STATS 770 Report

Melissa Bather
Department of Statistics, University of Auckland
Email: mbat755@aucklanduni.ac.nz

Abstract

Background

Alzheimer's Disease (AD) is the most common form of dementia, for which there is no cure.[1] While age is the most significant risk factor, it is unclear what the exact cause of AD is. It is thought to likely be a combination of age, environmental factors, and genes. The Alzheimer's Disease Neuroimaging Initiative (ADNI)[2] is a study that began in 2004 with the goal of improving early detection and prevention of pre-dementia Alzheimer's Disease. It is a multicentre longitudinal study that explores clinical, imaging, genetic, and biochemical biomarkers.

Objective

The primary objective of this analysis was to determine if the number of APOE4 gene alleles affects the chances of an elderly person having AD. The secondary objectives were: determine if the presence/absence of at least one APOE4 allele affects the chances of an elderly person having AD, determine if the number of APOE4 alleles is predictive of the various stages of cognitive decline that lead to AD, and determine if the presence/absence of at least one APOE4 allele is predictive of the various stages of cognitive decline.

Methods

A total of 1723 participants from three phases of ADNI were used in this analysis. Baseline measurements were used to construct logistic regression and ordinal logistic regression models to determine the odds of having AD and various stages of cognitive decline based on APOE4 gene allele frequency.

Results

The odds of having AD is greatly increased for every APOE4 gene allele a person has. The odds of having AD is also much higher for people with at least one APOE4 allele compared to people who have none. Having a greater number of APOE4 alleles increases the odds of an elderly person having AD.

Conclusion

There is a strong association between the presence and number of APOE4 gene alleles and Alzheimer's Disease.

Introduction

Background

The APOE gene provides instructions for making a protein called apolipoprotein E, which combines with lipids in the body and forms lipoproteins which transport cholesterol and other fats around the bloodstream. This is essential for preventing disorders of the blood vessels and heart.[3] The three major alleles of the APOE gene are e2, e3, and e4. e4, which is found in about 15-25% of the population, has been associated with late-onset AD[4]. No cures or prevention methods have been found[2].

ADNI is a multicentre longitudinal study with the overall goals of detecting the earliest stages of AD, and supporting advances in intervention, prevention, and treatment of AD.[2]

Methods

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

The dataset used in this analysis comprises the ADNI-1, ADNI-GO, and ADNI-2 phases from the wider ADNI study This includes a total of 417 control group participants who did not have any cognitive impairment at baseline (CN), 310 participants who had early mild cognitive impairment at baseline (EMCI), 562 participants who had late mild cognitive impairment at baseline, 106 participants who had serious memory concern at baseline (SMC), and 342 participants who had AD at baseline (AD). Participants had disease status and APOE4 gene allele levels recorded at baseline. Disease status in terms of dementia progression was recorded at each follow-up visit, as well as various other test, scan, and biomarker results.

Design

There have been four phases within ADNI: ADNI-1, ADNI-GO, ADNI-2, and ADNI-3.

ADNI-1 began in 2004 and ended in 2010. It included 400 participants who had been diagnosed with mild cognitive impairment (MCI), 200 participants who had early Alzheimer's Disease (AD), and 200 control participants who were at a comparable (elderly) age.

ADNI-GO (Grand Opportunities) then extended ADNI-1 until 2017 with the original cohort of participants in addition to 200 new participants who had been diagnosed with early mild cognitive impairment (EMCI). This allowed researchers to look at biomarkers at an earlier stage of AD.

In 2011, ADNI-2 extended ADNI-GO with an additional 150 elderly controls, 100 participants with EMCI, 150 participants with late mild cognitive impairment (LMCI), and 150 participants with mild AD.

ADNI-3 data was not included in the available dataset and so has been excluded from this analysis.

Particpants were recorded as either CN (control group), EMCI (early mild cognitive impairment), LMCI (late mild cognitive impairment), SMC (significant memory concern), or AD (Alzheimer's Disease) at baseline. This analysis treats these levels as ordinal in this order, as they represent various stages of disease progression/cognitive decline.

Statistical Analysis

Statistical analyses were completed using R programming language and RStudio 2022.07.1 software. In all analyses, a $p < 0.05$ was considered statistically significant. Given the very large number of missing follow-up observations, it was decided that modelling disease progression would not be reliable. Instead, participants were considered to have AD if they were diagnosed with AD at baseline. Having dementia in subsequent follow up sessions does not necessarily mean that the participant had AD, as there are other forms of dementia.

Logistic regression models were constructed to assess the effect of the number of APOE4 gene alleles on AD status at baseline, controlling for the potential effects of age, ethnicity, gender, and education. One model assessed the effect of each additional APOE4 allele on disease status, and another model assessed the effect of having at least one APOE4 allele on disease status. Variables were exponentiated and confidence intervals were calculated to determine the increase in odds ratio of developing dementia for every additional unit increase in the variable.

Ordinal logistic regression models were constructed to assess the effect of the number of APOE4 gene alleles on all baseline disease statuses (CN, EMCI, LMCI, SMC, and AD) using the same predictors as the logistic regression models.

Most participants had follow-up data for at least one observation following baseline, however very few participants had follow-up data for most or all time points. This is likely due to the old age of participants, as well as poor mental health. For this reason, baseline disease status was used as the outcome variable for this report.

Results

It is reasonable to believe that both age and sex can affect APOE4 gene expression[7], education level, and AD. Ethnicity can probably affect APOE4 gene expression and AD as well as education level. Education could also affect AD as people who 'use their brains' often as a result of high education levels could potentially have a lower risk of AD. There is some weak evidence for this[5]. I also think marital status could affect AD, as there is some weak evidence that being never-married or widowed can increase your risk of AD[6]. Lastly, there could be a site/centre effect, as some sites could be more or less likely to diagnose someone with MCI or dementia and different sites could also follow the testing procedure for APOE4 slightly differently. I think this is unlikely, but I have included it the causal diagram just in case.

Proposed causal diagram:

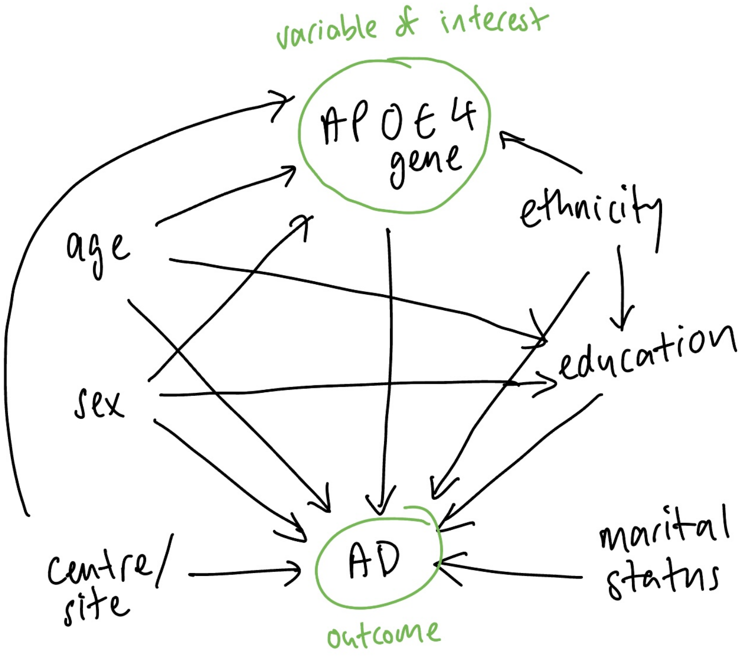


TABLE 1 Baseline Characteristics

	CN	EMCI	LMCI	SMC	AD	Overall
	(N=415)	(N=306)	(N=560)	(N=105)	(N=337)	(N=1723)
Gender						
Female	206 (49.6%)	136 (44.4%)	217 (38.8%)	62 (59.0%)	151 (44.8%)	772 (44.8%)
Male	209 (50.4%)	170 (55.6%)	343 (61.3%)	43 (41.0%)	186 (55.2%)	951 (55.2%)
Age						
Mean (SD)	74.8 (5.73)	71.2 (7.40)	74.0 (7.52)	72.2 (5.58)	75.0 (7.79)	73.8 (7.17)
Median [Min, Max]	74.1 [56.2, 89.6]	71.0 [55.0, 88.6]	74.4 [54.4, 91.4]	71.4 [59.7, 90.1]	75.6 [55.1, 90.9]	73.9 [54.4, 91.4]
Ethnicity						
Hisp/Latino	14 (3.4%)	14 (4.6%)	15 (2.7%)	2 (1.9%)	10 (3.0%)	55 (3.2%)
Not Hisp/Latino	399 (96.1%)	291 (95.1%)	542 (96.8%)	101 (96.2%)	324 (96.1%)	1657 (96.2%)
Unknown	2 (0.5%)	1 (0.3%)	3 (0.5%)	2 (1.9%)	3 (0.9%)	11 (0.6%)

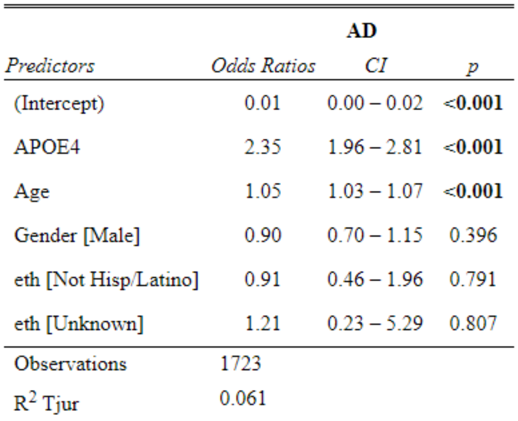
TABLE 2 Associations Between ADPOE4 Allele Frequency and Baseline Disease Status

	CN	EMCI	LMCI	SMC	AD	Overall
	(N=415)	(N=306)	(N=560)	(N=105)	(N=337)	(N=1723)
as.factor(APOE4)						
0	301 (72.5%)	175 (57.2%)	256 (45.7%)	70 (66.7%)	113 (33.5%)	915 (53.1%)
1	103 (24.8%)	110 (35.9%)	231 (41.3%)	34 (32.4%)	159 (47.2%)	637 (37.0%)
2	11 (2.7%)	21 (6.9%)	73 (13.0%)	1 (1.0%)	65 (19.3%)	171 (9.9%)

As this was an observational study, randomisation did not take place and confounders could be an issue. From the baseline table, the distribution of various covariates does not seem very even among different groups of disease status. Using logistic regression and adjusting for confounding variables based on the causal diagram, large p-values were calculated for site coefficients, and it was not strongly believed that site could cause confounding, therefore site was removed as a variable from the logistic regression models. All other covariates that were initially determined to need adjusting for remained in the logistic regression model due to the previously referenced existing body of evidence that suggests they are likely to be confounders.

As shown below, APOE4 allele frequency was shown to be a significant predictor for having AD at baseline. For every additional APOE4 allele a person has, the odds of having AD are multiplied by 2.35 (95% CI: 1.96-2.81).

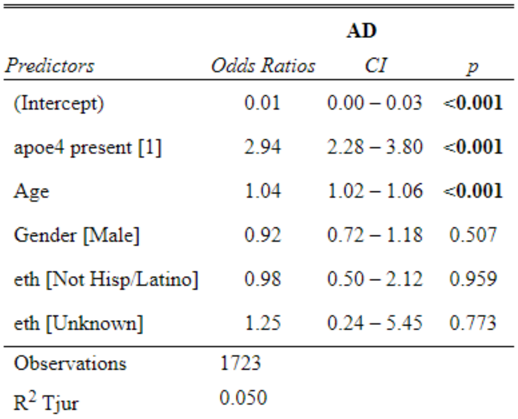
TABLE 3 Odds Ratios and P-Values From Logistic Regression Model for AD Based on Number of APOE4 Alleles



	AD		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.01	0.00 – 0.02	**<0.001**
APOE4	2.35	1.96 – 2.81	**<0.001**
Age	1.05	1.03 – 1.07	**<0.001**
Gender [Male]	0.90	0.70 – 1.15	0.396
eth [Not Hisp/Latino]	0.91	0.46 – 1.96	0.791
eth [Unknown]	1.21	0.23 – 5.29	0.807
Observations	1723		
R^2 Tjur	0.061		

Logistic regression was also used with the binary predictor of presence vs. absence of any APOE4 alleles. This predictor was also shown to be significant in predicting AD at baseline. The odds ratio for a person with at least one APOE4 allele having AD versus a person having no APOE4 alleles is 2.94 (95% CI: 2.28-3.8).

TABLE 4 Odds Ratios and P-Values From Logistic Regression Model for AD Based on Presence of any APOE4 Alleles



	AD		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.01	0.00 – 0.03	**<0.001**
apoe4 present [1]	2.94	2.28 – 3.80	**<0.001**
Age	1.04	1.02 – 1.06	**<0.001**
Gender [Male]	0.92	0.72 – 1.18	0.507
eth [Not Hisp/Latino]	0.98	0.50 – 2.12	0.959
eth [Unknown]	1.25	0.24 – 5.45	0.773
Observations	1723		
R^2 Tjur	0.050		

Ordinal logistic regression was used to assess how different numbers of APOE4 alleles could affect the probability of having all baseline disease statuses. Age was divided into five year age groups to make results more readable. As seen in the table below, the ordinal regression model shows that for each additional APOE4 allele a person has, the odds of progressing to a higher disease status are multiplied by 2.14 (95% CI 1.87-2.45) and this is statistically significant.

TABLE 5 Odds Ratios from Ordinal Logistic Regression Model for All Baseline Disease Statuses Based on Number of APOE4 Alleles

	OR	2.5 %	97.5 %
APOE4	2.1386092	1.8721041	2.445719
age_group55-59	0.6179891	0.0210426	18.154127
age_group60-64	0.4866087	0.0168239	14.079347
age_group65-69	0.4296868	0.0149143	12.383857
age_group70-74	0.3049462	0.0106091	8.768198
age_group75-79	0.4786679	0.0166507	13.766287
age_group80-84	0.5731320	0.0199029	16.510948
age_group85-89	0.9093097	0.0311819	26.531238
age_group90-95	5.9554427	0.1524247	246.574017
GenderMale	1.0720084	0.9006346	1.276069
ethNot Hisp/Latino	1.1366277	0.6992549	1.847929

	OR	2.5 %	97.5 %
ethUnknown	2.2258687	0.6692991	7.376226

Finally, ordinal logistic regression was used to assess how presence or absence of at least one APOE4 allele could affect the probability of having all baseline disease statuses. As seen in the table below, this ordinal regression model shows that the presence of at least one APOE4 allele means the odds of moving from one disease status to a higher disease status are multiplied by 2.58 (95% CI: 2.16-3.10) and this is statistically significant.

TABLE 6 Odds Ratios from Ordinal Logistic Regression Model for All Baseline Disease Statuses Based on Number of APOE4 Alleles

	OR	2.5 %	97.5 %
apoe4_present1	2.5802818	2.1643055	3.079457
age_group55-59	0.6407584	0.0218396	18.782052
age_group60-64	0.5029831	0.0174047	14.520906
age_group65-69	0.4380606	0.0152181	12.597498
age_group70-74	0.3036547	0.0105720	8.712589
age_group75-79	0.4680287	0.0162919	13.432626
age_group80-84	0.5534154	0.0192327	15.909732
age_group85-89	0.8796785	0.0301895	25.612308
age_group90-95	5.9707382	0.1528949	246.658608
GenderMale	1.0828059	0.9099323	1.288654
ethNot Hisp/Latino	1.1649305	0.7168773	1.895045
ethUnknown	2.2539994	0.6801696	7.442970

Discussion

While data for this analysis come from a longitudinal study, almost every participant in the study did not have observations recorded at every follow-up stage. This is expected given the old age and poor mental condition of many of the participants. This makes fitting a linear mixed-effects model quite unreliable. To model disease progression using the disease statuses at each visit (NL, NL to MCI, MCI, MCI to Dementia, Dementia) one would need to use multiple imputation with a categorical variable or Last Observation Carried Forward to fill in missing disease statuses. Given the huge number of missing observations, this does not sound sensible to me. Different phases of the ADNI study had different follow-up time points which also contributes to this, but even looking just at one of the study phases, e.g. ADNI-1, and looking only at the first three follow-up time points for example, there is still a lot of missing data, and in doing so the sample size is also greatly reduced. Additionally, not all ADNI phases recorded the various stages of cognition, e.g. EMCI or LMCI, that I was interested in analysing. Therefore, while it may have been interesting to follow disease progression, I did not make use of the follow-up data and instead stuck to baseline observations as in a case-control study.

Another issue with the study is that a lot of the baseline disease statuses are based on self report, which can be inaccurate or introduce recall bias. For example, patients considered to have significant memory concern (SMC) at baseline are included based on the following criteria from the ADNI website: *The key inclusion criteria that distinguish the SMC cohort are a self-report significant memory concern from the participant, quantified by using the Cognitive Change Index and the Clinical Dementia Rating (CDR) of zero. SMC participants score within the normal range for cognition, and the informant does not equate the expressed concern with progressive memory impairment.*[2] Moreover, “self report” suggests that there is a subjective nature to this “disease” category.

Overall, a more complete dataset would have been good for analysis, as Mixed Effects Models could have been constructed to track disease progression. However, given the nature of the disease and the fact that its relevant population is elderly, this may not be very feasible. Despite this, I think the models constructed in this report do a good job at demonstrating that APOE4 alleles are a clear risk factor developing Alzheimer’s Disease.

Conclusion

Having one or two APOE4 gene alleles greatly increases a person’s chance of having Alzheimer’s Disease or any of the stages of cognitive decline that lead up to it. To assess a patient’s risk of developing cognitive decline or AD at an old age, they should be screened for presence of APOE4 gene alleles. Absence of APOE4 alleles does not mean that a patient cannot have AD, but they are at lower risk than a patient with at least one APOE4 allele.

References

1. Alzheimer's disease. Health Navigator. Updated October, 2019. Accessed October, 2022. https://www.healthnavigator.org.nz/health-a-z/a/alzheimers-disease/
2. Alzheimer's Disease Neuroimaging Iniatiative. Accessed October, 2022. https://adni.loni.usc.edu/
3. APOE gene. Medline Plus, National Library of Medicine. Updated March, 2021. Accessed October, 2022. https://medlineplus.gov/genetics/gene/apoe/
4. Alzheimer's genes: Are you at risk? Mayo Clinic. Updated May, 2021. Accessed October, 2022. https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-genes/art-20046552
5. Sharp ES, Gatz M. Relationship between education and dementia: an updated systematic review. Alzheimer Dis Assoc Disord. 2011 Oct-Dec;25(4):289-304. doi: 10.1097/WAD.0b013e318211c83c.
6. Liu H, Zhang Z, Choi SW, Langa KM. Marital Status and Dementia: Evidence from the Health and Retirement Study. J Gerontol B Psychol Sci Soc Sci. 2020 Sep 14;75(8):1783-1795. doi: 10.1093/geronb/gbz087.
7. Hsu M, Dedhia M, Crusio WE, Delprato A. Sex differences in gene expression patterns associated with the APOE4 allele. F1000Res. 2019 Apr 5;8:387. doi: 10.12688/f1000research.18671.2.

Appendix

Code used in this report:

```
suppressPackageStartupMessages({
  library(dplyr)
  library(sjPlot)
  library(MASS)
})

#Read in and clean data - want one line for each participant
dat <- read.csv("phe.csv")

participants <- unique(dat$PTID)
visits <- unique(dat$VISCODE)
fields <- c("ptid", visits, "dx_bl", "Age", "Gender", "edu", "eth",
            "marry", "APOE4", "AD", "site", "origprot", "apoe4_present")
cleaned_data <- data.frame(matrix(ncol = length(fields), nrow = 0, dimnames = list(NULL, fields)))

#Number of participants in each baseline group
#Control
cn <- subset(dat, DX_bl == "CN")
length(unique(cn$PTID))

#EMCI
emci <- subset(dat, DX_bl == "EMCI")
length(unique(emci$PTID))

#LMCI
lmci <- subset(dat, DX_bl == "LMCI")
length(unique(lmci$PTID))

#SMC
smc <- subset(dat, DX_bl == "SMC")
length(unique(smc$PTID))

#AD
ad <- subset(dat, DX_bl == "AD")
length(unique(ad$PTID))


for(i in 1:length(participants)) {
```

```r
  #Look at one patient at a time
  patient_hist <- subset(dat, PTID == participants[i])
  #Fill in applicable visits
  all_visits <- rep(NA, length(visits))
  all_visits[which(visits %in% unique(patient_hist$VISCODE))] <- patient_hist$DX[1:nrow(patient_hist)]
  #Is AD present at all at any point?
  ad <- ifelse(patient_hist$DX_bl == "AD", 1, 0)
  #Is there at least one APOE4 allele?
  apoe4_present <- as.integer(as.numeric(patient_hist$APOE4) > 0)
  #Add everything to cleaned data frame
  cleaned_data[i,] <- c(patient_hist$PTID[1],
                        all_visits,
                        patient_hist$DX_bl[1],
                        patient_hist$AGE[1],
                        patient_hist$PTGENDER[1],
                        patient_hist$PTEDUCAT[1],
                        patient_hist$PTETHCAT[1], #unsure if eth or race should be used?
                        patient_hist$PTMARRY[1],
                        patient_hist$APOE4[1],
                        ad[1],
                        patient_hist$SITE[1],
                        patient_hist$ORIGPROT[1],
                        apoe4_present[1])
}

#Remove observations that have NA for APOE4 as they aren't helpful
cleaned_data <- cleaned_data[!is.na(cleaned_data$APOE4),]

#Change characters to numbers
cleaned_data$APOE4 <- as.numeric(cleaned_data$APOE4)
cleaned_data$Age <- as.numeric(cleaned_data$Age)
cleaned_data$AD <- as.numeric(cleaned_data$AD)

#Change baseline disease status to factors in desired order
cleaned_data$dx_bl <- factor(cleaned_data$dx_bl, levels = c("CN", "EMCI", "LMCI", "SMC", "AD"))

#Add five-year age groups for ordinal regression models
cleaned_data <- cleaned_data %>%
  mutate(age_group = case_when(
    Age < 55 ~ "50-54",
    Age < 60 ~ "55-59",
    Age < 65 ~ "60-64",
    Age < 70 ~ "65-69",
    Age < 75 ~ "70-74",
    Age < 80 ~ "75-79",
    Age < 85 ~ "80-84",
    Age < 90 ~ "85-89",
    TRUE ~ "90-95"))
cleaned_data$age_group <- as.factor(cleaned_data$age_group)

head(cleaned_data)

#Compare having AD at baseline with having APOE4 present
table(cleaned_data$APOE4, cleaned_data$AD)


#~~~~~~Logistic regression models~~~~~~
#Initial model
m0 <- glm(AD ~ APOE4 + Age + Gender + eth + site, data = cleaned_data, family = "binomial")
summary(m0)
tab_model(m0)
```

```r
#Only APOE4 and age are significant variables here,
#however gender and ethnicity are still likely to be confounders
#so I will keep gender and eth in the model but remove site as it
#is not strongly believed that site is a confounder

#Test how OR for having AD changes for every additional APOE4 allele
m1 <- glm(AD ~ APOE4 + Age + Gender + eth, data = cleaned_data, family = "binomial")
summary(m1)
tab_model(m1)

#Test OR for having AD based on presence/absence of any APOE4 alleles
m2 <- glm(AD ~ apoe4_present + Age + Gender + eth, data = cleaned_data, family = "binomial")
summary(m2)
tab_model(m2)

#~~~~~~Ordinal logistic regression models~~~~~~

#Test how OR for having different baseline DX changes for every additional APOE4 allele
#Age groups used here instead of age for more readable output
m3 <- polr(dx_bl ~ APOE4 + age_group + Gender + eth, data = cleaned_data, Hess=TRUE)
summary(m3)$coef
ci <- confint(m3)
x <- exp(cbind(OR = coef(m3), ci))
kable(as.data.frame(x))

#Test OR for having different baseline DX changes based on presence/absence of any APOE4 alleles
m4 <- polr(dx_bl ~ apoe4_present + age_group + Gender + eth, data = cleaned_data, Hess=TRUE)
summary(m4)$coef
ci <- confint(m4)
x <- exp(cbind(OR = coef(m4), ci))
kable(as.data.frame(x))
```