**EPI202: Fall 2022**

**Homework 3**

**To be uploaded as PDF to course website by 9:30 am on Thursday December 1, 2022.**

Please provide concise and precise answers.

We encourage collaborative learning in this course. You may discuss homework assignments with other students. However, all written work that you submit for grading must be your own, in your own words, reflecting your understanding of the homework assignment. Homework assignments should not be prepared by copying, paraphrasing, or summarizing someone else’s work.

**Proper notation should be used throughout the assignment and all calculations should be shown.**

**Part I. Matched Design and Matched Data Analysis**

A group of investigators conducted a case-control study to investigate the relationship between receiving the Measles, Mumps and Rubella (MMR) vaccination on the outcome of autism. The investigators believed that having an immediate family member with autism was an important potential confounder in their study. They reasoned that having an immediate family member with autism may influence whether or not a child receives MMR vaccination, and it also affects their risk of autism. They therefore chose to match cases and controls on whether they had an immediate family member with autism. In this study, cases were frequency matched to controls, i.e., if the investigators identified 20 cases with an immediate family member with autism, they selected 20 controls with an immediate family member with autism.

Raw data are presented in the 2x2 tables below, stratified by the matching factor:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Immediate family member with autism | | |  |  | No immediate family member with autism | | |
|  | MMR | No MMR |  |  |  | MMR | No MMR |  |
| Cases | 15 | 49 | 64 |  | Cases | 19 | 45 | 64 |
| Controls | 30 | 34 | 64 |  | Controls | 24 | 40 | 64 |
|  | 45 | 83 | 128 |  |  | 43 | 85 | 128 |

Reviewers of the study did not agree with the investigator’s reasoning regarding confounding and matching.

**Reviewer 1** believed that having an immediate family member with autism may be related to risk of autism, but will not influence the decision to receive MMR vaccination.

1. Draw a DAG of the relationship between the matching factor, exposure and outcome in the study base, as hypothesized by reviewer 1. Please label all variables you include on the DAG.

Diagram

Description automatically generated

* 1. Under this assumption, is matching in the case-control study an example of appropriate matching, overmatching, unnecessary matching, or matching on an intermediate?

This is the unnecessary matching, because it matches a factor (the family history of autism) which is unrelated with exposure (MMR vaccination).

1. Reviewer 1 wants the investigators to re-do their analysis using their existing data to best account (i.e., with least bias and most precision) for his view of the relationship between the family history of autism, MMR vaccination, and the incidence of autism:
   1. Calculate an appropriate measure of association.
   2. Explain your choice of analysis.

For unnecessary matching, the analysis of not stratified would obtain the least biased and most precision measure of association, no matter whether the design is matching or not. The existing data were collected after matching.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MMR | No MMR |  |
| Cases | 34 | 94 | 128 |
| Controls | 54 | 74 | 128 |
|  | 88 | 168 | 256 |

Let be the prevalence of MMR among the cases, and be the prevalence of MMR among the controls.

This is the crude (unstratified) analysis for case-control study, which did not take the unnecessary matching factor (the family history of autism) into account.

**Reviewer 2** believed that having an immediate family member with autism is not linked to an increased risk of autism in a child, but does believe that having an immediate family member with autism will influence whether or not a family chooses to opt their child out of a MMR vaccination requirement.

1. Draw a DAG of the relationship between family history of autism, MMR vaccination and the incidence of autism in the study base, as hypothesized by reviewer 2. Please label all variables you include on the DAG.



* 1. Under this assumption, is matching in this case-control study an example of appropriate matching, overmatching, unnecessary matching, or matching on an intermediate?

This is overmatching, because the matching factor (the family history of autism) is related to exposure (MMR vaccination).

1. Reviewer 2 wants the investigators to re-do their analysis using their existing data to best account (i.e., with least bias and most precision) for his/her view of the relationship between the matching factor, exposure and outcome.
   1. Calculate an appropriate measure of association.
   2. Explain your choice of analysis.

For overmatching, the least biased and most precision measure of association would be obtained by not matching design and unstratified analysis. However, the existing data were collected after matching. The second optimal (unbiased, but reduced precision) choice would be the stratified analysis; otherwise, it would be biased if the unstratified analysis was conducted.

**Part II. Matched Design and Matched Data Analysis (II)**

Another set of investigators conduct a case-control study, investigating the association between MMR vaccination versus no MMR vaccination and the risk of developing autism. In this study, cases were individually matched to neighborhood controls on gender in a 1:1 ratio. (Assume gender is a confounder of this relationship in the study base).

The following tables show data from the study:

Crude data from this study:

|  |  |  |  |
| --- | --- | --- | --- |
|  | MMR Vaccine | No MMR Vaccine |  |
| Cases | 30 | 34 | 64 |
| Controls | 29 | 35 | 64 |
|  | 59 | 69 | 128 |

Data from the pair-matched study:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Controls | |
|  |  | MMR Vaccine | No MMR Vaccine |
| Cases | MMR Vaccine | 17 | 13 |
| No MMR Vaccine | 12 | 22 |

1. Use the data above to:
   1. Calculate and interpret in words a valid measure of association

After accounting for the matching factor (gender), the odds of autism for those who had MMR vaccine is 1.0833 times the odds of autism for those who did not have MMR vaccine, assuming no residual confounding by gender, other unmeasured confounding, selection bias, or information bias.

* 1. Calculate and interpret in words the confidence interval for the measure in (a)

The 95% confidence interval for the is:

The 95% confidence interval for the is .

With 95% confidence, these data are consistent with the odds ratios ranging from 0.4943 to 2.3742 for the association between MMR vaccine and autism, after accounting for the matching factor (gender), assuming no residual confounding by gender, no other unmeasured confounding, no selection bias, and no information bias.

* 1. How many individuals (cases and controls) contribute information to this measure of association? Explain your answer.

Only the discordant pairs contribute information to the measure of the association between MMR vaccine and autism.

1. Is there statistical evidence for an association between MMR vaccination and the development of autism?
   1. State the null and alternate hypothesis.
   2. Calculate a valid test statistic.
   3. What is the p-value?

**.** There is no association between MMR vaccine and autism, after accounting for the matching factor (gender).

. There is an association between MMR vaccine and autism, after accounting for the matching factor (gender).

under the null,

**Chart, bar chart

Description automatically generated**

**Part III. Exploring the association between sedentary lifestyle at baseline and all-cause mortality using tabular analyses**

In this section, you will use data from a prospective cohort study based upon the first 10 years of follow-up of participants in the Myocardial Infarction Onset Study to examine the relationship between sedentary lifestyle defined as physical activity less than once per week compared to one or more times per week (at baseline) and the cumulative incidence of death from any cause within the first 10 years. The variables in this dataset are described below. [Note: In this study, conducted in the late 1980s, sex was recorded and categorized as female or male based on NIH reporting requirements in place at the time. No data on gender was recorded]:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| id | ID number |
| age | Age (continuous, years) |
| age\_cat | Age Category (1: <50yrs, 2: 50-64 yrs, 3: 65+ yrs) |
| Female | Female (1: female, 0: male) |
| Married | Married (1: yes, 0:no) |
| Educ | Educational Attainment (1: <HS, 2: HS, 3: >HS) |
| dm | Diabetes (1: yes, 0:no) |
| htn | Hypertension (1: yes, 0:no) |
| phys\_activity | Frequency of Physical Activity (0: <1/wk, 1: 1-3/wk, 2: 4+/wk) |
| follow\_up | Duration of follow-up (years) |
| dead | Death from any cause in the first 10 years (0: Alive, 1: Died) |

In homework 3, you will use tabular analysis methods to evaluate the association between **sedentary lifestyle at baseline** (as coded in homework 2) and the outcome of a**ll-cause mortality by ten years.** Next week, in homework 4, you will continue and extend this analysis using regression models.

You may use SAS, STATA, R or any other statistical analysis software package of your choosing. Include relevant lines of code. If you do any calculations by hand or using the Epi 202 calculator, identify the formulas you used and the values you included in the formulas, as in previous homework assignments

**Note: for all questions, please include all relevant formulas and define any variables that you use and have not already defined in previous questions to receive full credit.**

1. Use the data from the MI Onset study to complete Table 1 for a paper describing the results of the study. The rows of the table should provide information about the potential confounders and modifiers of interest in this study. Include number of subjects and percentages for categorical variables, means and standard deviations for continuous variables, as shown below. Include the appropriate p-values in your table (even though we may be skeptical about the validity and interpretability of these crude p-values).

**Table 1.** Baseline characteristics of individuals included in the analyses

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sedentary lifestyle  (N = 3199) | Not sedentary  (N = 513) | p-value |
| Age (years) | 62.524 (12.571) | 54.873 (11.644) | <0.001 |
| Age Category |  |  | <0.001 |
| < 50 years | 558 (17%) | 179 (35%) |  |
| 50-64 years | 1171 (37%) | 221 (43%) |  |
| 65+ years | 1470 (46%) | 113 (22%) |  |
| Gender |  |  | <0.001 |
| Male | 2075 (65%) | 425 (83%) |  |
| Female | 1124 (35%) | 88 (17%) |  |
| Marriage Status |  |  | <0.001 |
| Married | 2059 (64%) | 370 (72%) |  |
| Non-married | 1140 (36%) | 143 (28%) |  |
| Education Attainment |  |  | <0.001 |
| < HS | 746 (23%) | 64 (12%) |  |
| HS | 1354 (42%) | 204 (40%) |  |
| > HS | 1099 (34%) | 245 (48%) |  |
| Diabetes |  |  | <0.001 |
| Yes | 724 (23%) | 69 (13%) |  |
| No | 2465 (77%) | 444 (87%) |  |
| Hypertension |  |  | <0.001 |
| Yes | 1448 (45%) | 182 (35%) |  |
| No | 1751 (55%) | 331 (65%) |  |

1. Here, you will conduct crude analyses to evaluate the association between sedentary lifestyle at baseline and the cumulative incidence of death from any cause during the 10-year follow-up period.
   1. Calculate the cumulative incidence ratio and 95% confidence interval for the CIR for the association between sedentary lifestyle at baseline and death in the study population. **There is no need to interpret the finding in words.**

over the 10-year follow-up period

The 95% confidence interval for the ln CIR for the association between sedentary lifestyle at baseline and 10-year all-cause mortality is:

The 95% confidence interval for the CIR for the association between sedentary lifestyle at baseline and 10-year all-cause mortality is:

* 1. Is there an association between a sedentary lifestyle at baseline and 10-year all-cause mortality?
     1. State the null and alternative hypotheses
     2. Use software to compute the value of the test statistic
     3. Use software to compute the p-value

**There is no need to interpret the finding in words.**

. There is no association between a sedentary lifestyle at baseline and 10-year all-cause mortality.

. There is an association between a sedentary lifestyle at baseline and 10-year all-cause mortality.

under the null

Table

Description automatically generated

* 1. Compute the cumulative incidence ratio and 95% confidence interval for the association between sedentary lifestyle at baseline and death from any cause during the 10-year follow-up **among those with diabetes**. **There is no need to interpret the finding in words.**

over the 10-year follow-up period

The 95% confidence interval for :

The 95% confidence interval for the association between a sedentary lifestyle at baseline and 10-year all-cause mortality among those with diabetes is .

Table

Description automatically generated

* 1. Compute the cumulative incidence ratio and 95% confidence intervals for the association between sedentary lifestyle at baseline and death from any cause during the 10-year follow-up **among those without diabetes**. **There is no need to interpret the finding in words.**

over the 10-year follow-up period

The 95% confidence interval for :

The 95% confidence interval for the association between a sedentary lifestyle at baseline and 10-year all-cause mortality among those without diabetes is .

Table

Description automatically generated

* 1. Is there statistical evidence that diabetes modifies the association between sedentary lifestyle at baseline and 10-year all-cause mortality in the multiplicative scale?
     1. State null and alternative hypothesis,
     2. Compute a valid test statistic,
     3. Calculate the p-value associated with the test statistic,
     4. Interpret the p-value in words,
     5. State the outcome of a Neyman-Pearson hypothesis test.

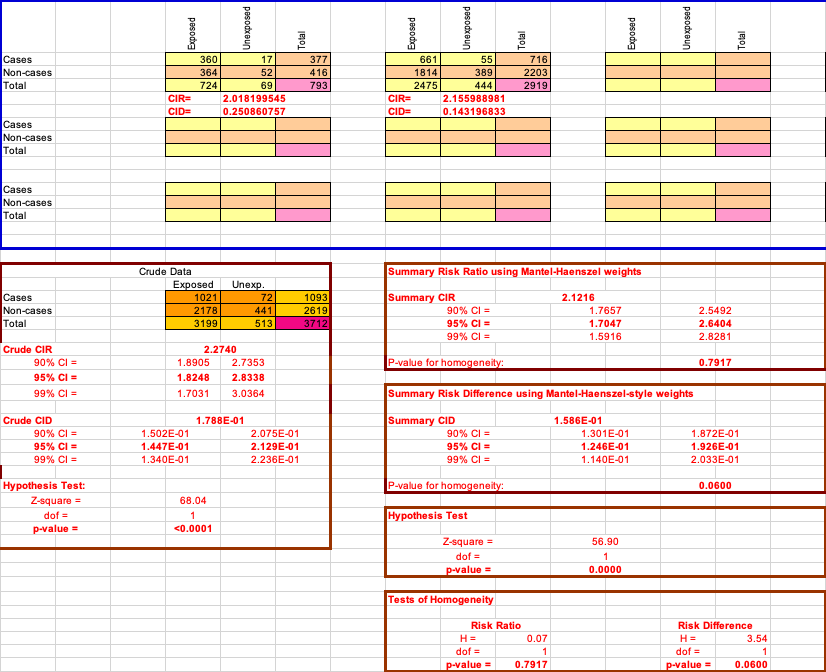
There is no effect measure modification of the association between sedentary lifestyle at baseline and 10-year all-cause mortality in the multiplicative scale.

There is effect measure modification of the association between sedentary lifestyle at baseline and 10-year all-cause mortality in the multiplicative scale.

under the null

The probability that a result is as or more extreme than 0.07, when the null is true, is 0.7917.

We fail to reject *H0* at the level. There is insufficient evidence to conclude that there is effect measure modification of the association between sedentary lifestyle at baseline and 10-year all-cause mortality in the multiplicative scale.



* 1. Based on your responses to question 2, complete the table below.

Table 2: Sedentary lifestyle at baseline and death during following 10-years

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Diabetes | |  | No diabetes | |
|  | Sedentary | Not sedentary |  | Sedentary | Not sednetary |
| Number of deaths | 360 | 17 |  | 661 | 55 |
| Number of participants | 724 | 69 |  | 2475 | 444 |
| Cumulative incidence ratio (95% CI) | 2.0182 | |  | 2.1560 | |
| p-value for the test of homogeneity of the cumulative incidence ratio: p=0.7917 | | | | | |

1. Next, you will adjust for diabetes at baseline and evaluate the association between sedentary lifestyle at baseline and 10-year all-cause mortality.
   1. Compute the Mantel-Haenszel summary cumulative incidence ratio and 95% confidence interval for the association between sedentary lifestyle at baseline and death during the follow-up period **after stratifying and adjusting for diabetes at baseline**. **There is no need to interpret the finding in words.**

over the 10-year follow-up period

The 95% confidence interval for :

The 95% confidence interval for , the association between sedentary lifestyle at baseline and the 10-year all-cause mortality is .

* 1. Is there an association between sedentary lifestyle at baseline and 10-year all-cause mortality **after adjusting for diabetes at baseline**?
     1. State the null and alternative hypotheses
     2. Use software to compute the value of the test statistic
     3. Use software to compute the p-value.

**There is no need to interpret the finding in words.**

. There is no association between sedentary lifestyle at baseline and 10-year all-cause mortality after adjusting for diabetes at baseline.

. There is an association between sedentary lifestyle at baseline and 10-year all-cause mortality after adjusting for diabetes at baseline.

under the null