#自变量：PFAS："log.PFOA", "log.PFOS", "log.PFHS", "log.PFNA"

#因变量：MCQ160N

#协变量：LBXSUA尿酸,RIAGENDR性别，RIDAGEYR年龄，BMI，RIDRETH1种族，DMDEDUC2教育程度，DMDMARTL婚姻状况，INDFMPIR年收入贫困线比数

#2. Materials and Methods

#2.1 Study population and design##########################################################################################################

#gout只有07-18年的数据，故以下全部统一用07-18年的数据

library(haven)

library(dplyr)

#(1)录入PFAS数据##############################################################################################

setwd("C:\\Users\\Administrator\\Desktop\\PFAS-Gout文章修改")

PFAS <- read.csv("PFAS\_07\_18.csv")

#(2)录入MCQ160数据###############################################################################

setwd("C:\\Users\\Administrator\\Desktop\\gout-尿酸-PFAS\\MCQ\_03\_18")

MCQ3<- read\_xpt("MCQ\_E.XPT") %>%

dplyr::select(SEQN,MCQ160N)

MCQ4<- read\_xpt("MCQ\_F.XPT") %>%

dplyr::select(SEQN,MCQ160N)

MCQ5<- read\_xpt("MCQ\_G.XPT") %>%

dplyr::select(SEQN,MCQ160N)

MCQ6<- read\_xpt("MCQ\_H.XPT") %>%

dplyr::select(SEQN,MCQ160N)

MCQ7<- read\_xpt("MCQ\_I.XPT") %>%

dplyr::select(SEQN,MCQ160N)

MCQ8<- read\_xpt("MCQ\_J.XPT") %>%

dplyr::select(SEQN,MCQ160N)

MCQ <- rbind(MCQ3,MCQ4,MCQ5,MCQ6,MCQ7,MCQ8)

#(3)录入协变量数据#############################################################################

#1\*尿酸数据导入##########################################################################33

setwd("C:\\Users\\Administrator\\Desktop\\gout-尿酸-PFAS\\LBXSUA")

UA3<- read\_xpt("BIOPRO\_E.xpt") %>%

dplyr::select(SEQN,LBXSUA)

UA4<- read\_xpt("BIOPRO\_F.xpt") %>%

dplyr::select(SEQN,LBXSUA)

UA5<- read\_xpt("BIOPRO\_G.xpt") %>%

dplyr::select(SEQN,LBXSUA)

UA6<- read\_xpt("BIOPRO\_H.xpt") %>%

dplyr::select(SEQN,LBXSUA)

UA7<- read\_xpt("BIOPRO\_I.xpt") %>%

dplyr::select(SEQN,LBXSUA)

UA8<- read\_xpt("BIOPRO\_J.xpt") %>%

dplyr::select(SEQN,LBXSUA)

UA <- rbind(UA3,UA4,UA5,UA6,UA7,UA8)

#2\*录入RIAGENDR性别，RIDAGEYR年龄，RIDRETH1种族，DMDEDUC2教育程度，DMDMARTL婚姻状况，INDFMPIR年收入贫困线比数数据

setwd("C:\\Users\\Administrator\\Desktop\\gout-尿酸-PFAS\\DEMO\_03\_18")

E2<- read\_xpt("DEMO\_E.XPT") %>%

dplyr::select(SEQN, RIAGENDR,RIDAGEYR,RIDRETH1,DMDEDUC2,DMDMARTL,INDFMPIR)

F2<- read\_xpt("DEMO\_F.XPT") %>%

dplyr::select(SEQN, RIAGENDR,RIDAGEYR,RIDRETH1,DMDEDUC2,DMDMARTL,INDFMPIR)

G2<- read\_xpt("DEMO\_G.XPT") %>%

dplyr::select(SEQN, RIAGENDR,RIDAGEYR,RIDRETH1,DMDEDUC2,DMDMARTL,INDFMPIR)

H2<- read\_xpt("DEMO\_H.XPT") %>%

dplyr::select(SEQN, RIAGENDR,RIDAGEYR,RIDRETH1,DMDEDUC2,DMDMARTL,INDFMPIR)

I2<- read\_xpt("DEMO\_I.XPT") %>%

dplyr::select(SEQN, RIAGENDR,RIDAGEYR,RIDRETH1,DMDEDUC2,DMDMARTL,INDFMPIR)

J2<- read\_xpt("DEMO\_J.XPT") %>%

dplyr::select(SEQN, RIAGENDR,RIDAGEYR,RIDRETH1,DMDEDUC2,DMDMARTL,INDFMPIR)

DEMO <- rbind(E2,F2,G2,H2,I2,J2)

#3\*录入BMI数据############################################################################################################

setwd("C:\\Users\\Administrator\\Desktop\\gout-尿酸-PFAS\\BMX\_03\_18")

BMX3<- read\_xpt("BMX\_E.xpt") %>%

dplyr::select(SEQN,BMXBMI)

BMX4<- read\_xpt("BMX\_F.xpt") %>%

dplyr::select(SEQN,BMXBMI)

BMX5<- read\_xpt("BMX\_G.xpt") %>%

dplyr::select(SEQN,BMXBMI)

BMX6<- read\_xpt("BMX\_H.xpt") %>%

dplyr::select(SEQN,BMXBMI)

BMX7<- read\_xpt("BMX\_I.xpt") %>%

dplyr::select(SEQN,BMXBMI)

BMX8<- read\_xpt("BMX\_J.xpt") %>%

dplyr::select(SEQN,BMXBMI)

BMI <- rbind(BMX3,BMX4,BMX5,BMX6,BMX7,BMX8)

#4.录入吸烟数据############################################################################################################

#SMQ020\_\_\_1:Y,2:F,7：拒绝,9：不知道，.:失踪

setwd("C:\\Users\\Administrator\\Desktop\\PFAS-Gout文章修改\\PFAS与骨质酥松关联第五阶段分析结果\\SMQ")

E1<- read\_xpt("SMQ\_E.XPT") %>%

dplyr::select(SEQN, SMQ020)

F1<- read\_xpt("SMQ\_F.XPT") %>%

dplyr::select(SEQN, SMQ020)

G1<- read\_xpt("SMQ\_G.XPT") %>%

dplyr::select(SEQN, SMQ020)

H1<- read\_xpt("SMQ\_H.XPT") %>%

dplyr::select(SEQN, SMQ020)

I1<- read\_xpt("SMQ\_I.XPT") %>%

dplyr::select(SEQN, SMQ020)

J1<- read\_xpt("SMQ\_J.XPT") %>%

dplyr::select(SEQN, SMQ020)

SMQ <- rbind(E1,F1,G1,H1,I1,J1)

#血清中的可替宁 （ng/mL）（大于0.05微克/升-为吸烟）#1:大于0.05#2：小于0.05#3：为空值

setwd("C:\\Users\\Administrator\\Desktop\\PFAS-Gout文章修改\\PFAS与骨质酥松关联第五阶段分析结果\\COTNAL")

COT3<- read\_xpt("COTNAL\_E.XPT") %>%

dplyr::select(SEQN, LBXCOT)

COT4<- read\_xpt("COTNAL\_F.XPT") %>%

dplyr::select(SEQN, LBXCOT)

COT5<- read\_xpt("COTNAL\_G.XPT") %>%

dplyr::select(SEQN,LBXCOT)

COT6 <- read\_xpt("COT\_H.XPT") %>%

dplyr::select(SEQN,LBXCOT)

COT7<- read\_xpt("COT\_I.XPT") %>%

dplyr::select(SEQN, LBXCOT )

COT8<- read\_xpt("COT\_J.XPT") %>%

dplyr::select(SEQN, LBXCOT )

COT <- rbind(COT3,COT4,COT5,COT6,COT7,COT8)

#5\*录入饮酒数据

#把喝酒次数（天数）大于等于3的定义为1：有喝过酒，其余为2：没喝过酒

setwd("C:\\Users\\Administrator\\Desktop\\PFAS-Gout文章修改\\PFAS与骨质酥松关联第五阶段分析结果\\ALQ")

ALQ\_E<- read\_xpt("ALQ\_E.XPT") %>%

dplyr::select(SEQN,ALQ120Q)

ALQ\_F<- read\_xpt("ALQ\_F.XPT") %>%

dplyr::select(SEQN,ALQ120Q)

ALQ\_G<- read\_xpt("ALQ\_G.XPT") %>%

dplyr::select(SEQN,ALQ120Q)

ALQ\_H<- read\_xpt("ALQ\_H.XPT") %>%

dplyr::select(SEQN,ALQ120Q)

ALQ\_I<- read\_xpt("ALQ\_I.XPT") %>%

dplyr::select(SEQN,ALQ120Q)

ALQ <- rbind(ALQ\_E,ALQ\_F,ALQ\_G,ALQ\_H,ALQ\_I)

ALQ <- ALQ %>%

mutate(ALQ120Q= case\_when(

ALQ120Q >= 3 & ALQ120Q < 776 ~ 1,

TRUE ~ 2

))

ALQ\_J <- read\_xpt("ALQ\_J.XPT") %>%

dplyr::select(SEQN,ALQ121)

ALQ\_J <- ALQ\_J %>%

mutate(ALQ121 = case\_when(

ALQ121 %in% c(10, 0) ~ 2,

ALQ121 %in% c(1, 2, 3, 4, 5, 6, 7, 8, 9) ~ 1,

TRUE ~ ALQ121

))

ALQ\_J <- rename(ALQ\_J, ALQ120Q=ALQ121)

ALQ <- rbind(ALQ,ALQ\_J)

#6\*导入运动数据

setwd("C:\\Users\\Administrator\\Desktop\\gout-尿酸-PFAS\\PAQ\_03\_18")

PAQ3<- read\_xpt("PAQ\_E.xpt") %>%

dplyr::select(SEQN,PAQ605,PAQ620,PAQ635)

PAQ4<- read\_xpt("PAQ\_F.xpt") %>%

dplyr::select(SEQN,PAQ605,PAQ620,PAQ635)

PAQ5<- read\_xpt("PAQ\_G.xpt") %>%

dplyr::select(SEQN,PAQ605,PAQ620,PAQ635)

PAQ6<- read\_xpt("PAQ\_H.xpt") %>%

dplyr::select(SEQN,PAQ605,PAQ620,PAQ635)

PAQ7<- read\_xpt("PAQ\_I.xpt") %>%

dplyr::select(SEQN,PAQ605,PAQ620,PAQ635)

PAQ8<- read\_xpt("PAQ\_J.xpt") %>%

dplyr::select(SEQN,PAQ605,PAQ620,PAQ635)

PAQ <- rbind(PAQ3,PAQ4,PAQ5,PAQ6,PAQ7,PAQ8)

PAQ <- PAQ %>%

mutate(PAQ180 = NA)

PAQ <- PAQ %>%

mutate(PAQ180 = case\_when(

PAQ605 == 1 ~ 4, # PAQ605为1的赋值为4

PAQ620 == 1 & is.na(PAQ180) ~ 2, # 除已赋值的行外，PAQ620为1的赋值为2

PAQ635 == 1 & is.na(PAQ180) ~ 3, # 除已赋值的行外，PAQ635为1的赋值为3

(PAQ605 == 2 & PAQ620 == 2 & PAQ635 == 2) & is.na(PAQ180) ~ 1,

# PAQ605，PAQ620和PAQ635列同时为2且未赋值的赋值为1

TRUE ~ PAQ180 # 保留之前赋的NA值

))

PAQ <- PAQ[, c("SEQN", "PAQ180")]

####2.合并数据#####################################################################################################

merge1 <- left\_join(PFAS,MCQ, by = "SEQN")

merge2 <- left\_join(merge1,UA, by = "SEQN")

merge3 <- left\_join(merge2,DEMO, by = "SEQN")

merge4 <- left\_join(merge3,BMI, by = "SEQN")

merge5 <- left\_join(merge4,SMQ, by = "SEQN")

merge6 <- left\_join(merge5,COT, by = "SEQN")

merge7 <- left\_join(merge6,ALQ, by = "SEQN")

merge <- left\_join(merge7,PAQ , by = "SEQN")

nrow(merge)#[1] 13160

write.csv(merge, "C:\\Users\\Administrator\\Desktop\\PFAS-Gout文章修改\\merge.csv")

####3.清洁数据######################################################################################

#(1)pfas:剔除log.PFOA,log.PFOS,log.PFHS，log.PFNA列为NA的行

cleaned\_merge <- merge %>%

filter(!is.na(log.PFOA) & !is.na(log.PFOS) & !is.na(log.PFHxS) & !is.na(log.PFNA))

# 计算指定列的均值和标准差

means <- sapply(cleaned\_merge[c("log.PFOA", "log.PFOS", "log.PFHxS", "log.PFNA")], mean, na.rm = TRUE)

sds <- sapply(cleaned\_merge[c("log.PFOA", "log.PFOS", "log.PFHxS", "log.PFNA")], sd, na.rm = TRUE)

# 创建一个逻辑向量，判断每行是否包含异常值

is\_outlier <- function(x, mean, sd) {

abs(x - mean) > 3 \* sd

}

outlier\_rows <- apply(cleaned\_merge[c("log.PFOA", "log.PFOS", "log.PFHxS", "log.PFNA")], 1, function(row) {

any(sapply(1:length(row), function(i) is\_outlier(row[i], means[i], sds[i])))

})

# 剔除包含异常值的行,以3倍标准差之外的值作为异常值的判断标准

cleaned\_merge <- cleaned\_merge[!outlier\_rows, ]

nrow(cleaned\_merge)#[1] 11846

#(2)痛风数据预处理

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(MCQ160N) )

cleaned\_merge <- cleaned\_merge %>%

dplyr::filter(MCQ160N != 9,MCQ160N != 7)

nrow(cleaned\_merge)#[1] 9774

#(3)UA

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(LBXSUA) )

nrow(cleaned\_merge)#[1] 9758

#(4)DEMO

#RIDAGEYR年龄

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(RIDAGEYR) )

nrow(cleaned\_merge)#[1] 9758

#RIAGENDR

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(RIAGENDR) )

nrow(cleaned\_merge)#[1] 9758

#RIDRETH1

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(RIDRETH1) )

nrow(cleaned\_merge)#[1] 9758

#DMDEDUC2

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(DMDEDUC2) )

cleaned\_merge <- cleaned\_merge %>%

dplyr::filter(DMDEDUC2!= 9,DMDEDUC2!= 7)

nrow(cleaned\_merge)#[1] 9741

#DMDMARTL

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(DMDMARTL) )

cleaned\_merge <- cleaned\_merge %>%

dplyr::filter(DMDMARTL!= 99,DMDMARTL!= 77)

nrow(cleaned\_merge)#[1]9737

#INDFMPIR

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(INDFMPIR) )

nrow(cleaned\_merge)#[1] 8766

#(5)BMI

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(BMXBMI) )

nrow(cleaned\_merge)#[1] 8677

#(6)SMQ

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(SMQ020) )

cleaned\_merge <- cleaned\_merge %>%

dplyr::filter(SMQ020!= 9,SMQ020!= 7)

nrow(cleaned\_merge)#[1] 8673

#(7)COT

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(LBXCOT) )

cleaned\_merge <- cleaned\_merge %>%

mutate(LBXCOT = case\_when(

LBXCOT >= 0.05 ~ 1,

LBXCOT < 0.05 ~ 2,

))#（大于0.05微克/升-为吸烟）#1:大于0.05#2：小于0.05

nrow(cleaned\_merge)#[1] 8671

#定义吸烟######################

cleaned\_merge <- cleaned\_merge %>%

mutate(SMOKE = case\_when(

SMQ020 == 1 | LBXCOT == 1 ~ 1,

TRUE ~ 2

))

#(8)ALQ

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(ALQ120Q) )

cleaned\_merge <- cleaned\_merge %>%

dplyr::filter(ALQ120Q != 999,ALQ120Q != 777,ALQ120Q != 77,ALQ120Q != 99)

nrow(cleaned\_merge)#[1] 8495

#(9)PAQ

cleaned\_merge <- cleaned\_merge %>%

filter(!is.na(PAQ180) )

nrow(cleaned\_merge)#[1] 8494

# 检查每一列是否存在空值

na\_check <- cleaned\_merge %>%

summarise(across(everything(), ~any(is.na(.))))

print(na\_check)

# 计算MCQ160N列中值为1和2的个数

sum(cleaned\_merge$MCQ160N == 1)#[1] 385-Yes

sum(cleaned\_merge$MCQ160N == 2)#[1] 8109-no

##3.1 Statistical description###########################################################################3

library(dplyr)

library(stats)

# 创建年龄分组变量

cleaned\_merge <- cleaned\_merge %>%

mutate(age\_group = case\_when(

RIDAGEYR >= 20 & RIDAGEYR < 40 ~ "20-39",

RIDAGEYR >= 40 & RIDAGEYR < 60 ~ "40-59",

RIDAGEYR >= 60 ~ "≥60"

))

# 创建 BMI 分组变量

cleaned\_merge <- cleaned\_merge %>%

mutate(bmi\_group = case\_when(

BMXBMI >= 25 & BMXBMI < 30 ~ "25-30.0",

BMXBMI < 25 ~ "<25",

BMXBMI >= 30 ~ "≥30.0"

))

# 创建教育程度分组变量

cleaned\_merge <- cleaned\_merge %>%

mutate(edu\_group = case\_when(

DMDEDUC2 < 3 ~ "<3",

DMDEDUC2 >= 3 ~ "≥3"

))

# 创建收入贫困比例分组变量

cleaned\_merge <- cleaned\_merge %>%

mutate(pir\_group = case\_when(

INDFMPIR <= 1.30 ~ "≤1.30",

INDFMPIR > 1.30 & INDFMPIR <= 3.50 ~ "1.31–3.50",

INDFMPIR > 3.50 ~ "> 3.50"

))

write.csv(cleaned\_merge, "C:\\Users\\Administrator\\Desktop\\PFAS-Gout文章修改\\cleaned\_merge.csv")

sum(cleaned\_merge$RIAGENDR == 1)#[1] 4112-Male

sum(cleaned\_merge$RIAGENDR == 2)#[1] 4382-Female

mean(cleaned\_merge$RIDAGEYR)#49.413

sd(cleaned\_merge$RIDAGEYR) / sqrt(length(cleaned\_merge$RIDAGEYR))#[1] 0.1912174

cleaned\_merge %>%

group\_by(MCQ160N) %>%

summarise(

Mean\_Age = sprintf("%.2f", mean(RIDAGEYR, na.rm = TRUE)),

SE\_Age = sprintf("%.2f", sd(RIDAGEYR, na.rm = TRUE) / sqrt(sum(!is.na(RIDAGEYR))))

)

sum(cleaned\_merge$age\_group == "20-39")

sum(cleaned\_merge$age\_group == "20-39")/nrow(cleaned\_merge)

sum(cleaned\_merge$age\_group == "40-59")

sum(cleaned\_merge$age\_group == "40-59")/nrow(cleaned\_merge)

sum(cleaned\_merge$age\_group == "≥60")

sum(cleaned\_merge$age\_group == "≥60")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N, age\_group) %>%

summarise(

Frequency = n(),

)

#检验cleaned\_merge数据框中的RIAGENDR列是否服从正态分布

ks.test(scale(cleaned\_merge$RIAGENDR), "pnorm")#D = 0.34953, p-value < 2.2e-16

#不符合正态分布用Wilcoxon秩和检验，符合正态分布用 Student’s t test ,

wilcox.test(cleaned\_merge %>%

filter(MCQ160N == 1) %>%

pull(RIAGENDR), cleaned\_merge %>%

filter(MCQ160N == 2) %>%

pull(RIAGENDR))#W = 1235582, p-value = 1.28e-15

#卡方检验用于评估各组分类变量的差异

chisq.test(table(cleaned\_merge$age\_group, cleaned\_merge$MCQ160N))#X-squared = 214.07, df = 2, p-value < 2.2e-16

mean(cleaned\_merge$BMXBMI)

sd(cleaned\_merge$BMXBMI) / sqrt(length(cleaned\_merge$BMXBMI))

cleaned\_merge %>%

group\_by(MCQ160N) %>%

summarise(

Mean\_Age = sprintf("%.2f", mean(BMXBMI, na.rm = TRUE)),

SE\_Age = sprintf("%.2f", sd(BMXBMI, na.rm = TRUE) / sqrt(sum(!is.na(BMXBMI))))

)

ks.test(scale(cleaned\_merge$BMXBMI), "pnorm")

wilcox.test(cleaned\_merge %>%

filter(MCQ160N == 1) %>%

pull(BMXBMI), cleaned\_merge %>%

filter(MCQ160N == 2) %>%

pull(BMXBMI))

sum(cleaned\_merge$bmi\_group == "<25")

sum(cleaned\_merge$bmi\_group == "<25")/nrow(cleaned\_merge)

sum(cleaned\_merge$bmi\_group == "25-30.0")

sum(cleaned\_merge$bmi\_group == "25-30.0")/nrow(cleaned\_merge)

sum(cleaned\_merge$bmi\_group == "≥30.0")

sum(cleaned\_merge$bmi\_group == "≥30.0")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N, pir\_group) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$pir\_group, cleaned\_merge$MCQ160N))

mean(cleaned\_merge$INDFMPIR)

sd(cleaned\_merge$INDFMPIR) / sqrt(length(cleaned\_merge$INDFMPIR))

cleaned\_merge %>%

group\_by(MCQ160N) %>%

summarise(

Mean\_Age = sprintf("%.2f", mean(INDFMPIR, na.rm = TRUE)),

SE\_Age = sprintf("%.2f", sd(INDFMPIR, na.rm = TRUE) / sqrt(sum(!is.na(INDFMPIR))))

)

ks.test(scale(cleaned\_merge$INDFMPIR), "pnorm")

wilcox.test(cleaned\_merge %>%

filter(MCQ160N == 1) %>%

pull(INDFMPIR), cleaned\_merge %>%

filter(MCQ160N == 2) %>%

pull(INDFMPIR))

sum(cleaned\_merge$pir\_group == "≤1.30")

sum(cleaned\_merge$pir\_group == "≤1.30")/nrow(cleaned\_merge)

sum(cleaned\_merge$pir\_group == "1.31–3.50")

sum(cleaned\_merge$pir\_group == "1.31–3.50")/nrow(cleaned\_merge)

sum(cleaned\_merge$pir\_group == "> 3.50")

sum(cleaned\_merge$pir\_group == "> 3.50")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N, pir\_group) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$pir\_group, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$RIAGENDR == "1")#Male

sum(cleaned\_merge$RIAGENDR == "1")/nrow(cleaned\_merge)

sum(cleaned\_merge$RIAGENDR == "2")#Female

sum(cleaned\_merge$RIAGENDR == "2")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N, RIAGENDR) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$RIAGENDR, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$RIDRETH1 == "1")

sum(cleaned\_merge$RIDRETH1 == "1")/nrow(cleaned\_merge)

sum(cleaned\_merge$RIDRETH1 == "2")

sum(cleaned\_merge$RIDRETH1 == "2")/nrow(cleaned\_merge)

sum(cleaned\_merge$RIDRETH1 == "3")

sum(cleaned\_merge$RIDRETH1 == "3")/nrow(cleaned\_merge)

sum(cleaned\_merge$RIDRETH1 == "4")

sum(cleaned\_merge$RIDRETH1 == "4")/nrow(cleaned\_merge)

sum(cleaned\_merge$RIDRETH1 == "5")

sum(cleaned\_merge$RIDRETH1 == "5")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N, RIDRETH1) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$RIDRETH1, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$DMDMARTL == "1")

sum(cleaned\_merge$DMDMARTL == "1")/nrow(cleaned\_merge)

sum(cleaned\_merge$DMDMARTL == "2")

sum(cleaned\_merge$DMDMARTL == "2")/nrow(cleaned\_merge)

sum(cleaned\_merge$DMDMARTL == "3")

sum(cleaned\_merge$DMDMARTL == "3")/nrow(cleaned\_merge)

sum(cleaned\_merge$DMDMARTL == "4")

sum(cleaned\_merge$DMDMARTL == "4")/nrow(cleaned\_merge)

sum(cleaned\_merge$DMDMARTL == "5")

sum(cleaned\_merge$DMDMARTL == "5")/nrow(cleaned\_merge)

sum(cleaned\_merge$DMDMARTL == "6")

sum(cleaned\_merge$DMDMARTL == "6")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N, DMDMARTL) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$DMDMARTL, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$SMQ020 == "1")#yes

sum(cleaned\_merge$SMQ020 == "1")/nrow(cleaned\_merge)

sum(cleaned\_merge$SMQ020 == "2")

sum(cleaned\_merge$SMQ020 == "2")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N,SMQ020) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$SMQ020, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$ALQ120Q== "1")#yes

sum(cleaned\_merge$ALQ120Q== "1")/nrow(cleaned\_merge)

sum(cleaned\_merge$ALQ120Q== "2")

sum(cleaned\_merge$ALQ120Q== "2")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N,ALQ120Q) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$ALQ120Q, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$PAQ180== "1")

sum(cleaned\_merge$PAQ180== "1")/nrow(cleaned\_merge)

sum(cleaned\_merge$PAQ180== "2")

sum(cleaned\_merge$PAQ180== "2")/nrow(cleaned\_merge)

sum(cleaned\_merge$PAQ180== "3")

sum(cleaned\_merge$PAQ180== "3")/nrow(cleaned\_merge)

sum(cleaned\_merge$PAQ180== "4")

sum(cleaned\_merge$PAQ180== "4")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N,PAQ180) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$PAQ180, cleaned\_merge$MCQ160N))

sum(cleaned\_merge$edu\_group== "<3")

sum(cleaned\_merge$edu\_group== "<3")/nrow(cleaned\_merge)

sum(cleaned\_merge$edu\_group== "≥3")

sum(cleaned\_merge$edu\_group== "≥3")/nrow(cleaned\_merge)

cleaned\_merge %>%

group\_by(MCQ160N,edu\_group) %>%

summarise(

Frequency = n(),

)

chisq.test(table(cleaned\_merge$edu\_group, cleaned\_merge$MCQ160N))

mean(cleaned\_merge$LBXSUA)

sd(cleaned\_merge$LBXSUA) / sqrt(length(cleaned\_merge$LBXSUA))

cleaned\_merge %>%

group\_by(MCQ160N) %>%

summarise(

Mean\_Age = sprintf("%.2f", mean(LBXSUA, na.rm = TRUE)),

SE\_Age = sprintf("%.2f", sd(LBXSUA, na.rm = TRUE) / sqrt(sum(!is.na(BMXBMI))))

)

ks.test(scale(cleaned\_merge$LBXSUA), "pnorm")

wilcox.test(cleaned\_merge %>%

filter(MCQ160N == 1) %>%

pull(LBXSUA), cleaned\_merge %>%

filter(MCQ160N == 2) %>%

pull(LBXSUA))

#参与者血清中全氟和多氟烷基物质（PFASs）的分布

library(dplyr)

library(broom)

exp(mean(cleaned\_merge$log.PFOA, na.rm = TRUE)) # 几何平均数

exp(mean(cleaned\_merge$log.PFOS, na.rm = TRUE))

exp(mean(cleaned\_merge$log.PFHxS, na.rm = TRUE))

exp(mean(cleaned\_merge$log.PFNA, na.rm = TRUE))

mean(cleaned\_merge$log.PFOA, na.rm = TRUE) # 算术平均数

mean(cleaned\_merge$log.PFOS, na.rm = TRUE)

mean(cleaned\_merge$log.PFHxS, na.rm = TRUE)

mean(cleaned\_merge$log.PFNA, na.rm = TRUE)

sd(cleaned\_merge$log.PFOA, na.rm = TRUE) / sqrt(sum(!is.na(cleaned\_merge$log.PFOA))) # 标准误差

sd(cleaned\_merge$log.PFOS, na.rm = TRUE) / sqrt(sum(!is.na(cleaned\_merge$log.PFOS)))

sd(cleaned\_merge$log.PFHxS, na.rm = TRUE) / sqrt(sum(!is.na(cleaned\_merge$log.PFHxS)))

sd(cleaned\_merge$log.PFNA, na.rm = TRUE) / sqrt(sum(!is.na(cleaned\_merge$log.PFNA)))

quantile(cleaned\_merge$log.PFOA, probs = c(0.25, 0.5, 0.75, 0.95), na.rm = TRUE) # 百分位数

quantile(cleaned\_merge$log.PFOS, probs = c(0.25, 0.5, 0.75, 0.95), na.rm = TRUE) # 百分位数

quantile(cleaned\_merge$log.PFHxS, probs = c(0.25, 0.5, 0.75, 0.95), na.rm = TRUE) # 百分位数

quantile(cleaned\_merge$log.PFNA, probs = c(0.25, 0.5, 0.75, 0.95), na.rm = TRUE) # 百分位数

####5.相关性分析############################################################################3

library(ggplot2)

library(GGally)

ggpairs(cleaned\_merge, columns = c("log.PFOA", "log.PFOS", "log.PFHxS", "log.PFNA"),

lower = list(continuous = "cor"),

upper = list(continuous = "cor"))

#3.2. Individual PFAS analysis#######################################

##6.individual PFAS association analysis via three logistic regressions####################################

library(gWQS)

library(ggplot2)

library(knitr)

library(kableExtra)

library(reshape2)

library(broom)

library(MASS)

cleaned\_merge<-cleaned\_merge%>%

mutate(MCQ160N= abs(MCQ160N-2))

calculate\_stats <- function(formula, data) {

model <- glm(formula, data = data, family = "binomial")

summary\_model <- summary(model)

or\_values <- exp(coef(model))

p\_values <- summary\_model$coefficients[, 4]

aic\_value <- AIC(model)

conf\_intervals <- exp(confint(model))

cat("Model:", deparse(formula), "\n")

cat("AIC:", round(aic\_value, 3), "\n")

cat("Coefficients:\n")

print(round(coef(summary\_model), 3))

cat("Odds Ratios:\n")

print(round(or\_values, 3))

cat("95% Confidence Intervals:\n")

print(round(conf\_intervals, 3))

cat("\n")

}

#Model1: 不控制协变量

calculate\_stats(MCQ160N ~ log.PFOA, cleaned\_merge)

calculate\_stats(MCQ160N ~ log.PFOS, cleaned\_merge) # PFOS

calculate\_stats(MCQ160N ~ log.PFHxS, cleaned\_merge) # PFHS

calculate\_stats(MCQ160N ~ log.PFNA, cleaned\_merge) # PFNA

#Model2: 控制年龄和性别

cleaned\_merge$RIAGENDR <- as.factor(cleaned\_merge$RIAGENDR)

calculate\_stats(MCQ160N~ log.PFOA + RIDAGEYR + RIAGENDR, data =cleaned\_merge)

calculate\_stats(MCQ160N~ log.PFOS + RIDAGEYR + RIAGENDR, data =cleaned\_merge)

calculate\_stats(MCQ160N~ log.PFHxS + RIDAGEYR + RIAGENDR, data =cleaned\_merge)

calculate\_stats(MCQ160N~ log.PFNA + RIDAGEYR + RIAGENDR, data =cleaned\_merge)

#Model3: 控制上述全部协变量

calculate\_stats(MCQ160N ~ log.PFOA + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge)

calculate\_stats(MCQ160N ~ log.PFOS + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge)

calculate\_stats(MCQ160N ~ log.PFHxS + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge)

calculate\_stats(MCQ160N ~ log.PFNA + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge)

## 6.绘制线性回归模型的 RCS 曲线#############################################################33

library(plotRCS)

rcsplot(data = cleaned\_merge,

outcome = "MCQ160N",

exposure = "log.PFOA",

covariates = c("LBXSUA","RIDAGEYR","RIAGENDR","BMXBMI","RIDRETH1","DMDEDUC2","DMDMARTL","INDFMPIR"))

rcsplot(data = cleaned\_merge,

outcome = "MCQ160N",

exposure = "log.PFOS",

covariates = c("LBXSUA","RIDAGEYR","RIAGENDR","BMXBMI","RIDRETH1","DMDEDUC2","DMDMARTL","INDFMPIR"))

rcsplot(data = cleaned\_merge,

outcome = "MCQ160N",

exposure = "log.PFHxS",

covariates = c("LBXSUA","RIDAGEYR","RIAGENDR","BMXBMI","RIDRETH1","DMDEDUC2","DMDMARTL","INDFMPIR"))

rcsplot(data = cleaned\_merge,

outcome = "MCQ160N",

exposure = "log.PFNA",

covariates = c("LBXSUA","RIDAGEYR","RIAGENDR","BMXBMI","RIDRETH1","DMDEDUC2","DMDMARTL","INDFMPIR"))

#3.3 Mixture PFAS analysis############################################33

##7.Mixture PFAS analysis########################################3

library(gWQS)

library(ggplot2)

library(knitr)

library(kableExtra)

library(reshape2)

CEP <-c("log.PFOA","log.PFOS","log.PFHxS","log.PFNA")

cleaned\_merge$ALQ120Q<-factor(cleaned\_merge$ALQ120Q)

cleaned\_merge$SMQ020<-factor(cleaned\_merge$SMQ020)

cleaned\_merge$PAQ180<-factor(cleaned\_merge$PAQ180)

results\_model1 <- gwqs(MCQ160N ~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =cleaned\_merge,q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial")

results\_model2 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180, mix\_name = CEP, data =cleaned\_merge, q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial",

covariables = ~ RIDAGEYR + RIAGEND)

results\_model3 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180, mix\_name = CEP, data =cleaned\_merge, q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

summary(results\_model1)

summary(results\_model2)

summary(results\_model3)

gwqs\_barplot(results\_model1)

gwqs\_barplot(results\_model2)

gwqs\_barplot(results\_model3)

#3.4 multiple linear regression analysis #######################################################################################

# 加载必要的库

library(dplyr)

library(broom)

# 定义数据框 cleaned\_merge

# 假设 cleaned\_merge 已经包含所需的变量

# 定义模型

model1 <- glm(MCQ160N ~ log.PFOA + log.PFOS + log.PFHxS + log.PFNA, family = "binomial", data = cleaned\_merge)

model2 <- glm(MCQ160N ~ log.PFOA + log.PFOS + log.PFHxS + log.PFNA + RIAGENDR + RIDAGEYR, data = cleaned\_merge, family = "binomial")

model3 <- glm(MCQ160N ~ log.PFOA + log.PFOS + log.PFHxS + log.PFNA + LBXSUA + RIAGENDR + RIDAGEYR + BMXBMI + RIDRETH1 + DMDEDUC2 + DMDMARTL + INDFMPIR, data = cleaned\_merge, family = "binomial")

# 定义计算统计量的函数

calculate\_stats <- function(model) {

summary\_model <- summary(model)

coefficients <- round(summary\_model$coefficients[, 1:4], 3)

or\_values <- round(exp(coefficients[, "Estimate"]), 3)

aic\_value <- round(AIC(model), 3)

ci\_values <- confint(model, level = 0.95) # 计算 95% 置信区间

ci\_values <- round(exp(ci\_values), 3) # 转换为 OR 的置信区间

list(

Estimates = coefficients[, "Estimate"],

StdErrors = coefficients[, "Std. Error"],

ZValues = coefficients[, "z value"],

PValues = coefficients[, "Pr(>|z|)"],

OddsRatios = or\_values,

CI\_Lower = ci\_values[, 1],

CI\_Upper = ci\_values[, 2],

AIC = aic\_value

)

}

# 计算每个模型的统计量

results\_model1 <- calculate\_stats(model1)

results\_model2 <- calculate\_stats(model2)

results\_model3 <- calculate\_stats(model3)

# 打印结果

print("Model 1 Results:")

print(results\_model1)

print("\nModel 2 Results:")

print(results\_model2)

print("\nModel 3 Results:")

print(results\_model3)

#3.5 Subgroup analysis############################################################################################################

#3.5.1 Individual PFAS analysis####################################

library(MASS)

library(rms)

library(ggplot2)

library(gridExtra)

cleaned\_merge$ age\_group <- as.factor(cleaned\_merge$ age\_group)

dd <- datadist(cleaned\_merge)

options(datadist = "dd")

#PFOA

model <-lrm(MCQ160N~rcs(log.PFOA,4)+age\_group+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,

log.PFOA,

age\_group = levels(cleaned\_merge$ age\_group),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

summary(AGE)

#使用 gplot 绘制图形

library(ggplot2)

ggplot() +

geom\_line(data = AGE, aes(log.PFOA, yhat, color = age\_group),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFOA, ymin = lower, ymax = upper, fill = age\_group),

alpha = 0.2) +

scale\_color\_manual(values = c("red", "#9Ec4be","#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

scale\_fill\_manual(values = c("pink", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFOA level", y = "OR(95%CI)",color = "age\_group",fill = "age\_group" )

#PFOS

model <-lrm(MCQ160N~rcs(log.PFOS,4)+age\_group+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,

log.PFOS,

age\_group = levels(cleaned\_merge$ age\_group),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

summary(AGE)

ggplot() +

geom\_line(data = AGE, aes(log.PFOS, yhat, color = age\_group),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFOS, ymin = lower, ymax = upper, fill = age\_group),

alpha = 0.2) +

scale\_color\_manual(values = c("red", "#9Ec4be","#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

scale\_fill\_manual(values = c("pink", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFOS level", y = "OR(95%CI)",color = "age\_group",fill = "age\_group" )

#PFHS

model <-lrm(MCQ160N~rcs(log.PFHxS,4)+age\_group+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,

log.PFHxS,

age\_group = levels(cleaned\_merge$ age\_group),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

summary(AGE)

ggplot() +

geom\_line(data = AGE, aes(log.PFHxS, yhat, color = age\_group),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFHxS, ymin = lower, ymax = upper, fill = age\_group),

alpha = 0.2) +

scale\_color\_manual(values = c("red", "#9Ec4be","#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

scale\_fill\_manual(values = c("pink", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFHxS level", y = "OR(95%CI)",color = "age\_group",fill = "age\_group" )

#PFNA

model <-lrm(MCQ160N~rcs(log.PFNA,4)+age\_group+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,

log.PFNA,

age\_group = levels(cleaned\_merge$ age\_group),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

summary(AGE)

#使用 gplot 绘制图形

ggplot() +

geom\_line(data = AGE, aes(log.PFNA, yhat, color = age\_group),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFNA, ymin = lower, ymax = upper, fill = age\_group),

alpha = 0.2) +

scale\_color\_manual(values = c("red", "#9Ec4be","#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

scale\_fill\_manual(values = c("pink", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("20-39", "40-59","≥60")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFNA level", y = "OR(95%CI)",color = "age\_group",fill = "age\_group" )

#性别：

cleaned\_merge$ RIAGENDR <- as.factor(cleaned\_merge$ RIAGENDR)

# 使用lrm()函数拟合模型

dd <- datadist(cleaned\_merge)

options(datadist = "dd")

#PFOA

model <-lrm(MCQ160N~rcs(log.PFOA,4)+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,log.PFOA,

RIAGENDR = levels(cleaned\_merge$ RIAGENDR),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

#使用 gplot 绘制图形

library(ggplot2)

ggplot() +

geom\_line(data = AGE, aes(log.PFOA, yhat, color = RIAGENDR),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFOA, ymin = lower, ymax = upper, fill = RIAGENDR),

alpha = 0.2) +

scale\_color\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

scale\_fill\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFOA level", y = "OR(95%CI)",color = "Sex",fill = "Sex" )

#PFOS

model <-lrm(MCQ160N~rcs(log.PFOS,4)+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,log.PFOS,

RIAGENDR = levels(cleaned\_merge$ RIAGENDR),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

#使用 gplot 绘制图形

library(ggplot2)

ggplot() +

geom\_line(data = AGE, aes(log.PFOS, yhat, color = RIAGENDR),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFOS, ymin = lower, ymax = upper, fill = RIAGENDR),

alpha = 0.2) +

scale\_color\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

scale\_fill\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFOS level", y = "OR(95%CI)",color = "Sex",fill = "Sex" )

#PFHxS

model <-lrm(MCQ160N~rcs(log.PFHxS,4)+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,log.PFHxS,

RIAGENDR = levels(cleaned\_merge$ RIAGENDR),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

#使用 gplot 绘制图形

library(ggplot2)

ggplot() +

geom\_line(data = AGE, aes(log.PFHxS, yhat, color = RIAGENDR),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFHxS, ymin = lower, ymax = upper, fill = RIAGENDR),

alpha = 0.2) +

scale\_color\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

scale\_fill\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFHxS level", y = "OR(95%CI)",color = "Sex",fill = "Sex" )

#PFNA

model <-lrm(MCQ160N~rcs(log.PFNA,4)+LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR,data= cleaned\_merge)

summary(model)

AGE <- Predict(model,log.PFNA,

RIAGENDR = levels(cleaned\_merge$ RIAGENDR),

fun = exp,

type = "predictions",

ref.zero = TRUE,

conf.int = 0.95)

#使用 gplot 绘制图形

library(ggplot2)

ggplot() +

geom\_line(data = AGE, aes(log.PFNA, yhat, color = RIAGENDR),

linetype = "solid", size = 1, alpha = 0.9) +

geom\_ribbon(data = AGE,

aes(log.PFNA, ymin = lower, ymax = upper, fill = RIAGENDR),

alpha = 0.2) +

scale\_color\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

scale\_fill\_manual(values = c("#43978F", "#9Ec4be", "#abd0f1", "#dce9f4"),

labels = c("Male", "Female")) +

theme\_classic() +

geom\_hline(yintercept = 1, linetype = 2, size = 1) +

labs(title = "Risk", x = "log-transformed PFNA level", y = "OR(95%CI)",color = "Sex",fill = "Sex" )

#3.5.2 Mixture PFAS analysis#########################################################################################3

library(gWQS)

library(ggplot2)

library(knitr)

library(kableExtra)

library(reshape2)

library(dplyr)

CEP <- c("log.PFOA", "log.PFOS", "log.PFHxS", "log.PFNA")

cleaned\_merge$ALQ120Q <- factor(cleaned\_merge$ALQ120Q)

cleaned\_merge$SMQ020 <- factor(cleaned\_merge$SMQ020)

cleaned\_merge$PAQ180 <- factor(cleaned\_merge$PAQ180)

cleaned\_merge$age\_group

first\_age\_group\_data <- cleaned\_merge%>%

filter(age\_group == "20-39")

model3\_age1 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180,

mix\_name = CEP,

data = first\_age\_group\_data,

q = 4,

validation = 0.6,

b = 10,

b1\_pos = TRUE,

rh = 10,

family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

first\_age\_group\_data2 <- cleaned\_merge %>%

filter(age\_group == "40-59")

model3\_age2 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180,

mix\_name = CEP,

data = first\_age\_group\_data2,

q = 4,

validation = 0.6,

b = 10,

b1\_pos = TRUE,

rh = 10,

family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

first\_age\_group\_data3 <- cleaned\_merge %>%

filter(age\_group == "≥60")

model3\_age3 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180,

mix\_name = CEP,

data = first\_age\_group\_data3,

q = 4,

validation = 0.6,

b = 10,

b1\_pos = TRUE,

rh = 10,

family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

summary(model3\_age1)

summary(model3\_age2)

summary(model3\_age3)

#性别

data1 <- cleaned\_merge%>%

filter(RIAGENDR == 1)

model3\_SEX1 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180,

mix\_name = CEP,

data = data1,

q = 4,

validation = 0.6,

b = 10,

b1\_pos = TRUE,

rh = 10,

family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

data2 <- cleaned\_merge%>%

filter(RIAGENDR == 2)

model3\_SEX2 <- gwqs(MCQ160N ~ wqs + ALQ120Q + SMQ020 + PAQ180,

mix\_name = CEP,

data = data2,

q = 4,

validation = 0.6,

b = 10,

b1\_pos = TRUE,

rh = 10,

family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

summary(model3\_SEX1)

summary(model3\_SEX2)

gwqs\_barplot(model3\_age1)

gwqs\_barplot(model3\_age2)

gwqs\_barplot(model3\_age3)

gwqs\_barplot(model3\_SEX1)

gwqs\_barplot(model3\_SEX2)

#3.6 Sensitivity analysis##########################################################################################################3

pfas\_vars <- c("log.PFOA", "log.PFOS", "log.PFHxS", "log.PFNA") # 假设这些是PFAS含量的变量名

pfas\_99th\_percentiles <- apply(cleaned\_merge[, pfas\_vars], 2, function(x) quantile(x, 0.99))

merge\_clear\_filtered <- cleaned\_merge[apply(cleaned\_merge[, pfas\_vars], 1, function(row) all(row <= pfas\_99th\_percentiles)), ]

#Model1: 不控制协变量

#PFOA

model\_PFOA\_0I <- glm(MCQ160N ~ log.PFOA, data =merge\_clear\_filtered, family = "binomial")

#PFOS

model\_PFOS\_0I <-glm(MCQ160N ~ log.PFOS, data =merge\_clear\_filtered, family = "binomial")

#PFHS

model\_PFHS\_0I <-glm(MCQ160N ~ log.PFHxS, data =merge\_clear\_filtered, family = "binomial")

#PFNA

model\_PFNA\_0I <-glm(MCQ160N ~ log.PFNA, data =merge\_clear\_filtered, family = "binomial")

merge\_clear\_filtered$RIAGENDR <- as.factor(merge\_clear\_filtered$RIAGENDR)

#Model2: 协变量 + RIDAGEYR + RIAGENDR

#PFOA

model\_PFOA\_2I <- glm(MCQ160N~ log.PFOA + RIDAGEYR + RIAGENDR, data =merge\_clear\_filtered , family = "binomial")

#PFOS

model\_PFOS\_2I <-glm(MCQ160N~ log.PFOS + RIDAGEYR + RIAGENDR, data =merge\_clear\_filtered, family = "binomial")

#PFHS

model\_PFHS\_2I <-glm(MCQ160N~ log.PFHxS + RIDAGEYR + RIAGENDR, data =merge\_clear\_filtered, family = "binomial")

#PFNA

model\_PFNA\_2I <-glm(MCQ160N~ log.PFNA + RIDAGEYR + RIAGENDR, data =merge\_clear\_filtered, family = "binomial")

##Model3: 控制上述全部协变量

#PFOA

model\_PFOA\_7I <- glm(MCQ160N ~ log.PFOA + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =merge\_clear\_filtered, family = "binomial")

#PFOS

model\_PFOS\_7I <-glm(MCQ160N ~ log.PFOS + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =merge\_clear\_filtered, family = "binomial")

#PFHS

model\_PFHS\_7I <-glm(MCQ160N ~ log.PFHxS + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =merge\_clear\_filtered, family = "binomial")

#PFNA

model\_PFNA\_7I <-glm(MCQ160N ~ log.PFNA + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =merge\_clear\_filtered, family = "binomial")

summary(model\_PFOA\_0I)

summary(model\_PFOS\_0I)

summary(model\_PFHS\_0I)

summary(model\_PFNA\_0I)

summary(model\_PFOA\_2I)

summary(model\_PFOS\_2I)

summary(model\_PFHS\_2I)

summary(model\_PFNA\_2I)

summary(model\_PFOA\_7I)

summary(model\_PFOS\_7I)

summary(model\_PFHS\_7I)

summary(model\_PFNA\_7I)

model1\_0I<- gwqs(MCQ160N ~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =merge\_clear\_filtered,q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial")

model2\_2I<- gwqs(MCQ160N ~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =merge\_clear\_filtered, q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial",

covariables = ~ RIDAGEYR + RIAGEND)

model3\_7I<- gwqs(MCQ160N~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =merge\_clear\_filtered, q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

summary(model1\_0I)

summary(model2\_2I)

summary(model3\_7I)

gwqs\_barplot(model1\_0I)

gwqs\_barplot(model2\_2I)

gwqs\_barplot(model3\_7I)

#敏感性分析补充：

# 定义不同周期对应的最低可测浓度

min\_detect\_conc <- data.frame(

cycle = c("2007-2008", "2009-2010", "2011-2012", "2013-2014", "2015-2016", "2017-2018"),

log\_PFOA = c(0.1, 0.1, 0.1, 0.1, 0.1, 0.1),

log\_PFOS = c(0.2, 0.2, 0.2, 0.1, 0.1, 0.1),

log\_PFHxS = c(0.1, 0.1, 0.1, 0.1, 0.1, 0.1),

log\_PFNA = c(0.08, 0.08, 0.08, 0.1, 0.1, 0.1)

)

cleaned\_merge <- merge(cleaned\_merge, min\_detect\_conc, by = "cycle")

set.seed(123) # 设置随机种子以便结果可复现

cleaned\_merge$log\_PFOA\_new <- ifelse(cleaned\_merge$log.PFOA == cleaned\_merge$log\_PFOA,

runif(nrow(cleaned\_merge), min = 0, max = cleaned\_merge$log\_PFOA),

cleaned\_merge$log.PFOA)

cleaned\_merge$log\_PFOS\_new <- ifelse(cleaned\_merge$log.PFOS == cleaned\_merge$log\_PFOS,

runif(nrow(cleaned\_merge), min = 0, max = cleaned\_merge$log\_PFOS),

cleaned\_merge$log.PFOS)

cleaned\_merge$log\_PFHxS\_new <- ifelse(cleaned\_merge$log.PFHxS == cleaned\_merge$log\_PFHxS,

runif(nrow(cleaned\_merge), min = 0, max = cleaned\_merge$log\_PFHxS),

cleaned\_merge$log.PFHxS)

cleaned\_merge$log\_PFNA\_new <- ifelse(cleaned\_merge$log.PFNA == cleaned\_merge$log\_PFNA,

runif(nrow(cleaned\_merge), min = 0, max = cleaned\_merge$log\_PFNA),

cleaned\_merge$log.PFNA)

#Model1: 不控制协变量

#PFOA

a1 <- glm(MCQ160N ~ log\_PFOA\_new, data =cleaned\_merge, family = "binomial")

summary(a1)

#PFOS

a2<- glm(MCQ160N ~ log\_PFOS\_new, data =cleaned\_merge, family = "binomial")

summary(a2)

#PFHS

a3 <- glm(MCQ160N ~ log\_PFHxS\_new, data =cleaned\_merge, family = "binomial")

summary(a3)

#PFNA

a4 <- glm(MCQ160N ~ log\_PFNA\_new, data =cleaned\_merge, family = "binomial")

summary(a4)

cleaned\_merge$RIAGENDR <- as.factor(cleaned\_merge$RIAGENDR)

#Model2: 协变量 + RIDAGEYR + RIAGENDR

#PFOA

b1 <- glm(MCQ160N~ log\_PFOA\_new + RIDAGEYR + RIAGENDR, data =cleaned\_merge , family = "binomial")

summary(b1)

#PFOS

b2 <-glm(MCQ160N~ log\_PFOS\_new + RIDAGEYR + RIAGENDR, data =cleaned\_merge, family = "binomial")

summary(b2)

#PFHS

b3 <-glm(MCQ160N~ log\_PFHxS\_new + RIDAGEYR + RIAGENDR, data =cleaned\_merge, family = "binomial")

summary(b3)

#PFNA

b4 <-glm(MCQ160N~ log\_PFNA\_new + RIDAGEYR + RIAGENDR, data =cleaned\_merge, family = "binomial")

summary(b4)

##Model3: 控制上述全部协变量

#PFOA

C1 <- glm(MCQ160N ~ log\_PFOA\_new + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge, family = "binomial")

summary(C1)

#PFOS

C2 <-glm(MCQ160N ~ log\_PFOS\_new + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge, family = "binomial")

summary(C2)

#PFHS

C3 <-glm(MCQ160N ~ log\_PFHxS\_new + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge, family = "binomial")

summary(C3)

#PFNA

C4 <-glm(MCQ160N ~ log\_PFNA\_new + LBXSUA+RIDAGEYR + RIAGENDR+BMXBMI+RIDRETH1+DMDEDUC2+DMDMARTL+INDFMPIR, data =cleaned\_merge, family = "binomial")

summary(C4)

#WQS

CEP <- c("log\_PFOA\_new", "log\_PFOS\_new", "log\_PFHxS\_new", "log\_PFNA\_new")

WQS1<- gwqs(MCQ160N ~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =cleaned\_merge,q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial")

WQS2<- gwqs(MCQ160N ~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =cleaned\_merge, q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial",

covariables = ~ RIDAGEYR + RIAGEND)

WQS3<- gwqs(MCQ160N~ wqs+ALQ120Q+SMQ020+PAQ180, mix\_name = CEP, data =cleaned\_merge, q = 4,

validation = 0.6, b = 10, b1\_pos = TRUE, rh = 10, family = "binomial",

covariables = ~ LBXSUA+RIDAGEYR + RIAGENDR + RIDRETH1 + BMXBMI + DMDEDUC2 + DMDMARTL + INDFMPIR)

summary(WQS1)

summary(WQS2)

summary(WQS3)

gwqs\_barplot(WQS1)

gwqs\_barplot(WQS2)

gwqs\_barplot(WQS3)

#3.7 Analysis of intermediary effect###################################################

library(mediation)

#log.PFOA, , ,

model.m <- lm(LBXSUA ~ log.PFOA, cleaned\_merge)

model.y <- lm(MCQ160N ~ log.PFOA + LBXSUA, cleaned\_merge)

med <- mediate(model.m,model.y,treat="log.PFOA",mediator = "LBXSUA", boot = TRUE, sims = 5000)

summary(med)

#log.PFOS

model.m <- lm(LBXSUA ~ log.PFOS, cleaned\_merge)

model.y <- lm(MCQ160N ~ log.PFOS + LBXSUA, cleaned\_merge)

med <- mediate(model.m,model.y,treat="log.PFOS",mediator = "LBXSUA", boot = TRUE, sims = 5000)

summary(med)

#log.PFHS

model.m <- lm(LBXSUA ~ log.PFHS, cleaned\_merge)

model.y <- lm(MCQ160N ~ log.PFHS + LBXSUA, cleaned\_merge)

med <- mediate(model.m,model.y,treat="log.PFHS",mediator = "LBXSUA", boot = TRUE, sims = 5000)

summary(med)

#log.PFNA

model.m <- lm(LBXSUA ~ log.PFNA, cleaned\_merge)

model.y <- lm(MCQ160N ~ log.PFNA + LBXSUA, cleaned\_merge)

med <- mediate(model.m,model.y,treat="log.PFNA",mediator = "LBXSUA", boot = TRUE, sims = 5000)

summary(med)