Question 1 (7 marks)

It has been established that there is an association between sedentary behaviours

(including sitting for extended periods of time, TV viewing time, and screen time) and all-cause mortality.

However, it is not clear whether the association is causal.

a) (5 marks) On the next page are the (lightly edited) abstracts from two past cohort studies examining the association between sedentary behaviours and all-cause mortality (the papers are also linked in Learn under the Assignment 2 folder). Consider the five Bradford Hill guidelines/criteria for judging whether an association is causal listed in the table below. For each criteria and each study indicate whether there is:
   - Evidence for causality;
   - Inconclusive evidence for causality; or
   - No evidence for causality (either evidence is null or there is lack of evidence).

Justify your selection with direct reference to the design and/or results of the study.

Note: For the purposes of this question you may treat HR as equivalent to RR.

You should recreate the table below in your solutions filling in the blank cells with your assessments

Criteria	Hidde P. van der Ploeg et al. (2012)	Katzmarzyk et al. (2009)
Temporal Relationship	There is evidence for causality as the outcome of death would have to occur after the exposure: hours/day sitting.	Likewise, for the Katzmarzyk study there is also evidence for causality as the outcome happens after the exposure. Where the outcome is death, and the exposure is self reported daily time of sitting.
Strength of Association	There is inconclusive evidence for causality as the strength of association although is present as RR (HR) is lying between 1.02 and 1.40 it is not high enough to clearly indicate that there is evidence for causality.	There is inconclusive evidence for causality as the strength of association although is present as RR (HR) is lying between 1.00 and 1.54 it is not high enough to clearly indicate that there is evidence for causality.
Dose-Response Relationship	There is evidence for causality as higher/longer sitting times were associated with a larger all cause HR. From the Hidde Study there was a recorded 1.02 RR (HR) for 4-8 hours of sitting, 1.14 for 8-11 hours of sitting and 1.40 for 11 or more hours a day	There is evidence for causality as higher/longer sitting times were associated with a larger all cause RR (HR). From the Katzmarzyk Study there was a recorded RR of 1.00 for a daily sitting time of “almost none of the time”, 1.01 for “one fourth of the time”, 1.22 for “half of the time”, 1.47 for “three

		fourths of the time" and 1.54 for almost all of the time.
Cessation of Exposure	There is evidence for causality since the RR decreases as the levels of sitting decrease and the RR is statistically significant at all levels except for the baseline level of 4 or less hours a day (this can be overlooked as it is the baseline measurement) as can be seen by the Confidence Interval minimum being always larger than 1.	Although the RR decreases as the levels of sitting decrease there is no indication of these findings to be statistically significant. There is a lack of evidence for causality due to a lack of statistically significant data.
Consideration of Alternative Explanations	There is inconclusive evidence regarding whether sitting is associated with higher all cause mortality. Although there are some considerations such as amount of physical activity there are other factors such as diet (i.e. snacking on unhealthy foodstuffs while sitting, replacing exercise with sitting, etc.), genetic predisposition, etc. which may provide an alternative explanation.	There is inconclusive evidence regarding whether sitting is associated with higher all cause mortality. Although there are some considerations such as amount of physical activity there are other factors such as diet (i.e. snacking on unhealthy foodstuffs while sitting, replacing exercise with sitting, etc.), genetic predisposition, etc. which may provide an alternative explanation.

b) (2 marks) For whichever of the above criteria you feel has the least supporting evidence, suggest how a medical study could be designed to provide evidence for this criterion. Provide a few sentences of detail on how the design and/or analysis of your proposed study would address the criteria you selected

    a. Because the Cessation of Exposure has the most polarity between two very similar studies it would be the wise to ascertain more evidence for this criterion. This can be accomplished with a Randomized Control Trial where the treatment group would be offered a standing desk at work and the control group will be given a more ergonomic chair (placebo) to use at work. The study population would be eligible adults (adults that work at a desk) in Canada. These participants would be followed over multiple years with a questionnaire that would self assess their health, hours sitting on an average day, etc. Since this would encourage the treatment group to stand more/sit less while the control group would be sitting the same amount it can be used to determine whether the strength of association is meaningful or not.

Question 2 (20 marks)

Import the datafile 'phbirths.txt' into SAS. This file can be found in Learn under the "Assignment 2" folder. The dataset is based on a 5% sample of all births that occurred in Philadelphia in 1990. The sample consists of 1115 observations on five variables: black = Mother is black (True vs. False), educ= Mother's years of education, smoke = Whether mother smoked during pregnancy (True vs. False), gestate = Gestational age in weeks, and grams= Baby's birth weight in grams. Interest lies in investigating whether the baby's birth weight in grams is associated with whether the mother smoked during pregnancy or not. (3 marks) Produce univariate statistics using Proc Univariate, examining the possible association between the mother's smoking status and the baby's birth weight in grams. Comment on any differences on see.

a. We would expect baby birth weights to be normally distributed due to the large sample size of 1115 which would mean that the mean, median and mode being roughly equal, the kurtosis to be 3 and the skewness to be 0. But the output given from PROC UNIVARIATE reveals that the mean is ≈ 3220, the median is ≈ 3267 and the mode is ≈ 3494 and since the mode is larger than the mean we can already see that the distribution will be negatively (left) skewed. This is confirmed by the skewness value of ≈ -1.12. The kurtosis value on the other hand of ≈ 3.13 is very close to the expected value.

b) (3 marks) Produce boxplots of the baby's weight in grams by the mother's smoking status. Do the plots suggest any association between the baby's weight and the mother's smoking status? Explain.
   a. The boxplot for weights of newborns whose mothers did not smoke have roughly the same maximum of roughly 4800 and the roughly the same minimum of around 300 compared to the boxplot for weights of newborns whose mothers did smoke. The largest difference comes from IQR, mean and mode of weight for non-smoking mothers being higher for each variable mentioned than smoking mothers. This could indicate that smoking can lead to an average decrease of birth weight of newborns.

c) (3 marks) Fit the simple linear regression of the form grams ~ smoke, i.e. the response outcome is baby's birth weight in grams and the explanatory variable of interest is the mum's smoking status. Provide an interpretation of your slope estimate.
   a. The intercept estimate of $\hat{\beta}_1 \approx -337$ meaning that for an increase in smoking (i.e. if the mother smoked) the expected weight of the baby is decreased by 337 grams.

d) (7 marks) The investigators suspect that gestational age in weeks is a confounder for the association between baby's birth weight and the mother's smoking status. Using the formal definition of confounding determine whether their suspicion holds. Note for this question you will need to fit two models as described on pages 46 and 47 of the course notes.
   a. First, we fit a linear regression model with response outcome baby's birth weight in grams against explanatory variable gestational age in weeks. Using the formal definition of confounding we need to check Is Gestational Age a risk factor for Birth Weight. We notice that the intercept is negative which may seem odd but since we are extrapolating data of babies born in weeks as low as 0, we can safely continue with the investigation. The variable θ_1 (gestational age in weeks) increases the baby's weight by 166 grams per increase in x_2. The p-value calculated was < 0.001 which indicates statistical significance.

b. Next, we check Is Gestational Age associated with Smoking Status but not a direct result of it. Fitting the response outcome mothers smoking status against the same explanatory variable we are testing for confounding which is gestation age in weeks. γ_1 (gestational age in weeks) decreases the baby's weight by 0.023 per x_2 (I.e. weeks of gestation). The association with Gestational Age and Smoking status of the mother is statistically significant since the p-value is < 0.001.
c. Therefore, we conclude that the Gestational Age is a confounder for the association between Birth Weights and Smoking status of the mother.

e) (4 marks) The investigators suspect that the mother's ethnicity is an effect modifier. Fit an appropriate multiple linear regression model to investigate whether their suspicion holds.
   a. In the equation $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$, The interaction term $\hat{\beta}_3 x_1 x_2$ was estimated to be 110.56 meaning that when Black is equal to 0 (birth given by non-black mother) we are left with $y = \hat{\beta}_0 + \hat{\beta}_1 x_1$. When x_2 = 1 (a baby is birthed from a black mother) $y = (\hat{\beta}_0 + \hat{\beta}_1) + x_2(\hat{\beta}_2 + \hat{\beta}_3)$. $\hat{\beta}_3 = 110.56$ is the added change of the average response as SmokeB increases by 1 when we go from BlackB=0 to BlackB=1. The p-value is > 0.05 ($p = 0.2015$) which is not enough to conclude the investigators suspicion.

Question 3 (10 marks)

(3 marks) Refer to Table 1 in the paper. Focusing on the variable "Animal in the House", create a 2x2 table and calculate the crude Odds Ratio (OR) of Asthma for homes that have animals present versus absent, its 95% confidence interval, and give a one sentence interpretation of your estimate in the context of the data.

a. The Odds ratio for this data is ≈ 0.613 with a confidence interval of [0.3636,1.0339] meaning that asthma is 0.613 times as likely to affect preschool children who have an animal present in their home compared to children who have an absence of animal in their house. Since the confidence interval contains 1 the OR is not statistically significant.

$$OR = \frac{32 \cdot 104}{118 \cdot 46} \approx 0.613, \ln(OR) \approx -0.4892,$$

$$var(\ln(OR)) = \frac{1}{118} + \frac{1}{46} + \frac{1}{32} + \frac{1}{104} \approx 0.0711$$

$$\left[e^{\ln(OR)-1.96\sqrt{var(\ln(OR))}}, e^{\ln(OR)+1.96\sqrt{var(\ln(OR))}}\right] = [0.3636, 1.0339]$$

b) (4 marks) Consider the potential confounding effect of location of residence (Urban vs. Rural). The table below gives the hypothetical distribution of the data in the paper. Calculate the two stratum-specific OR for the association between Asthma and "Animal in the House". Use the Mantel-Haenszel method to calculate the pooled OR. Discuss any differences you observed between the unadjusted and this adjusted/ pooled OR. What does this tell you about the potential confounder?

E+: Animal Present in Residence

D+: Case –Child with Asthma

	Urban		Rural		
	D+	D-	D+	D-	Row total
E+	30	31	10	7	78
E-	90	92	20	20	222
Col total	120	123	30	27	300

a. Urban stratum specific OR = 0.700
b. Rural stratum specific OR = 1.011
c. Pooled OR = 0.9375
d. Adjusting for location of residence (i.e. Pooled OR) shows a weaker association between animal presence and the likelihood of having asthma compared to the Rural specific OR ($OR_{POOLED} < OR_{RURAL}$) although the association increases compared to the Urban specific OR ($OR_{URBAN} < OR_{POOLED}$). The pooled OR result is still not statistically significant with a p-value of 0.3144 (>0.05). Thus, the potential confounder Location of Residence is not statistically significant / not a confounder.

c) (3 marks) Under the "Patients and Methods" section the paper describes that: "The sample size was calculated, considering the power of 80%, alpha of 5% and confidence level (CI) of 95%, while taking proportions equal to 0.57, and 0.73". Perform a sample size calculation in SAS using this information to show the sample size suggested for the investigators. Recall this study did 1 to 1 matching. How many additional children would be needed in the study if they wanted to achieve a power of 85?

a. Using group proportions with 0.57 and 0.73, a power of 80% and an alpha of 0.05 the investigators would need at least a sample size of 278 to reach reliable results. If instead the investigators would like to use a power of 85% keeping all other variables constant, they would need a sample size of at least 318. Meaning that they would need 40 more samples than recommended from the 80% power study.

Question 4 (10 marks)

Goldman et al (2008) conducted a matched case-control study investigating the association between Raynaud's syndrome (RS) and past treatment with central nervous system (CNS) stimulants1. All patients seen in a pediatric rheumatology practice during a 5-year period who had signs and symptoms of RS and met diagnostic criteria for RS were studied as cases. Controls were randomly selected patients at the same clinic who did not have signs or symptoms of RS. They were matched to the cases for age, sex, and year of initial evaluation. Sixty-four patients were enrolled in the study (32 cases with RS [23 female, 9 male] and 32 control patients). Charts of both cases and controls were reviewed to ascertain all medications taken at the time of the initial visit (including CNS stimulants). A summary of the data collected is given below.

History of CNS Stimulation Medication	Number of Pairs
Use by Case and Control	1
Use by Case and Not Control	4
Use by Not Case and Control	8
Use by Not Case and not Control	19

a) (3 marks) Construct the appropriate 2x2 table for this study using the information given above. Calculate the matched pair Odds Ratio (OR), its 95% confidence interval, and give a one sentence interpretation of your estimate.
    a. The odds of exposure in cases was found to be 0.5 compared to the odds of exposure in controls controlling for matched factor. The 95% confidence interval was calculated to be [0.1506, 1.6604] Since the confidence interval contains 1 the OR is not statistically significant.

b) (4 marks) Use McNemar's Test to test the significance of the association between Raynaud's syndrome (RS) and past treatment with central nervous system (CNS) stimulants. Be sure to clearly state the null, the alternative hypothesis, give the formula for the test statistic, calculate its value and find the p-value. What is the conclusion of the test?
    a. The null hypothesis H_0: no association between exposure and disease (i.e. $OR_m = 1$) while the alternative hypothesis H_A: $OR_m \neq 1$. We test this with the matched case-control test statistic $\frac{(f-g)^2}{(f+g)} \sim \chi^2_{(1)} \rightarrow \frac{(8-4)^2}{(8+4)} = \frac{16}{12} = \frac{4}{3} \sim \chi^2_{(1)}$. With a corresponding p-value = 0.2482. In conclusion the findings about the association between RS and CNS medication is not significant due to the p-value > 0.05.

c) (3 marks) Show that the formula for the matched pair Odds Ratio can be derived from the Mantel-Haenszel Odds Ratio where the data is stratified by pair. Hint: Consider the four types of possible matched pairs (and their corresponding unmatched 2x2 tables) and determine what each contributes to the Mantel-Haenszel Odds Ratio. Then sum over the number of each type of matched pair (i.e. e, f, g, h from a general matched 2x2 table).
    a. Recall the formula for the Mantel-Haenszel Pooled $\mathrm{OR}_{\mathrm{MH}} = \sum_{i=1}^{k} \frac{\frac{a_i d_i}{n_i}}{\frac{b_i c_i}{n_i}}$

	Cases	Controls	Total
Exposed	1	1	2
Unexposed	0	0	0
Total	1	1	2

$$\frac{a_i d_i}{n_i} = \frac{1 \cdot 0}{2} = 0, \frac{b_i c_i}{n_i} = \frac{1 \cdot 0}{2} = 0$$

	Cases	Controls	Total
Exposed	1	0	1
Unexposed	0	1	1
Total	1	1	2

$$\frac{a_i d_i}{n_i} = \frac{1 \cdot 1}{2} = \frac{1}{2}, \frac{b_i c_i}{n_i} = \frac{0 \cdot 0}{2} = 0$$

	Cases	Controls	Total
Exposed	0	1	1
Unexposed	1	0	1
Total	1	1	2

$$\frac{a_i d_i}{n_i} = \frac{0 \cdot 0}{2} = 0, \frac{b_i c_i}{n_i} = \frac{1 \cdot 1}{2} = \frac{1}{2}$$

	Cases	Controls	Total
Exposed	0	0	0
Unexposed	1	1	2
Total	1	1	2

$$\frac{a_i d_i}{n_i} = \frac{0 \cdot 1}{2} = 0, \frac{b_i c_i}{n_i} = \frac{0 \cdot 1}{2} = 0$$

Then summing these over the number of each type of matched pair (i.e. e,f,g,h) we obtain

$$OR_i = \frac{e \cdot 0 + f \cdot \frac{1}{2} + g \cdot 0 + h \cdot 0}{e \cdot 0 + f \cdot 0 + g \cdot \frac{1}{2} + h \cdot 0} = \frac{\frac{1}{2} f}{\frac{1}{2} g} = \frac{f}{g}$$

Question 5 (13 marks)

A (hypothetical) clinical trial tests whether scheduled exercise can prevent the development of diabetes among obese adults. Researchers establish a multi-site consortium that enrolls 3500 participants from clinics in 8 US cities. Inclusion criteria are a body mass index (BMI) > 30 kg/m2

and no previous history of diabetes. Exclusion criteria include any physical or medical condition that would preclude regular exercise or a previous history of heart failure. Characteristics of enrolled participants are 62% female, mean age of 44 years, and mean BMI of 34 kg/m2. Participants are randomized in a 1:1 ratio to receive either a scheduled exercise program or no such treatment. Participants in the exercise group receive a gym membership within close proximity of their residence and a suggested workout routine prescribing 150 min of aerobic activity per week. Study personnel contact participants in the exercise group every 2 months to encourage compliance with the program. Participants in the no treatment group receive educational materials describing the importance of exercise at the start of the study.

All participants complete annual study visits to assess the development of diabetes, which is defined by a fasting blood glucose level >126 mg/dL or the new use of a medication for diabetes. The researchers are prevented from knowing which treatment was administered throughout trial. Study results over a median of 5 years of follow-up are presented below.

Group	Number of participants	Number of diabetes cases	Diabetes incidence per 100 people	Diabetes Incidence per 100 person-years
Assigned to exercise	1750	53	3.0	6.5
Assigned to no treatment	1750	77	4.4	8.7

a) (2 marks) Calculate the relative risk of diabetes, comparing participants who were assigned to the exercise program with participants who were assigned to no treatment. Give a one sentence interpretation of your estimate. [Note: Use the diabetes incidences per 100 person-years for this calculation.]
    a. $RR = \frac{6.5}{8.7} = 0.747$ meaning that participants assigned to the exercise group were 0.747 times as likely to develop diabetes compared to participants who were assigned to the group without treatment.
b) (4 marks) For each of the following statements determine whether it is more likely to compromise the internal or external validity of the study. Give a one sentence justification of your choice for each.
    i. The possibility of non-adherence with the scheduled exercise program.
        a. This will compromise the study's internal validity since if the participants are not actively engaging in the exercise program, we would not be able to conclude whether exercise (or lack of) or some alternative reason is associated with the development of diabetes.
    ii. Frequent contacts from study personnel to encourage compliance with exercise.

a. This could compromise internal validity by causing some participants in the exercise group to behave in a way that was not expected from the initial study behaviour such as greatly increasing the workout routine which would mean the study conclusions do not reflect the true effect of the intervention.

iii. Exclusion of people with a history of heart failure.
   a. This could compromise external validity as excluding too many groups may have the opposite effect as intended and instead make the study conclusions not applicable to the general population.

iv. Inadvertent unblinding of research physicians by participants reporting their randomization status during annual study visits.
   a. this would compromise the internal validity of the study as physicians could unknowingly change results or introduce bias to results by behaving in a certain way towards people in different groups which would impact the true effect of the study.

c) (2 marks) Approximately how many people similar to those in the trial would need to be treated with the exercise program to prevent one instance of diabetes over a median of 5 years of follow-up? Note: Use the diabetes incidences per 100 people for this calculation.
   a. To number of subjects who need to be given the treatment to prevent a single occurrence is the Number Needed to Treat (NNT)

$$NNT = \frac{1}{Risk\ Difference} = \frac{1}{P[D+|E-] - P[D+|E+]} = \frac{1}{4.4 - 3.0} = 0.714$$

d) (5 marks) The investigators are concerned about the possibility of low adherence in the exercise group. A secondary analysis of the trial data reveals that only 1105 (63%) of participants assigned to the exercise program maintained compliance with this program during the trial. There were 28 diabetes cases among these compliers. Moreover, among the 1750 participants assigned to no treatment, 260 reported initiating a regular exercise program on their own during the trial period. There were 8 diabetes cases diagnosed among these noncompliers.

Based on the information given above, estimate relative risk of diabetes using both an intention to treat (ITT) and a per protocol (PP) analyses. Discuss reasons for differences between the two estimates. Note: Use the diabetes incidences per 100 people for this calculation.

Assigned to Exercise		Assigned to No Treatment	
Compliers	Non-compliers	Compliers	Non-compliers
1105	645	1490	260
28 developed diabetes		8 developed diabetes	

ITT: $RR_{ITT} = \frac{58/1750}{77/1750} = \frac{3.0}{4.4} = 0.682$

PP: $RR_{PP} = \frac{28/1105}{8/260} = 0.823529412$

Using a Per-Protocol analysis meaning only those who adhered to their assigned groups in a randomized control trial are analyzed introduces bias into the analysis because the 1:1 (or any n:k) randomization is broken. Meanwhile in an Intention to Treat analysis the data may be underestimated because of the inclusion of participants that did not take the treatment the analysis will guarantee unbiased data on the effect of the treatment. This inclusion of participants who did not take the treatment (exercise) would explain why $RR_{ITT} < RR_{PP}$.