a lower prevalence of coronary heart disease compared with other smokers who already passed away. Therefore, the "healthiest" smokers are being compared with "ordinary" non-smokers, which is biasing the association between smoking and prevalent coronary heart disease.

Further research should be conducted on the association between the variables in this dataset and heart disease using GEE and/or generalized linear mixed effects models. For example, one outcome of interest is the binary variable prevalent coronary heart disease status, which was measured repeatedly over time (correlated measurements). In addition, we would want to address omitted variables bias/confounding. For instance, the dataset contains 38 total variables, many of which were never analyzed and may confound the relationship between the predictors cholesterol, smoking status, and age and the primary outcome heart disease status. In addition, the results are not particularly generalizable to the general population since the data was collected from a single region with a lack of diversity among study participants.

References

Ambrose, John A, and Rajat S Barua. "The Pathophysiology of Cigarette Smoking and Cardiovascular Disease: An Update." *Journal of the American College of Cardiology* 43, no. 10 (May 19, 2004): 1731–37. https://doi.org/10.1016/j.jacc.2003.12.047.

Després, J P, I Lemieux, G R Dagenais, B Cantin, and B Lamarche. "HDL-Cholesterol as a Marker of Coronary Heart Disease Risk: The Québec Cardiovascular Study." *Atherosclerosis* 153, no. 2 (December 2000): 263–72. https://doi.org/10.1016/s0021-9150(00)00603-1.

Fernandez, Maria Luz, and Densie Webb. "The LDL to HDL Cholesterol Ratio as a Valuable Tool to Evaluate Coronary Heart Disease Risk." *Journal of the American College of Nutrition* 27, no. 1 (February 2008): 1–5. https://doi.org/10.1080/07315724.2008.10719668.

Lipid Panel. Lipid Panel | Johns Hopkins Medicine. (2020, December 4). https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel

National Center for Health Statistics (NCHS). "FastStats - Leading Causes of Death." Centers for Disease Control and Prevention, January 18, 2023. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

Wadhera, Rishi K, Dylan L Steen, Irfan Khan, Robert P Giugliano, and JoAnne M Foody. "A Review of Low-Density Lipoprotein Cholesterol, Treatment Strategies, and Its Impact on Cardiovascular Disease Morbidity and Mortality." *Journal of Clinical Lipidology* 10, no. 3 (2016): 472–89. https://doi.org/10.1016/j.jacl.2015.11.010.

Appendix

Figure A1: Odds of Prevalent Coronary Heart Disease vs. Total Cholesterol Category Stratified by Current Smoking Status for Examination Cycle 1

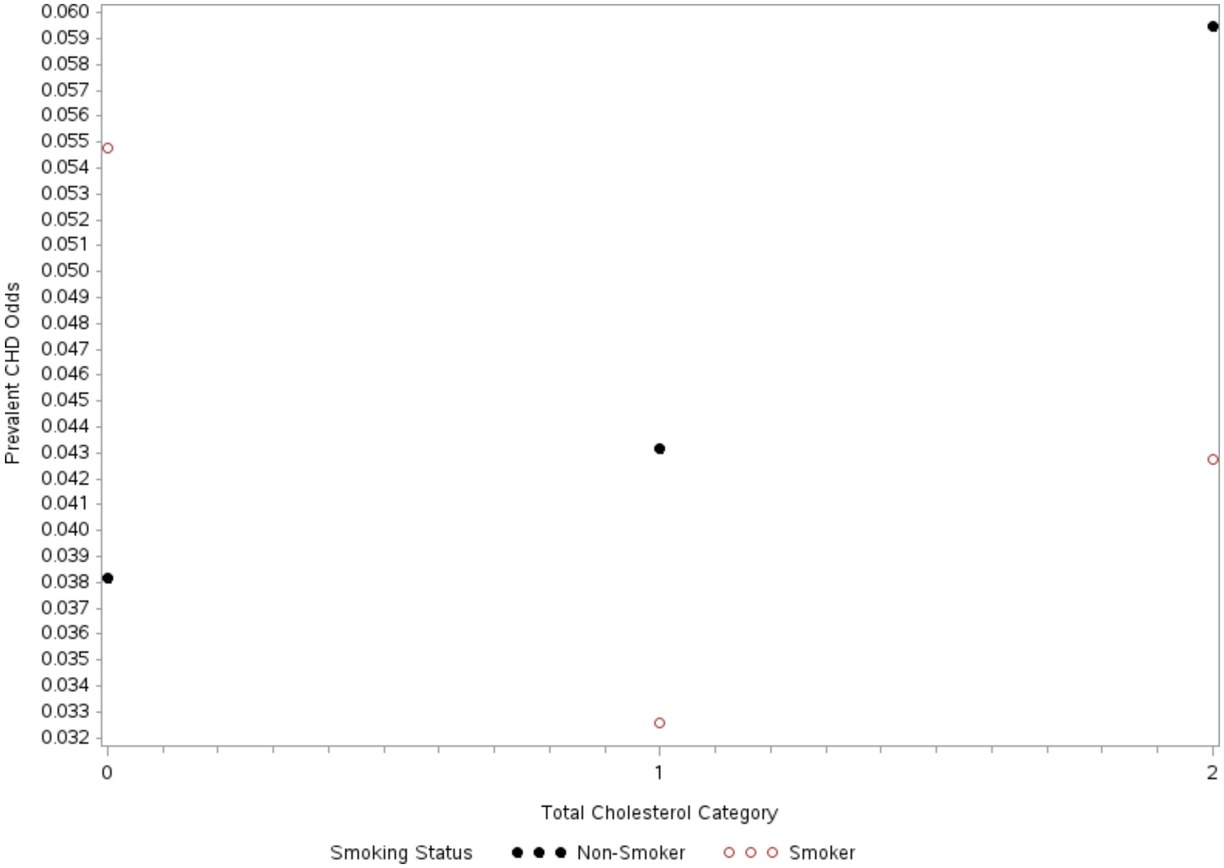


Figure A2: Odds of Prevalent Coronary Heart Disease vs. Total Cholesterol Category Stratified by Current Smoking Status for Examination Cycle 2

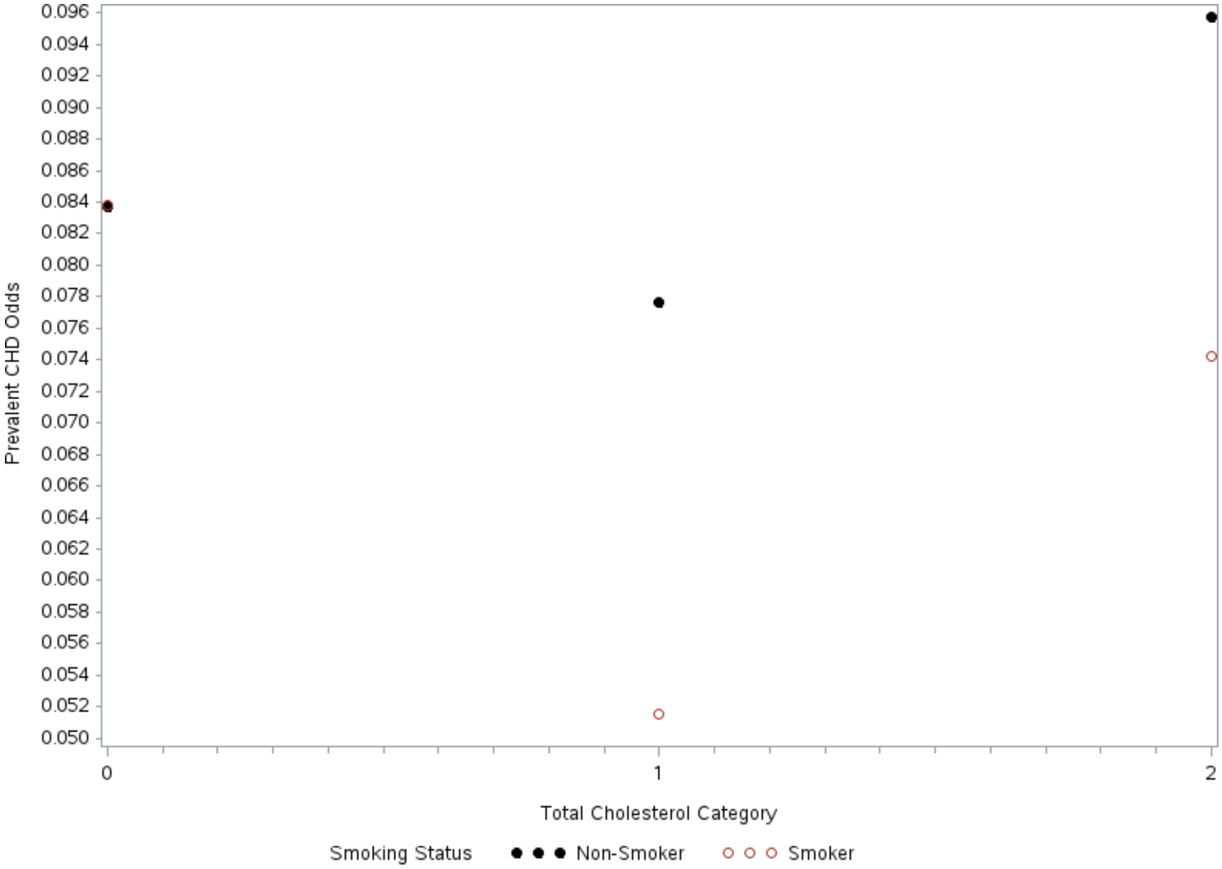


Figure A3: Odds of Prevalent Coronary Heart Disease vs. Total Cholesterol Category Stratified by Current Smoking Status for Examination Cycle 3

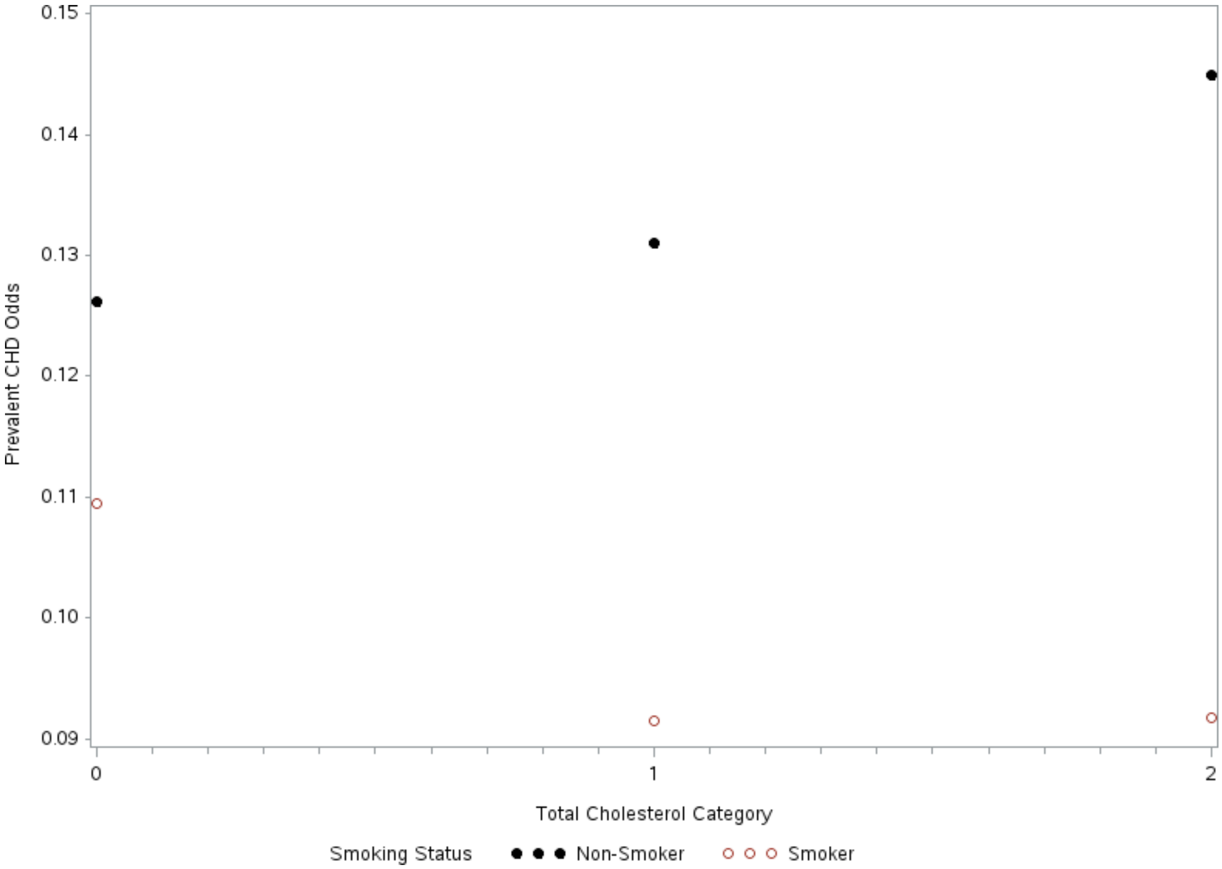
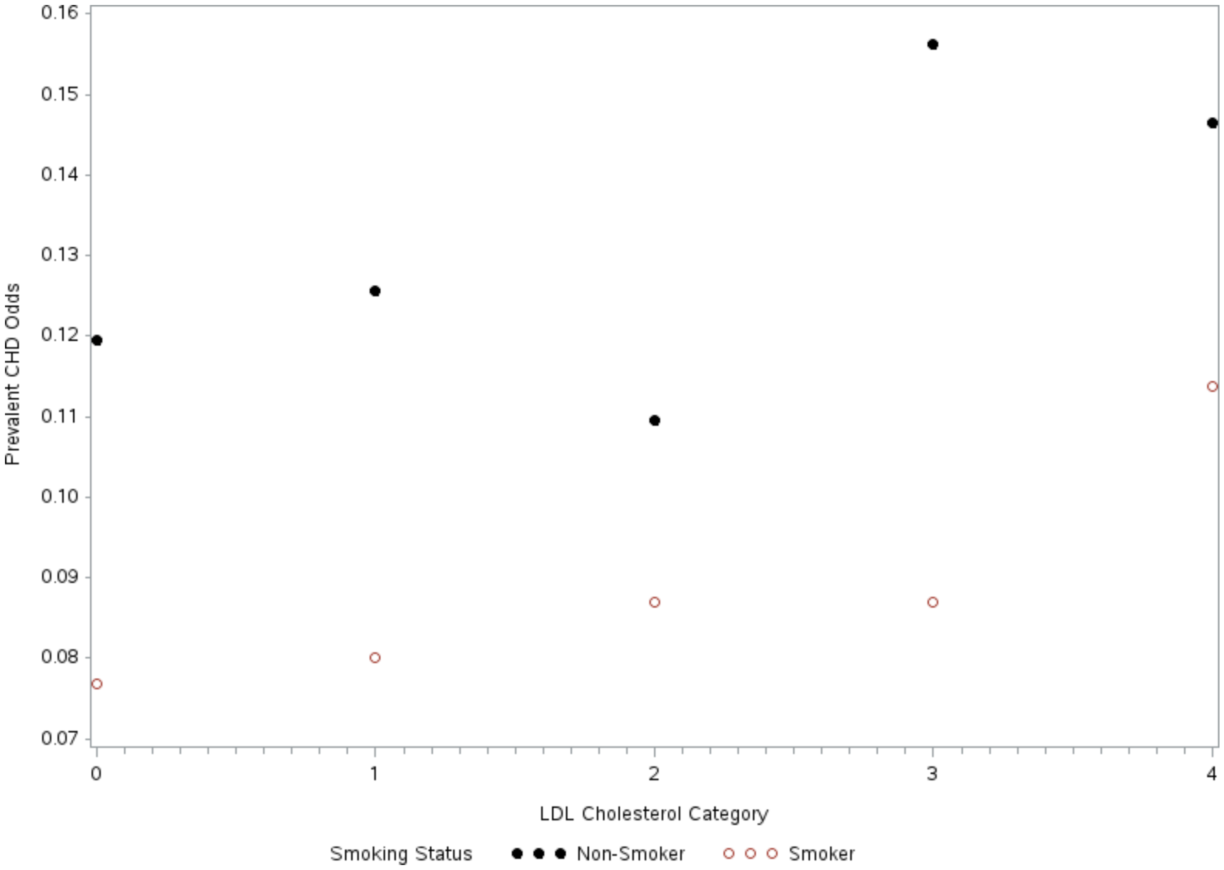


Figure A4: Odds of Prevalent Coronary Heart Disease vs. LDL Cholesterol Category Stratified by Current Smoking Status for Examination Cycle 3


Prevalent CHD Odds
LDL Cholesterol Category
Smoking Status
● ● ● Non-Smoker
○ ○ ○ Smoker