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		
	MMR	No MMR	
Cases	15	49	64
Controls	30	34	64
	45	83	128

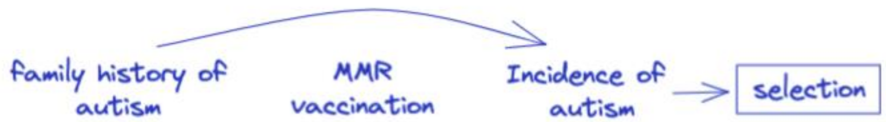
	No immediate family member with autism		
	MMR	No MMR	
Cases	19	45	64
Controls	24	40	64
	43	85	128

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

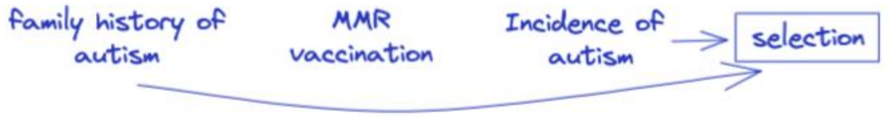
Reviewer 1 believed that having an immediate family member with autism may be related to risk of autism, but will not influence the decision to receive MMR vaccination.

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.

This case-control study, before matching:



This case-control study, after matching:



a. 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).

2. 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:
   a. Calculate an appropriate measure of association.
   b. 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

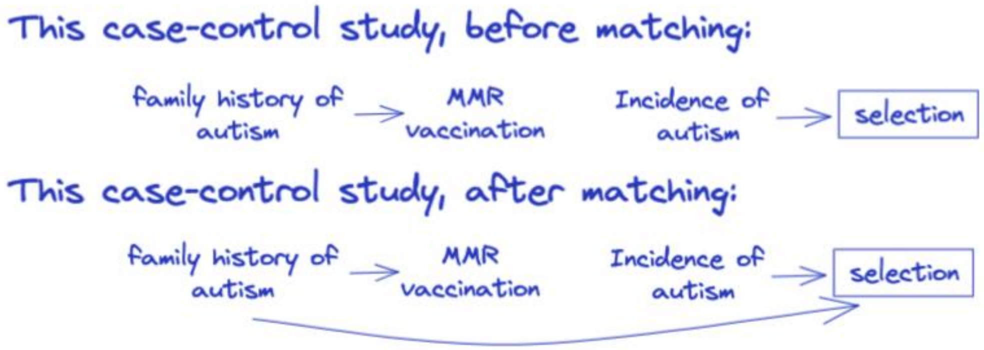
Let p_1 be the prevalence of MMR among the cases, and p_2 be the prevalence of MMR among the controls.

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{34 \times 74}{54 \times 94} = 0.4957$$

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.

3. 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.



a. 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).

4. 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.
   a. Calculate an appropriate measure of association.
   b. 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.

$$OR_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i d_i}{T_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{T_i}} = \frac{\frac{15\times34}{128} + \frac{19\times40}{128}}{\frac{49\times30}{128} + \frac{45\times24}{128}} = 0.498$$

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:
   a. Calculate and interpret in words a valid measure of association

$$OR_{MH} = \frac{f_{10}}{f_{01}} = \frac{13}{12} = 1.0833$$

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.

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

The 95% confidence interval for the $ln(OR_{MH})$ is:

$$ln\left(\widehat{OR_{MH}}\right) \pm 1.96\sqrt{Var\left(ln\left(\widehat{OR_{MH}}\right)\right)} = ln\left(\widehat{OR_{MH}}\right) \pm 1.96\sqrt{\frac{1}{f_{10}} + \frac{1}{f_{01}}}$$

$$= ln(1.0833) \pm 1.96\sqrt{\frac{1}{13} + \frac{1}{12}} = (-0.7046, 0.8647)$$

The 95% confidence interval for the OR_{MH} is $e^{(-0.7046,0.8647)} = (0.4943, 2.3742)$.

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.

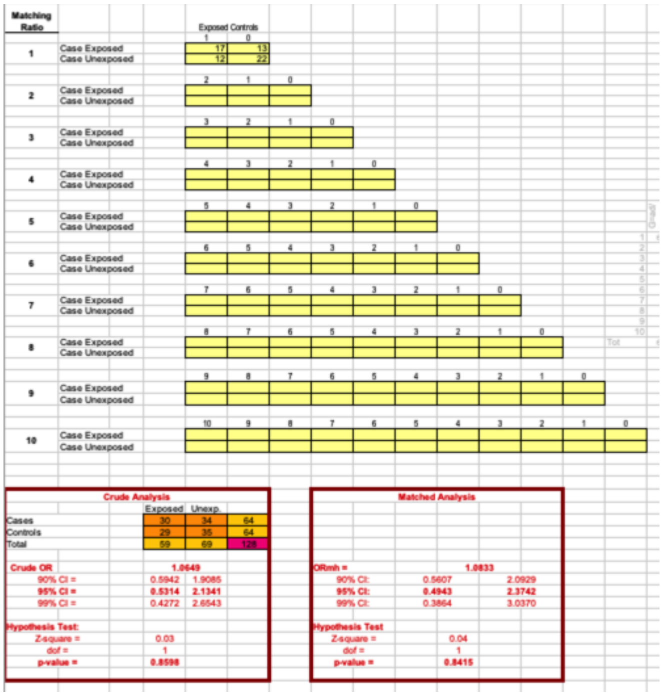
c. How many individuals (cases and controls) contribute information to this measure of association? Explain your answer.

$$(f_{10} + f_{01}) \times 2 = (13 + 12) \times 2 = 50$$

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

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

$\boldsymbol{H_0: OR_{MH} = 1.}$ There is no association between MMR vaccine and autism, after accounting for the matching factor (gender).
$H_A: OR_{MH} \neq 1.$ There is an association between MMR vaccine and autism, after accounting for the matching factor (gender).
$Z^2 = \frac{(f_{10}-f_{01})^2}{f_{10}+f_{01}} = 0.04 \sim \chi_1^2$ under the null, $p = P(\chi_1^2 \geq 0.04) = 0.8415$



Matching Ratio		Exposed Controls										
		1	0									
1	Case Exposed	17	13									
	Case Unexposed	12	22									
		2	1	0								
2	Case Exposed											
	Case Unexposed											
		3	2	1	0							
3	Case Exposed											
	Case Unexposed											
		4	3	2	1	0						
4	Case Exposed											
	Case Unexposed											
		5	4	3	2	1	0					
5	Case Exposed											
	Case Unexposed											
		6	5	4	3	2	1	0				
6	Case Exposed											
	Case Unexposed											
		7	6	5	4	3	2	1	0			
7	Case Exposed											
	Case Unexposed											
		8	7	6	5	4	3	2	1	0		
8	Case Exposed											
	Case Unexposed											
		9	8	7	6	5	4	3	2	1	0	
9	Case Exposed											
	Case Unexposed											
		10	9	8	7	6	5	4	3	2	1	0
10	Case Exposed											
	Case Unexposed											

Crude Analysis

	Exposed	Unexp.	
Cases	30	34	64
Controls	29	35	64
Total	59	69	128

Crude OR	**1.0649**	
90% CI =	0.5942	1.9085
95% CI =	**0.5314**	**2.1341**
99% CI =	0.4272	2.6543
Hypothesis Test:		
Z-square =	0.03	
dof =	1	
p-value =	**0.8598**	

Matched Analysis

ORmh =	**1.0833**	
90% CI:	0.5607	2.0929
95% CI:	**0.4943**	**2.3742**
99% CI:	0.3864	3.0370
Hypothesis Test		
Z-square =	0.04	
dof =	1	
p-value =	**0.8415**	

NAME: Diego Liang

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.

NAME: Diego Liang

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%)	

2. 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.
   a. 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.**

$$CIR = \frac{1021/(1021+2178)}{72/(72+441)} = 2.274 \text{ over the 10-year follow-up period}$$

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

$$\ln\left(\frac{\mathrm{a}}{\mathrm{N}_1}/\frac{\mathrm{b}}{\mathrm{N}_0}\right) \pm 1.96\sqrt{\frac{\mathrm{c}}{\mathrm{aN}_1}+\frac{\mathrm{d}}{\mathrm{bN}_0}} = (0.6015, 1.0416)$$

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

$$e^{(0.6015,0.0416)} = (1.8248,2.8338)$$

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

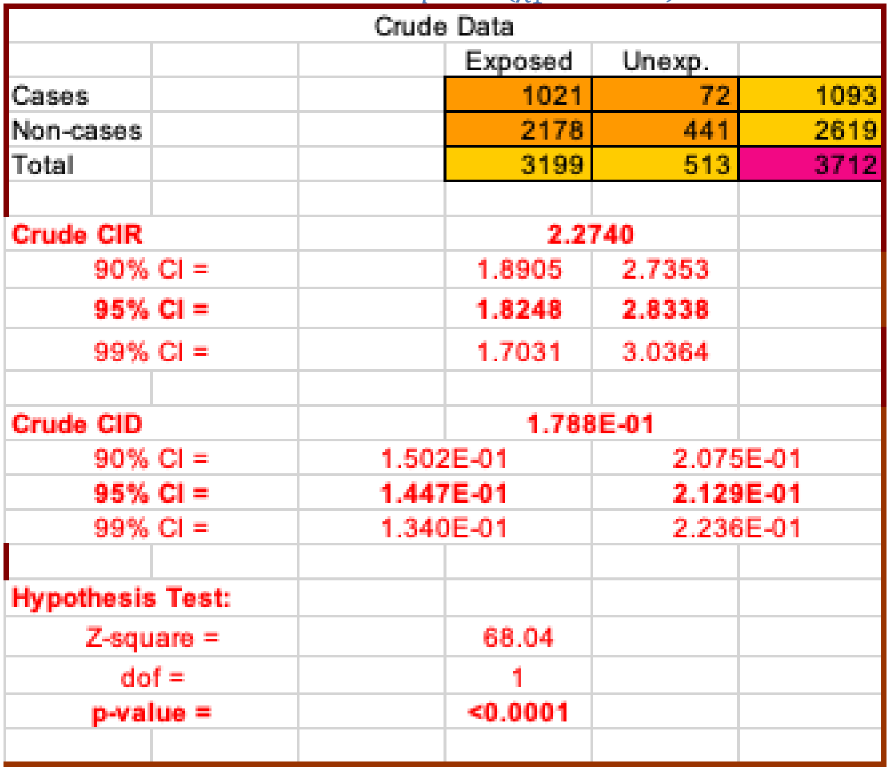
There is no need to interpret the finding in words.

$H_0: CIR = 1$. There is no association between a sedentary lifestyle at baseline and 10-year all-cause mortality.

$H_A: CIR \neq 1$. There is an association between a sedentary lifestyle at baseline and 10-year all-cause mortality.

$$Z^2 = \frac{\left(x-\hat{E}(X|H_0)\right)^2}{\widehat{Var}(X|H_0)} = 68.04 \sim \chi_1^2 \text{ under the null}$$

$$p = P(\chi_1^2 \geq 68.04) < 0.0001$$



Crude Data

			Exposed	Unexp.	
Cases			1021	72	1093
Non-cases			2178	441	2619
Total			3199	513	3712
Crude CIR			2.2740		
90% CI =			1.8905	2.7353	
95% CI =			**1.8248**	**2.8338**	
99% CI =			1.7031	3.0364	
Crude CID			1.788E-01		
90% CI =		1.502E-01		2.075E-01	
95% CI =		**1.447E-01**		**2.129E-01**	
99% CI =		1.340E-01		2.236E-01	
Hypothesis Test:					
Z-square =			68.04		
dof =			1		
p-value =			<0.0001		

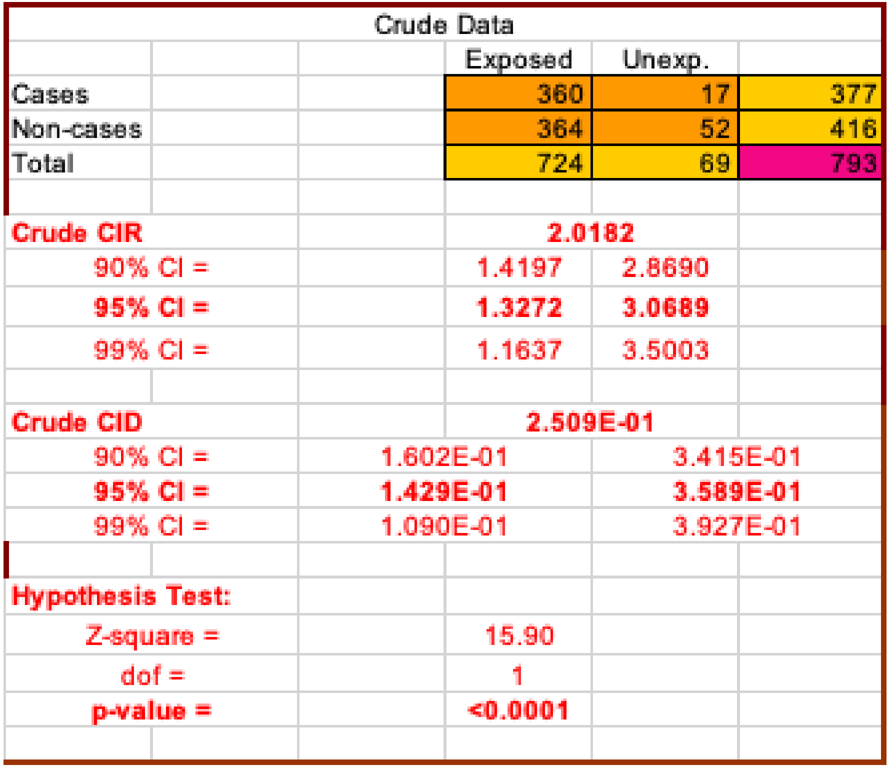
c. Compute the cumulative incidence ratio and 95% confidence interval for the association between sedentary lifestyle at baseline and death from any cause during the 10-year follow-up **among those with diabetes. There is no need to interpret the finding in words.**

$CIR = \frac{360/(360+364)}{17/(17+52)} = 2.018$ over the 10-year follow-up period

The 95% confidence interval for $ln(CIR)$:

$$ln(CIR) \pm 1.96\sqrt{Var(ln(CIR))} = ln(CIR) \pm 1.96\sqrt{\frac{c}{aN_1} + \frac{d}{bN_0}} = (0.2831,1.1213)$$

The 95% confidence interval for the association between a sedentary lifestyle at baseline and 10-year all-cause mortality among those with diabetes is $e^{(0.2831,1.1213)} = (1.3272,3.0689)$.



			Crude Data		
			Exposed	Unexp.	
Cases			360	17	377
Non-cases			364	52	416
Total			724	69	793
Crude CIR			2.0182		
	90% CI =		1.4197	2.8690	
	95% CI =		1.3272	3.0689	
	99% CI =		1.1637	3.5003	
Crude CID			2.509E-01		
	90% CI =	1.602E-01		3.415E-01	
	95% CI =	1.429E-01		3.589E-01	
	99% CI =	1.090E-01		3.927E-01	
Hypothesis Test:					
	Z-square =		15.90		
	dof =		1		
	p-value =		<0.0001		

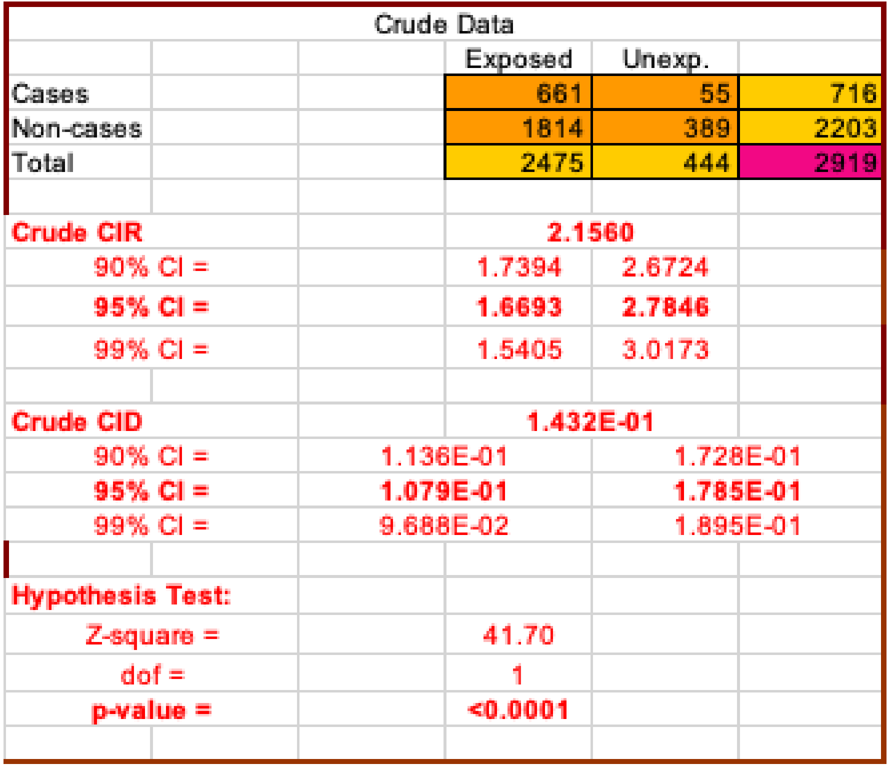
d. Compute the cumulative incidence ratio and 95% confidence intervals for the association between sedentary lifestyle at baseline and death from any cause during the 10-year follow-up **among those without diabetes. There is no need to interpret the finding in words.**

$CIR = \frac{661/(661+1814)}{55/(55389)} = 2.156$ over the 10-year follow-up period

The 95% confidence interval for $ln(CIR)$:

$$ln(CIR) \pm 1.96\sqrt{Var(ln(CIR))} = ln(CIR) \pm 1.96\sqrt{\frac{c}{aN_1} + \frac{d}{bN_0}} = (0.5124, 1.0241)$$

The 95% confidence interval for the association between a sedentary lifestyle at baseline and 10-year all-cause mortality among those without diabetes is $e^{(0.5124,1.0241)} = (1.6693, 2.7846)$.



			Crude Data		
			Exposed	Unexp.	
Cases			661	55	716
Non-cases			1814	389	2203
Total			2475	444	2919
Crude CIR			2.1560		
90% CI =			1.7394	2.6724	
95% CI =			**1.6693**	**2.7846**	
99% CI =			1.5405	3.0173	
Crude CID			1.432E-01		
90% CI =		1.136E-01		1.728E-01	
95% CI =		**1.079E-01**		**1.785E-01**	
99% CI =		9.688E-02		1.895E-01	
Hypothesis Test:					
Z-square =			41.70		
dof =			1		
p-value =			<0.0001		

e. Is there statistical evidence that diabetes modifies the association between sedentary lifestyle at baseline and 10-year all-cause mortality in the multiplicative scale?

i. State null and alternative hypothesis,

ii. Compute a valid test statistic,

iii. Calculate the p-value associated with the test statistic,

iv. Interpret the p-value in words,

v. State the outcome of a Neyman-Pearson hypothesis test.

H_0: There is no effect measure modification of the association between sedentary lifestyle at baseline and 10-year all-cause mortality in the multiplicative scale.

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

$$CIR_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i N_{0i}}{T_i}}{\sum_{i=1}^{I} \frac{b_i N_{1i}}{T_i}}$$

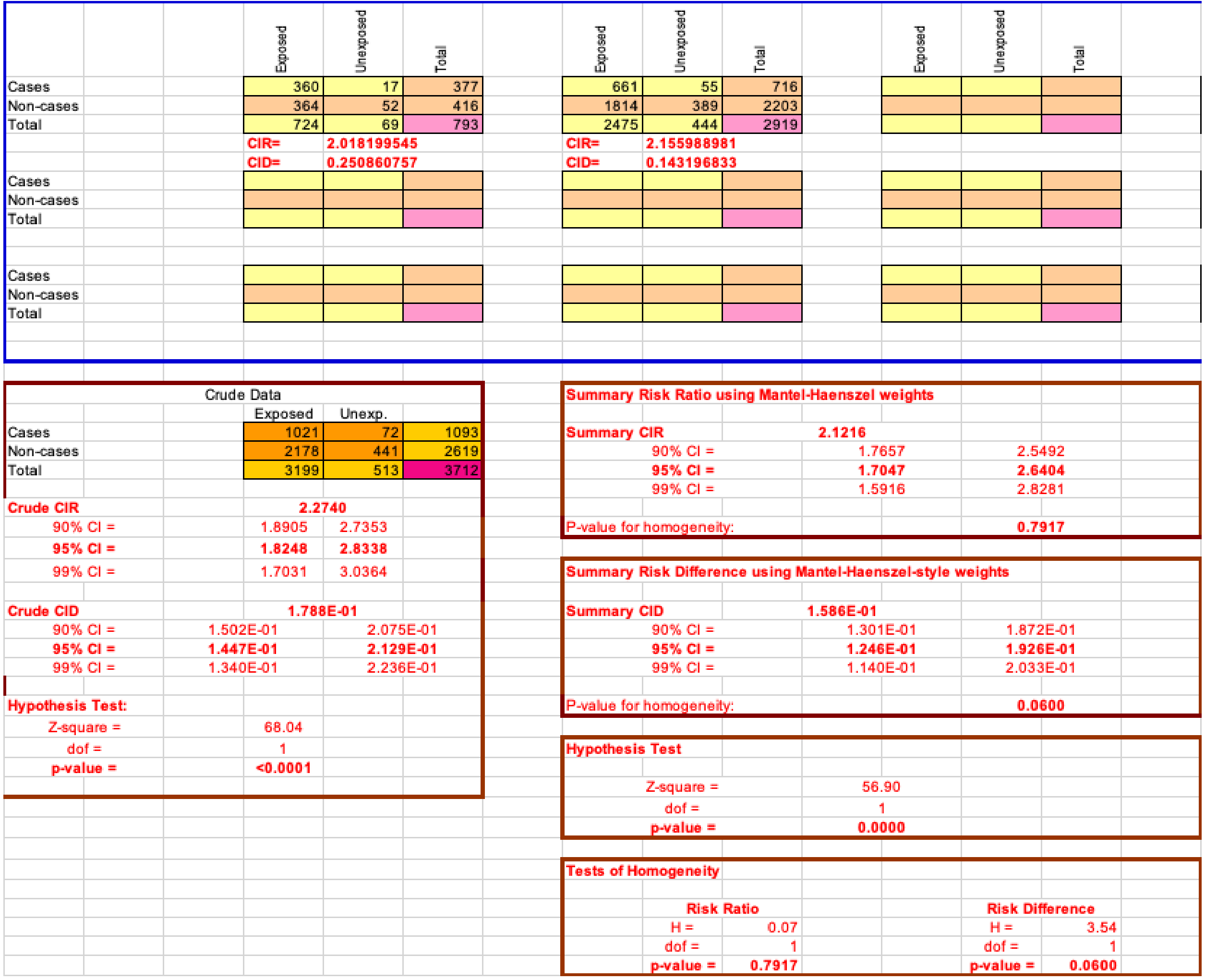
$$Var(CIR_i) = \frac{c_i}{a_i N_{1i}} + \frac{d_i}{b_i N_{0i}}$$

$$H = \sum_{i=1}^{I} \frac{\left(ln(CIR_i) - ln(CIR_{MH})\right)^2}{Var(CIR_i)} = 0.07 \sim \chi_1^2 \text{ under the null}$$

$$p = P(\chi_1^2 \geq 0.07) = 0.7917$$

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 α = 0.05 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.



	Exposed	Unexposed	Total	Exposed	Unexposed	Total	Exposed	Unexposed	Total
Cases	360	17	377	661	55	716			
Non-cases	364	52	416	1814	389	2203			
Total	724	69	793	2475	444	2919			
	CIR=	2.018199545		CIR=	2.155988981				
	CID=	0.250860757		CID=	0.143196833				
Cases									
Non-cases									
Total									
Cases									
Non-cases									
Total									

Crude Data

	Exposed	Unexp.	
Cases	1021	72	1093
Non-cases	2178	441	2619
Total	3199	513	3712

Crude CIR	2.2740	
90% CI =	1.8905	2.7353
95% CI =	**1.8248**	**2.8338**
99% CI =	1.7031	3.0364
Crude CID	1.788E-01	
90% CI =	1.502E-01	2.075E-01
95% CI =	**1.447E-01**	**2.129E-01**
99% CI =	1.340E-01	2.236E-01
Hypothesis Test:		
Z-square =	68.04	
dof =	1	
p-value =	**<0.0001**	

Summary Risk Ratio using Mantel-Haenszel weights

Summary CIR	2.1216	
90% CI =	1.7657	2.5492
95% CI =	**1.7047**	**2.6404**
99% CI =	1.5916	2.8281
P-value for homogeneity:		**0.7917**

Summary Risk Difference using Mantel-Haenszel-style weights

Summary CID	1.586E-01	
90% CI =	1.301E-01	1.872E-01
95% CI =	**1.246E-01**	**1.926E-01**
99% CI =	1.140E-01	2.033E-01
P-value for homogeneity:		**0.0600**

Hypothesis Test

Z-square =	56.90
dof =	1
p-value =	**0.0000**

Tests of Homogeneity

Risk Ratio		Risk Difference	
H =	0.07	H =	3.54
dof =	1	dof =	1
p-value =	**0.7917**	**p-value =**	**0.0600**

f. 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 (1.3272,3.0689)		2.1560 (1.6693, 2.7846)	

p-value for the test of homogeneity of the cumulative incidence ratio: p=0.7917

3. Next, you will adjust for diabetes at baseline and evaluate the association between sedentary lifestyle at baseline and 10-year all-cause mortality.

   a. 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.**

$CIR_{MH} = \frac{\sum_{i=1}^{I}\frac{a_i N_{0i}}{T_i}}{\sum_{i=1}^{I}\frac{b_i N_{1i}}{T_i}} = 2.1216$ over the 10-year follow-up period

The 95% confidence interval for $ln(CIR_{MH})$:

$$ln(CIR_{MH}) \pm 1.96\sqrt{Var\big(ln(CIR_{MH})\big)} = ln(CIR_{MH}) \pm 1.96\sqrt{\frac{\sum_{i=1}^{I}(M_{1i}N_{1i}N_{0i} - a_i b_i T_i)/T_i}{(\sum_{i=1}^{I}\frac{a_i N_{0i}}{T_i})(\sum_{i=1}^{I}\frac{b_i N_{1i}}{T_i})}}$$

$$= (0.5334, 0.9709)$$

The 95% confidence interval for CIR_{MH}, the association between sedentary lifestyle at baseline and the 10-year all-cause mortality is $e^{(0.5334,0.9709)} = (1.7047,2.6404)$.

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

There is no need to interpret the finding in words.

$H_0: CIR_{MH} = 1$. There is no association between sedentary lifestyle at baseline and 10-year all-cause mortality after adjusting for diabetes at baseline.

$H_A: CIR_{MH} \neq 1$. There is an association between sedentary lifestyle at baseline and 10-year all-cause mortality after adjusting for diabetes at baseline.

$$Z^2 = \frac{\left(X - \hat{E}(X|H_0)\right)^2}{\widehat{Var}(X|H_0)} = \frac{\left(\sum_{i=1}^{I} a_i - \sum_{i=1}^{I}\frac{N_{1i}M_{0i}}{T_i}\right)^2}{\sum_{i=1}^{I}\frac{M_{1i}M_{0i}N_{1i}N_{0i}}{T_i^3}} = 56.90 \sim \chi_1^2 \text{ under the null}$$

$$p = P(\chi_1^2 \geq 56.90) < 0.001$$