Epi 204 Project Report

**Introduction**

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 examined the association between alcohol consumption and risk of death from cancer.

**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 have been described elsewhere[[1]](#footnote-1). The cross-sectional survey collected demographic, socioeconomic, and biomarker variables. Death occurrence during the follow-up was administered. The original impute dataset contains 11,258 observations. After excluding 2008 observations with missing values in death status, 4 observations without baseline myocardial infarction information, 6 observations without baseline hypertension information, and excluding unreasonable and age information (there are 32 observations whose age at baseline is greater or equal to the age at last check), a total 0f 9190 observations were included in this study.

The exposure in this study is alcohol consumption (drinks/week) measured at baseline. The original exposure is recorded as a continuous variable with a large range (0 to 77 per week), and we considered continuous, categorical, and ordinal versions. During exploratory analysis, we found about 43.7% of people didn’t take alcohol per week. After excluding people with 0 alcohol consumption, the minimum, median, and maximum alcohol consumption were 0.5, 2.0, and 77.0 per week. Thus, we divided people into four groups based on their alcohol consumption: 0 per week, 0-0.5 drinks per week, 0.5-2 drinks per week, and >2 drinks per week. Categorical and continuous version of the exposure were included in the models to test the potential non-linear relationship, as well as to mitigate the influence of leverage values of alcohol consumption. Ordinal versions of the exposure were used to compute the p-value for trend of categorical variables. For the 4 categories, we assigned the numerical value of 1 to 4 to each group, i.e., the ordinal version of the alcohol consumption. Then this ordinal version of the exposure was included in the model and a p-value for the ordinal variable was reported, i.e., p-value for trend.

We also included following variables as confounders in this study: sex, age (continuous), race (White, Black, and others), martial status (categorical), average smoking cigarettes per day (continuous), hypertension status at baseline (yes vs. no), average red blood cells at baseline (continuous), diagnosed myocardial infarction at baseline (yes vs. no), diabetes at baseline (continuous), size of residence place (categorical), and urban region (yes vs. no). Descriptive statistics were conducted to describe the distribution of these covariates by alcohol consumption groups. Chi-squared test and analysis of variance were used for categorical and continuous variable respectively. The imbalance distribution indicated potential confounding which would be considered when modeling the outcome.

The outcome in this study is the risk of death during the follow-up. The event time is recorded as time to death of cancer, or time to censoring, or time to death from other causes, whichever comes first. In our analysis, death from other causes served as competing risks and we treated them as censoring. This simplified procedure to deal with competing risks yields valid cause-specific hazard in Cox models, but its unbiasedness in Poisson and logistic models is not clear. Thus, we decided to use Cox models as the main analysis and to conduct Poisson and logistic models as sensitivity analysis. In the main analysis, we conducted Cox proportional hazard models for our main research question. 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 acknowledge that the alcohol consumption might happen before the baseline, we chose age as the time scale. set the age of 21 to be the time origin. Following this approach, we conducted a crude model and a Multivariate-adjusted (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, or disease status at baseline since they might be the consequence of the prior alcohol intake.
2. The second approach is setting the survey baseline 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.

We explored the potential effect modifications by sex through three approaches: (1) including a product term by sex and alcohol consumption in the MV-adjusted Cox model without stratification; (2) conducting stratified Cox regressions by sex and including an interaction term by sex and alcohol intake; and (3) performing Cox regressions in males and females separately (as subgroup analysis). The difference between approaches (1) and (2) is that in approach (1) we allowed the hazard ratio (HR) to vary with sex while in approach (2) we additionally allowed the baseline hazard function to vary with sex. Wald-tests were utilized to test the significance of effect modification.

We checked the proportional hazard assumption using Schofield residuals. Test results for the MV-adjusted models were reported. We also explored the nonlinear dose-response relationship between alcohol consumption and death from cancer. We first compared the category-specific hazard ratio. Based on the results from the main analysis, we replaced the linear term of alcohol consumption by a natural cubic spline term in the MV-adjusted model following the second approach (setting year as the time scale).

In the sensitivity analyses, Poisson and logistic regressions were conducted with the same exposure configurations as well as confounder adjustments as the Cox model following the second approach (setting year as the time scale).

The statistical significance level was 0.05. R version 4.1.0 was used to clean data and conduct model analysis.

**Results**

The baseline characteristics were reported in the *Table 1*. The proportion of males monotonically increased from the group with 0 drinks per week (35.55%) to the group with >2 drinks per week (66.37%). Additionally, we observed a correlation between average smoking cigarettes and alcohol consumption. Contrary to our expectation, the educational years also monotonically increased from the group with 0 drinks per week (9.98 years) to the group with >2 drinks per week (12.06 years). Hypertension, myocardial infarction, and diabetes prevalence at baseline decreased with alcohol consumption level. Overall, in the study population, alcohol consumption showed strong correlation with sex, educational level, smoking habit, and baseline disease status. Participants with basic morbidities tend to consume less alcohol at baseline. Residence features for the participants were shown in *Table S1 (appendix).*

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Alcohol consumption | | | |  |
|  | 0/week | 0-0.5/week | 0.5-2/week | >2/week | p-value |
|  | N=4022 (43.76%) | N=935 (10.17%) | N=1720 (18.72%) | N=2513 (27.34%) |
| Sex, Male | 1430 (35.55%) | 365 (39.04%) | 851 (49.48%) | 1668 (66.37%) | <0.001 |
| Mean Age at entry (SD) | 57.93 (12.80) | 54.26 (13.37) | 51.54 (13.53) | 51.66 (13.17) | <0.001 |
| Race |  |  |  |  | 0.008 |
| White | 3469 (86.25%) | 821 (87.81%) | 1509 (87.73%) | 2239 (89.10%) |  |
| Black | 474 (11.79%) | 93 (9.95%) | 191 (11.10%) | 233 (9.27%) |  |
| Other | 79 (1.96%) | 21 (2.24%) | 20 (1.16%) | 41 (1.63%) |  |
| Mean year of schooling (SD) | 9.98 (3.59) | 11.03 (3.30) | 11.40 (3.41) | 12.06 (3.30) | <0.001 |
| Marital status |  |  |  |  | <0.001 |
| Married | 2865 (71.23%) | 679 (72.62%) | 1284 (74.65%) | 1961 (78.03%) |  |
| Widowed | 664 (16.51%) | 125 (13.37%) | 169 (9.83%) | 168 (6.69%) |  |
| Divorced | 190 (4.72%) | 67 (7.17%) | 100 (5.81%) | 168 (6.69%) |  |
| Separated | 94 (2.34%) | 20 (2.14%) | 59 (3.43%) | 70 (2.79%) |  |
| Never married | 200 (4.97%) | 40 (4.28%) | 103 (5.99%) | 139 (5.53%) |  |
| Unknown | 9 (0.22%) | 4 (0.43%) | 5 (0.29%) | 7 (0.28%) |  |
| Mean cigarettes per day (SD) | 4.81 (10.85) | 6.75 (12.65) | 8.27 (13.78) | 9.10 (13.92) | <0.001 |
| Hypertension at baseline, Yes | 1630 (40.53%) | 307 (32.83%) | 508 (29.53%) | 718 (28.57%) | <0.001 |
| Mean RBC (SD) | 4.72 (0.72) | 4.76 (0.72) | 4.74 (0.70) | 4.78 (0.70) | 0.712 |
| Diagnosed MI at baseline, Yes | 280 (6.96%) | 55 (5.88%) | 82 (4.77%) | 116 (4.62%) | <0.001 |
| Diabetes at baseline, Yes | 368 (9.15%) | 53 (5.67%) | 57 (3.31%) | 62 (2.47%) | <0.001 |

In the main analysis, we conducted Cox proportional hazard models for the risk of death from cancer with a homogeneity effect assumption. Table 2 presents results of five Cox regression models through two settings. Considering alcohol consumption as a categorical variable, we observed that there were 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 yielded 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 was 1.15 (95% CI, 0.94, 1.39) in the crude model and was 1.24 (95% CI, 1.00, 1.55) in the MV-adjusted model. Consuming >2 drinks per week showed detrimental effect on the cancer death than consuming no drink. All the five models yielded similar results for the association between continuous alcohol consumption and the hazard of cancer death: the point estimates were around 1.03 and were of marginal significance.

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 (*Table S2* in appendix). The results were consistent with the findings obtained using Cox models. 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 was 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 but 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). To sum up, alcohol consumption showed a detrimental effect for death from cancer only with >2 drinks per week as a categorical variable, and a weak but significant harmful effect as a continuous variable.

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. 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., we adjusted for hypertension at baseline, RBC, diagnosed myocardial infarction, and diabetes.

With Scofield’s residuals, we checked the proportional hazard assumption in the MV-adjusted Cox models (*Table S3* in appendix). Based on the corresponding chi-squared statistics and p-values for the exposure and for the whole model, we concluded that the assumption satisfied for the alcohol consumption in all MV-adjusted Cox models.

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. We chose this comparison because this is the only one showing statistical significance with alcohol consumption as a categorical variable. The HRs estimated in the females were greater than in the males in all the models, indicating that alcohol consumption’s harmful effects were more pronounced in the females than in the males. However, none of these differences was statistically significant across all the six models. Then we treated alcohol consumption as a continuous variable. All the six models yielded similar HR for males and females, very close to the estimates in the overall sample analysis in Table 2. The differences in HR by sex were of no statistical significance. Overall, no matter focusing on categorical or continuous alcohol consumption, we detected no statistically significant effect modification by sex.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Alcohol consumption | | | |
|  | >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 | p-valueHeterogeneity | MV-adjusted | p-valueHeterogeneity |
| 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 | p-valueHeterogeneity | MV-adjusted | p-valueHeterogeneity |
| 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 | p-valueHeterogeneity | MV-adjusted | p-valueHeterogeneity |
| 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 | p-valueHeterogeneity | MV-adjusted | p-valueHeterogeneity |
| 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 | p-valueHeterogeneity | MV-adjusted | p-valueHeterogeneity |
| 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 | p-valueHeterogeneity | MV-adjusted | p-valueHeterogeneity |
| 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) |  |

a. Wald test were used to test the Heterogeneity in the effect of alcohol consumption in male and females.

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.

**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**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Alcohol consumption | | | |  |
|  | 0/week | 0-0.5/week | 0.5-2/week | >2/week | p-value |
| Size of residence place |  |  |  |  | <0.001 |
| Urbanized are with 3, 000, 000 or more | 310 (7.71%) | 110 (11.76%) | 282 (16.40%) | 444 (17.67%) |  |
| Urbanized area with  1,000,000 to 2,999,999 | 358 (8.90%) | 134 (14.33%) | 252 (14.65%) | 439 (17.47%) |  |
| Urbanized area with 250,000 to 999,999 | 428 (10.64%) | 102 (10.91%) | 218 (12.67%) | 349 (13.89%) |  |
| Urbanized area under 250,000 | 303 (7.53%) | 83 (8.88%) | 165 (9.59%) | 234 (9.31%) |  |
| Urban place 25,000 or more outside urbanized area | 148 (3.68%) | 39 (4.17%) | 79 (4.59%) | 108 (4.30%) |  |
| Urban place 10,000 to 24,999 outside urbanized area | 172 (4.28%) | 43 (4.60%) | 59 (3.43%) | 96 (3.82%) |  |
| Urban place 2,500 to 9,999 outside urbanized area | 395 (9.82%) | 73 (7.81%) | 145 (8.43%) | 164 (6.53%) |  |
| Rural | 1908 (47.44%) | 351 (37.54%) | 520 (30.23%) | 679 (27.02%) |  |
| Standard metropolitan statistical Area |  |  |  |  | <0.001 |
| In central city | 915 (22.75%) | 259 (27.70%) | 524 (30.47%) | 754 (30.00%) |  |
| Not in central | 741 (18.42%) | 244 (26.10%) | 570 (33.13%) | 979 (38.96%) |  |
| Not in SMSA | 2366 (58.83%) | 432 (46.20%) | 626 (36.40%) | 780 (31.04%) |  |
| Resides in urban area, Yes | 2118 (52.66%) | 584 (62.55%) | 1202 (69.88%) | 1836 (73.06%) | <0.001 |

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)