MAS6061 Project 1

student number 170138286

Question 1

(i)

The code calculating and displaying the UK age specific organ donation rates per 1,000,000 for the whole period is shown below.

```r
knitr::opts_chunk$set(echo = TRUE)
Age_band <- c("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79")
UK_don <- c(12,78,83,90,165,206,90,14)
UK_pop <- c(7395033, 7524570, 7677822, 9524953, 7895663, 7262395, 5429785, 4293963)
Scot_don <- c(1,8,11,6,12,17,6,1)
Scot_pop <- c(610385, 647952, 662645, 818884, 707375, 620934, 493845, 369438)
cav.don.uk <- data.frame(Age_band,UK_don,UK_pop,Scot_don,Scot_pop)
# rates are per 1,000,000 unless otherwise states
cav.don.uk$UK_rates_1000000 <- round(1000000*cav.don.uk$UK_don/cav.don.uk$UK_pop, digits=2)
```

Age_band	UK_don	UK_pop	UK_rates_1000000
0-9	12	7395033	1.62
10-19	78	7524570	10.37
20-29	83	7677822	10.81
30-39	90	9524953	9.45
40-49	165	7895663	20.90
50-59	206	7262395	28.37
60-69	90	5429785	16.58
70-79	14	4293963	3.26

(ii)

The code to determine the Standardised donation (Incidence) Rate (SDR) for organ donation for Scotland is displayed below. The SDR is calculated as the sum of the expected cases in Scotland multiplied by 1,000,000 divided by the product of the Scottish population and the UK donor rate per 1,000,000, for each age specific group.

```r
# Scottish expected cases expressed per 1000000
cav.don.uk$Scot_exp <- round((cav.don.uk$UK_rates_1000000*cav.don.uk$Scot_pop)/1000000, digits=0)
# Standardised Donation (Incidence) Rate =
    #obs events in study pop / (study pop at risk * standard pop age rate per person)
SDR <- sum(cav.don.uk$Scot_don)/sum(cav.don.uk$Scot_exp)
```

The relevant data for SDR is displayed in the table below where UK_rates_1000000 represents the UK donar rate per 1,000,000 poputation, Scot_rates_1000000 represents the rate of Scottish donors per 1,000,000 and Scot_exp represents the Expected cases per 1,000,000 in the Scottish population.

Age_band	UK_don	UK_pop	Scot_don	Scot_pop	UK_rates_1000000	Scot_exp
0-9	12	7395033	1	610385	1.62	1

Age_band	UK_don	UK_pop	Scot_don	Scot_pop	UK_rates_1000000	Scot_exp
10-19	78	7524570	8	647952	10.37	7
20-29	83	7677822	11	662645	10.81	7
30-39	90	9524953	6	818884	9.45	8
40-49	165	7895663	12	707375	20.90	15
50-59	206	7262395	17	620934	28.37	18
60-69	90	5429785	6	493845	16.58	8
70-79	14	4293963	1	369438	3.26	1

The SDR is calculated as 0.95.

(iii)

95% C.I. for the SDR is calculated as

```
SE_logSIR <- sum(1/sqrt(cav.don.uk$Scot_don))
CISDR <- exp(1.96/SE_logSIR)
lowerCI <- SDR/CISDR
upperCI <- SDR*CISDR
```

where CISDR is one tail of the 95% SDR confidence interval.

The calculated 95% SDR confidence interval is 0.58 to 1.56.

(iv)

As the 95% confidence interval for the SDR contains the value 1, there is no evidence to suggest that organ donations appear particularly high or low compared to the UK.

Question 2

(i)

The study design is a prospective cohort study. Referencing "$BMI < 30$" as the unexposed risk group, it can be seen that according to the classic definition of confounding, Age is not a risk factor for Disease in the unexposed group.
Where $Age \leq 45, P(CVD|BMI < 30) = \frac{120}{1920} = \frac{1}{16}$.
Where $Age > 45, P(CVD|BMI < 30) = \frac{60}{960} = \frac{1}{16}$.

However, Age is associated with exposure ($BMI \geq 30$).

For $Age \leq 45, P(BMI \geq 30) = \frac{555}{2475} \approx 0.22$.
For $Age > 45, P(BMI \geq 30) = \frac{840}{1800} \approx 0.47$.

According to the classic definition of confounding, both the above conditions must be met. Age must be a risk factor for Disease and Age must be associated with exposure. As only one of these conditions is met, there is no evidence of confounding in the relationship between BMI and being diagnosed with CVD.

(ii)

Assuming that the probability of having CVD is independent of BMI groups, we can model the outcome of having the disease using binomial distributions.
For $BMI \geq 30$ group, the probability of having CVD can be modelled as $Y_1 \sim Bin(555, \frac{105}{555})$. Within the $BMI < 30$ group the probability of having CVD can be modelled as $Y_2 \sim Bin(1920, \frac{1}{16})$.
The risk difference for CVD by BMI exposure for the subgroup aged 45 and below at the study onest is

$$RiskDiff(Age \leq 45) = \frac{105}{555} - \frac{1}{16} = \frac{75}{592} \approx 0.127$$

For the group $Age > 45$, with $BMI \geq 30$ having CVD can be modelled as $X_1 \sim Bin(840, \frac{240}{600})$.
With $BMI < 30$ having CVD can be modelled as $X_2 \sim Bin(960, \frac{60}{960})$.

Therefore the risk difference for CVD by BMI exposure for the subgroup aged over 45 at the study onest is

$$RiskDiff(Age > 45) = \frac{2}{5} - \frac{1}{16} = \frac{27}{80} = 0.3375$$

The risk difference over the crude (collapsed) table is calculated as follows.
For the subgroup $BMI \geq 30$ the probability of having CVD can be modelled as $C_1 \sim Bin(1395, \frac{345}{1395})$.

For the subgroup $BMI < 30$ the probability of having CVD can be modelled as $C_2 \sim Bin(2880, \frac{180}{2880})$.

Therefore the crude risk difference for CVD by BMI exposure at the study onest is

$$RiskDiff = \frac{345}{1395} - \frac{180}{2880} = \frac{275}{1488} \approx 0.185$$

(iii)

The value of the risk difference for the crude table is 0.185, which is between the value of the risk difference for $Age \leq 45 (0.127)$ and the risk difference for $Age > 45$. According to the collapsibility defintion, as the crude risk difference is located between the risk differences of the subgroups, there is no evidence of confounding in the risk difference measure.

(iv)

Confounding was not seen in the risk difference measure using the classic and collapsibility definitions. However, the "Age" factor can be classified as an effect modifier, because no evidence of confounding was present, and the inclusion of "Age" in the model did change the estimate of the risk difference in disease between the two levels of Age.

Results should be reported as strata specific risk differences to demonstrate the effect modification of Age on the relationship between BMI and CVD in secondary school teachers.

Question 3

(i)

The contigency tables for Lung Ca and exposure to passive smoking are displayed below, stratified by study.

```
Ctab <- array(c(105,30, 277,125, 32,11, 160,36,11,13,16,9,17,5,78,54),
      dim=c(2,2,4),
      dimnames=list(c("Exposed", "Unexposed"), c("Lung Ca", "control"),
                    c("Study1", "Study2", "Study3", "Study4")))
Ctab
```

```
## , , Study1
##
##             Lung Ca control
## Exposed         105     277
## Unexposed        30     125
##
## , , Study2
##
##             Lung Ca control
## Exposed          32     160
## Unexposed        11      36
##
## , , Study3
##
##             Lung Ca control
## Exposed          11      16
## Unexposed        13       9
##
## , , Study4
##
##             Lung Ca control
## Exposed          17      78
## Unexposed         5      54
```

Odds ratios of lung cancer associated with passive smoking exposure for each study is as follows:

```
OR.Study <- function(x,y){
  #calculates OR for each study strata
 OR = (x[1,1,]*x[2,2,])/(x[1,2,]*x[2,1,])
 return(OR)
}
round(OR.Study(Ctab,1:4), digits = 2)
```

```
## Study1 Study2 Study3 Study4
##   1.58   0.65   0.48   2.35
```

The study which shows the highest odds of having lung cancer associated with passive smoking exposure is Study 4, with an $OR = 2.35$.

(b)

The Mantel-Haenzel estimate is given by the function below

```
ormh <- function(x,y){
nitotal = apply(x, 3, sum)
# components of contingency table for stratum y
ai = x[1,1,y]
bi = x[2,1,y]
ci = x[1,2,y]
```

```r
di = x[2,2,y]

numer = (ai*di)/nitotal[y]
denom = (bi*ci)/nitotal[y]
OR_MH = sum(numer)/sum(denom)

#sum of E(a_i) over strata
sum.E = sum(((ai+bi)*(ai+ci))/nitotal[y])

#sum Var(a_i) over strata
sum.Var = sum(((ai+bi)*(ci+di)*(ai+ci)*(bi+di))/(nitotal[y]^2*(nitotal[y]-1)))

#chi_sq test statistic
chi2 = (sum(ai) - sum.E)^2/sum.Var
p.value = 1-pchisq(chi2, 1)

#log(OR_MH)
log.OR = log(OR_MH)

#SE(ln(OR_MH))
SE.ln.OR = log(OR_MH)

output =
list(c("The MH estimate is" , round(OR_MH, digits = 2),
"The chi squared test statistic is" , round(chi2, digits = 2),
"The p-value of chi squared statistic is", round(p.value, digits = 2),
"The sqrt chi squared test statistic is" , round(sqrt(chi2), digits = 2),
"The ln(OR_MH) value is ", round(log.OR, digits = 2),
"The SE(ln(OR_MH) value is ", round(log.OR/sqrt(chi2), digits = 2),
"The approximate lower 95% CI for log(OR_MH) is", round(log.OR-(1.96*SE.ln.OR), digits = 2),
"The approximate upper 95% CI for log(OR_MH) is", round(log.OR+(1.96*SE.ln.OR), digits = 2),
"The approx. lower Confidence limit for OR_MH is the exponential",
round(exp(log.OR-(1.96*SE.ln.OR)), digits = 2),
"The approx. upper Confidence limit for OR_MH is the exponential",
round(exp(log.OR+(1.96*SE.ln.OR)), digits = 2)))

return(output)
}
ormh(Ctab, 1:4)
```

```
## [[1]]
##  [1] "The MH estimate is"
##  [2] "1.26"
##  [3] "The chi squared test statistic is"
##  [4] "1.71"
##  [5] "The p-value of chi squared statistic is"
##  [6] "0.19"
##  [7] "The sqrt chi squared test statistic is"
##  [8] "1.31"
##  [9] "The ln(OR_MH) value is "
## [10] "0.23"
## [11] "The SE(ln(OR_MH) value is "
## [12] "0.18"
## [13] "The approximate lower 95% CI for log(OR_MH) is"
## [14] "-0.22"
## [15] "The approximate upper 95% CI for log(OR_MH) is"
## [16] "0.68"
## [17] "The approx. lower Confidence limit for OR_MH is the exponential"
## [18] "0.8"
## [19] "The approx. upper Confidence limit for OR_MH is the exponential"
## [20] "1.97"
```

As can be seen from the above results, the $OR_{MH} = 1,26$ and the 95% confidence interval is 0.80 to 1.97.

(c)

From the results in Q3(b), the confidence interval contains the value 1. Therefore from these studies, there is no evidence to reject the null hypothesis that female lung cancer is associated with passive smoking. However, before accepting these results, we need to test whether the OR for all strata are homogeneous. This could be done using the Breslow-Day test, where the expected mean cell values > 5 in at least 80% of the cells. From the OR calculated in Q2(i) for the different studies, it can be seen that the OR are in different directions, which suggest that the non-significant result of the Mantel-Haenszel test may be misleading. Infact, there may be associations between passive smoking and female lung cancer, conditional on the study.

Other factors that need to be borne in mind when generalising the results from the MH analysis is the presence of small sample sizes in the strata of a table. The presence of small numbers in a strata can adversely affect results. In addition, the test is only useful if there are only a few confounding variables presence. Infact, it is often impossible to stratify by more than one variable at a time. Finally, the MH test of OR is only useful if the confounder is a categorical variable.

Question 4

(a)

The matched table for examining the relationship between physical activity in early pregnancy and the risk of pre-term birth in new mothers is shown below

```
matched <- matrix(c(193, 39, 63, 8), nrow = 2,
                  dimnames = list(c("standing_case", "non_standing_case"),
                              c("standing_Control", "non_standing_control")))
matched
```

```
##                     standing_Control non_standing_control
## standing_case                    193                   63
## non_standing_case                 39                    8
```

```
#continuity corrected McNemar's test for correlated data
OR_m = matched[1,2]/matched[2,1]

chi2_m = ((abs(matched[1,2]-matched[2,1]))-1)^2/(matched[1,2]+matched[2,1])
p_value_chi_m = round(1-pchisq(chi2_m, 1), digits = 3)
#Confidence interval
lower.CI.m = OR_m^(1-1.96/sqrt(chi2_m))
upper.CI.m = OR_m^(1+1.96/sqrt(chi2_m))
list(c("OR continuity corrected using McNemar's test is", round(OR_m, digits=2),
     "p-value of McNemars test is", round(p_value_chi_m, digits=3),
     "95% OR_lower is", round(lower.CI.m, digits = 2),
     "95% OR_upper is", round(upper.CI.m, digits = 2)))
```

```
## [[1]]
## [1] "OR continuity corrected using McNemar's test is"
## [2] "1.62"
## [3] "p-value of McNemars test is"
## [4] "0.023"
## [5] "95% OR_lower is"
## [6] "1.07"
## [7] "95% OR_upper is"
## [8] "2.44"
```

The chi-squared test has a p-value of 0.02 and the 95% confidence interval spans values above one. Hence there is moderate evidence to reject the null hypothesis that there is no relationship between physical activity in early pregnancy and the risk of pre-term birth in new mothers. Therefore it appears that a relationship between physical activity in early pregnancy and the risk of pre-term birth in new mothers exits.

(b)

If we ignored matching for Q3(i), then the data tables would appear as below

```r
non_matched <- matrix(c(256, 47, 232, 71), nrow = 2,
                  dimnames = list(c("standing", "non_standing"),
                                  c("Case", "Control")))

non_matched
```

```
##               Case Control
## standing      256     232
## non_standing   47      71
```

```r
OR_not_m = (non_matched[1,1]*non_matched[2,2])/(non_matched[1,2]*non_matched[2,1])
c("the OR for the non matched data is ",round(OR_not_m, digits = 2))
```

```
## [1] "the OR for the non matched data is "
## [2] "1.67"
```

```r
chisq.test(non_matched)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  non_matched
## X-squared = 5.5671, df = 1, p-value = 0.0183
```

Using Miettinen's test based confidence limits,

```r
OR_l = OR_not_m^(1-1.96/sqrt(5.5671))
OR_u = OR_not_m^(1+1.96/sqrt(5.5671))
list(c("The lower 95% C.L is", round(OR_l, digits = 2)),
     c("The upper 95% C.L is", round(OR_u, digits = 2)))
```

```
## [[1]]
## [1] "The lower 95% C.L is" "1.09"
##
## [[2]]
## [1] "The upper 95% C.L is" "2.55"
```

The OR for the non_matched data is slightly higher (1.67) compared to the matched data (1.62) and the confidence limits are slightly wider for the non-matched data (1.09-2.55) compared to the matched data

7-2.44). However, there appears little difference between these figures.

(c)

Adjusting the figures so that the concordent values for standing are 143 and the concordant values for non-standing are 58. The mathced pairs table now appears below

```
matched2 <- matrix(c(143, 39, 63, 58), nrow = 2,
                  dimnames = list(c("standing_case", "non_standing_case"),
                              c("standing_Control", "non_standing_control")))
matched2
```

```
##                     standing_Control non_standing_control
## standing_case                    143                   63
## non_standing_case                 39                   58
```

```
#continuity corrected McNemar's test for correlated data
OR_m2 = matched2[1,2]/matched2[2,1]

chi2_m2 = ((abs(matched2[1,2]-matched2[2,1]))-1)^2/(matched2[1,2]+matched2[2,1])
p_value_chi_m2 = round(1-pchisq(chi2_m2, 1), digits = 3)
#Confidence interval
lower.CI.m2 = OR_m2^(1-1.96/sqrt(chi2_m2))
upper.CI.m2 = OR_m2^(1+1.96/sqrt(chi2_m2))
list(c("OR continuity corrected using McNemar's test is", round(OR_m, digits=2),
     "p-value of McNemars test is", round(p_value_chi_m2, digits=3),
     "95% OR_lower is", round(lower.CI.m2, digits = 2),
     "95% OR_upper is", round(upper.CI.m2, digits = 2)))
```

```
## [[1]]
## [1] "OR continuity corrected using McNemar's test is"
## [2] "1.62"
## [3] "p-value of McNemars test is"
## [4] "0.023"
## [5] "95% OR_lower is"
## [6] "1.07"
## [7] "95% OR_upper is"
## [8] "2.44"
```

```
non_matched2 <- matrix(c(206, 97, 182, 121), nrow = 2,
                  dimnames = list(c("standing", "non_standing"),
                                 c("Case", "Control")))

non_matched2
```

```
##              Case Control
## standing      206     182
## non_standing   97     121
```

```
OR_not_m2 = (non_matched2[1,1]*non_matched2[2,2])/(non_matched2[1,2]*non_matched2[2,1])
c("the OR for the non matched data is ",round(OR_not_m2, digits = 2))
```

```
## [1] "the OR for the non matched data is "
## [2] "1.41"
```

```
chisq.test(non_matched2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  non_matched2
## X-squared = 3.79, df = 1, p-value = 0.05156
```

Using Miettinen's test based confidence limits,

```
OR_l2 = OR_not_m2^(1-1.96/sqrt(3.79))
OR_u2 = OR_not_m2^(1+1.96/sqrt(3.79))
list(c("The lower 95% C.L is", round(OR_l2, digits = 3)),
     c("The upper 95% C.L is", round(OR_u2, digits = 3)))
```

```
## [[1]]
## [1] "The lower 95% C.L is" "0.998"
##
## [[2]]
## [1] "The upper 95% C.L is" "1.998"
```

The change in concordant values does not affect the calculations and results of the matched pairs data. However, the unmatched results are no longer significant at the 95% confidence level. The confidence limits for the unmatched data now include 1 (0.998-1.998), and the OR for the unmatched data has reduced from 1.67 to 1.41. So it appears that when the two sets of concordant pairs approach each other (by keeping the overall number of subjects constant and only changing the values of concordant pairs), unmatched paired data test becomes less powerful in finding a difference between the relationship of physical activity in early pregnancy and the risk of pre-term birth in new mothers. In contrast, changing concordant pairs in a matched study has no influence on the results as concordant paired data is not used for this type of analysis.