Epi 204 Project Report

**Introduction**

According to CDC, cancer is one of the leading causes of death, so it is important to understand how our lifestyle is associated with the risk of dying from cancer. Since people enjoy alcohol for various reasons, we are interested in how alcohol relates to cancer deaths. In this project, we used Cox models to examine the association between alcohol intake and death from cancer during follow-up.

**Methods**

The data used in this study is from an imputed version of the NHANES II mortality follow-up study. The details of the NHANES II survey are described elsewhere[[1]](#footnote-1). The cross-sectional survey has information about demographic, socioeconomic, and biomarker variables. Then a longitudinal follow-up study of the participants in the survey was conducted, and death occurred after the survey was administered.

The exposure variable for our project is alcohol consumption (drinks/week) measured at the baseline. Since alcohol intake is a continuous variable ranging from xx to xx, we consider continuous, categorical, and ordinal versions of the alcohol intake for result robustness. After exploring the distribution of alcohol intake, we divided the continuous alcohol intake into the categorical variable by the following cutoff values to reduce the influence of outliers. We found about 43.7% of people didn’t take alcohol per week. After excluding people with 0 alcohol intake, the minimum, median, and maximum alcohol intake were 0.5, 2.0, and 77.0 per week. Thus, we divided people into four groups based on their alcohol intake: 0 per week, 0-0.5 drinks per week, 0.5-2 drinks per week, and >2 drinks per week.

We are interested in the time to death from cancer during follow-up. There was some loss of follow-up in the study and death from other causes. In our analysis, we treated competing risks as censoring. So, the outcome is time to death of cancer during the follow-up, or time to censoring, or time to death from other causes, whichever comes first.

With additivity, linearity, and proportional hazard assumptions, we conducted the Cox proportional hazard models for our main research question. Cox regression will account for time-to-event and the death rate over an interval of time is unspecified. In contrast, Poisson regression assumes a constant death rate, and logistic regression cannot account for time. So, we then ran logistic regression and Poisson regression as sensitivity analysis.

In the Cox models, we considered two settings for time scale (The conceptual framework is shown in Appendix 1):

1. Given the cross-sectional nature of the data, alcohol intake can be perceived as a prevalent intake, which could induce the “prevalent user” bias. To test the robustness of the findings given this potential bias, we first set the age of 21 to be the time origin and chose age as the time scale. In this approach, we conducted a crude model and a MV-adjusted model. The MV-adjusted model only adjusted for sex and covariates related to socio-economic status. Baseline age was not adjusted since age was set as the time scale, and we did not include any biomarkers, biometrics, and disease status at baseline since they might be the consequence of the past alcohol intake.
2. The second approach is setting baseline time as the time origin, and time since entry is the time scale. We conducted a crude model, a model adjusting for baseline age, and a MV-adjusted model. In the MV-adjusted model, we adjusted for age, sex, and some biomarkers and disease status at baseline; we considered these variables as confounders since they are reasonable common causes of prevalent alcohol intake and future risk of death because of cancer.

For MV-adjusted Cox models, we checked the proportional hazard assumption using Schofield residuals and the linearity assumption for alcohol consumption by comparing the category-specific hazard ratio and fitting a natural cubic spline . Next, we checked the effect modification of sex by (1) including the product term in the multivariable-adjusted (MV-adjusted) Cox model, (2) running stratified Cox regression by sex with an interaction term of strata sex and alcohol intake, and (3) performing Cox regressions in males and females separately (as subgroup analysis). Wald-tests were utilized to test the significance of effect modification.

Note that the imputed raw data size is xxxx, and there are missing data in the dataset. After deciding the models, we removed all missing data for our selected variables. The resulted complete case sample size we used in this project is xxxx.

Table 1 Baseline characteristics of study population by alcohol intake(N=9190).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Alcohol intake per week | | | |  |
|  | 0/week | 0-0.5/week | 0.5-2/week | >2/week | p-value |
|  | N=1052 | N=209 | N=319 | N=530 |
| Sex, Male | 515 (49.0%) | 110(52.6%) | 205(64.3%) | 407(76.8%) | <0.001 |
| Mean Age at entry (SD) | 65.2 (7.81) | 64.9 (8.32) | 62.2 (9.43) | 62.1 (9.67) | <0.001 |
| Race |  |  |  |  | <0.001 |
| white | 109 (10.4%) | 21 (10.0%) | 51 (16.0%) | 87 (16.4%) |  |
| black | 388 (36.9%) | 93 (44.5%) | 121(37.9%) | 234(44.2%) |  |
| other | 555 (52.8%) | 95 (45.5%) | 147(46.1%) | 209(39.4%) |  |
| Mean year of schooling (SD) | 9.24 (3.69) | 10.4 (3.42) | 10.3 (3.57) | 10.9 (3.35) | <0.001 |
| Marital status |  |  |  |  | . |
| Married | 699 (66.4%) | 143 (68.4%) | 225 (70.5%) | 379 (71.5%) |  |
| Widowed | 231 (22.0%) | 36 (17.2%) | 49 (15.4%) | 67 (12.6%) |  |
| Divorced | 44 (4.18%) | 12 (5.74%) | 13 (4.08%) | 12 (9.76%) |  |
| Separated | 24 (2.28%) | 7 (3.35%) | 10 (3.13%) | 11 (2.08%) |  |
| Never married | 53 (5.04%) | 9 (4.31%) | 21 (6.58%) | 32 (6.04%) |  |
| Blank | 1 (0.10%) | 2 (0.96%) | 1 (0.31%) | 5 (0.94%) |  |
| Mean BMI (SD) | 26.3 (5.58) | 25.8 (4.51) | 26.1 (5.21) | 25.1 (4.17) | <0.001 |
| Mean smoke per day (SD) | 5.60 (11.8) | 6.74 (11.9) | 9.78 (15.2) | 12.0 (15.0) | <0.001 |
| Size of place | 5.70 (2.57) | 5.07 (2.70) | 4.53 (2.69) | 4.44 (2.69) | <0.001 |
| Standard Metropolitan  Statistical Area |  |  |  |  | <0.001 |
| In central city | 265 (25.2%) | 59 (28.2%) | 105(32.9%) | 186(35.1%) |  |
| Not in central | 201 (19.1%) | 57 (27.3%) | 102(32.0%) | 183(34.5%) |  |
| Not in SMSA | 586 (55.7%) | 93 (44.5%) | 112(35.1%) | 161(30.4%) |  |
| Resides in urban area, Yes | 578 (54.9%) | 578(54.9%) | 235(73.7%) | 387(73.0%) | <0.001 |

Table 2. Association between alcohol intake and hazard of death from cancer in overall NHANES II (N=9190).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0/week | 0-0.5/week | 0.5-2/week | >2/week | Continuous | Ptrend |
| First approach, defining the exposure as alcohol intake age 21-entry, time scale as age | | | | | | |
| Crude model | Ref | 1.10 (0.82, 1.48) | 1.13 (0.88, 1.44) | 1.57 (1.29, 1.91) | 1.04 (1.03, 1.05) | <0.001 |
| Age-adjusted | Ref |  |  |  |  |  |
| MV-adjusteda | Ref | 1.06 (0.78, 1.43) | 0.98 (0.76, 1.27) | 1.25 (1.00, 1.55) | 1.02 (1.01, 1.04) | 0.084 |
| Second approach, defining the exposure as prevalent alcohol intake at baseline, time scale as years since entry | | | | | | |
| Crude model | Ref | 0.95 (0.70, 1.28) | 0.83 (0.65, 1.06) | 1.15 (0.94, 1.39) | 1.03 (1.02, 1.04) | 0.335 |
| Age-adjusted | Ref | 1.11 (0.82, 1.45) | 1.13 (0.88, 1.45) | 1.59 (1.30, 1.94) | 1.04 (1.03, 1.05) | <0.001 |
| MV-adjustedb | Ref | 1.06 (0.79, 1.44) | 0.98 (0.76, 1.26) | 1.24 (1.00, 1.55) | 1.03 (1.01, 1.04) | 0.092 |

a. Here we adjusted for sex, race (categorical), education years, marital status (categorical), size of place (categorical), average cigarettes per day, Standard Metropolitan Statistical Area (categorical), and residence in urban region.

b. In addition to the covariates in a., here we also adjusted for hypertension at baseline, RBC, diagnosed MI, and diabetes.

Table 3. Associations between alcohol intake and hazard of death from cancer in males and females from NHANES II (N=9190).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | >2/week vs 0/week | | Continuous | |
| First approach, defining the exposure as alcohol intake age 21-entry, time scale as age | | | | |
| Adding interaction term | | | | |
|  | MV-adjusted | Heterogeneity | MV-adjusted | Heterogeneity |
| Female | 1.06 (0.82, 1.36) | 0.053 | 1.02 (1.01, 1.04) | 0.712 |
| Male | 0.69 (0.38, 1.27) |  | 1.02 (0.97, 1.06) |  |
| Stratified Cox | | | | |
|  | MV-adjusted | Heterogeneity | MV-adjusted | Heterogeneity |
| Female | 1.52 (1.06, 2.18) | 0.122 | 1.03 (0.99, 1.07) | 0.868 |
| Male | 1.08 (0.84, 1.40) |  | 1.02 (1.01, 1.04) |  |
| Subgroup analysis | | | | |
|  | MV-adjusted | Heterogeneity | MV-adjusted | Heterogeneity |
| Female | 1.35 (0.92-1.96) | 0.231 | 1.02 (0.97-1.06) | 0.293 |
| Male | 1.13 (0.87-1.48) |  | 1.03 (1.01-1.04) |  |
| Second approach, defining the exposure as prevalent alcohol intake at baseline, time scale as years since entry | | | | |
| Adding interaction term | | | | |
|  | MV-adjusted | Heterogeneity | MV-adjusted | Heterogeneity |
| Female | 1.59 (1.11-2.28) | 0.063 | 1.03 (0.99-1.07) | 0.758 |
| Male | 1.06 (0.82-1.37) |  | 1.02 (1.01-1.04) |  |
| Stratified Cox | | | | |
|  | MV-adjusted | Heterogeneity | MV-adjusted | Heterogeneity |
| Female | 1.58 (1.10-2.27) | 0.067 | 1.03 (0.99-1.07) | 0.772 |
| Male | 1.06 (0.82-1.38) |  | 1.02 (1.01-1.04) |  |
| Subgroup analysis | | | | |
|  | MV-adjusted | Heterogeneity | MV-adjusted | Heterogeneity |
| Female | 1.35 (0.93-1.98) | 0.240 | 1.01 (0.97-1.06) | 0.717 |
| Male | 1.15 (0.88-1.49) |  | 1.03 (1.01-1.04) |  |

**Results**

Table 1 shows the baseline characteristics of the complete case data. The number and percentages for categorical variables and mean and standard deviation for continuous variables are reported. Through these descriptive statistics, we measured the unbalanced distribution for the covariates between alcohol consumption groups and thus, explored potential confounding in the association between alcohol assumption and cancer mortality. Highlight any observations from Table 1?

First, we assume that the contributions of each covariate to the log hazard rate of cancer death are additive. Table 2 presents results of five Cox regression models through two settings. Considering alcohol consumption as a categorical variable, we observe that there are no significant associations for alcohol consumption of 0-0.5 drinks per week and 0.5-2 drinks per week compared to 0 drinks per week. For the category >2 drinks per week, the first approach yields an HR of 1.57 (95% CI, 1.29, 1.91) in the crude model and 1.25 (95% CI, 1.00, 1.54) in the MV-adjusted model. Following the second approach, the estimated HR is 1.15 (95% CI, 0.94, 1.39) in the crude model and is 1.24 (95% CI, 1.00, 1.55) in the MV-adjusted model. So consuming >2 drinks per week has detrimental effect on the cancer death than consuming no drink. All the five models yield similar results for the association between continuous alcohol consumption and the hazard of cancer death: the point estimates are around 1.03 and are of marginal significance. So, 1-unit increasing in alcohol intake per week increases the hazard rate of cancer death by about 3%, holding other covariates constant.

To test the sensitivity of the findings to the model specification, we conducted logistic and Poisson regression with time since baseline as the time scale (Appendix 2). The results are consistent with the Cox regressions results. Only the group with >2 drinks per week has a statistically significant association with death from cancer. The odds ratio from the MV-adjusted logistic regression is 1.29 (95% CI, 1.02, 1.63), and the incidence rate ratio from the MV-adjusted Poisson regression is 1.24 (95% CI 1.00, 1.55). Consistent with the beforementioned Cox regressions, the continuous alcohol consumption demonstrated statistically significant and weak association with cancer death in the MV-adjusted logistic regression (OR = 1.03, 95% CI, 1.01, 1.04) and MV-adjusted Poisson regression (IRR = 1.02, 95% CI, 1.01, 1.09).

Using Scofield’s residuals, we checked the proportional hazard assumption in the MV-adjusted Cox models (Appendix 3). Based on the corresponding chi-squared statistics and p-values for the exposure and for the whole model, we conclude that the assumption satisfied the alcohol consumption in all MV-adjusted Cox models.

In the Cox models in Table 2, we also assumed that the relationships between continuous variables and the log hazard rate of cancer death are linear, conditional on the other covariates. Based on the results from the primary analysis, the non-linear relationship analysis was conducted for the MV-adjusted model following the second approach. We compared the category-specific hazard ratio, and we found that xxx. Additionally, we fit a natural cubic spline with 3 degrees of freedom, and we found no clear curvature pattern. The estimated 95% CI for hazard ratio covers 1 before alcohol consumption of 5 drinks per week. This finding can partially explain why we found non-significance for the results from groups 0-0.5 and 0.5-2 drinks per week, but the significant association for the group >2 drinks per week.

With the findings from the primary analysis, we examined the potential modification of sex in MV-adjusted Cox models (Table 3). First, we compared the estimated hazard rate (HR) comparing >2 drinks per week to 0 drinks per week across males and females. The HRs estimated in the females are greater than in the males in all the models, indicating that alcohol consumption’s harmful effects are more pronounced in the females than in the males. However, the differences are not statistically significant across all the six models. For other alcohol intake categories, why don’t we present the result/what is the conclusion? Then we treated alcohol consumption as a continuous variable. All the six models yield similar HR for males and females, which are very close to the estimates in the overall sample analysis in Table 2, and the difference in HR by sex are of no statistical significance. Overall, no matter focusing on categorical or continuous alcohol consumption, we detected no statistically significant effect modification by sex.

**Discussion**

In this study with a cross-sectional dataset from NHANES II, we found a significant harmful association for alcohol consumption and the risk of death from cancer. Taking alcohol less than 2 drinks per week does not significantly increase the risk. The exploratory non-linear analysis demonstrates that such a relationship is close to linear with a significant harmful effect only after 5 drinks per week. No significant sex difference is found in the relationship.

In the primary analysis, we employed two different approaches to assess the association of alcohol consumption on hazard of cancer-specific death. The first approach takes the age as the time scale and takes the temptation into account, which has a more reasonable causal structure and explicit causal assumption. However, the second approach takes the whole dataset singly as a cross-sectional study and utilizes the time from entry as the time scale. The two approaches yield different results in the crude model but get similar results in the MV-adjusted model.

Given the data at hand, there are some explanations for the very similar findings in the two approaches. Since the estimated association for alcohol intake and cancer-specific death (about 1.03 HR per 1 drink/week increment) is very small, it suggests both approaches, even including the logistic and Poisson regression, don’t adjust for enough confounding. So, there could be residual confounding or other unknown confounders. Another possible explanation is that as the diseases’ distribution is pretty balanced across alcohol consumption groups, additional adjusting for these diseases and biomarkers such as baseline hypertension would not substantially influence the estimates. Lastly, it also could be the case that the true association is weak, and both approaches output the correct estimand.

In this study, we didn’t detect significant effect modification by sex, though the point estimates for the >2 drinks per week group show a more harmful effect among women than men. However, we cannot totally exclude the presence of potential effect modification since the model is of low power to detect such differences. We also explored some potential non-linear dose-response relationship, and we find the curve is pretty linear. This analysis also tells that no significant harmful effects manifest with a low dose of alcohol consumption, and the harmful effect becomes significant when taking alcohol greater than 5 drinks per week.

There are several limitations to our analysis. First, as we mentioned before, the estimated association would be very sensitive to unmeasured confounders. Although the findings from different models are consistent without additional confounder information, we must admit that the observed association is very likely to be attributed to an unmeasured confounder. Second, the original data is in the cross-sectional form, which limits the survival analysis. It is hard to tell which exposure or confounder happens first, making the model adjustment difficult and obscure. Third, we did not consider selection bias in the study. If censoring is dependent to cancer death, then there will be an open non-causal path between alcohol intake and cancer death, and causal methods will be needed to account for this selection bias. Last, we must notify that the event of interest in this study is the death from cancer. Deaths by other causes are competing risks but are treated as censored in this study. Although the Cox model is still valid, the logistic and Poisson regression are invalid, and their estimations are biased.

**Appendix**

图示

Description automatically generated

Appendix 1. Conceptual framework for Cox proportional hazard models in the primary analysis

Appendix 2. Association between alcohol intake and hazard of death from cancer in overall NHANES II (N=9190).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 0/week | 0-0.5/week | 0.5-2/week | >2/week | Continuous | Ptrend |
| Logistics regressiona | | | | | | |
| Crude model | Ref | 0.97 (0.72, 1.32) | 0.87 (0.68, 1.12) | 1.23 (1.00, 1.50) | 1.03 (1.01, 1.05) | 0.127 |
| Age-adjusted | Ref | 1.12 (0.82, 1.53) | 1.17 (0.90, 1.51) | 1.67 (1.36, 2.06) | 1.04 (1.03, 1.06) | <0.001 |
| MV-adjusted | Ref | 1.06 (0.77, 1.45) | 1.00 (0.77, 1.31) | 1.29 (1.02, 1.63) | 1.03 (1.01, 1.04) | 0.054 |
| Poisson regressiona | | | | | | |
| Crude model | Ref | 0.95 (0.70,1.27) | 0.84 (0.66, 1.07) | 1.16 (0.95, 1.41) | 1.03 (1.02, 1.04) | 0.309 |
| Age-adjusted | Ref | 1.11 (0.83, 1.50) | 1.14 (0.89, 1.45) | 1.60 (1.31, 1.95) | 1.04 (1.03, 1.05) | <0.001 |
| MV-adjusted | Ref | 1.06 (0.79, 1.44) | 0.98 (0.76, 1.27) | 1.24 (1.00, 1.55) | 1.02 (1.01, 1.09) | 0.088 |

1. Odds ratios are reported for logistic regressions and incidence rate ratios are reported for Poisson regressions. Log of follow-up years are set as the offset in Poisson regressions.

Appendix 3. Tests for the proportional hazard assumption throughout the Cox models in the primary analysis (N=9190).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Chi-squared statistic | p-value for alcohol consumption | Global p-value |
| First approach, defining the exposure as alcohol intake age 21-entry, time scale as age | | | |
| MV-adjusted categorical model | 6.93 | 0.074 | 0.178 |
| MV-adjusted continuous model | 0.15 | 0.697 | 0.2326 |
| Second approach, defining the exposure as prevalent alcohol intake at baseline, time scale as years since entry | | | |
| MV-adjusted categorical model | 2.92 | 0.404 | 0.653 |
| MV-adjusted continuous model | 0.53 | 0.465 | 0.672 |

图表, 折线图

Description automatically generated

Appendix 4. title? // not discussed in text?

1. https://wwwn.cdc.gov/nchs/nhanes/nhanes2/default.aspx [↑](#footnote-ref-1)