IBMS3 ICA2: data analysis report

2020-03-27

Contents

1 Introduction

Osteoporosis is a systematic skeletal disease characterised by low bone density and deterioration of bone architecture which leads to an increased risk of fracture (Compston *et al.*, 2019). It is related to ageing and is more commonly reported in females than in males (Pietschmann *et al.*, 2009). Worldwide, it is a major public health issue that affects 1 in 3 women aged over 50 years old. Great importance should be attached to understanding how to treat and prevent the disease (Sözen *et al.*, 2017).

In females, osteoporosis is primarily caused by oestrogen decline, which can be effectively reversed and prevented by hormone replacement therapy (HRT) (Gambacciani and Levancini, 2014). However, the safety concerns of the therapy, especially its link to cancer onset (Rossouw *et al.*, 2002; Million Women Study Collaborators, 2003; Gambacciani and Levancini, 2014), decreases its use, which consequently increases the risk of osteoporosis for women (Karim *et al.*, 2011). The use of HRT is still controversial due to a lack of comprehensive evaluation of the effect of HRT (Shulman, 2011).

To find out whether and how HRT and/or Age can affect the risk of osteoporosis, our study analysed data from a case-control study involving 4500 females. We used R to assess how the odds of osteoporosis are influenced by HRT and/or age so that we can explain their relationships to the risk of the disease.

2 Text analytic

Understanding the data is the first step of our analysis.

First, the result of the case-control study is the proportions of subjects with osteoporosis, called cases, and those without osteoporosis, called controls in each group of certain age ranged between 50-90 and given certain treatment, either HRT or no HRT. By definition, the proportions are prevalence as prevalence is defined as the proportion of a population with certain condition among a population-at-risk (PAR). (Schoenbach and Rosamond, 2000) Elderly females in our study are PAR. We can know the prevalence of osteoporosis/no osteoporosis within each experimental group.

Second, with increasing sample size, according to the law of large numbers, the observed frequency of a certain condition, should converge to the probability of the condition (Siegmund, 2018). Therefore, prevalence almost equals to possibility when the sample size is large enough. Considering that the dataset given to us is a summative result involving 4,500 individuals in the case-control study, the sample size (n=4,500) for the whole study, and the respective sample size for each age group should be large enough to make the values of the observed prevalence almost equals to those of the probability.

Understanding the aims is the second step. We are required to discover the relationship between Age, HRT and the risk of osteoporosis. Statistically, risk is defined as the incidence proportion of an event, which refers to the incidence of osteoporosis in this report, during a period of time(Ranganathan *et al.*, 2015). However, risk is unable to be calculated in our study since the study design does not use disease-free populations (Dicker *et al.*, 2006).

Nevertheless, odds is more often used in such a study to infer the risk. Odds is defined as the ratio of the possibility of the occurrence of an event to the possibility that the event does not happen. In this report, it means the ratio of the possibility of osteoporosis, whose values equal to those of the prevalence, to the possibility of no osteoporosis, equals to the prevalence of no osteoporosis within the group (Equation (1)).

$$Odds = \frac{Probability(osteoporosis)}{Probability(no\,osteoporosis)} = \frac{Prevalence(osteoporosis)}{Prevalence(no\,osteoporosis)} \tag{1}$$

To compare the change of risk, odds ratio (OR) and the relative risk (RR) are usually calculated as measures of association between groups. OR correlates with RR and is well recognised as an indicator of risk for disease. (Chen *et al.*, 2010) If the ratio is larger than 1, it means the exposure is associated with a higher chance of disease incidence, i.e. the risk, while if it is less than 1, it means the exposure is associated with a lower chance of disease incidence. OR can be calculated by Equation (2):

$$OR = \frac{Odds(exposed)}{Odds(unexposed)} = \frac{Odds(HRT)}{Odds(no\,HRT)} \tag{2}$$

Therefore, to find the relationship between the effect of Age and HRT on osteoporosis risk, we need to find the relationship between the two factors and the odds of osteoporosis. Then we should examine whether the correlation between these variables and the odds is significant and we should also determine the effect size indicated by the OR. Besides, considering that the data set is multivariate, the interaction between variables should be described, if it is found significant.

3 Exploratory analysis

3.1 Data preparation

To analyse the results, we need to import them with R, along with the packages we will use for our analysis:

```r
suppressPackageStartupMessages(library(emmeans))# for: emmean, emmip
library(ggplot2)                    # ggplot
theme_set(theme_bw())               # ggplot theme
library(reshape2)                   # for: melt, dcast
library(knitr)                      # for: kable
options(knitr.table.format = "pandoc")  # change all kables to pandoc format
results <- read.csv("osteo.csv") # Import the dataset
```

We can exhibit the first 6 lines of the dataset with the following code, but we may not exhibit the output of the code, since it could be duplicate information after we visualise them in a well-labelled graph later.

```r
head(results)
```

Then, we will notice that HRT is both the name of the variable and its values, which may lead to confusion when we refer to the variable or its values, so we change the name of the variable to treatment:

```r
names(results)[names(results)=="HRT"] <- "Treatment"
```

3.2 Data visualisation

Note that the prevalence of two contradicting medical conditions, which sums up to 100%. It should be better if we can visualise them in a percentage stacked plot or a stacked plot as the data have been normalised to percentages, as shown in Figure 1 below. This gives us several benefits: 1) two different treatment groups are visualised in two graphs; 2) two different medical conditions, osteoporosis and no osteoporosis, are visualised in different colours. In short, it is more informative than presenting a table alone.

```r
# Use a backup dataset to avoiding mess up with results from other chunks
results4plot <- results
# Change the name of the column before using ggplot::facet_wrap()
names(results4plot)[names(results4plot)=="NoOsteoporosis"] <- "No Osteoporosis"
# Number all rows with unique IDs
results4plot$ID <- 1:nrow(results4plot)
```

```r
# Remain the first 3 columns and reshape the latter 3 columns
results4plot.headers <- results4plot[c("ID","Age","Treatment")]
results4plot.values<- melt(results4plot[c("ID","Osteoporosis","No Osteoporosis")],
                           variable.name ="Outcome",value.name="Proportion",id.vars="ID")
# Merge the 2 subsets according to preencoded ID
results4plot <- merge(by="ID",results4plot.headers,results4plot.values)
# Visualise the outcome table in percentage stacked barplots
ggplot(results4plot, aes(fill=Outcome, y=Proportion, x=Age)) +   # Define variables
  geom_bar(position="stack", stat="identity")+                  # Stacked plot
  scale_x_continuous(breaks=seq(50,90,5))+ylab("% proportion")+  # Axis scale & label
  theme(legend.position="bottom",legend.direction="horizontal",  # Legend options
        legend.title = element_blank())+
  geom_text(aes(x = Age, y = Proportion, label = paste0(Proportion,"%")), # Data label
            position = position_stack(vjust = 0.5), size = 2.5)+
  facet_wrap(vars(Treatment))  # Divide the plot according to Treatment
```

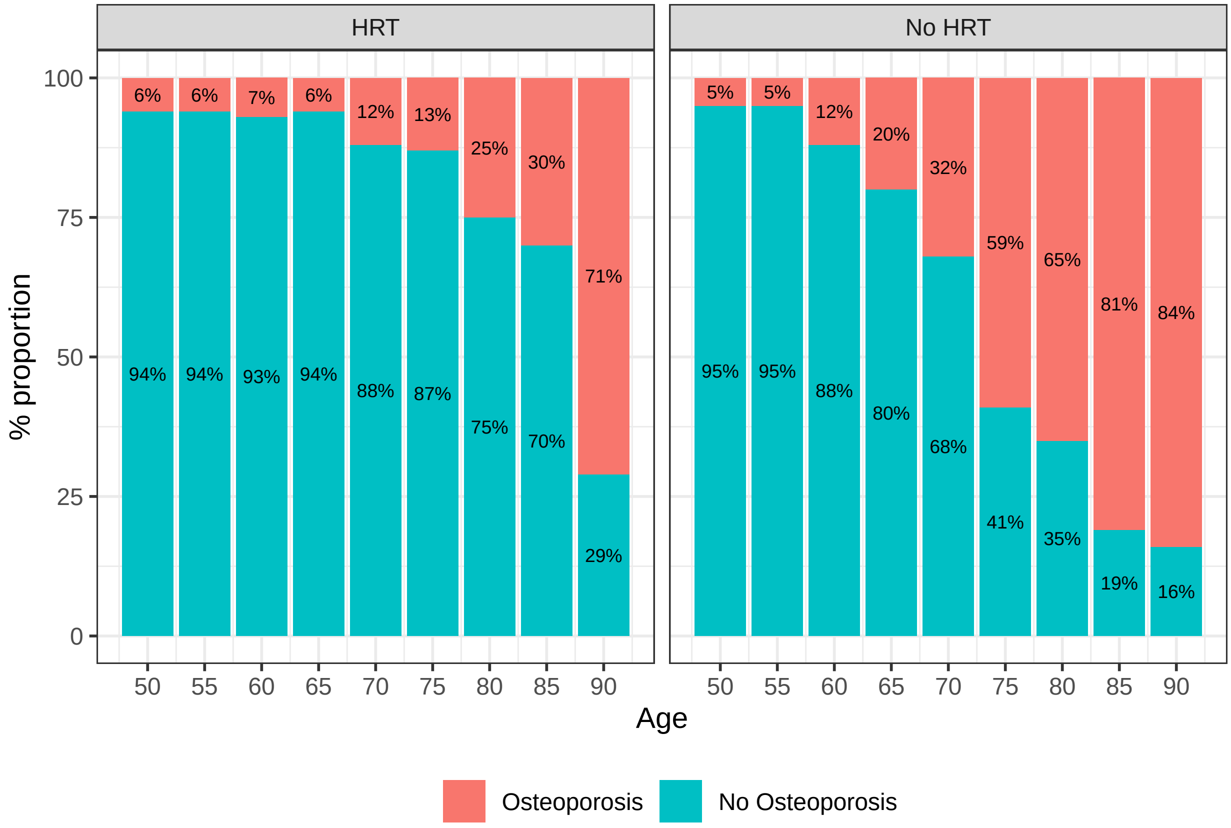

Figure 1: (Percentage) stacked plots of the proportion of osteoporotic and non-osteoporotic subjects

By looking at Figure 1, there are two categories which refer to two different types of treatment, HRT and no HRT. The proportions of subjects with and without osteoporosis are calculated for 9 ages and every 5 years between 50 and 90 years, which makes 9 pairs of data. Although 4,500 is a large sample size, the 9 pairs of data for analysis based on the results from 4,500 subjects is a small sample size for most analyses. When using the statistic methods that require a normal distribution, a normality test is needed.

Interestingly, the data are not organised as typical results from a case-control study. In each group at a given age, the proportions of the two conditions, no osteoporosis and osteoporosis, instead of the proportions of different exposures, HRT or no HRT, are shown. The two proportions seem to be different dimensions, yet they are not independent, because they always sum up to 100 in a given treatment group at any age. Thus, the dataset can be simplified to 3 dimensions (i.e. age, treatment, and osteoporosis prevalence) as Table 1:

```
results.wide <- # Transform to a wide data showing the prevalence of osteoporosis
  dcast(results, Treatment~Age, value.var = "Osteoporosis")
names(results.wide)[1] <- "Age" # Change the first column name to "Age"
kable(results.wide, caption = "Osteoporosis prevalence by age and treatment groups")
```

Table 1: Osteoporosis prevalence by age and treatment groups

Age	50	55	60	65	70	75	80	85	90
HRT	6	6	7	6	12	13	25	30	71
No HRT	5	5	12	20	32	59	65	81	84

Table 1 shows how osteoporosis prevalence changes with age, which can be visualised by Figure 2 below:

```
ggplot(results, aes(x=Age,y =Osteoporosis, colour=Treatment)) +
  geom_line(size=1)+
  ylab("% prevalence of osteoporosis")+
  scale_colour_manual(name="Treatment",values=c("darkred","steelblue"))
```

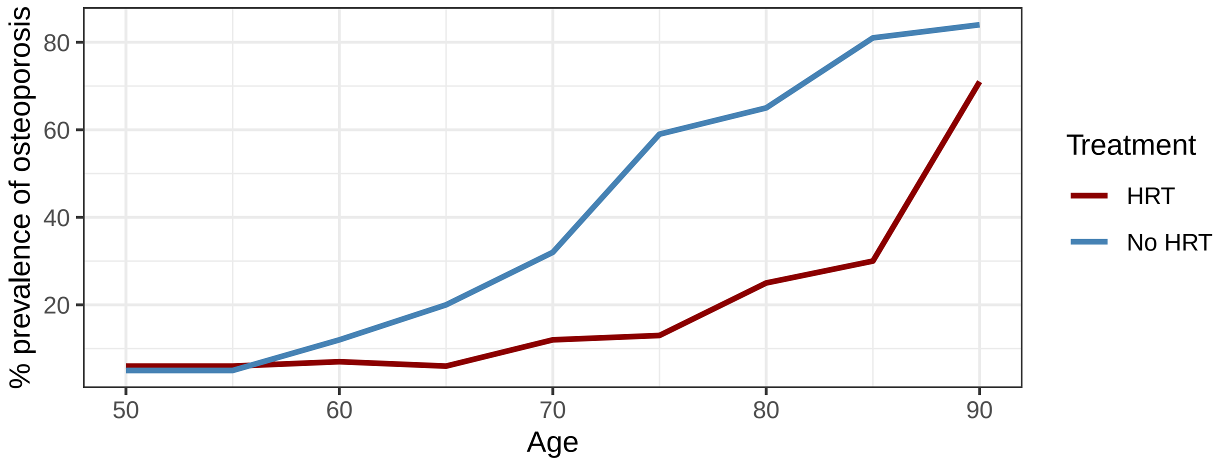

Figure 2: Change of the prevalence of osteoporosis with age

Note that in the HRT group, the osteoporosis prevalence at the age of 65 is lower than at the age of 60 (Table 1 and Figure 2). Considering that there is no cure for osteoporosis (Lewiecki *et al.*, 2019), the decrease suggests that there were different groups of people of the two ages. Also, no country has a life expectancy longer than 85 years in the world (Roser *et al.*, 2013), making it unlikely to track the 4,500 subjects throughout the years. Tracking those who can live until 90 years old often makes the statistics lack representativeness of the general public who are mostly unable to live that long. In conclusion, each age group should be distinct from each other, making it possible that the study be a cross-sectional case-control study that uses, as described in a previous study (Klossek *et al.*, 2005), but we are unable to know whether this conjecture is right unless given further details.

3.3 Measures of central tendency

Assuming each age group and treatment are weighted the same, we can see the summary of the data set to observe its central tendency:

```r
kable(summary(results),caption = "Measures of central tendency")
```

Table 2: Measures of central tendency

Age	Treatment	Osteoporosis	NoOsteoporosis
Min. :50	HRT :9	Min. : 5.00	Min. :16.00
1st Qu.:60	No HRT:9	1st Qu.: 6.25	1st Qu.:47.75
Median :70	NA	Median :16.50	Median :83.50
Mean :70	NA	Mean :29.94	Mean :70.06
3rd Qu.:80	NA	3rd Qu.:52.25	3rd Qu.:93.75
Max. :90	NA	Max. :84.00	Max. :95.00

Table 2 above that describes measures of the central tendency in our data set. Note that the measures are only meaningful when the data from each group is of the same weight, or else there might be a base rate fallacy[1]. Table 2 shows major deviance of the median from the mean in both HRT and no HRT groups, which could suggest a non-normal distribution. We can examine this with Shapiro–Wilk test, as shown below:

```r
# Normality test for osteoporosis prevalence in HRT group
shapiro.test(results[results$Treatment=="HRT",]$Osteoporosis)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  results[results$Treatment == "HRT", ]$Osteoporosis
## W = 0.70622, p-value = 0.001678
```

```r
# Normality test for osteoporosis prevalence in no HRT group
shapiro.test(results[results$Treatment=="No HRT",]$Osteoporosis)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  results[results$Treatment == "No HRT", ]$Osteoporosis
## W = 0.87791, p-value = 0.1493
```

The test results show that the prevalence data of osteoporosis in no HRT groups are not normally distributed ($p = 0.1493 > 0.05$), while those in HRT groups is normally distributed ($p = 0.001678 < 0.05$).

[1] For example, we calculate the mean of age because we assume each age group has the same sample size. However, if among the 4,500 individuals involved in the study, there are 3,700 aged 50, and for the rest of the ages, there are only 100 subjects. Then the mean of age in the HRT group should be equal to $(3700 \times 50 + 100 \times 55 + 100 \times 60 + 100 \times 65 + \ldots + 100 \times 90) \div 4500 = 54$ instead of the value of arithmetic mean of age which equals to 70.

4 Aims and hypotheses

We are aimed to assess whether Age and HRT affect the risk of osteoporosis and describe how each of the terms can influence the risk of osteoporosis. Figure 2 can be considered as an interaction plot[2] that shows how Age and HRT interact with each other on the effect on osteoporosis prevalence. Since the odds of osteoporosis depends on prevalence, the interaction, which is statistically defined as a situation where the effect of a fixed factor is changed due to the values of another variable (Darwin, 2005), affects the odds and the risk. Assume that $p = 0.05$ is the significant level in this report. We can have three hypotheses based on our aim and what we have observed in Figure 2:

1. Age has a significant impact on the odds of osteoporosis. More specifically, the odds of osteoporosis increases with age, regardless of treatment.
2. HRT has a significant impact on the odds of osteoporosis. More specifically, HRT can significantly reduce the odds of HRT.
3. There is a significant interaction between age and treatment.

To prove the hypotheses involving more than two terms and potentially interrelated variables, a regression model is recommended. With the regression model, we should test whether the terms, including Age, Treatment and the interaction term are significant and describe the effect of Age and HRT on the odds osteoporosis and if there is a significant interaction between Age and Treatment, we should also describe how the effect of HRT changes with age and how the effect of age differ when there is and is not HRT.

5 Model fitting

5.1 Rationale

5.1.1 Consideration of the interaction term

The presence of interactions in a multivariate model, such as our data, makes the main effects no longer of interest, because the main effects alone fail to describe the relationship between variables. (Darwin, 2005, p. 146) Although not all interactions may have biological plausibility and reproducibility (Altman and Matthews, 1996), it is reasonable to analyse the interaction between Age and HRT and its effects on the risk of osteoporosis. First, ageing is associated with the increased risk of osteoporosis onset (Sözen *et al.*, 2017), whereas HRT has been known to reduce the risk (Gambacciani and Levancini, 2014). Thus, the preventive effect of HRT might be reduced or even reversed by the increase of the age. Second, previous studies have shown such interactions. For example, Port *et al.* (2003) reported that HRT was a more effective method to prevent osteoporosis for women aged 80 years or more. Therefore, finding out both what are the main effects and whether there is a significant interaction between the age and the treatment is important.

5.1.2 Model choice

To study our multivariate data set that includes non-normally distributed data (see *Measures of Central Tendency* section), we can use a generalised linear model (GLM) that requires no normality test. In the *Text Analytic* session, we learn that we can calculate the odds to estimate the risk. In GLM, we can use the link function of the *logit* function, as shown by Equation (3), to create a logistic regression.

$$log(odds(Y)) = log(\frac{p(Y)}{1-p(Y)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{3}$$

[2] A similar but essentially the same interaction plot can be made with R script: `interaction.plot(results$Age, results$Treatment, results$Osteoporosis)`

In Equation (3), Y represents the event of interest, i.e. developing osteoporosis, X represents the factors that influence the log-odds of Y. For $log(\frac{p(Y)}{1-p(Y)}) > 0$, $0 < p(Y) < 1$. This feature makes the logistic regression very suitable for the data ranging between 0 and 1, such as the prevalence of osteoporosis. So, a logistic regression model could be more suitable than others for us to find the relationships between variables.

5.2 Logistic regression model

To establish a logistic regression model, we need suitable parameters and a well-defined formula. Since the smallest age is 50, we will deduce all the ages by 50 to create a variable called *AgeAdj*. Using such a variable does not affect other coefficients in the model but the intercept. The intercept shows the log(odds) when $Age = 50$ instead of $Age = 0$, which simplifies our calculation. Treatment is regarded as a dummy variable where HRT equals to 1 and No HRT equals to 0. Then we can establish a model as below:

$$log(odds) = \beta_0 + \beta_1 \times AgeAdj + \beta_2 \times Treatment + \beta_3 \times AgeAdj \times Treatment \tag{4}$$

The R script to build the model accordingly is shown below:

```r
results4GLM <- results
# To change the order of factors for GLM
results4GLM$Treatment <- factor(results4GLM$Treatment, levels = c("No HRT","HRT"))
# Deduce age by 50 to make it start from 0
results4GLM$AgeAdj<- results$Age -50
GLM <- glm(cbind(Osteoporosis,NoOsteoporosis)~AgeAdj*Treatment,
           family = binomial(link = "logit"),
           data= results4GLM)
```

Then we can visualise the model parameters in Table 3 below:

```r
kable(cbind(
  coef(GLM),             # beta
  exp(coef(GLM)),        # OR
  format(as.data.frame(suppressMessages(exp(confint(GLM))),digit=3)),  ## CIs
  format(as.data.frame(summary(GLM)$coefficients)$"Pr(>|z|)",digit=3)),
  caption = "GLM parameters",  ## Caption
  col.names=c("beta","OR", "2.5%CI", "97.5%CI", "p-value"))
```

Table 3: GLM parameters

	beta	OR	2.5%CI	97.5%CI	p-value
(Intercept)	-3.2502787	0.0387634	0.02498216	0.05840025	5.32e-51
AgeAdj	0.1307565	1.1396902	1.12139219	1.15951033	3.47e-53
TreatmentHRT	-0.6058650	0.5456023	0.27167272	1.07454073	8.34e-02
AgeAdj:TreatmentHRT	-0.0314041	0.9690839	0.94569047	0.99319607	1.19e-02

5.3 Model evaluation

5.3.1 Hypothesis testing

```r
kable(cbind(GLM[["deviance"]],GLM[["null.deviance"]], GLM[["aic"]]),
      col.names = c("Deviance","Null deviance", "AIC"),
      caption = "Parameters of goodness-of-fit")
```

Table 4: Parameters of goodness-of-fit

Deviance	Null deviance	AIC
37.47361	676.4883	122.6511

Table 4 shows three parameters related to the goodness-of-fit. The null deviance is the deviance of the null model that supposes that either Treatment or AgeAdj affects the odds of disease, while the residual deviance is that of our model. Apparently, the deviance of our model is much smaller than that of the null model, suggesting a bad fitting for the null model, which hypothesise that none of the age and the treatment influences the odds of disease. Furthermore, the p-value for the intercept is significant ($p = 5.3 \times 10^{-51} < 0.05$; Table 3), meaning that our model is significantly different from the null model. Thus, we can reject the null model. Either/both of Age and Treatment has a significant impact.

5.3.2 In-sample forecast

After we have developed a regression model based on our analysed data, an in-sample forecast can help us to evaluate the predictive capabilities of the model. To do this, we should calculate the values and the CIs of the predicted prevalence, or called possibility, at different ages for females with or without HRT based on our model, as shown in Figure 3.

```r
#Build a function to calculate HRT prediction and CIs
Pred_with_CI <- function(model, category){
  newdata <- list(AgeAdj = 50:90-50, Treatment = rep(category,length(50:90)))
  model <- predict(model, type = "link", newdata, se.fit = TRUE)
  model$loCI <- plogis(model$fit - qnorm(1 - (1 - 0.95)/2)*model$se.fit)*100
  model$hiCI <- plogis(model$fit + qnorm(1 - (1 - 0.95)/2)*model$se.fit)*100
  model$fit <-  plogis(model$fit)*100
  model$Age <-  50:90
  return(model)}
# Data preparation for visualisation
pred <- list(Age=50:90,NoHRT=Pred_with_CI(GLM, "No HRT"),HRT=Pred_with_CI(GLM, "HRT"))
prev_by_treatment <-reshape2::dcast(results,Age~Treatment,value.var ="Osteoporosis")
# Not changeing the name will disable ggplot to choose the column for y.
names(prev_by_treatment)[names(prev_by_treatment)=="No HRT"] <- "NoHRT"
ggplot() +ylab("%")+#plot all data: geom_point for observed, geom_smooth for predicted
  geom_point(aes(x=Age,y=NoHRT,colour="No HRT",shape="No HRT"),prev_by_treatment)+
  geom_point(aes(x=Age,y=HRT,colour="HRT",shape="HRT"), prev_by_treatment)+
  geom_smooth(aes(x=Age,y = HRT.fit, ymin = HRT.loCI, ymax = HRT.hiCI,
                  colour="HRT"), stat = "identity", as.data.frame(pred))+
  geom_smooth(aes(x=Age,y = NoHRT.fit,ymin = NoHRT.loCI, ymax = NoHRT.hiCI,
                  colour="No HRT"), stat = "identity", as.data.frame(pred))+
  # Manually add legends based on shape and colour
  scale_colour_manual(values = c("No HRT"="steelblue", "HRT"="darkred"),
                      name = "Predicted possibility \n(with 95% CI)")+
  scale_shape_manual(values = c("No HRT"=16, "HRT"=17), name = "Observed prevalence")
```

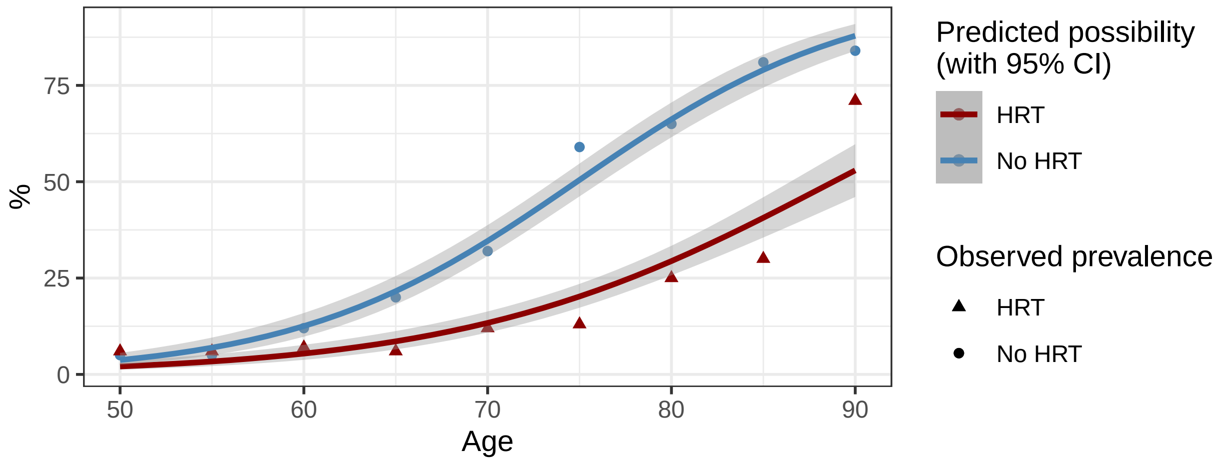

Figure 3: Predicted proablity and observed prevalence of osteoporosis[3]

Figure 3 actually shows a poor predictive capability of our model, since most observations do not fall in the confidence band where 95% of data fall in, which could suggest a bad fitting. So we will analyse the goodness of fit of the model compared to other models in the next section.

5.3.3 Comparison between possible models

Although the in-sample forecast provides visible information about the model fitting, considering that the potential bad fitting shown in Figure 3, it is better to compare our model with other possible models. Comparison of the models with and without the interaction is shown in Table 5 below:

```
kable(drop1(GLM), caption = "Goodness-of-fit change  after dropping
      the interaction term")
```

Table 5: Goodness-of-fit change after dropping the interaction term

	Df	Deviance	AIC
	NA	37.47361	122.6511
AgeAdj:Treatment	1	43.72252	126.9000

Table 5 compares the models with and without drop one term, which is AgeAdj:Treatment. It shows the coefficients indicating goodness-of-fit of the two models. After dropping the interaction, AIC rises from 122.6511 to 126.9000, suggesting that the model with the interaction has a better goodness-of-fit. Also, failure to include the interaction in the model leads to the increase of deviance from 37.47361 to 43.72252, which also supports the better goodness-of-fit of the model with the interaction.

We can drop more terms, as shown in Table 6:

[3] Why we use different terms is because although the prevalence is close to the probability, but it is not probability.

```r
GLM.no_interaction <-  # Creating a model with the interaction term
  glm(cbind(Osteoporosis,NoOsteoporosis)~AgeAdj+Treatment,
      family = binomial(link = "logit"),
      data= results4GLM)
kable(drop1(GLM.no_interaction), digits = 3,
      caption = "Goodness-of-fit change after dropping more terms")
```

Table 6: Goodness-of-fit change after dropping more terms

	Df	Deviance	AIC
	NA	43.723	126.900
AgeAdj	1	582.408	663.586
Treatment	1	173.625	254.802

Table 6 shows that the model that only consider Age or Treatment has much higher deviance (Treatment-only model: 582.40849, Age-only model: 173.62469) and a much higher AICs ((Treatment-only model: 663.5860, Age-only model: 254.8022), compared to the model which considers the two factors without interaction (AIC: 126.9000, deviance: 43.72252), which has proven to be no better fitting than the model considering the interaction term. In conclusion, the model with the interaction is the best-fitting among all these models.

6 Interpretations and further analyses

Table 3 is a summary of logistic regression model parameters, including β, the odds calculated according to β, 95% confidence intervals, p values for each term. We fill the $/beta$ values we have acquired according to the logistic model into Equation (4) that we use to build our model:

$$log(odds) = -3.2502787 + 0.1307565 \times AgeAdj - 0.605865 \times Treatment - 0.0314041 \times AgeAdj \times Treatment$$

Apparently, according to Table 3, the interaction term ($p = 1.19e - 02$) is significant. Therefore, when describing the effect of either of HRT or Age, it is necessary to describe the value of the other variable. Then we will begin with analysing the effect of age.

6.1 The effect of age

Our observations in Figure 2 lead to the hypothesis that the risk of osteoporosis increases significantly with age, which is confirmed in our model. The effect of age has a p-value of 3.5×10^{-53} which is far less than 0.05, indicating the significance of the effect. The increase of age can have an impact on the odds of disease, since every single year, the OR of disease increases to 1.14 times (95% CI: [1.12, 1.16]; Table 3) of the OR of the previous year, which suggests the increased risk owing to the increasing age.

To appreciate the effect of age, we need to figure out what is the reference level. First, the intercept is -3.25, which means that the basal log(odds) of getting osteoporosis for an individual who does not use HRT (i.e. in the reference group) at the age of 50 (alternatively, $AgeAdj = 0$). Thus, the odds for the person is $e^{-3.25} \approx 0.03877421 = 3.9\%$ (Table 3). The CIs of the odds is between 0.02498 and 0.0584 (Table 3), which is significant, suggesting only 3.9% (95%CI: [2.5, 5.8]) of females have osteoporosis at the age of 50.

Taking the interaction into account, for Treatment is a dummy variable where HRT=1 and No HRT = 0,

$$log(odds) = \begin{cases} -3.856144 + 0.0993524 \times AgeAdj & (with\ HRT) \\ -3.2502787 + 0.1307565 \times AgeAdj & (without\ HRT) \end{cases}$$

When HRT is used, the reference level of the odds of osteoporosis (i.e. $Age = 50 \ or\ AgeAdj = 0$) equals to $e^{-3.856144} = 0.02114939$, which means it is estimated that only around 2.1% of the subjects who use HRT has osteoporosis when they are 50 years old. Similarly, when HRT is not used, the reference level of the odds of the disease equals to $e^{-3.2502787} = 0.0387634$, which means only 3.9% of the subjects without HRT is expected to have osteoporosis when they are 50. For both groups, the odds are fairly low at the age of 50, but HRT already lowers the odds at the age, as our model shows.

What makes a difference is that the rate of odds increasing with every single age. For HRT users, log(odds) increases by 0.0993524 each year after they are 50 years old, which means the odds increases to $e^{0.0993524} = 1.104455$ times of the previous year, while for non-HRT users, log(odds) increases by 0.1307565 each year and odds increases to $e^{0.1307565} = 1.13969$ times of the previous year every year. The yearly odds ratio is much larger for non-HRT users. Although both HRT and non-HRT users are exposed to increased risk, as our model suggests, the odds increase faster for non-HRT users.

6.2 The effect of HRT

6.2.1 Regression model-based interpretation

For HRT, $p = 0.08 > 0.05$, which means its effect is not significant if we compare two groups as a whole (Table 3). Yet, for the effect of HRT, $\hat{\beta} = -0.6059$. This means that compared to the references, which refers to no HRT group, the subjects with HRT have an odds of 54.6%, which suggests that the odds equals to $e^{-0.6059} \approx 0.5455832 = 54.6\%$ (Table 3). The 95% CI is between 0.2717 and 1.075 (Table 3), indicating a broad range of possibility where the true effect size could be. Therefore, in this sense, it is also reasonable to see that HRT increases the chance of osteoporosis when OR is larger than 1 but smaller than 1.075, the upper boundary of the CI. Although there is a large effect size, the effect itself is not significant.

One of the reasons to cause the insignificance of the effect could be that the effect is not always significant at all ages. Recall that the interaction term is significant, so we should evaluate the age-specific effect of HRT with pairwise comparison, as shown by Table 7 below:

```
marginals <- emmeans(GLM, ~AgeAdj*Treatment, cov.reduce = unique)
pairwise_comparison <-as.data.frame(pairs(marginals, by = "AgeAdj", type = "response"))
pairwise_comparison$p.value <- pairwise_comparison$p.value
kable(pairwise_comparison, caption="Pairwise comparison of HRT effect")
```

Table 7: Pairwise comparison of HRT effect

contrast	AgeAdj	odds.ratio	SE	df	z.ratio	p.value
No HRT / HRT	0	1.832837	0.6413950	Inf	1.731307	0.0833970
No HRT / HRT	5	2.144455	0.6282495	Inf	2.604018	0.0092138
No HRT / HRT	10	2.509054	0.5989433	Inf	3.853608	0.0001164
No HRT / HRT	15	2.935641	0.5564503	Inf	5.681493	0.0000000
No HRT / HRT	20	3.434757	0.5170907	Inf	8.196447	0.0000000
No HRT / HRT	25	4.018733	0.5268858	Inf	10.609363	0.0000000
No HRT / HRT	30	4.701995	0.6566634	Inf	11.084259	0.0000000
No HRT / HRT	35	5.501426	0.9466357	Inf	9.908744	0.0000000
No HRT / HRT	40	6.436775	1.4029033	Inf	8.543321	0.0000000

Table 7 shows a significant effect of HRT ($p < 0.05$) for those aged between 55 ($AgeAdj = 5$) and 90 ($AgeAdj = 40$), while it shows that the effect of HRT is not significant for those aged 50 ($AgeAdj = 0$). However, the odds ratio increases steadily with age, from 1.83283 at the age of 50 to 6.43677 at the age of 90, which suggests an increasing effect of HRT when people grow older. Also, note that the odds ratio of No

HRT/HRT increases with age.

In conclusion, our results show that not taking HRT can significantly increase the odds of osteoporosis, which suggests a protective effect of HRT against osteoporosis. The effect increases with age and becomes significant after the age of 55. Therefore, our hypothesis that HRT can lower the odds of osteoporosis is not completely statistically correct, since there is no significant effect at the age of 50, while there is increasing protective effect after 55 years old. The increasing effect can be visualised in Figure 4, as the slope rate increases with age.

```
emmip(GLM, AgeAdj~Treatment, type = "response",cov.reduce = FALSE)
```

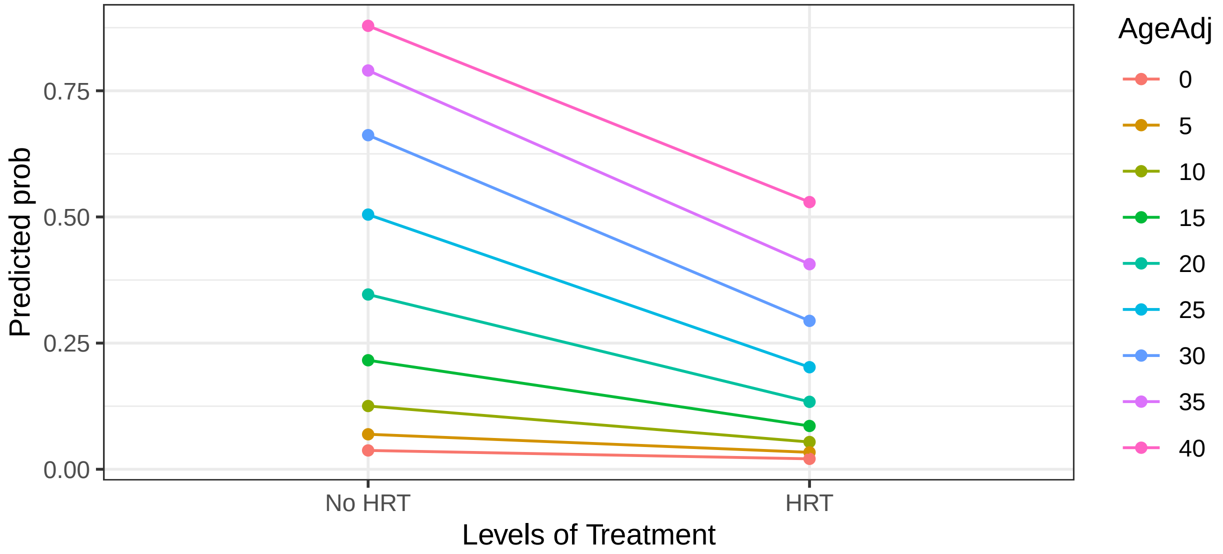

Figure 4: The effect of HRT by age

6.2.2 Disparity in age threshold

Interestingly, however, Fisher's exact test shows a different result for the effect of HRT at different ages. For example, for Age=50 (AgeAdj=0), we can have a 2 × 2 table like Table 8:

```
results.Age50<-results[which(results$Age==50), ]
results.Age50$Age<-NULL
kable(results.Age50, col.names = c("","% Osteoporosis","% No osteoporosis"),
      caption = "Exemplar 2×2 Table for Fisher's exact test ($Age=50$)")
```

Table 8: Exemplar 2×2 Table for Fisher's exact test ($Age = 50$)

	% Osteoporosis	% No osteoporosis
HRT	6	94
No HRT	5	95

Then, for all different ages, we can have 9 2 × 2 tables, which can be used for the calculation of Fisher's exact test. The test can tell us he OR and their CIs and p-values, as shown in Table 9 below:

```
OR_results<-NULL          # create an empty dataframe
for (age in unique(results$Age)){
  results.age<-results[which(results$Age==age), ]
  results.age$Age<-NULL
  fisher<-fisher.test(data.matrix(results.age[,-1]))
  new_results<-data.frame(  # Define values in dataframe
    "Age"=age,
    "p.value"=format(fisher[["p.value"]],digits=3),
    "OR"=fisher[["estimate"]][["odds ratio"]],
    "loCI"=as.vector(fisher[["conf.int"]])[1],
    "hiCI"=as.vector(fisher[["conf.int"]])[2])
  OR_results<-rbind(OR_results, new_results)}
kable(OR_results, digits = 3,
      caption="Result of Fisher's exact test",
      col.names = c("Age","p-value", "OR", "2.5%","97.5%"))
```

Table 9: Result of Fisher's exact test

Age	p-value	OR	2.5%	97.5%
50	1	1.212	0.297	5.201
55	1	1.212	0.297	5.201
60	0.335	0.554	0.176	1.607
65	0.00543	0.257	0.080	0.705
70	0.00103	0.292	0.127	0.633
75	9.51e-12	0.105	0.047	0.220
80	1.92e-08	0.181	0.093	0.345
85	3.18e-13	0.102	0.049	0.203
90	0.0414	0.468	0.219	0.974

According to Table 9, when the age equals to 50, 55 or 60, the effect of HRT is not significant, while when the age equals to 65, 70, 75, 80, 85, 90, the effect is significant. This may suggest that the effect of HRT only becomes significant after 60 years old, while before that the drug does not significantly protect the patients from osteoporosis. The difference in the age threshold for significant HRT effect could be due to the conservative nature of the Fisher's exact test, since the actual rejection rate of the test is actually lower than the nominal significance level of 0.05%.

6.2.3 Disparity in odds ratios

Nevertheless, the odds ratio calculated from the samples, as Fisher's exact test shows, are different from those based on logistic regression model. The disparity between the observed and predicted OR can be visualised in Figure 5 below:

```
# Calculating odds ratios for GLM
Odds.pred<-data.frame(cbind(pred$Age, ((pred$HRT$fit/100)/(1-pred$HRT$fit/100))/
                              (pred$NoHRT$fit/100/(1-pred$NoHRT$fit/100))))
colnames(Odds.pred)<-c("Age", "OR")
# Plotting the OR changes with age
ggplot()+
  geom_line(aes(x = Age, y = OR, colour = "Observed"), OR_results, size=1)+ #Observed OR
  geom_line(aes(x = Age, y = OR, colour = "Predicted"), Odds.pred, size=1)+ #Prediceted OR
  geom_hline(yintercept=1, linetype="dashed", colour = "red")+ # OR = 1 dashed line
  labs(x ="Age", y = "OR")+  # y axis label
  # Manual legend
  scale_colour_manual(name=NULL, values=c("Predicted"="slategray3", "Observed"="steelblue4"))
```

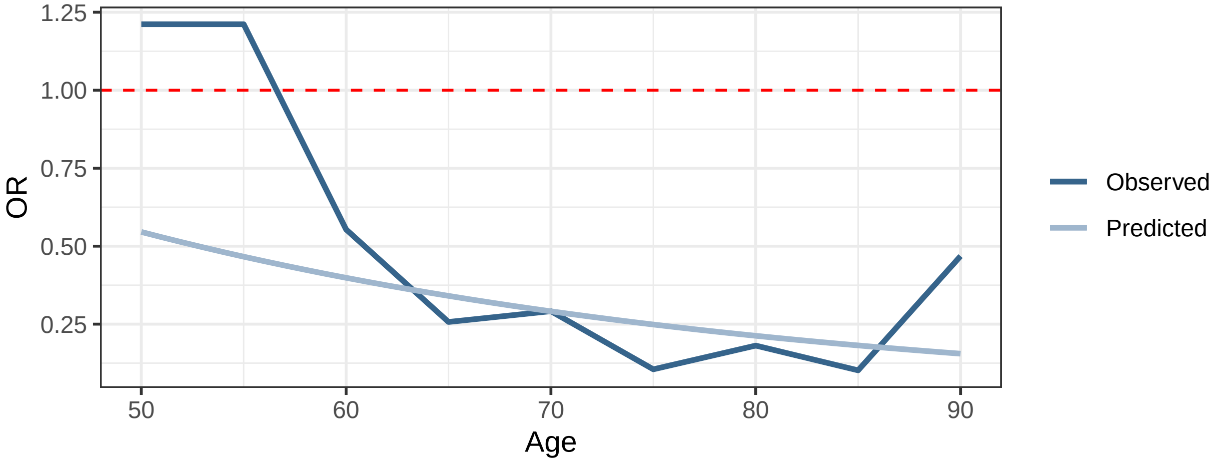

Figure 5: The change of observed and predicted OR for HRT with age

Figure 5 shows the age-dependent change of HRT effect on osteoporosis that supports the presence of the interaction term, since no matter whether it is according to our observations and our predictive model, the effect of HRT is different at different ages, but note that Figure 5 itself indicates nothing about the significance of the effect but the odds ratios, which indicates the effect size. According to our predictive model, the ORs are always below 0, which means HRT have a protective effect at all ages, whereas according to our observed ORs, there was not any protective effect when the age was no less 60 years. This could be a result of wide confidence interval during the ages, because there is no significant effect of HRT at before 55 based on either of our model or Fisher's exact test.

7 Limitations

A major drawback in our analysis is that our predictive model does not have enough predictiveness. According to our in-sample forecast (in Figure 3), most observations we have so far acquired do not fall in the confidence intervals. One reason for this is that our measure is a single experiment with random effects that need to be duplicated for more times to acquire the correct result. The second reason is that we could have failed to consider some factors that affect the model. Thus, we should take them into consideration and introduces more terms in the model. The third reason is that there are biases and major errors that disadvantage the study, which will soon be discussed.

7.1 Overestimating risk with OR

We are unable to calculate the risk of osteoporosis but to infer the association between age and treatment through risk-related variable, the odds and odds ratio. However, our results could overestimate the effect of age and underestimate the effect of HRT. The comparative risk of one group with certain exposure compared to another without the exposure can be shown as the relative risk (RR). OR and relative risk (RR) are different measures of association that correlate with each other. OR tends to exaggerate the association: when RR is larger than 1, OR is larger than RR; when RR is smaller than 1, OR is smaller than 1. The disparity between the two parameters is small, only when there is no association (i.e. $OR = RR = 1$), or when the event is rare (normally smaller than 10%). However, in our study, osteoporosis is not at all a rare disease; in fact, more than two-thirds of the people got the disease when they were 90, which means that our OR could greatly exaggerate the risk. Since the ORs for every single year older (Table 3) and no HRT (Table 7) are greater than 1, **the protective effect of HRT could be underestimated while the relative risk of age could be overestimated.**

7.2 Opaque experimental design and records

The data set we acquired does not report the protocol and records of data collection, as well as possible alterations or transformation to the data. How the study was carried out and how the data were recorded may have a significant impact on the results. We can assume that the study is a cross-sectional case-control study that uses different age groups because of 1) a reduction in prevalence which is unlikely due to no cure for osteoporosis and 2) the fact that most people hardly stay alive until their 90s. However, our assumption could be wrong, if the study could use a diagnostic protocol that allows the reduction to happen or the study really measures 4,500 consistently for 40 years, or the prevalence data at that age could be wrong.

If the diagnostic protocol is different from what is recognised today, then what were the criteria for admission of subjects and their grouping? Were they consistent across different doctors and different medical centres? If the criteria were inconsistent, a patient with osteoporosis might be diagnosed by one doctor and be considered healthy by another. Then our results of osteoporosis prevalence could be inaccurate. If the study really measures 4,500 consistently for 40 years, then will the result still be useful for ordinary people who normally cannot live that long? Besides, how do the researcher define the HRT users? How frequent should one take HRT to be defined in the study? These differences in **unreported background information** could all lead to changes in how to interpret the results and should be reported in detail.

Besides, since we cannot ensure all experimental groups in a study to have an equal number of subjects and cannot expect them all to agree to provide their data, especially when the experiment is a large-scale one involving thousands of people. However, we are not reported the **sample size** of each age group. Since people aged older are relatively rarer, it is likely that the researchers fail to recruit a sample population of the same size for older age groups, or they fail to acquire a sample population with large enough sample sizes. In that case, we can not calculate the probablity because we cannot use the law of large numbers. Also, unbalanced group design could lead to base ratio fallacy that diminished the predictiveness of our logistic regression model.

7.3 Age-related selection bias

First, there could **survivorship bias** where osteoporotic patients who die from other diseases are not taken into consideration. The cumulative death unavoidably increases with age and most people die before 85 years old even in the areas with the world-leading medical care systems. For example, Japan has the highest life expectancy in the world, which is 83.98 years. So, people aged 90 could be less representative of ordinary people. The results from them could be wrong when applying to the general public. The number of available subjects will be reduced as age increases. Thus, a higher prevalence of osteoporosis among the older population could be a result of cumulative mortality from other diseases which are more deadly than osteoporosis. In this case, the association between osteoporosis prevalence and age may be biologically false although statistically true and the effect of HRT could be overestimated for those of older ages.

The increase of osteoporosis prevalence could be due to its relatively low risk of death compared to other diseases, which could fail to predict the risk of osteoporosis. For each age group we select, they were all the survivors from all the other causes of death whose risk could increase year by year. Many of those causes of death, e.g. breast cancer, can cause osteoporosis, but those osteoporotic patients who die from breast cancer are more and more unlikely to be examined by older age groups in the study, because age is also a major risk factor for cancer. So, those who happened to be in the experiment could be just those "lucky" survivors, whereas a major proportion of osteoporosis could be missed due to the increasing deaths associated with ageing-related diseases. Therefore, a comprehensive study and analysis that considers more ageing-related disease should be done.

Second, there could be **sampling bias** for younger subjects, where people with the awareness of early diagnosis of osteoporosis are more likely to be selected. The awareness could arise from many factors including a family history and genetic testing that could suggest a higher chance of osteoporosis. It is reasonable to guess that those HRT users who come to clinics for osteoporosis screening have a higher frequency and willingness to take the examinations, because taking HRT could be associated with a good awareness of osteoporosis prevention and treatment, which encouraged them to get diagnosed earlier than most people. Thus, the odds of disease positively correlates with HRT for the females of younger age, and therefore, the presumedly significant effects of HRT are insignificant during the younger ages (Table 7; Table 9) and the odds ratio is even larger than 1 in some cases (Figure 5).

8 Conclusion

Based on a case-control study involving 4500 females, our report analyses and describes the effects of two major risk factors for osteoporosis including failure to take hormonal replacement therapy (HRT) and ageing on the prevalence of osteoporosis among females aged between 50-90. We find that **both age and HRT can affect the risk of osteoporosis for the population.** First, **age is significantly associated with the increasing risk** of osteoporosis, because for every year between 50-90, for HRT users, the odds of osteoporosis increase to 1.104455 times of the odds in the previous year, while for non-HRT users, the odds only increases to 1.13969 times of that in the previous year. Second, **HRT can significantly reduce the odds of osteoporosis after an age threshold.** The logistic model shows that the effect of HRT becomes significant after 55 years old, while Fisher's exact test, which is a more conservative statistical test, shows that the threshold is 65 or above. The effect of HRT varies according to age. The model predicts that the protective effect of HRT increases with age. The odds ratio increases steadily each year, from 1.83283 at the age of 50 to 6.43677 at the age of 90. Lastly, more comprehensive research considering more ageing-related conditions, such as breast cancer, should be studied to reduce the potential bias and errors in the study and to provide more evidence-based approaches for using HRT.

References

Altman, D. G. and Matthews, J. N. S. (1996) Statistics Notes: Interaction 1: heterogeneity of effects, *BMJ*, 313(7055), p. 486.

Chen, H., Cohen, P. and Chen, S. (2010) How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies, *Communications in Statistics - Simulation and Computation.* Taylor & Francis, 39(4), pp. 860–864.

Compston, J. E., McClung, M. R. and Leslie, W. D. (2019) Osteoporosis, *The Lancet.* London: Elsevier, 393(10169), pp. 364–376.

Darwin, C. (2005) Continuous Variables: Analysis of Variance. (Wiley online books).

Dicker, R. C., Coronado, F., Koo, D. and Parrish, R. G. (2006) Lesson 3: Measures of Risk, in *Principles of epidemiology in public health practice: An introduction to applied epidemiology and biostatistics.* Third Edit. Atlanta, Georgia, U.S.A.: U.S. Centers for Disease Control; Prevention (CDC).

Gambacciani, M. and Levancini, M. (2014) Hormone replacement therapy and the prevention of post-menopausal osteoporosis, *Przeglad menopauzalny = Menopause review.* 2014/09/09. Termedia Publishing House, 13(4), pp. 213–220.

Karim, R., Dell, R. M., Greene, D. F., Mack, W. J., Gallagher, J. C. and Hodis, H. N. (2011) Hip fracture in postmenopausal women after cessation of hormone therapy: results from a prospective study in a large health management organization, *Menopause (New York, N.Y.)*, 18(11), pp. 1172–1177.

Klossek, J. M., Neukirch, F., Pribil, C., Jankowski, R., Serrano, E., Chanal, I. and El Hasnaoui, A. (2005) Prevalence of nasal polyposis in France: a cross-sectional, case–control study, *Allergy.* John Wiley & Sons, Ltd, 60(2), pp. 233–237.

Lewiecki, E. M., Binkley, N. and Bilezikian, J. P. (2019) Treated Osteoporosis Is Still Osteoporosis, *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research.* New Mexico Clinical Research & Osteoporosis Center, Albuquerque, NM, USA., 34(4), pp. 605–606.

Million Women Study Collaborators (2003) Breast cancer and hormone-replacement therapy in the Million Women Study, *The Lancet.* Elsevier, 362(9382), pp. 419–427.

Pietschmann, P., Rauner, M., Sipos, W. and Kerschan-Schindl, K. (2009) Osteoporosis: An Age-Related and Gender-Specific Disease – A Mini-Review, *Gerontology.* S. Karger AG, 55(1), pp. 3–12.

Port, L., Center, J., Briffa, N. K., Nguyen, T., Cumming, R. and Eisman, J. (2003) Osteoporotic fracture: missed opportunity for intervention, *Osteoporosis International*, 14(9), pp. 780–784.

Ranganathan, P., Aggarwal, R. and Pramesh, C. S. (2015) Common pitfalls in statistical analysis: Odds versus risk, *Perspectives in clinical research.* Medknow Publications & Media Pvt Ltd, 6(4), pp. 222–224.

Roser, M., Ortiz-Ospina, E. and Ritchie, H. (2013) Life expectancy, *Our World in Data.*

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A., Howard, B. V., Johnson, K. C., Kotchen, J. M., Ockene, J. and Writing Group for the Women's Health Initiative Investigators (2002) Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal WomenPrincipal Results From the Women's Health Initiative Randomized Controlled Trial, *JAMA*, 288(3), pp. 321–333.

Schoenbach, V. J. and Rosamond, W. D. (2000) *Understanding the Fundamentals of Epidemiology: an evolving text.* Chapel Hill, North Carolina: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, p. 91.

Shulman, L. P. (2011) The adverse impact of hormone therapy discontinuation on bone health: Initiating a more balanced and comprehensive assessment of the impact and role of postmenopausal hormone therapy, *Menopause*, 18(11), pp. 1152–1153.

Siegmund, D. O. (2018) Probability theory. Encyclopædia Britannica, inc.

Sözen, T., Özişik, L. and Başaran, N. Ç. (2017) An overview and management of osteoporosis, *European journal of rheumatology.* 2016/12/30. Medical Research; Education Association, 4(1), pp. 46–56.