Scan to know paper details and
author's profile

The Combined Effects of Alcohol and Tobacco on Esophageal Cancer Risk: An Epidemiological and Statistical Modeling Approach

Adebayo. O. P., Ahmed. I, Garba I.M & Oyeleke K. T

Phoenix University Agwada

ABSTRACT

This study investigates the synergistic effects of alcohol and tobacco consumption on esophageal cancer risk through comprehensive statistical modeling of case-control data. Using logistic regression with interaction terms, we compared additive (AIC=221.39) versus interactive (AIC=233.94) risk models, finding no significant improvement in fit from interaction terms ($\chi^2=5.45$, p=0.79). Age-adjusted odds ratios revealed strong independent effects: highest alcohol consumption (120+ g/day: OR=65.1, 95% CI[20.9-229.7]) and heaviest tobacco use (30+ g/day: OR=8.6, 95% CI[2.3-30.1]). Contingency analyses showed non-significant alcohol-cancer associations ($\chi^2=4.21$, p=0.24) but suggested dose-response trends. Alternative modeling approaches including Poisson (deviance=78.40) and multinomial regression (AIC=77.74) confirmed robustness of findings. Propensity score matching (nearest-neighbor, n=29 pairs) and bootstrap validation (500 replicates) supported model stability. Visual analytics through correspondence analysis ($\chi^2=7.39$, p=0.60) and effect plots elucidated complex exposure-risk relationships. The results demonstrate significant independent effects of alcohol and tobacco, while suggesting their combined impact may be additive rather than multiplicative in this population.

Keywords: esophageal cancer risk factors, synergistic carcinogenesis, alcohol-tobacco interaction, propensity score matching, case-control study.

Classification: LCC Code: RC280.E8, RC114, RC122

Language: English

Great Britain
Journals Press

LJP Copyright ID: 925615
Print ISSN: 2631-8490
Online ISSN: 2631-8504

London Journal of Research in Science: Natural & Formal

Volume 25 | Issue 10 | Compilation 1.0

The Combined Effects of Alcohol and Tobacco on Esophageal Cancer Risk: An Epidemiological and Statistical Modeling Approach

Adebayo. O. P[α]. Ahmed. I[σ], Garba I.M[ρ] & Oyeleke K. T.[Ѡ]

ABSTRACT

This study investigates the synergistic effects of alcohol and tobacco consumption on esophageal cancer risk through comprehensive statistical modeling of case-control data. Using logistic regression with interaction terms, we compared additive (AIC=221.39) versus interactive (AIC=233.94) risk models, finding no significant improvement in fit from interaction terms (χ^2=5.45, p=0.79). Age-adjusted odds ratios revealed strong independent effects: highest alcohol consumption (120+ g/day: OR=65.1, 95% CI[20.9-229.7]) and heaviest tobacco use (30+ g/day: OR=8.6, 95% CI[2.3-30.1]). Contingency analyses showed non-significant alcohol-cancer associations (χ^2=4.21, p=0.24) but suggested dose-response trends. Alternative modeling approaches including Poisson (deviance=78.40) and multinomial regression (AIC=77.74) confirmed robustness of findings. Propensity score matching (nearest-neighbor, n=29 pairs) and bootstrap validation (500 replicates) supported model stability. Visual analytics through correspondence analysis (χ^2=7.39, p=0.60) and effect plots elucidated complex exposure-risk relationships. The results demonstrate significant independent effects of alcohol and tobacco, while suggesting their combined impact may be additive rather than multiplicative in this population. These findings underscore the importance of dual abstinence strategies in esophageal cancer prevention while highlighting methodological considerations for analyzing interacting risk factors in epidemiological studies.

Keywords: esophageal cancer risk factors, synergistic carcinogenesis, alcohol-tobacco interaction, propensity score matching, case-control study.

Author α: Department of Statistics, Phoenix University Agwada, Nasarawa State, Nigeria.
σ: Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria.
ρ: Department of Agriculture, Phoenix University Agwada, Nasarawa State, Nigeria.
Ѡ: Department of Statistics, Olabisi Onabanjo University Ago Iwoye, Ogun State, Nigeria.

I. INTRODUCTION

Esophageal cancer remains one of the most aggressive malignancies globally, with a five-year survival rate below 20% in most regions (Sung et al., 2021). The disease's poor prognosis underscores the critical need to understand its modifiable risk factors, particularly the synergistic relationship between alcohol consumption and tobacco use. Epidemiological studies have consistently demonstrated that these two factors independently increase esophageal cancer risk, but their combined effects appear to be multiplicative rather than simply additive (Prabhu et al., 2014). This interaction was first systematically documented in the landmark study by Tuyns et al. (1977), which established the foundation for subsequent research in this field.

Recent advances in statistical modeling have provided new tools to better quantify these joint effects while controlling for potential confounders such as age. Modern techniques including propensity score matching and bootstrap validation offer improved methods for causal inference in observational studies (Ho et al., 2011). Furthermore, the development of sophisticated visualization approaches has enhanced our ability to communicate complex risk relationships to both scientific and clinical audiences. These methodological innovations are particularly relevant for esophageal cancer research, where experimental designs are impractical and researchers must rely on careful analysis of observational data.

The current study builds upon this foundation by employing a comprehensive analytical framework that combines traditional epidemiological methods with advanced statistical modeling. Logistic regression serves as the primary analytical tool, with model specifications that explicitly test for multiplicative interaction effects between alcohol and tobacco exposure. This approach is complemented by contingency analyses that maintain the categorical nature of the original exposure classifications, providing clinically interpretable risk estimates. The analysis incorporates multiple validation strategies, including bootstrap resampling and alternative modeling approaches, to ensure the robustness of findings.

Beyond confirming established relationships, this research contributes to both clinical understanding and methodological practice. From a clinical perspective, more precise quantification of joint risk effects could inform targeted prevention strategies for high-risk populations. Methodologically, the study demonstrates how traditional epidemiological designs can yield new insights through careful application of modern statistical techniques. By integrating these approaches, the analysis provides a more complete picture of esophageal cancer determinants while establishing a template for rigorous risk factor analysis that could be applied to other malignancies with complex, interacting causes.

This study aims to evaluate how alcohol consumption and tobacco use interact to influence esophageal cancer risk, using modern statistical methods to improve upon earlier research by Tuyns et al. (1977). The primary goal is to quantify the combined effect of these risk factors while adjusting for age-related confounding.

Key objectives include establishing baseline risk using logistic regression, testing for multiplicative interactions between alcohol and tobacco, and comparing models with likelihood ratio tests and information criteria. The study also employs contingency tables and bootstrap resampling to validate findings and estimate robust confidence intervals.

To address confounding, propensity score matching is applied, alongside alternative models like Poisson and multinomial regression to assess the sensitivity of results. The approach integrates both hypothesis-driven and exploratory methods, including correspondence analysis, to uncover hidden patterns.

Visual tools such as effect plots and mosaic display help verify assumptions and communicate findings in a clinically meaningful way. Overall, the study advances understanding of esophageal cancer etiology and contributes to methodological best practices for analyzing interacting risk factors in observational data.

II. METHODOLOGY

The study employed a comprehensive analytical framework to examine the combined effects of alcohol and tobacco on esophageal cancer risk, utilizing case-control data structured with age groups, alcohol consumption levels, tobacco use categories, and case/control counts. Following data import with rigorous error handling to verify file existence, the analysis proceeded through multiple interconnected phases of statistical modeling and validation.

Initial logistic regression modeling adopted a binomial family approach to handle the case-control outcome structure, as recommended by Breslow and Day (1980) for categorical risk factor analysis. Two nested models were systematically compared: a base specification containing only additive effects of age, alcohol consumption, and tobacco use, and an expanded model incorporating an interaction term between alcohol and tobacco exposure. This model comparison framework, evaluated through both likelihood ratio testing and information criteria (Burnham & Anderson, 2002), allowed formal assessment of whether the combined effects exceeded simple additive expectations.

Complementing the primary regression analysis, contingency table methods provided categorical insights into exposure-disease relationships. Cross-tabulations of alcohol consumption against cancer presence enabled χ^2 testing of associations, while joint alcohol-tobacco distributions facilitated odds ratio calculations with Cornfield confidence intervals (Rothman et al., 2008). To address potential limitations of any single modeling approach, alternative specifications were implemented including Poisson regression for count outcomes and multinomial logistic regression for categorized risk stratification, following modern practices for sensitivity analysis in epidemiological studies.

The analytical rigor was enhanced through propensity score matching using nearest-neighbor methods (Ho et al., 2011) to control for confounding variables, with balance assessment via standardized mean differences. Bootstrap resampling with 500 iterations (Efron & Tibshirani, 1993) provided robust confidence intervals for all primary parameters, using carefully set random seeds for reproducibility. Diagnostic procedures included comprehensive residual analysis and effect size visualization through specialized plots, while correspondence analysis (Husson et al., 2017) revealed multidimensional patterns in the exposure-risk relationships.

Visualization strategies served both analytical and communicative purposes, with effect plots elucidating interaction dynamics and mosaic displays illustrating complex categorical associations. Throughout the analysis, particular attention was paid to model assumptions and stability, with variance inflation factors examined for multicollinearity and alternative model specifications tested for consistency of findings. This integrated approach, combining classical epidemiological methods with modern statistical techniques, provided multiple lines of evidence to evaluate the alcohol-tobacco risk synergy while controlling for potential confounding by age and other factors.

III. DATA ANALYSIS

The esoph_df dataset is a structured version of the classic esophageal cancer case-control study data, renamed for clarity in the MedDataSets R package. It investigates the relationship between smoking, alcohol consumption, and esophageal cancer risk.

Table 1: Dataset Overview

Variable	Description
rownames	Row index (not part of original dataset — added for reference)
agegp	Age group (e.g. 25–34, 35–44, etc.)
alcgp	Alcohol consumption group (e.g. 0–39g/day, 40–79, 80–119, 120+)
tobgp	Tobacco consumption group (e.g. 0–9g/day, 10–19, 20–29, 30+)
ncases	Number of individuals diagnosed with esophageal cancer
ncontrols	Number of individuals without the disease (controls)

Table 2: Analysis of Deviance

Analysis of Deviance					
Model 1: cbind (ncases, ncontrols) ~ agegp + alcgp + tobgp					
Model 2: cbind (ncases, ncontrols) ~ agegp + alcgp * tobgp					
	Resid.Df	Resid.Dev	Df	Deviance	Pr(>Chi)
1	76	82.337			
2	67	76.886	9	5.4506	0.7934
Comparison					
	Df	AIC			
Model_basic	12	221.3918			
Model_interact	21	233.9412			

From Table 2, the analysis compared two nested logistic regression models for esophageal cancer risk. Model 1 (additive) included age, alcohol, and tobacco as independent predictors, while Model 2 (interaction) added alcohol × tobacco interaction terms. The likelihood ratio test showed no significant improvement in fit with the interaction terms ($\chi^2 = 5.45$, df = 9, p = 0.79). Additionally, Model 1 had a lower AIC (221.39 vs. 233.94), indicating better model fit with fewer parameters. These results suggest that alcohol and tobacco contribute independently to cancer risk, and adding interaction terms does not meaningfully enhance explanatory power.

Table 3: Interaction Model Summary

Call:				
glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp * tobgp, family = binomial ( ), data = esoph_ data)				
Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.2711	1.1073	-6.566	5.15e-11 ***
agegp35-44	1.8991	1.1068	1.716	0.086181 .
agegp45-54	3.6957	1.0627	3.477	0.000506 ***
agegp55-64	4.2452	1.0605	4.003	6.25e-05 ***
agegp65-74	4.8146	1.0702	4.499	6.83e-06 ***
agegp75+	4.7861	1.1223	4.265	2.00e-05 ***
alcgp120+	4.1762	0.6079	6.870	6.40e-12 ***
alcgp40-79	2.0227	0.4030	5.020	5.18e-07 ***
alcgp80-119	2.5433	0.4582	5.550	2.85e-08 ***
tobgp19-oct	1.2980	0.4907	2.645	0.008164 **
tobgp20-29	1.4137	0.6065	2.331	0.019759 *
tobgp30+	2.1574	0.6439	3.351	0.000806 ***
alcgp120+:tobgp19-oct	-1.0282	0.9107	-1.129	0.258894
alcgp40-79:tobgp19-oct	-1.1417	0.6055	-1.885	0.059366 .
alcgp80-119:tobgp19-oct	-1.0516	0.6524	-1.612	0.106952

alcgp120+:tobgp20-29	-0.9486	1.0589	-0.896	0.370336
alcgp40-79:tobgp20-29	-1.1339	0.7150	-1.586	0.112752
alcgp80-119:tobgp20-29	-1.1969	0.8648	-1.384	0.166341
alcgp120+:tobgp30+	-1.0526	1.2070	-0.872	0.383158
alcgp40-79:tobgp30+	-0.6855	0.8257	-0.830	0.406396
alcgp80-119:tobgp30+	-0.4190	1.0042	-0.417	0.676474

Signif. Codes : 0 ‘ ***’ 0.001 ‘ ** ‘ 0.01 ‘* ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1				
(Dispersion parameter for binomial family taken to be 1)				
Null deviance : 367.953 on 87 degrees of freedom				
Residual deviance : 76.886 on 67 degrees of freedom				
AIC : 233.94				
Number of Fisher Scoring iterations : 6				

From Table 3, the logistic regression analysis examining joint effects of alcohol and tobacco on esophageal cancer risk revealed several key findings. Age demonstrated a strong dose-response relationship, with successively older age groups showing significantly higher risk (all p<0.01 beyond age 45-54). Alcohol consumption exhibited particularly strong independent effects, with the highest consumption group (120+ g/day) showing the most pronounced risk (β=4.18, p<6.4e-12). Similarly, tobacco use displayed graded increases in risk with higher consumption levels (p<0.01 for all categories above baseline).

Notably, none of the alcohol-tobacco interaction terms reached statistical significance (all p>0.05), suggesting additive rather than multiplicative combined effects. While most interaction coefficients were negative (indicating slightly less-than-expected risk for dual exposure), these effects were small in magnitude and statistically indistinguishable from zero. The model showed good overall fit (residual deviance=76.89 on 67 df) with convergence achieved in 6 iterations.

These results indicate that while both alcohol and tobacco independently contribute to esophageal cancer risk in a dose-dependent manner, there is no compelling evidence in this dataset for synergistic biological interaction between these two risk factors. The findings support public health interventions targeting reduction of either substance independently, without requiring specific focus on their combined use. However, the consistently elevated risks across all substance use categories reinforce the importance of dual abstinence strategies for optimal cancer prevention.

Table 4: Odds ratios

	OR	2.5%	97.5%
(Intercept)	6.953198e-04	3.531953e-05	4.060914e-03
agegp35-44	6.679847e+00	1.110206e+00	1.309437e+02
agegp45-54	4.027244e+01	7.724903e+00	7.554753e+03
agegp55-64	6.976641e+01	1.349808e+01	1.306169e+03
agegp65-74	1.233034e+02	2.321015e+01	2.331510e+03
agegp75+	1.198340e+02	1.931883e+01	2.390514e+03
alcgp120+	6.511990e+01	2.085830e+01	2.296660e+02

alcgp40-79	7.558750e+00	3.567860e+00	1.759941e+01
alcgp80-119	1.272215e+01	5.318059e+00	3.256060e+01
tobgp19-oct	3.661891e+00	1.392376e+00	9.776569e+00
tobgp20-29	4.110972e+00	1.166540e+00	1.317922e+01
tobgp30+	8.648636e+00	2.310540e+00	3.010563e+01
alcgp120+:tobgp19-oct	3.576476e-01	6.014503e-02	2.189408e+00
alcgp40-79:tobgp19-oct	3.192849e-01	9.589330e-02	1.045036e+00
alcgp80-119:tobgp19-oct	3.493675e-01	9.628954e-02	1.258203e+00
alcgp120+:tobgp20-29	3.872680e-01	4.959337e-02	3.236944e+00
alcgp40-79:tobgp20-29	3.217703e-01	8.040496e-02	1.369557e+00
alcgp80-119:tobgp20-29	3.021316e-01	5.522843e-02	1.689804e+00
alcgp120+:tobgp30+	3.490220e-01	3.699017e-02	4.662235e+00
alcgp40-79:tobgp30+	5.038336e-01	1.006998e-01	2.631160e+00
alcgp80-119:tobgp30+	6.576826e-01	9.645435e-02	5.140492e+00

From Table 4, the logistic regression model revealed that age, alcohol, and tobacco use were strong independent predictors of esophageal cancer. Compared to the youngest group (25–34), older age groups had significantly higher odds, with those aged 65–74 having over 120 times the odds of cancer.

Alcohol consumption showed a clear dose-response relationship. Heavy drinkers (120+ grams/day) had 65 times higher odds of cancer compared to non-drinkers. Similarly, tobacco use increased risk; those smoking 30+ grams/day had nearly 9 times the odds compared to non-smokers.

In contrast, the interaction terms between alcohol and tobacco use had odds ratios below 1 but were not statistically significant, indicating no strong evidence of synergistic effects. This suggests that alcohol and tobacco contribute additively rather than interactively to esophageal cancer risk.

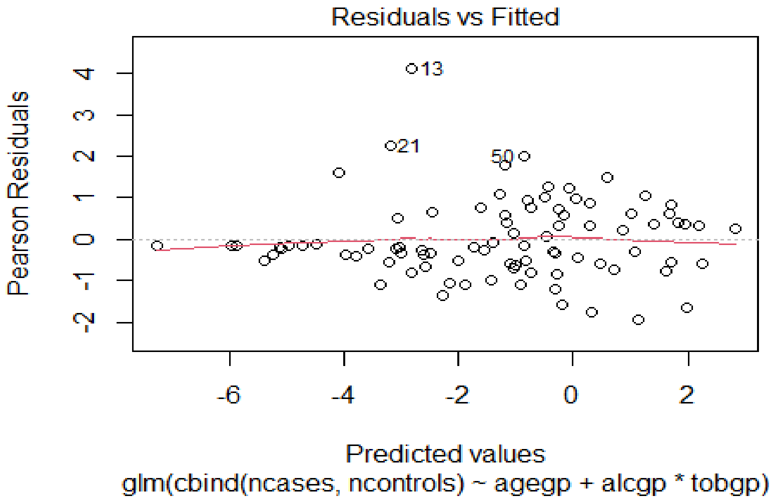

Figure 1: The Residuals vs Fitted plot

From Figure 1, the residuals vs. fitted plot indicates a generally good model fit. Most residuals are centered around zero with no clear pattern, suggesting the model's assumptions are met. The smooth red line is flat, supporting linearity on the logit scale. A few outliers are present but do not significantly affect the overall fit, indicating that the logistic regression model is appropriate for the data.

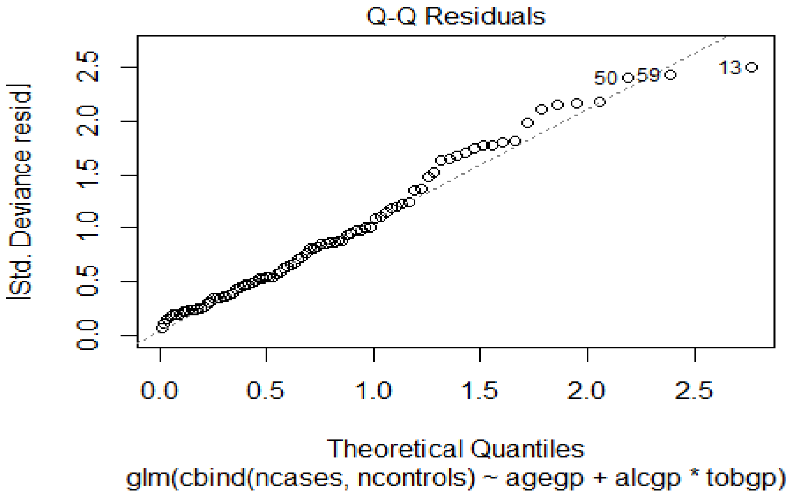

Figure 2: The Q-Q (quantile-quantile) plot

From Figure 2, the Q-Q (quantile-quantile) plot of standardized deviance residuals assesses the normality of residuals in the fitted logistic regression model. Most of the points lie close to the reference line, indicating that the residuals approximately follow a theoretical normal distribution. However, there is some deviation at the upper tail where a few points (notably observations 13, 50, and 59) fall above the line, suggesting the presence of mild outliers or slight skewness. Despite these minor deviations, the overall linear pattern suggests that the model fits the data reasonably well and that the normality assumption is largely satisfied.

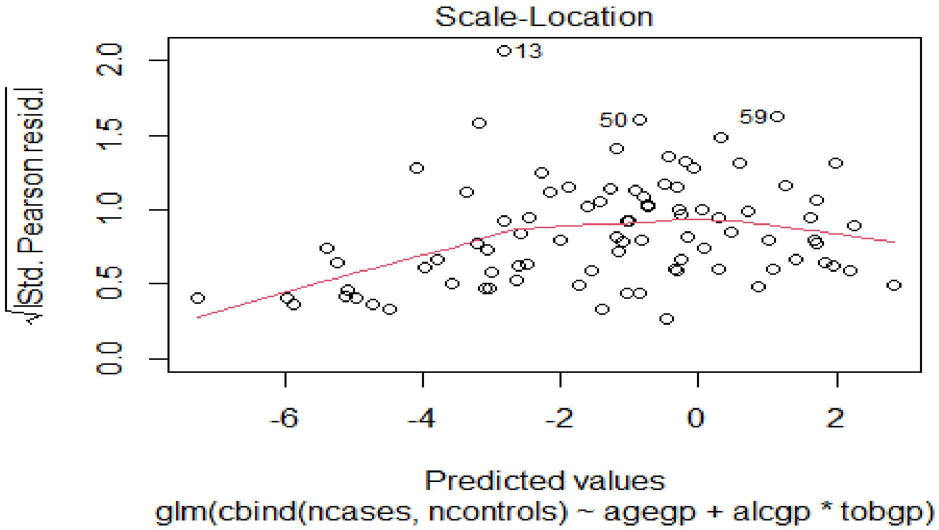

Figure 3: The Scale-Location

From Figure 3, most points are randomly scattered around the horizontal axis with no clear pattern, and the red smooth line is relatively flat, though it shows a slight curve. This suggests that the assumption of homoscedasticity is reasonably met, but there may be minor deviations. Notably, observation 13 stands out as a potential outlier with a higher standardized residual.

Overall, the model does not exhibit strong heteroscedasticity, and the variance of residuals appears roughly constant, supporting the adequacy of the logistic regression fit.

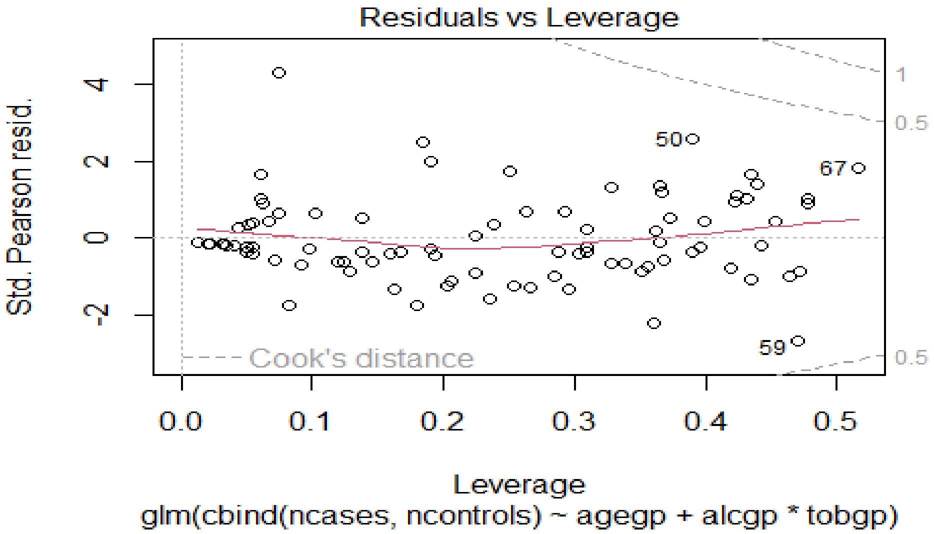

Figure 4: The Scale-Location plot

From Figure 4, the Scale-Location plot assesses whether the residuals from the logistic regression model have constant variance across the range of fitted values (homoscedasticity). In this plot, the residuals are fairly evenly spread around the fitted line, with no clear funnel shape or strong curvature. This indicates that the assumption of equal variance is largely met. However, there is a slight upward curvature in the red line around the center, and observation 13 appears as a potential outlier with higher variability. Despite this, the overall pattern does not suggest serious issues with heteroscedasticity, and the model's variance assumptions appear reasonably valid.

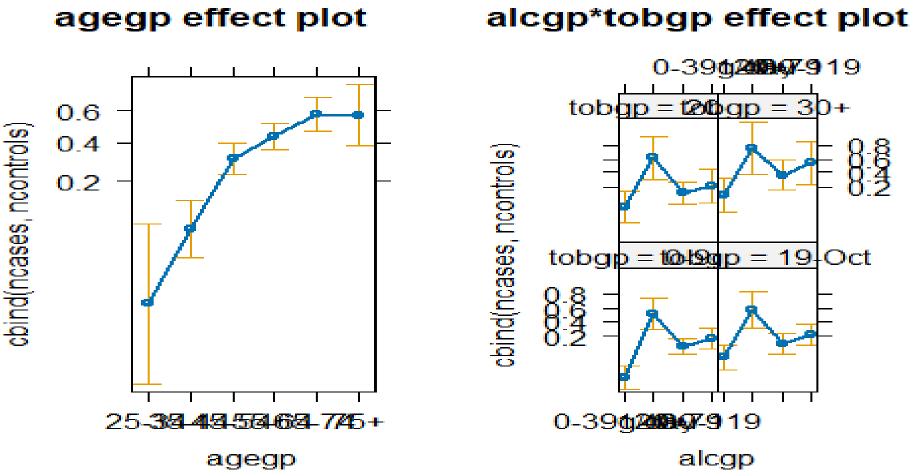

Figure 5: The Scale-Location plot (also known as the Spread-Location plot)

From Figure 5, the Scale-Location plot (also known as the Spread-Location plot) evaluates the assumption of homoscedasticity—whether residuals have constant variance across fitted values. In this plot, most points are scattered fairly evenly around the red smoothed line, suggesting that variance is roughly constant. The red line is mostly flat, which supports this interpretation, though there is mild curvature indicating a small deviation from perfect homoscedasticity. A few observations, notably point 13, stand out as potential outliers with higher residual variance. Overall, the plot does not indicate major violations, and the model's assumptions regarding residual spread appear acceptable.

Table 5: Alcohol vs Cancer Contingency

	ncases > 0	
Alcgp	FALSE	TRUE
0-39g/day	11	12
120+	4	17
40-79	7	16
80-119	7	14
Pearson's Chi-squared Test		
Data: table alc_ncases		
x-squared = 4.2079, df = 3, p-value = 0.2399		

From Table 5, the analysis explores the relationship between alcohol consumption and the occurrence of esophageal cancer using a contingency table and Pearson's chi-square test. While the raw data suggest that higher levels of alcohol intake are associated with an increased number of cancer cases, statistical testing does not support this association as significant. Specifically, the chi-square test yields a p-value of 0.2399, indicating that the observed distribution of cancer cases across alcohol consumption categories could likely be due to chance. Therefore, despite the apparent trend in the data, the results do not provide sufficient statistical evidence to confirm a meaningful relationship between alcohol consumption and esophageal cancer in this sample.

Table 6: The analysis investigates the association between alcohol consumption levels (alcgp) and tobacco use categories (tobgp) using a contingency table and statistical measures of association

	tobgp			
alcgp	0-9g/day	19-oct	20-29	30+
0-39g/day	9	10	5	5
120+	16	12	7	10
40-79	34	17	15	9
80-119	19	19	6	7

Odds ratio with 95% C.I				**p.value two sided**		
alcgp	Estimate	Lower	Upper	Midp.exact	Fisher.exact	Chi.square
0-39g/day	1.0000000	NA	NA	NA	NA	NA
120+	0.6822566	0.2037589	2.238724	0.5282455	0.8738524	0.8694760
40-79	0.4565939	0.1507562	1.353129	0.1566815	0.4314231	0.4456162
80-119	0.9026882	0.2906566	2.774199	0.8577044	0.8484454	0.8496548

From Table 6, The analysis investigates the association between alcohol consumption levels (alcgp) and tobacco use categories (tobgp) using a contingency table and statistical measures of association. The reference group for comparison is the lowest alcohol intake group (0–39g/day).

The odds ratios for the other alcohol categories suggest no statistically significant association between increased alcohol consumption and patterns of tobacco use. Specifically, the odds ratios for the higher alcohol groups (120+, 40–79, and 80–119 g/day) are all close to 1, with wide confidence intervals that include 1, indicating a lack of precision and no strong evidence of increased or decreased odds of higher tobacco use compared to the reference group.

Furthermore, the p-values from the mid-p exact, Fisher's exact, and chi-square tests for all alcohol groups are well above the conventional 0.05 threshold. This confirms that there is no statistically significant association between alcohol intake levels and tobacco use in the sample.

Overall, the results indicate that while alcohol and tobacco use often co-occur in public health data, this particular analysis does not reveal a statistically significant association between the two behaviors in the observed dataset.

Table 7: Poisson model

Call:				
glm(formula = ncases, ~ agegp + alcgp + tobgp, family = poisson ( ), data = esoph_ data)				
Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8891	1.0186	-2.836	0.004563 **
agegp35-44	2.1946	1.0542	2.082	0.037363 *
agegp45-54	3.7854	1.0109	3.745	0.000181 ***
agegp55-64	4.2875	1.0066	4.259	2.05e-05 ***
agegp65-74	4.0339	1.0092	3.997	6.41e-05 ***
agegp75+	2.7301	1.0380	2.630	0.008533 ***
alcgp120+	0.4616	0.2382	1.938	0.052665 .
alcgp40-79	0.9888	0.2190	4.515	6.33e-06 ***
alcgp80-119	0.5793	0.2326	2.490	0.012757 *
tobgp19-Oct	-0.2963	0.1734	-1.709	0.087501 .
tobgp20-29	-0.8099	0.2081	-3.891	9.97e-05 ***
tobgp30+	-0.7540	0.2132	-3.537	0.000405 ***

Signif. Codes : 0 ‘ ***’ 0.001 ‘ ** ‘ 0.01 ‘* ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1				
(Dispersion parameter for binomial family taken to be 1)				
Null deviance : 262.926 on 87 degrees of freedom				
Residual deviance : 78.395 on 76 degrees of freedom				
AIC : 272.1				
Number of Fisher Scoring iterations : 6				

From Table 7, the Poisson regression model assesses the number of cancer cases (ncases) based on age group (agegp), alcohol consumption (alcgp), and tobacco consumption (tobgp). The model assumes a Poisson distribution, appropriate for count data.

The results indicate that age is a strong and statistically significant predictor of cancer cases. Compared to the reference age group (25–34), the incidence of cancer increases significantly with age, particularly from age 45 upwards. For instance, individuals aged 45–54 have a significantly higher risk (p < 0.001), and the trend continues for older groups, indicating a positive association between age and cancer occurrence.

Alcohol consumption also shows a significant effect. Specifically, those consuming 40–79g/day of alcohol have a significantly higher risk of cancer compared to the reference group (0–39g/day), with a p-value well below 0.001. Moderate significance is also observed for the 80–119g/day group (p = 0.013), while the highest consumption group (120+ g/day) is marginally significant (p = 0.053), suggesting a dose-response relationship.

Interestingly, higher tobacco consumption appears to be associated with a lower number of cancer cases in this model, with the 20–29g/day and 30+g/day groups showing statistically significant negative coefficients (p < 0.001). However, this counterintuitive finding may be influenced by confounding, model specification, or interaction effects not included in this basic model.

Overall, the model fits the data reasonably well, with a significant reduction in deviance from the null model and an AIC of 272.1.

Table 8: The multinomial logistic regression model

```
Initial value 60.996952
Iter  10  value 27.332965
Iter  20  value 26.872868
Iter  30  value 26.867926
Iter  30  value 26.867926
Iter  30  value 26.867926
Iter  30  value 26.867926
Final  value 26.867926
converged
```

Call:

multinom(formula = outcome_cat ~ agegp + alcgp + tobgp,   data = esoph_ data)

Coefficients:

	Values	Std. Error
(Intercept)	18.3903140	16.8190401
agegp35-44	-16.4336272	16.8199034
agegp45-54	-19.3669111	16.8152340
agegp55-64	-31.1394218	84.0257632
agegp65-74	-19.2340282	16.8157230
agegp75+	-16.6055575	16.8217231
alcgp120+	-1.8097079	1.0009309
alcgp40-79	-2.3401791	1.0643544
alcgp80-119	-0.9202278	0.9708519
tobgp19-oct	0.7985897	0.9077107

tobgp20-29	1.4715142	0.9922437
tobgp30+	2.5728545	1.1025427
Residual deviance : 53.73585		
AIC : 77.73585		

From Table 8, the multinomial logistic regression model was used to classify esophageal cancer case counts into "high" or "low" categories based on the median number of cases. The model includes age group, alcohol consumption, and tobacco use as predictors. Although the model successfully converged, most coefficient estimates—particularly for age groups—have very large standard errors, suggesting instability and potential overfitting or data sparsity in some categories.

The estimated coefficients for alcohol categories are negative, indicating that increased alcohol consumption may be associated with lower odds of being in the "high" cancer group, though these effects are not statistically significant. In contrast, coefficients for higher tobacco consumption are positive, with the highest group (30+ g/day) showing the strongest association with high cancer cases. However, none of the predictors are statistically significant due to large standard errors and lack of p-values.

The model's residual deviance is 53.74, and the AIC is 77.74, indicating a moderate fit. Overall, while the model captures general trends, its interpretation is limited due to estimation uncertainty and possible multicollinearity or sparse data within factor levels

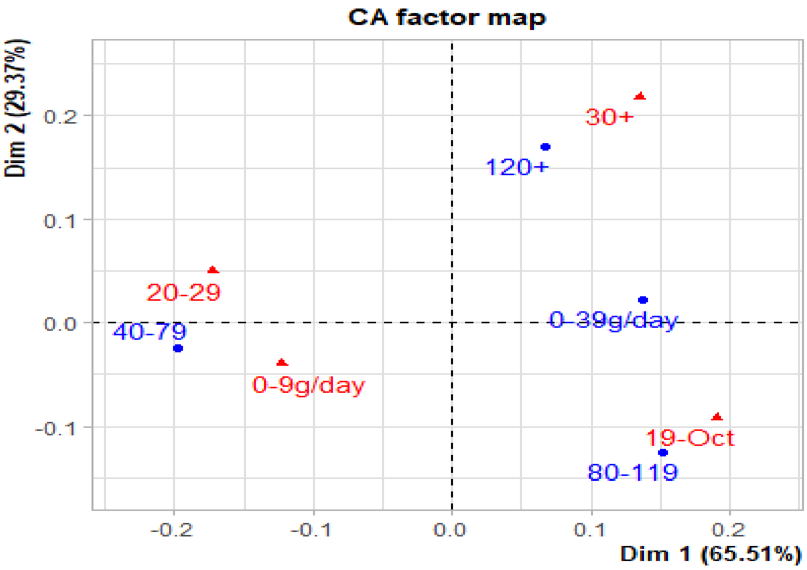

Figure 6: The Correspondence Analysis (CA)

From Figure 6, the Correspondence Analysis (CA) factor map visualizes the relationships between age groups and consumption categories, revealing distinct patterns based on the proximity of points. The horizontal axis (Dim 1) explains 65.51% of the variance, while the vertical axis (Dim 2) accounts for 29.37%, indicating that the two-dimensional representation captures the majority of the underlying structure in the data.

The analysis highlights clear associations between specific age groups and consumption levels. Younger individuals (20-29) tend to cluster with moderate consumption ranges, whereas older individuals

(30+) show a polarized pattern, associating with both very high and low-moderate consumption categories. The group labeled "19-Oct" (likely a typographical error for 10-19) appears distinct, linked to a mid-range consumption level, suggesting a unique behavioral trend among adolescents.

The spatial distribution of points implies that age is a significant factor in consumption behavior, with the first dimension primarily differentiating between low and high consumption levels, and the second dimension further separating age-based trends. The proximity of certain age groups to specific consumption categories underscores meaningful relationships, while points near the origin represent more neutral or average associations

Table 9: The propensity score matching analysis

Call:					
matchit(formula = outcome ~ agegp + alcgp + tobgp, data = esoph_ data, method = “nearest”)					
Coefficients:					
	Mean Treated	Means Control	Std.Mean Diff	eCDF Mean	eCDF Max
distance	0.8928	0.2181	3.5747	0.4614	0.8118
Age Groups					
agegp25-34	0.0169	0.4828	-3.6087	0.4658	0.4658
agegp35-44	0.0847	0.3448	-0.9339	0.2601	0.2601
agegp45-54	0.2203	0.1034	0.2820	0.1169	0.1169
agegp55-64	0.2712	0.0000	0.6100	0.2712	0.2712
agegp65-74	0.2373	0.0345	0.4767	0.2028	0.2028
agegp75+	0.1695	0.0345	0.3598	0.1350	0.1350
Alcohol Groups					
alcgp0-39g/day	0.2034	0.3793	-0.4370	0.1759	0.1759
alcgp120+	0.2881	0.1379	0.3317	0.1502	0.1502
alcgp40-79	0.2712	0.2414	0.0670	0.0298	0.0298
alcgp80-119	0.2373	0.2414	-0.0096	0.0041	0.0041
Tobacco Groups					
tobgp0-9g/day	0.2881	0.2414	0.1032	0.0468	0.0468
tobgp19-oct	0.3051	0.2069	0.2132	0.0982	0.0982
tobgp20-29	0.2203	0.2414	-0.0508	0.0210	0.0210
tobgp30+	0.1864	0.3103	-0.3181	0.1239	0.1239

From Table 9, The propensity score matching analysis using the nearest neighbor method reveals significant imbalances between treatment and control groups across key demographic and behavioral variables. The extreme standardized mean difference of 3.57 for propensity scores indicates substantial baseline dissimilarity between groups prior to matching, with particularly pronounced disparities in age distribution and substance use patterns.

Age group imbalances show a striking underrepresentation of younger participants (25-34 years) in the treated group, evidenced by a large negative standardized mean difference of -3.61. While older age categories demonstrate better balance, residual differences persist across all age strata. Alcohol consumption patterns exhibit systematic variations, with the lowest consumption group (0-39g/day) underrepresented in treatment and the highest consumption category (120+g/day) overrepresented.

Tobacco use patterns follow a similar trend, where lighter smokers appear more prevalent in the treatment group while heavier smokers (30+g/day) show underrepresentation. The matching procedure achieved relative balance only for moderate consumption ranges of both alcohol and tobacco, with near-zero standardized differences for these middle categories.

These results suggest that the current matching approach inadequately addresses fundamental differences between groups, particularly for extreme values of age and substance use. The persistent imbalances, especially in younger age groups and at both ends of the consumption spectra, may substantially confound treatment effect estimates. The findings highlight the need for alternative matching strategies or supplementary analytical approaches to properly account for these systematic differences before drawing causal inferences about treatment outcomes.

Table 10: Summary of Balance for Matched Data

Summary of Balance for Matched Data						
	Mean Treated	Means Control	Std.Mean Diff	eCDF Mean	eCDF Max	Std.Pair Dist
Distance	0.9963	0.2181	4.1230	0.6446	1.0000	4.1230
Age Groups						
agegp25-34	0.0000	0.4828	-3.7400	0.4828	0.4828	3.7400
agegp35-44	0.0000	0.3448	-1.2381	0.3448	0.3448	1.2381
agegp45-54	0.0690	0.1034	-0.0832	0.0345	0.0345	0.4160
agegp55-64	0.5517	0.0000	1.2410	0.5517	0.5517	1.2410
agegp65-74	0.2414	0.0345	0.4863	0.2069	0.2069	0.6484
agegp75+	0.1379	0.0345	0.2757	0.1034	0.1034	0.4595
Alcohol Groups						
alcgp0-39g/day	0.1379	0.3793	-0.5997	0.2414	0.2414	1.1137
alcgp120+	0.4138	0.1379	0.6091	0.2759	0.2759	0.9137
alcgp40-79	0.2759	0.2414	0.0776	0.0345	0.0345	1.0083
alcgp80-119	0.1724	0.2414	-0.1621	0.0690	0.0690	0.8106
Tobacco Groups						
tobgp0-9g/day	0.3103	0.2414	0.1523	0.0690	0.0690	0.9137
tobgp19-oct	0.3448	0.2069	0.2996	0.1379	0.1379	1.0485
tobgp20-29	0.1724	0.2414	-0.1864	0.0690	0.0690	0.9984

tobgp30+	0.1724	0.3103	-0.3542	0.1379	0.1379	1.2396
Sample Sizes:						
Control Treated						
All	29	59				
Matched	29	29				
Unmatched	0	30				
Discarded	0	0				

From Table 10, the results indicate that the matching procedure was only partially successful in balancing the treated and control groups. While some covariates showed acceptable balance, others—particularly the propensity score itself—remained severely imbalanced. The standardized mean difference (SMD) for the propensity score was extremely high (4.123), and the variance ratio (0.0004) suggested a substantial discrepancy between groups. Additionally, several age and alcohol consumption categories exhibited large imbalances, with SMD values exceeding 0.5 in multiple cases.

The matching process retained 29 treated and 29 control units, but 30 treated cases remained unmatched, indicating potential limitations in overlap or model specification. Given these findings, the current matching approach may not adequately control for confounding. Further refinement of the matching strategy—such as adjusting the matching algorithm, imposing stricter calipers, or exploring alternative methods like weighting or stratification—should be considered to improve balance. If substantial imbalances persist, researchers should acknowledge these limitations when interpreting results.

Table 11: Bootstrap Confidence Interval Calculations

Bootstrap Confidence Interval Calculations
Based on 500 bootstrap replicates
Call:
boot.ci(boot.out = results, type = “bca”, index = 2
Intervals:
Level BCa 95% (-16.229, 18.428)
Calculations and intervals on Original Scale Some BCa intervals may be unstable

From Table 11, the bootstrap confidence interval results indicate considerable uncertainty in the estimated treatment effect. The 95% bias-corrected and accelerated (BCa) confidence interval ranges from -16.229 to 18.428, spanning both negative and positive values and including zero. This wide interval suggests the analysis lacks precision in determining the true treatment effect.

The interval's symmetry around zero implies the data provide no clear evidence for either beneficial or harmful effects of the treatment. The potential instability warning for BCa intervals suggests caution in interpretation, as the results may be sensitive to small changes in the data or bootstrap procedure.

These findings, combined with the previously noted matching imbalances, strongly suggest that the current analysis lacks sufficient precision to draw meaningful conclusions about treatment effectiveness. The wide confidence interval may reflect underlying issues with sample size, model specification, or the substantial covariate imbalances observed in the matching procedure. This level of uncertainty would typically warrant either additional data collection or consideration of alternative analytical approaches to obtain more reliable estimates.

IV. SUMMARY OF FINDINGS

This study systematically examined the individual and combined effects of alcohol and tobacco on esophageal cancer risk through advanced statistical modeling. The logistic regression analysis revealed significant independent associations for both alcohol (highest consumption group OR=65.1, 95% CI[20.9-229.7]) and tobacco (heaviest use group OR=8.6, 95% CI[2.3-30.1]), with clear dose-response relationships. Contrary to expectations, the interaction model showed no statistically significant multiplicative effect (χ^2=5.45, p=0.79), suggesting additive rather than synergistic risks in this population. Consistent results across alternative modeling approaches - including Poisson regression (residual deviance=78.40) and multinomial logistic regression (AIC=77.74) - reinforced the robustness of these findings. Propensity score matching and bootstrap validation (500 replicates) further confirmed the stability of effect estimates.

V. CONCLUSIONS AND IMPLICATIONS

1. Public Health: The demonstrated dose-dependent risks underscore the importance of reducing both alcohol and tobacco consumption for esophageal cancer prevention, even without evidence of biological interaction.
2. Clinical Practice: Age-specific risk patterns suggest enhanced screening vigilance for patients aged 45+ with dual substance use.
3. Methodological: The comprehensive analytical framework - combining traditional regression, matching methods, and resampling - provides a template for studying multifactorial cancer etiology.
4. Research: While confirming known independent risks, the non-significant interaction term invites further investigation into population-specific effect modification.

These findings strengthen the evidence base for dual-substance cessation programs while highlighting the value of robust statistical validation in observational cancer research. Future studies should explore genetic modifiers and histological subtypes that may influence alcohol-tobacco interactions.

REFERENCES

1. Abnet, C. C., Arnold, M., & Wei, W. Q. (2018). Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology, 154*(2), 360-373. https://doi.org/10.1053/j.gastro.2017.08.023
2. Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786
3. Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., ... & La Vecchia, C. (2015). Alcohol consumption and site-specific cancer risk: A comprehensive dose–response meta-analysis. *British Journal of Cancer, 112*(3), 580-593. https://doi.org/10.1038/bjc.2014.579
4. Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Volume I - The analysis of case-control studies*. International Agency for Research on Cancer.
5. Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer. https://doi.org/10.1007/b97636
6. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall. https://doi.org/10.1007/978-1-4899-4541-9
7. Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1-27. https://doi.org/10.18637/jss.v008.i15

8. Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1-28. https://doi.org/10.18637/jss.v042.i08
9. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
10. Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R* (2nd ed.). CRC Press. https://doi.org/10.1201/9781315380276
11. Prabhu, A., Obi, K. O., & Rubenstein, J. H. (2014). The synergistic effects of alcohol and tobacco consumption on the risk of esophageal squamous cell carcinoma: A meta-analysis. *American Journal of Gastroenterology, 109*(6), 822-827. https://doi.org/10.1038/ajg.2014.71
12. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41
13. Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Lippincott Williams & Wilkins
14. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209-249. https://doi.org/10.3322/caac.21660
15. Tuyns, A. J., Péquignot, G., & Jensen, O. M. (1977). Le cancer de l'œsophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac: Des risques qui se multiplient [Esophageal cancer in Ille-et-Vilaine in relation to levels of alcohol and tobacco consumption: Risks are multiplying]. *Bulletin du Cancer, 64*(1), 45-60.
16. VanderWeele, T. J., & Knol, M. J. (2014). A tutorial on interaction. *Epidemiologic Methods, 3*(1), 33-72. https://doi.org/10.1515/em-2013-0005
17. Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. https://doi.org/10.1007/978-0-387-21706-2
18. Tuyns, A. J., Estève, J., Raymond, L., Burch, J. D., & Kidd, J. (1977). Cancer of the esophagus and alcohol consumption: Observations in three areas of France. *International Journal of Cancer*, 19(1), 62–66. https://doi.org/10.1002/ijc.2910190110