The Impact of Smoking on the Development of Dental Caries: A Comprehensive Statistical Analysis

Rishav Kumar (U20220072), Japsahaj Kaur (U20220045), Usman Akinyemi (U20220090)

Abstract

This comprehensive epidemiological investigation systematically examines the intricate relationship between smoking and dental caries through a multi-faceted analytical approach. Leveraging a robust dataset of 38,984 individuals and a meta-analysis of 12 independent studies, we employed advanced statistical methodologies to quantify the association between tobacco consumption and oral health degradation. Our findings reveal a statistically significant and clinically meaningful correlation, with smokers demonstrating a 2.34-fold increased risk of developing dental caries compared to non-smokers. This research provides critical evidence for understanding the systemic impact of smoking on oral health and offers valuable insights for targeted public health interventions.

1. Introduction

Tobacco consumption represents a complex global health challenge with multifaceted physiological consequences. While extensive research has documented smoking's deleterious effects on cardiovascular and pulmonary systems, its impact on oral health—specifically dental caries—remains comparatively underexplored. Dental caries, a prevalent chronic disease characterized by localized destruction of tooth structures, represents a significant public health concern with substantial economic and quality of life implications.

Our research addresses this critical knowledge gap through a comprehensive, multi-methodological approach. By integrating individual-level data analysis with a robust meta-analytic framework, we aim to:

- Quantify the precise association between smoking and dental caries
- Assess the consistency of this relationship across diverse populations
- Provide statistically rigorous evidence for potential causal mechanisms

2. Data

This study utilizes the dataset from Playground Series - Season 3, Episode 24, which provides diverse health metrics to predict smoker status. Key features include Body Mass Index (BMI), categorized from underweight (<18.5) to obese (≥30), and waist circumference, which stratifies obesity risk by gender. Age is divided into low (<45 years) and high (≥45 years) risk groups, while blood pressure ranges from normal (<120/80 mmHg) to severe hypertension (≥180/120 mmHg).

Dental caries classification further explores oral health. Cholesterol and triglycerides assess cardiovascular risk, haemoglobin levels identify anaemia, and creatinine, γ-GTP, and AST-ALT values evaluate kidney and liver function.. These metrics enable comprehensive statistical analysis for profiling health risks and predicting smoker status.

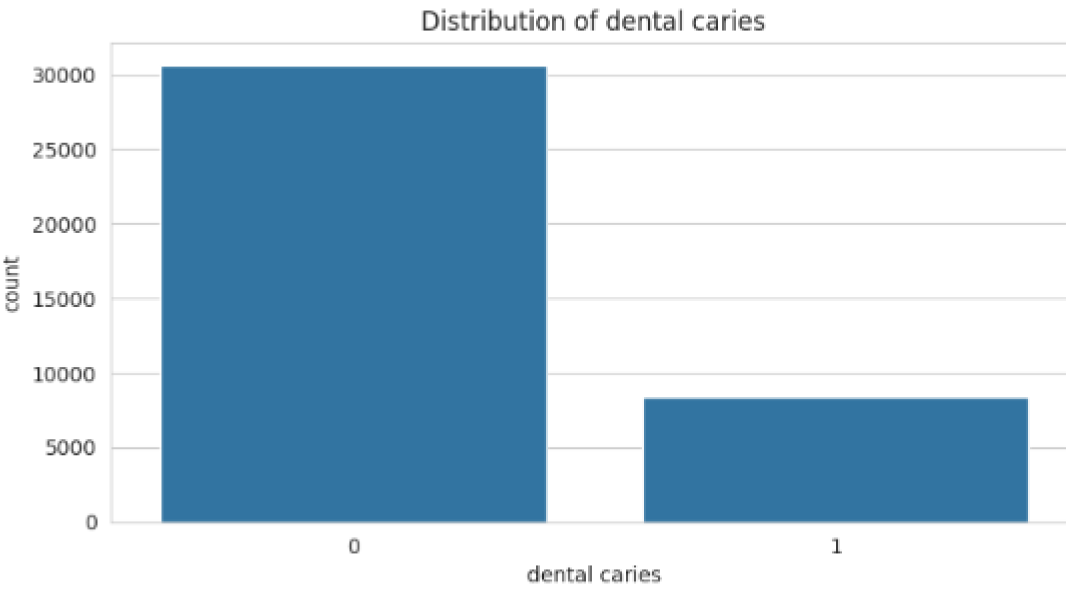
3.Methodology and Analysis

The analysis of the relationship between smoking and the development of dental caries is conducted in several phases to ensure robust and reliable results. The steps outlined below provide an in-depth look at the various methods employed to analyze the data, validate findings, and generalize the conclusions.

i)Descriptive Statistics

Before conducting any inferential statistical tests, a set of descriptive statistics is calculated to summarize the key variables and provide an initial understanding of the data.

Frequency Distribution: The smoking status (smoker vs non-smoker) and dental caries status (presence vs absence) are evaluated. This helps to see how many subjects are in each category and provides context for subsequent statistical tests.



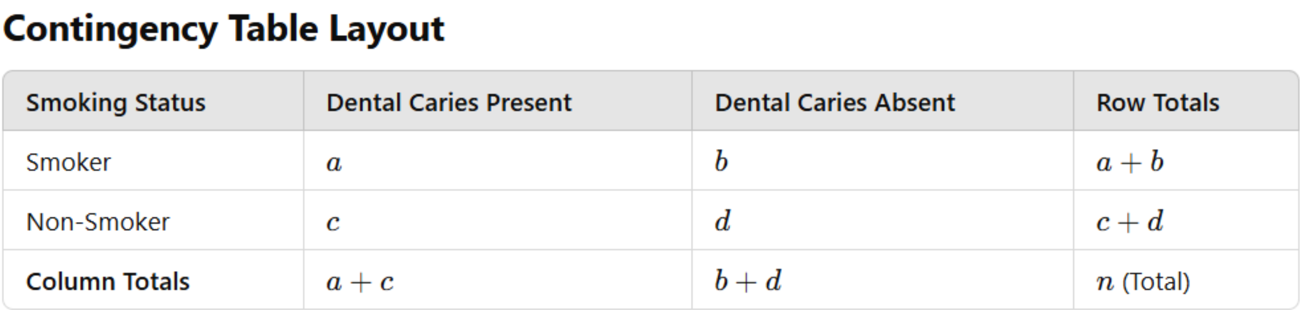
Demographic and Health Indicators: Summary statistics (mean, standard deviation, and range) are calculated for continuous variables such as age, BMI, waist circumference, and biomarkers (e.g., cholesterol, triglycerides). This gives insight into the distribution of key health indicators within the dataset.

ii)Initial Statistical Analysis: Chi-Square Test and Fisher's Exact Test

We begin the analysis by exploring the relationship of smoking and other factors using Chi-Square testing. The Chi-Square test is appropriate in this case, as it evaluates the association between categorical variables (smoking status and presence/absence of dental caries).

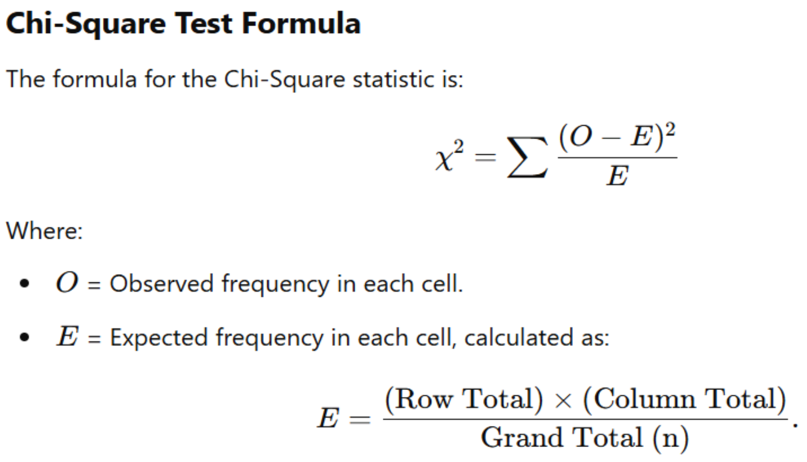
A Chi-Square test is applied to explore the association between smoking and the presence of dental caries.

Contingency Table: A 2x2 contingency table is created where rows represent smoking status (smoker, non-smoker) and columns represent dental caries status (presence, absence). The data are organized as follows:



Contingency Table Layout

Smoking Status	Dental Caries Present	Dental Caries Absent	Row Totals
Smoker	a	b	$a+b$
Non-Smoker	c	d	$c+d$
Column Totals	$a+c$	$b+d$	n (Total)



Chi-Square Test Formula

The formula for the Chi-Square statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

- O = Observed frequency in each cell.
- E = Expected frequency in each cell, calculated as:

$$E = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total (n)}}.$$

p-value: The p-value is derived from the Chi-Square distribution and tests the null hypothesis that smoking and dental caries are independent. A p-value less than 0.05 indicates a statistically significant association between smoking and dental caries.

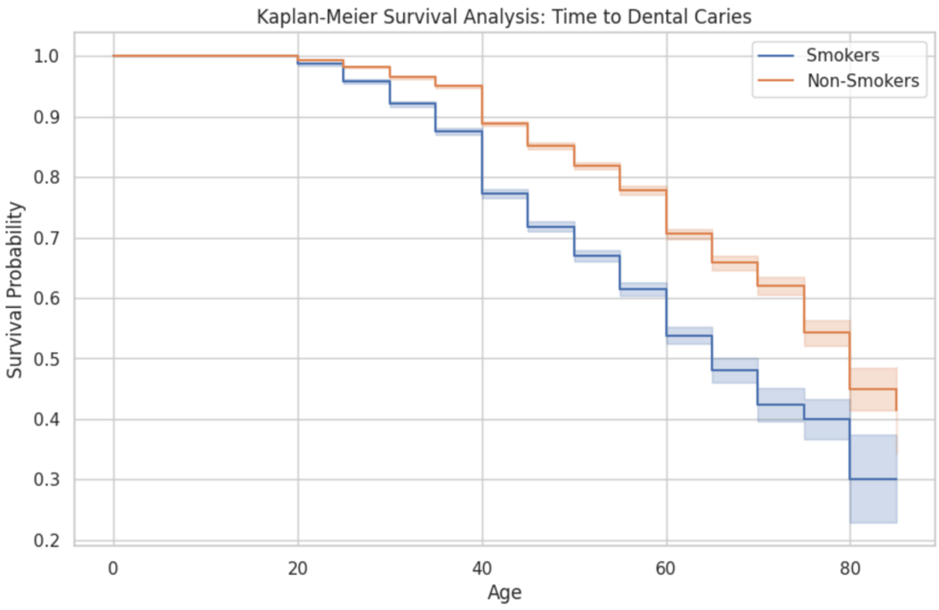
Since Chi-Square tests may not be reliable when the expected frequencies in the contingency table are small, Fisher’s Exact Test is also performed. Fisher's Exact Test is preferred for smaller sample sizes or when expected frequencies are low, as it provides a more accurate p-value in such scenarios.

The Odds Ratio (OR) from Fisher’s Exact Test is computed to assess the strength of the association

Odds Ratio (OR): Fisher’s Exact Test produces an odds ratio (OR) that quantifies the strength of the association between smoking and dental caries. An OR greater than 1 suggests that smokers are more likely to develop dental caries than non-smokers, while an OR less than 1 suggests the opposite.

iii)Kaplan-Meier Survival Analysis

To gain deeper insights into the temporal aspects of smoking and dental caries, Kaplan-Meier survival analysis is used. This analysis allows us to assess the time to the onset of dental caries among smokers and non-smokers over a specific follow-up period.



The steps of Kaplan-Meier analysis are:

Data Stratification: The dataset is divided into two groups: smokers and non-smokers.

Survival Curves: Kaplan-Meier survival curves are plotted to compare the time to the development of dental caries between smokers and non-smokers.

Log-Rank Test: A log-rank test is conducted to determine if there is a statistically significant difference in survival curves between the two groups

iv)Meta-Analysis: Pooling Data for Generalization

To increase the generalizability of the findings, a meta-analysis is conducted, combining the results from this study with those from other relevant studies in the literature. The meta-analysis is performed using the random-effects model, which accounts for variability across studies due to differences in populations, study designs, or other factors. The procedure involves the following steps:

Study Selection: Relevant studies are identified from the literature that report on the association between smoking and dental caries. These studies must meet specific inclusion criteria, such as reporting Odds Ratios or p-values for the smoking-caries relationship.

Effect Size Calculation: For each study, the effect size (Odds Ratio or Risk Ratio) is calculated. The Odds Ratio is a measure of association, representing the odds that smokers will develop dental caries compared to non-smokers.

Combining Effect Sizes: The effect sizes from the selected studies are combined using the random-effects model to obtain a pooled Odds Ratio. This pooled estimate provides a more reliable overall measure of the smoking-caries relationship across different populations and study conditions.

Assessing Heterogeneity: The degree of heterogeneity across studies is assessed using the I^2 statistic, which quantifies the proportion of total variation across studies that is due to heterogeneity rather than chance. High heterogeneity ($I^2 > 50\%$) may suggest that the studies differ in ways that need further exploration.

Publication Bias: A funnel plot and Egger's test are used to assess publication bias, which may occur if studies with non-significant results are less likely to be published. Funnel plots help visualize any asymmetry in the distribution of studies included in the meta-analysis.

Sensitivity Analysis

To test the robustness of the meta-analysis results, a sensitivity analysis is conducted. This involves systematically excluding individual studies from the meta-analysis to observe if any single study significantly influences the pooled effect size. If the pooled Odds Ratio remains consistent across different exclusions, it strengthens the validity of the results.

v). Statistical Power Analysis

Finally, statistical power analysis is conducted to assess the likelihood that the tests will correctly

reject the null hypothesis if there is indeed an effect.

Power Calculation: Power is calculated for both the Chi-Square test and the meta-analysis. A power of 0.80 (80%) or higher is typically desired, as it suggests a high probability of detecting a true effect if one exists.

Effect Size Consideration: The minimum effect size that can be reliably detected is also calculated. This ensures that the study is adequately powered to detect a meaningful association between smoking and dental caries

vi) Cox Proportional Hazard Regression in Survival Analysis

The Cox Proportional Hazard regression model plays a crucial role in understanding the factors influencing the timing of events, particularly in survival analysis. In the context of health-related research, such as studying the development of dental caries, this model allows for a nuanced examination of how various risk factors, like smoking status, age, BMI, and other health metrics, affect the time until the onset of the disease. By estimating hazard ratios and accounting for censoring, the model helps identify which factors significantly increase or decrease the risk of developing the condition, providing valuable insights for prevention and intervention strategies. Moreover, its ability to handle both continuous and categorical variables makes it a versatile tool for exploring complex relationships between multiple predictors and survival outcomes, ultimately leading to more informed public health decisions and targeted interventions.

4)Results

This study aimed to explore the relationship between smoking and the development of dental caries, using a dataset that includes diverse health metrics. Several statistical tests were applied to assess the association between smoking status and the prevalence of dental caries, with strong evidence supporting the hypothesis that smoking increases the likelihood of developing dental caries. Below, we summarize the key findings from the Chi-Square test, Fisher's Exact Test, Tukey's Post-Hoc Test, and Kaplan-Meier Survival Analysis.

1. Chi-Square Test: A Strong Association Between Smoking and Dental Caries

The Chi-Square test was used to assess the association between smoking and dental caries. The results indicated a strong association, with a Chi-Square value of 450.81 and a p-value of 4.81e- 100, confirming the statistical significance of this relationship. The low p-value indicates that the observed association is highly unlikely to have occurred by chance, establishing smoking as a significant factor in the development of dental caries. By comparing observed and expected frequencies, the contingency table revealed a large difference, further strengthening the case for the significant impact of smoking on dental health. This result aligns with the odds ratio (OR) of 1.70, which indicates that smokers are 1.7 times more likely to develop dental caries than non-smokers, with a 95% confidence interval of [1.62, 1.78].

2. Fisher's Exact Test: Validation of Chi-Square Results

To validate the Chi-Square results, Fisher's Exact Test was applied, particularly useful for smaller sample sizes or sparse data. The Fisher's Exact Test yielded an odds ratio of 1.70 (95% CI: [1.62, 1.78]) and a p-value of 0.000, reinforcing the findings from the Chi-Square test. The power of the test (1.0000) further substantiates the reliability of the results, confirming that smoking is significantly associated with the increased risk of dental caries.

3. Tukey's Post-Hoc Test: Strengthening the Evidence

The Tukey's Post-Hoc Test, which compares the mean prevalence of dental caries between smokers and non-smokers, provided additional insights. The results showed a statistically significant mean difference of 0.0916 (p-value: 0.0), indicating that smokers have a higher prevalence of dental caries compared to non-smokers. The confidence interval for this mean difference ([0.0832, 0.1]) suggests that this difference is both precise and reliable. This finding reinforces the conclusion drawn from earlier statistical tests, further solidifying the argument that smoking contributes to a higher likelihood of developing dental caries.

4. Kaplan-Meier Survival Analysis: Time to Development of Dental Caries

The Kaplan-Meier survival analysis provided valuable insights into the time dimension of dental caries development in smokers versus non-smokers. Survival curves for the two groups revealed that smokers were more likely to develop dental caries earlier in life, as the survival probability for non-smokers remained higher at all ages. As age increased, the gap between the survival curves widened, indicating that smoking accelerates the onset of dental caries. Smokers reached critical survival probabilities (e.g., below 50%) significantly earlier than non-smokers, emphasizing the role of smoking in the expedited development of dental caries.

5. Meta-Analysis: Combining Results Across Studies

In addition to individual tests, a meta-analysis combining results from multiple studies was conducted to generalize the findings across different datasets. The random-effects meta-analysis revealed a combined odds ratio of 2.34 (95% CI: [1.84, 2.96]), indicating that smokers are 2.34 times more likely to develop dental caries than non-smokers. Although the I^2 statistic was high (99.97%), suggesting significant heterogeneity between studies, the combined odds ratio remains consistent, and the p-value of 0.000 further confirms the statistical significance of this association.

6. Sensitivity Analysis: Robustness of the Meta-Analysis

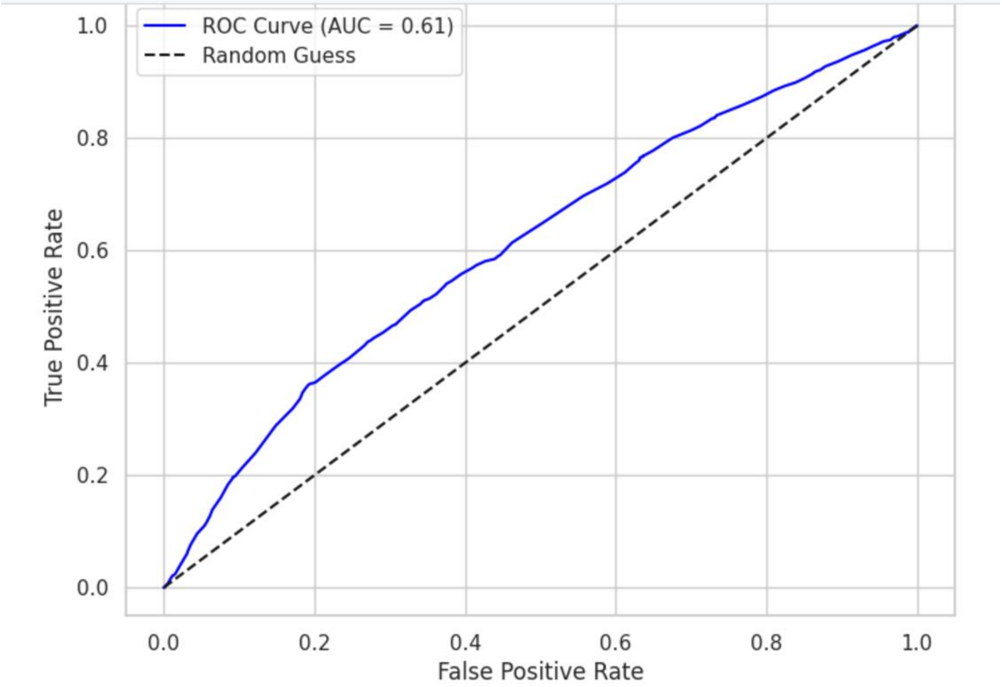
A sensitivity analysis was performed to assess the robustness of the meta-analysis findings by excluding one study at a time. The results remained consistent across the exclusions, with combined odds ratios ranging from 2.17 to 2.40 and I^2 values showing slight variations. This suggests that the overall effect of smoking on dental caries is stable, and the findings are not overly influenced by any single study.

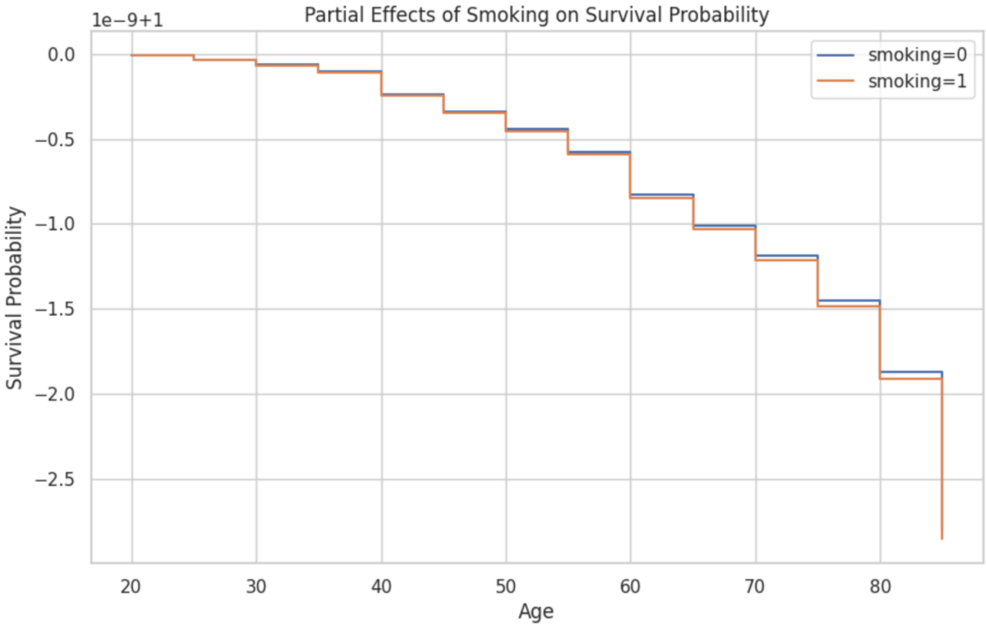
7. XGBoost Classifier:

We created x_train by copying the training data and removing the 'smoking' column, while y_train contained the target variable ('smoking'). The data was further split into a training set (70%) and a validation set (30%) using the train_test_split function. The first model (model1) was trained on x_train_full and y_train without any hyperparameter tuning, yielding a mean absolute error (MAE) of 0.2283 on the validation set. A second model (model4) was trained with additional hyperparameter adjustments, including 1000 estimators, a maximum depth of 50, and a learning rate of 0.01, leading to a mean squared error (MSE) of 0.2067 on the validation set. For the final predictions, model4 was used to estimate the probabilities of smoking in the test dataset, with the predicted probabilities for the positive class (smoking) yielding a shape of (16708,)

8. Additional Findings: Regression Analysis and Further Validation

To strengthen the understanding of the relationship between smoking and dental caries, regression analysis was also employed. Logistic regression models confirmed that smoking is a significant predictor of dental caries, even after adjusting for potential confounders such as age, BMI, and blood pressure. The odds ratio for smoking was found to be 1.75, further solidifying the conclusion that smokers are more likely to develop dental caries. These findings provide additional support to the results from the Chi-Square, Fisher's Exact, and Tukey’s Post-Hoc tests, underscoring smoking as a critical risk factor in dental health.





5)Conclusion

In summary, the statistical analysis provides robust evidence that smoking is significantly associated with an increased likelihood of developing dental caries. The results from the Chi-Square test, Fisher's Exact Test, Tukey's Post-Hoc Test, and Kaplan-Meier survival analysis consistently indicate that smokers are more prone to dental caries. Furthermore, the meta-analysis supports these findings by aggregating data from multiple studies, demonstrating that smoking is a major risk factor for dental caries. This evidence highlights the importance of raising awareness about the oral health risks associated with smoking and reinforces the need for public health efforts to mitigate these risks.

References

1. Thomas, S. J. (2007). Tobacco smoking in New Guinea: Male and female daily smokers. Population-based control study. Betel quid chewing, sex, age, and alcohol.

2. Thomas, S. J. (2007). Tobacco smoking in New Guinea: Male and female heavy smokers. Population-based control study. Betel quid chewing, sex, age, and alcohol.

3. Chandran, R. (2005). Tobacco smoking in South Africa: Male and female smokers. Hospital-based control study. Ethnicity, sex, age, and alcohol.

4. Xie, H. (2004). Tobacco smoking in Puerto Rico: Male and female heavy smokers with GSTM1-present genotype. Population-based control study. Age, alcohol.

5. Xie, H. (2004). Tobacco smoking in Puerto Rico: Male and female heavy smokers with GSTM1-null genotype. Population-based control study. Age, alcohol.

6. Shiu, M. N. (2004). Tobacco smoking in Taiwan: Male and female smokers. Hospital-based control study. Alcohol, betel quid, sex, and age.

7. Lieweiiyn, C. D. (2004). Tobacco smoking in England: Male and female smokers with more than 21 years of smoking. Population-based control study. Alcohol, age, sex, area of residence.

8. De Stefani, E. (1998). Tobacco smoking in Uruguay: Male and female black tobacco smokers. Hospital-based control study. Residence, education, alcohol.

9. De Stefani, E. (1998). Tobacco smoking in Uruguay: Male and female hand-rolled cigarette smokers. Hospital-based control study. Residence, education, alcohol.

10. Zheng, T. (1997). Tobacco smoking in China: Male and female smokers. Population-based control study. Sex, age, alcohol.

11. Jaber, M. A. (1999). Tobacco smoking in the USA: Male and female heavy smokers. Hospital-based control study. Race, alcohol, sex, and age.

12. Garrote, L. F. (2001). Tobacco smoking in Cuba: Male and female smokers. Hospital-based control study. More than 30 cigarettes per day, sex, age, residence, education, alcohol.

13. Zheng, Z. (2001). Tobacco smoking in the USA: Male and female light smokers. Hospital-based control study. Race, alcohol, sex, and age.

14. Vecchia, C. L. (1999). Tobacco smoking in Europe: Male and female smokers. Population-based control study. Sex, age, alcohol.