探討非線性關係

醫學研究部 基礎醫學科 生統小組：陳韻伃 博士

授課日期：113年6月18日

探討臨床試驗中的線性關係

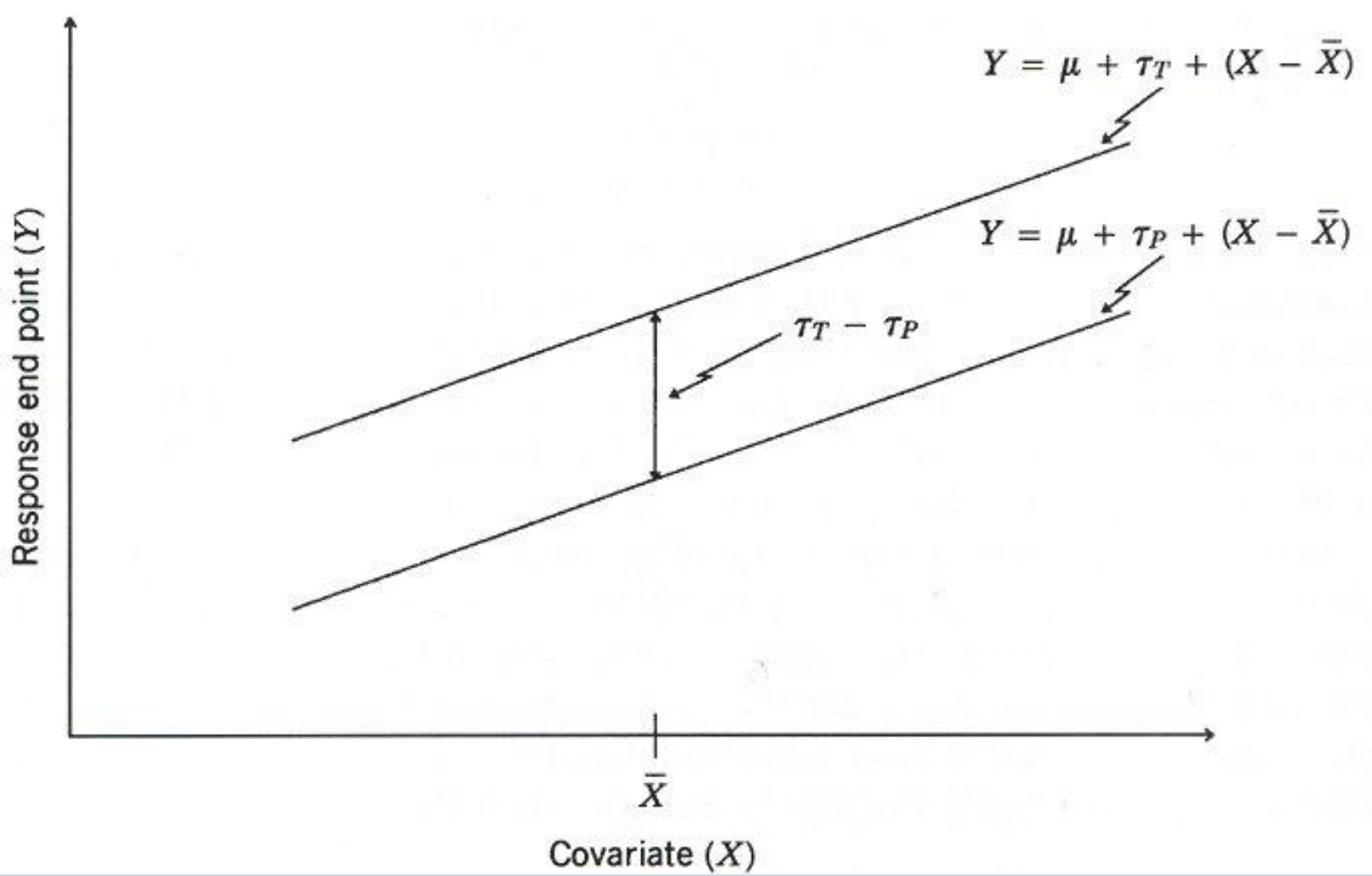


Figure 11.4.1 Adjustment for covariate in estimation of treatment effect.
Case I: **Common slope.**

探討臨床試驗中的線性關係

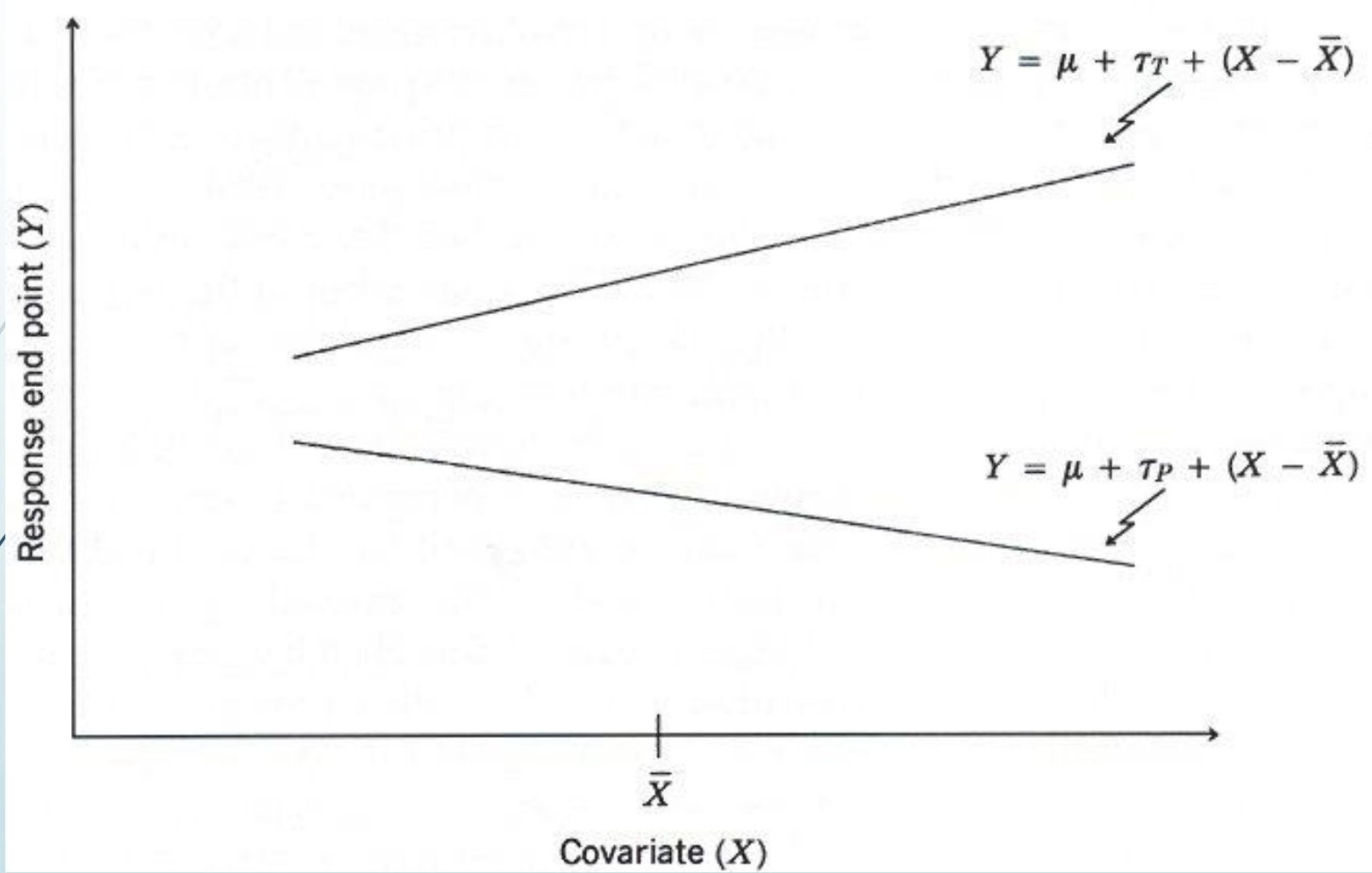


Figure 11.4.2 Adjustment for covariate in estimation of treatment effect.
Case I: **Different slope, different magnitude, same direction**

探討臨床試驗中的線性關係

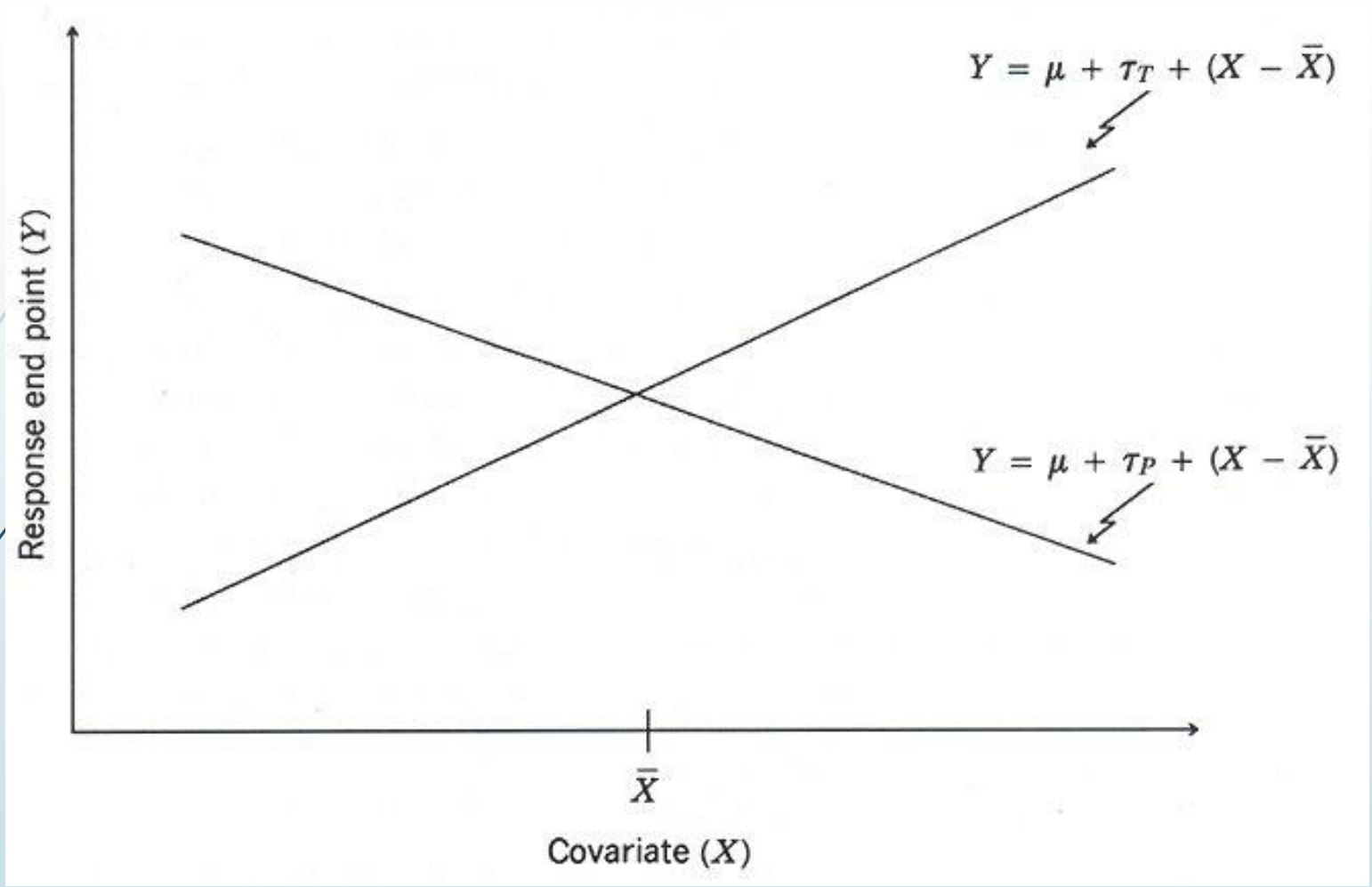
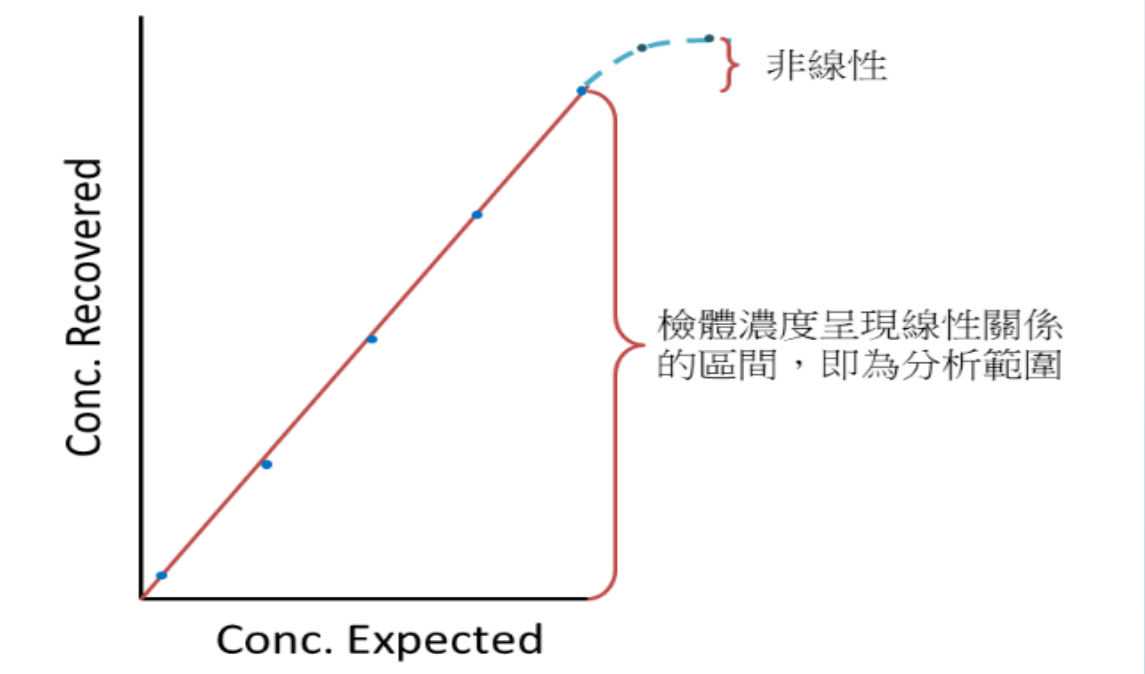


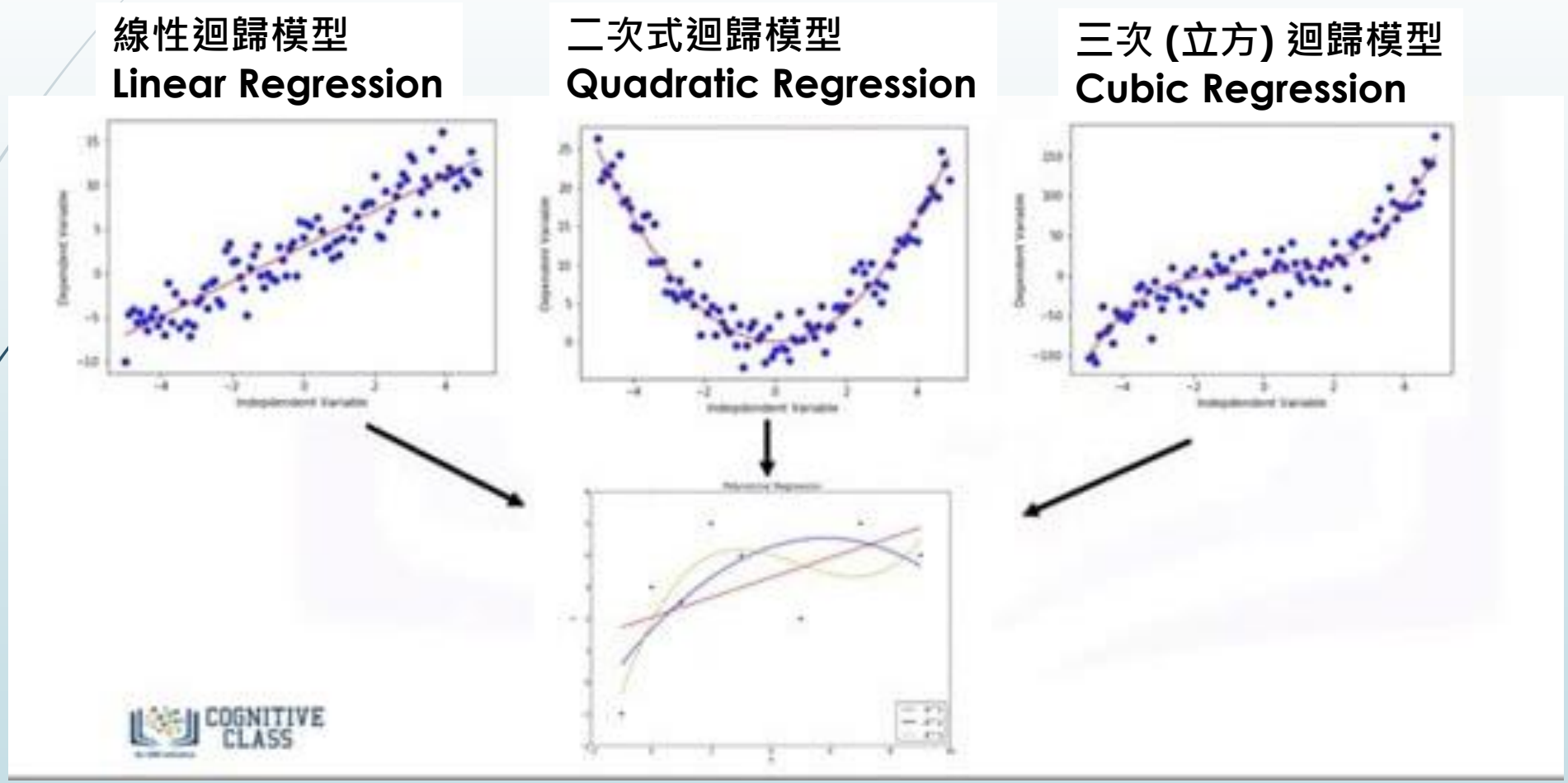
Figure 1 1.4.3 Adjustment for covariate in estimation of treatment effect.
Case III:**Different slopes, different direction with different magnitude.**

臨床試驗的效能評估項目：線性關係 v.s.非線性關係

- 以定量檢驗為例，**實際值與測量值**呈**線性關係**的區間為分析範圍。
- 臨床檢體經稀釋後量測，符合線性關係的區間則稱為**臨床可報告範圍**。
- **測量範圍 (Measuring Interval Range)** 通常以分析範圍 (Analytical Measurement Range, AMR) 或臨床可報告範圍 (Clinical Reportable Range, CRR)表示。



Different Types of Regression



Linear Model v.s. Non-Linear (Smooth) Model

- 傳統迴歸分析模型的重要假設
  - 自變項和依變項呈線性關係，然實際收案資料在分析時甚少能符合此假設條件
  - 若能以模式模擬自變量與依變項之間的非線性關係，將可更忠實呈現自變項和依變項的相關
- 非線性模型
  - **Restricted Cubic Spline (RCS) Model v.s. Fractional Polynomials (FP)**

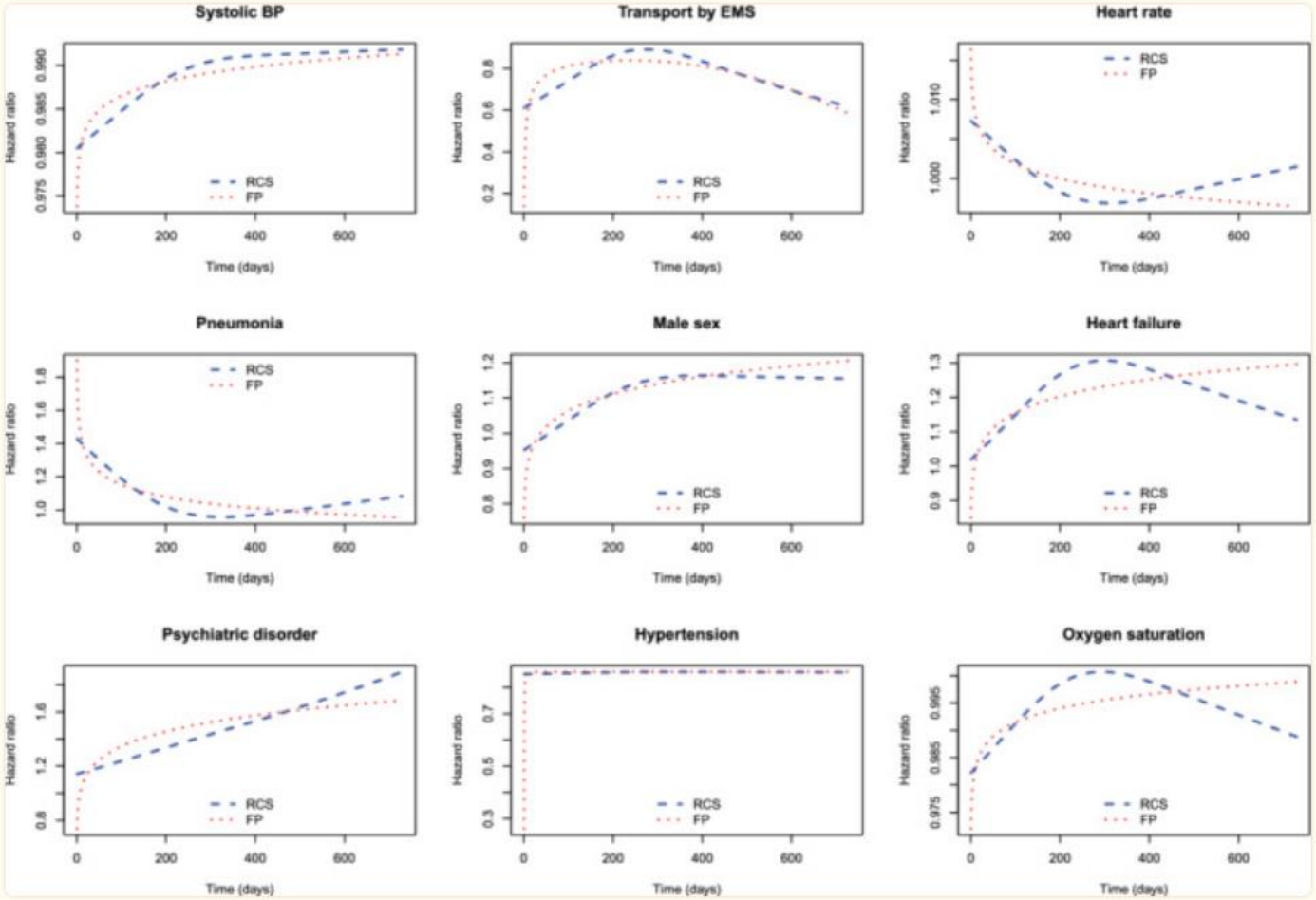
    v.s. Generalized Additive Model (GAM)
  - 可清楚描述自變項與依變項之間的關係（無論線性或非線性關係）
  - 適用於各分析模式（包括線性迴歸、邏輯斯迴歸、存活)

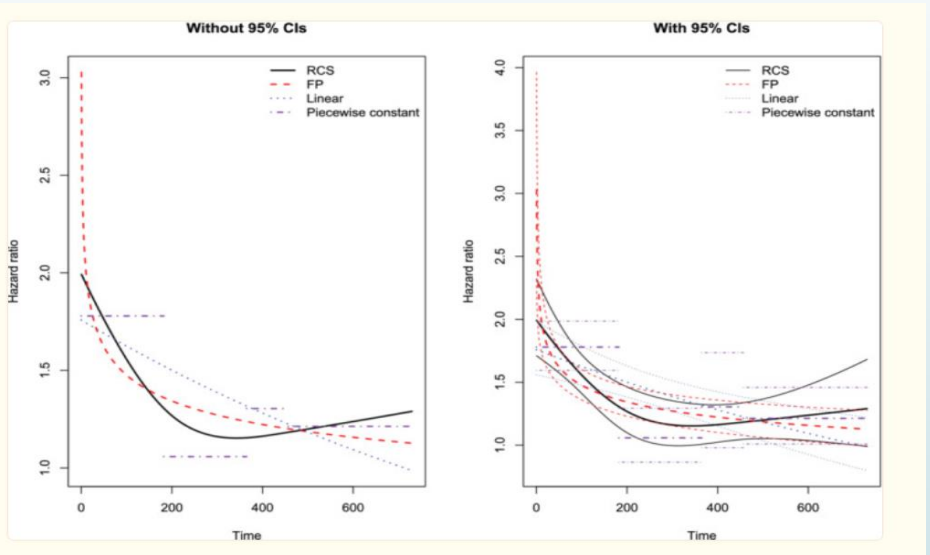
非線性模型的比較

特性	限制性立方樣條（RCS）	分數多項式（FP）	廣義加法模型（GAM）
說明	使用立方平滑函數來近似數據，通過選擇適當的節點和平滑度來擬合數據。	使用多項式函數來近似數據，選擇最適合的多項式次數和係數來擬合數據。	使用平滑函數來建模關係，允許非線性關係的形式未被指定。
非線性建模方式	使用立方平滑函數	使用多項式函數	使用平滑函數
多項式次數選擇	不適用	根據模型配度選擇	不適用
節點數選擇	需要選擇適當的節點	不適用	不適用
適用性	在大型數據集上工作較好	在較小數據集上工作較好	在複雜數據集上工作較好
彈性	較有彈性，可以擬合不同形式的非線性關係	較有彈性，可以擬合不同形式的非線性關係	高度彈性，可以擬合複雜的非線性關係
選擇最佳模型	通過比較不同節點和平滑度來選擇最佳模型	通過比較不同多項式次數選擇最佳模型	通過比較不同平滑函數來選擇最佳模型
適合執行的統計軟體	R, SAS, Stata	R, Stata	R, Python, SAS
SAS 使用的模組	PROC TRANSREG	不適用	PROC GAM
R 使用的模組	splines	mfp	mgcv
Stata 使用的模組	mkspline	fp	不適用

當代主流的非線性關係

- Fractional polynomials (FP)
- Restricted cubic spline (RCS)





Bae *et al. BMC Endocrine Disorders* (2022) 22:123
https://doi.org/10.1186/s12902-022-01041-3

BMC Endocrine Disorders

RESEARCH Open Access

Body mass index at baseline directly predicts new-onset diabetes and to a lesser extent incident cardio-cerebrovascular events, but has a J-shaped relationship to all-cause mortality

Yoon-Jong Bae[1], Sang-Jun Shin[1] and Hee-Taik Kang[2,3*]

Table 2 Cox-proportional hazards regression model for the incidence of diabetes, cardio-cerebrovascular diseases, or all-cause mortality according to body mass index category

		Underweight (BMI < 18.5 kg/m^2)	Normal (18.5–< 23 kg/m^2)	Overweight (23–< 25 kg/m^2)	Grade 1 Obesity (25–< 30 kg/m^2)	Grade 2 Obesity (30–< 35 kg/m^2)	Grade 2 Obesity (≥35 kg/m^2)
Primary Outcomes[a]	**Men**	1.293 (1.224–1.365)	1	1.101 (1.073–1.129)	1.320 (1.288–1.353)	1.789 (1.689–1.897)	2.376 (2.019–2.857)
	Women	1.084 (1.010–1.163)	1	1.150 (1.116–1.185)	1.385 (1.346–1.425)	1.865 (1.725–2.019)	2.472 (2.025–3.028)
Diabetes	**Men**	0.843 (0.717–0.990)	1	1.603 (1.524–1.685)	2.431 (2.321–2.546)	4.784 (4.371–5.237)	5.592 (3.845–6.624)
	Women	0.744 (0.607–0.911)	1	1.624 (1.528–1.725)	2.567 (2.43–2.712)	4.372 (3.991–4.788)	4.808 (3.624–6.378)
CCVDs	**Men**	0.974 (0.894–1.060)	1	1.109 (1.073–1.146)	1.200 (1.162–1.241)	1.351 (1.226–1.489)	1.002 (0.551–1.821)
	Women	0.907 (0.827–0.995)	1	1.099 (1.060–1.140)	1.169 (1.128–1.211)	1.264 (1.158–1.380)	1.462 (1.101–2.215)
All-cause mortality	**Men**	1.710 (1.596–1.832)	1	0.780 (0.744–0.817)	0.730 (0.694–0.767)	0.948 (0.807–1.114)	1.029 (0.613–1.863)
	Women	1.698 (1.524–1.893)	1	0.801 (0.744–0.862)	0.830 (0.773–0.891)	1.123 (0.949–1.331)	1.368 (0.869–2.306)

Adjusted for age, systolic blood pressure, glucose, total cholesterol, tobacco smoking, alcohol intake, physical activity, and economic status

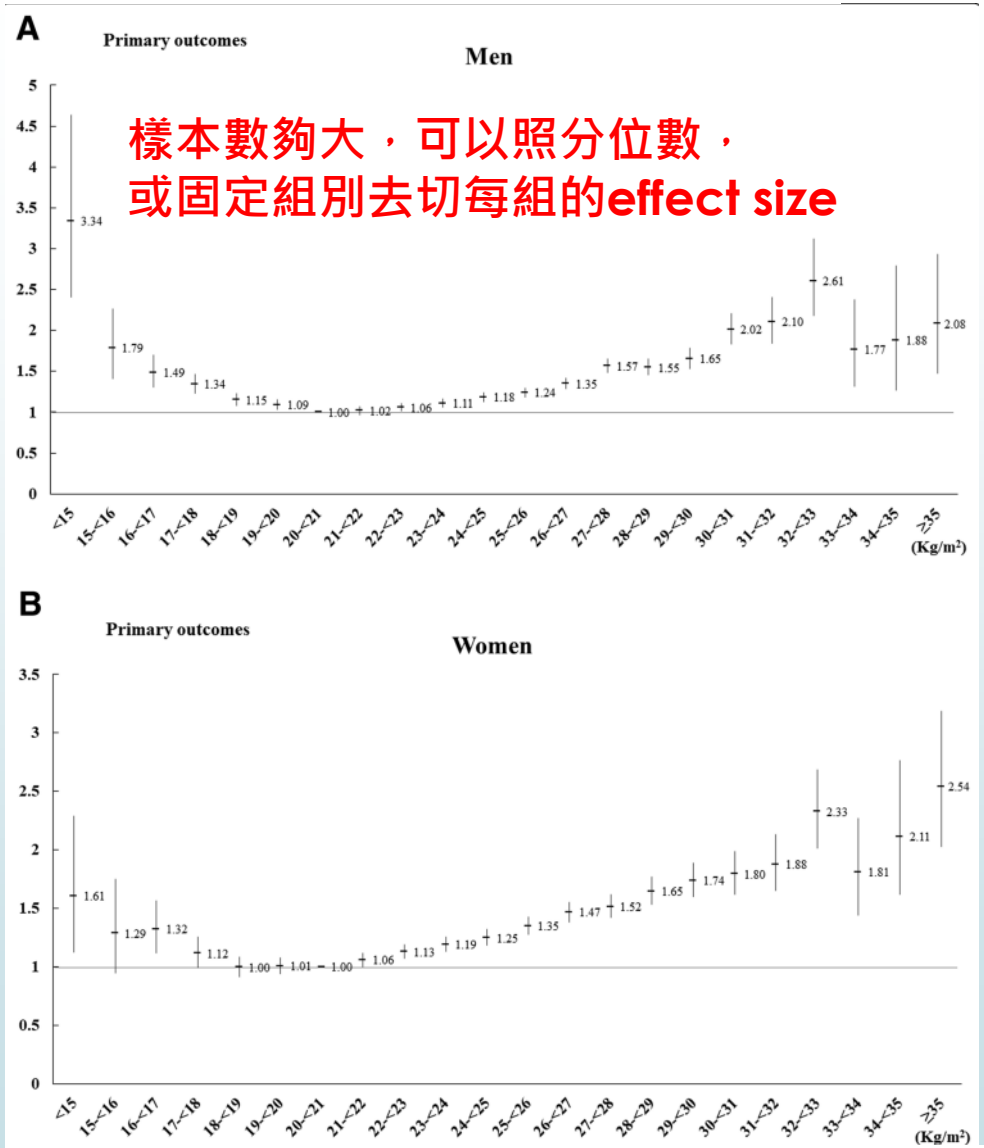
Abbreviation: CCVDs Cardio-cerebrovascular diseases

[a] Primary outcomes include diabetes mellitus, cardio-cerebrovascular diseases, or all-cause mortality

Bae *et al. BMC Endocrine Disorders* (2022) 22:123

Table 1 Baseline characteristics according to sex

	Men	Women
Number (N)	167,500	143,916
Age, years	50.4 ± 8.6	51.4 ± 9.0
Body mass index, kg/m^2	23.8 ± 2.8	23.5 ± 2.9
Systolic blood pressure, mmHg	125.3 ± 15.6	120.5 ± 16.3
Glucose, mg/dL	91.3 ± 12.9	89.5 ± 11.9
Total cholesterol, mg/dL	197.0 ± 36.1	198.1 ± 36.5
Body mass index category, N (%)		
Underweight (< 18.5 kg/m^2)	4137 (2.5)	3861 (2.7)
Normal (18.5–< 23 kg/m^2)	61,916 (37.0)	61,708 (42.9)
Overweight (23–< 25 kg/m^2)	47,585 (28.4)	37,763 (26.2)
Grade 1 Obesity (25–< 30 kg/m^2)	51,220 (30.6)	37,137 (25.8)
Grade 2 obesity (30–< 35 kg/m^2)	2531 (1.5)	3220 (2.2)
Grade 3 obesity (≥ 35 kg/m^2)	111 (0.07)	227 (0.16)
Smoking status, N (%)		
Never smokers	66,415 (39.7)	138,671 (96.4)
Ever smokers	101,085 (60.3)	5245 (3.6)
Drinking status, N (%)		
Rare	55,825 (33.3)	116,095 (80.7)
Sometimes	79,507 (47.5)	25,058 (17.4)
Often	32,168 (19.2)	2763 (1.9)
Physical activity, N (%)		
Rare	83,575 (49.9)	95,767 (66.5)
Sometimes	69,568 (41.5)	36,091 (25.1)
Regular	14,357 (8.6)	12,058 (8.4)
Economic status, N (%)		
Low	27,488 (16.4)	39,624 (27.5)
Middle	55,054 (32.9)	46,155 (32.1)
High	84,958 (50.7)	58,137 (40.4)



樣本數夠大，可以照分位數，
或固定組別去切每組的**effect size**

Fig. 2 Cox-proportional hazards regression models for the primary outcomes (diabetes mellitus, cardio-cerebrovascular diseases, or all-cause mortality) according to body mass index. Adjusted for age, systolic blood pressure, glucose, total cholesterol, tobacco smoking, alcohol intake, physical activity, and economic status

Yuan *et al. Journal of Translational Medicine* (2022) 20:417
https://doi.org/10.1186/s12967-022-03621-2

Journal of Translational Medicine

RESEARCH Open Access

Body mass index, genetic susceptibility, and Alzheimer's disease: a longitudinal study based on 475,813 participants from the UK Biobank

Shiqi Yuan[1†], Wentao Wu[2†], Wen Ma[2†], Xiaxuan Huang[1], Tao Huang[3], MIn Peng[1], Anding Xu[1*] and Jun Lyu[3,4*]

Yuan *et al. Journal of Translational Medicine* (2022) 20:417

使用RCS方法

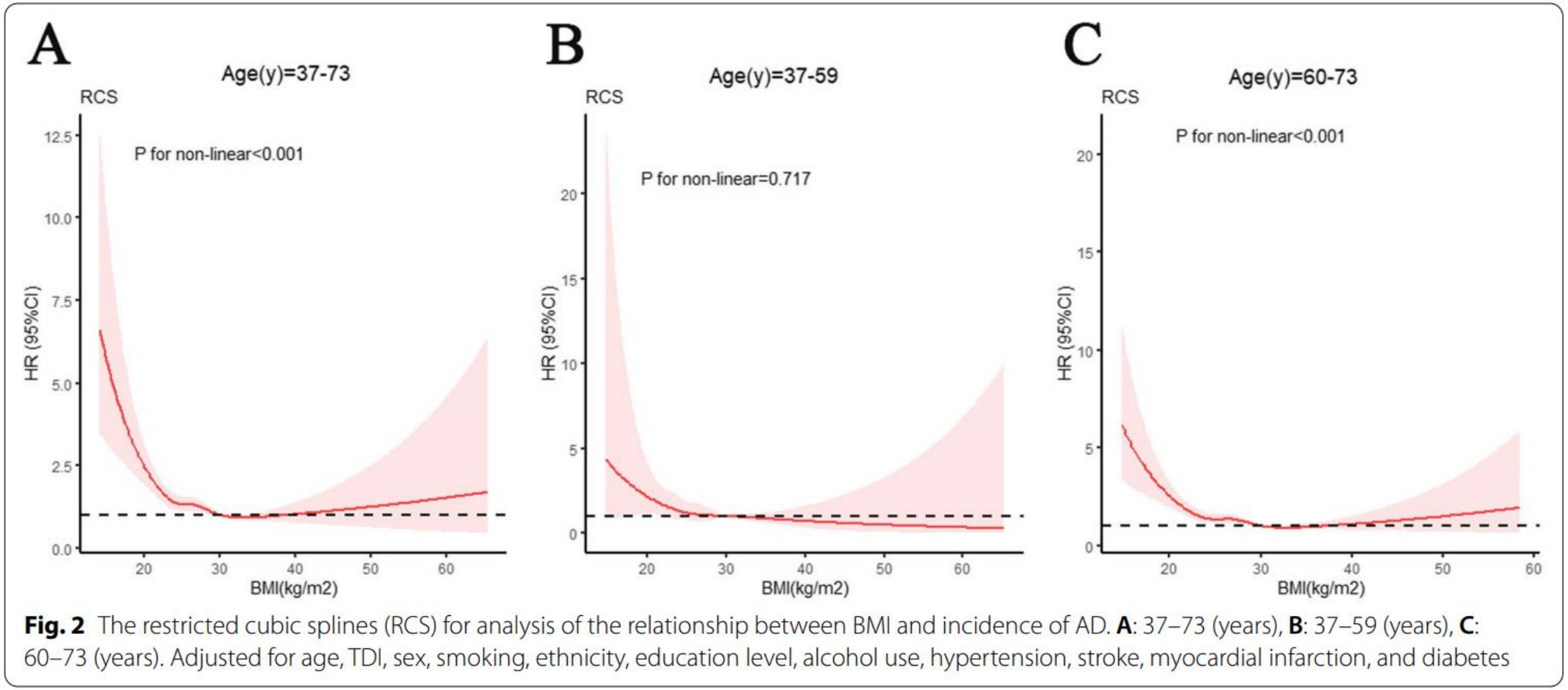


Fig. 2 The restricted cubic splines (RCS) for analysis of the relationship between BMI and incidence of AD. **A**: 37–73 (years), **B**: 37–59 (years), **C**: 60–73 (years). Adjusted for age, TDI, sex, smoking, ethnicity, education level, alcohol use, hypertension, stroke, myocardial infarction, and diabetes

Wong *et al. BMC Medical Research Methodology* 2011, **11**:175
http://www.biomedcentral.com/1471-2288/11/175

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Examining the BMI-mortality relationship using fractional polynomials

Edwin S Wong[1,2*], Bruce CM Wang[1], Louis P Garrison[1], Rafael Alfonso-Cristancho[1], David R Flum[3], David E Arterburn[4] and Sean D Sullivan[1]

使用FP方法

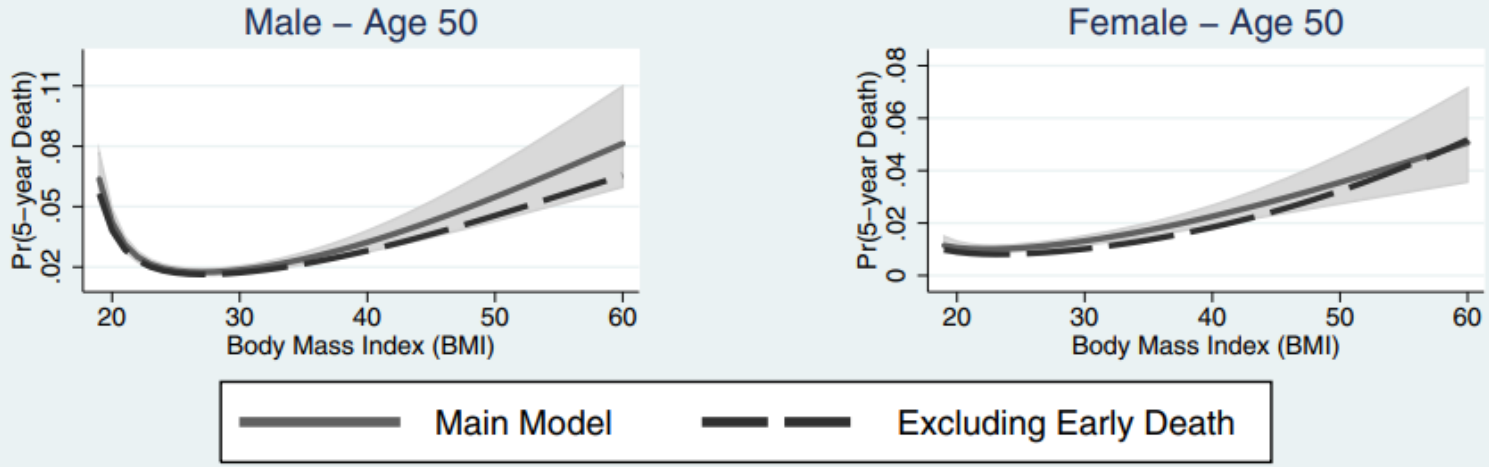
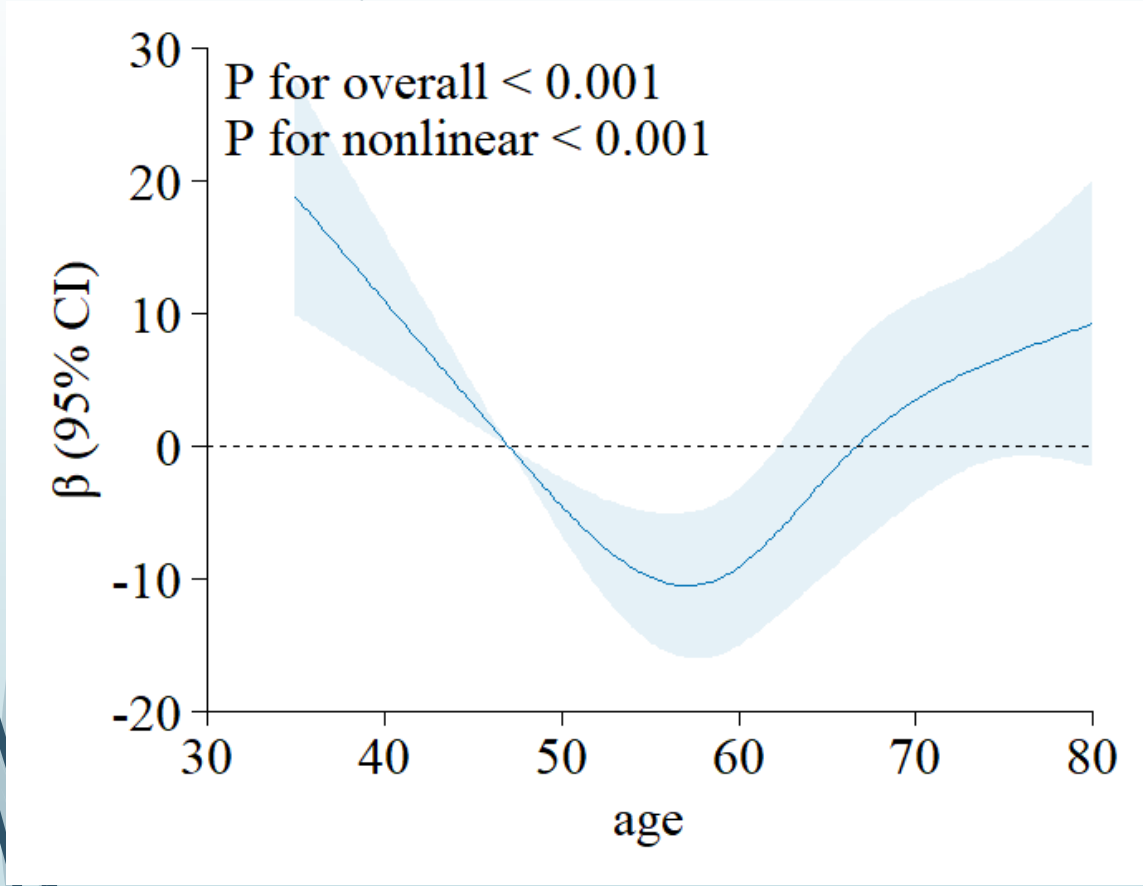


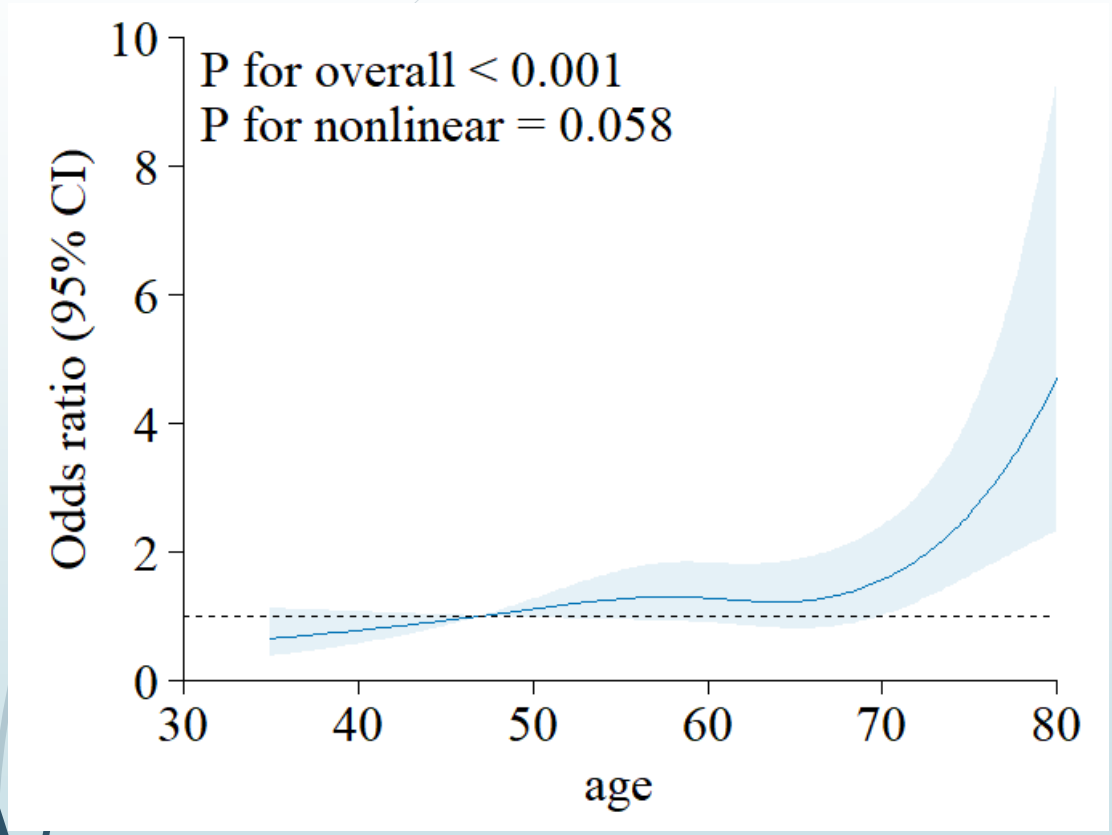
Figure 1 Comparison of the main fractional polynomial model with BMI categorized into narrow bins (top row), the fractional polynomial model excluding extreme BMI values (middle row) and the fractional polynomial model excluding early deaths (bottom row). Shaded regions denote 95% confidence interval for the fractional polynomial model.

RCS Curves for A Linear Regression Model




```
install.packages("plotRCS")
install.packages("rms")
install.packages("ggplot2")
install.packages("readxl")
# install.packages("Hmisc")

library(plotRCS)
library(rms)
library(ggplot2)
# library(Hmisc)


data(cancer)
head(cancer)

# RCS curves for a linear regression model
rcsplot(data = cancer,
outcome = "size",
exposure = "age",
covariates = c("sex", "race", "metastasis"))
```

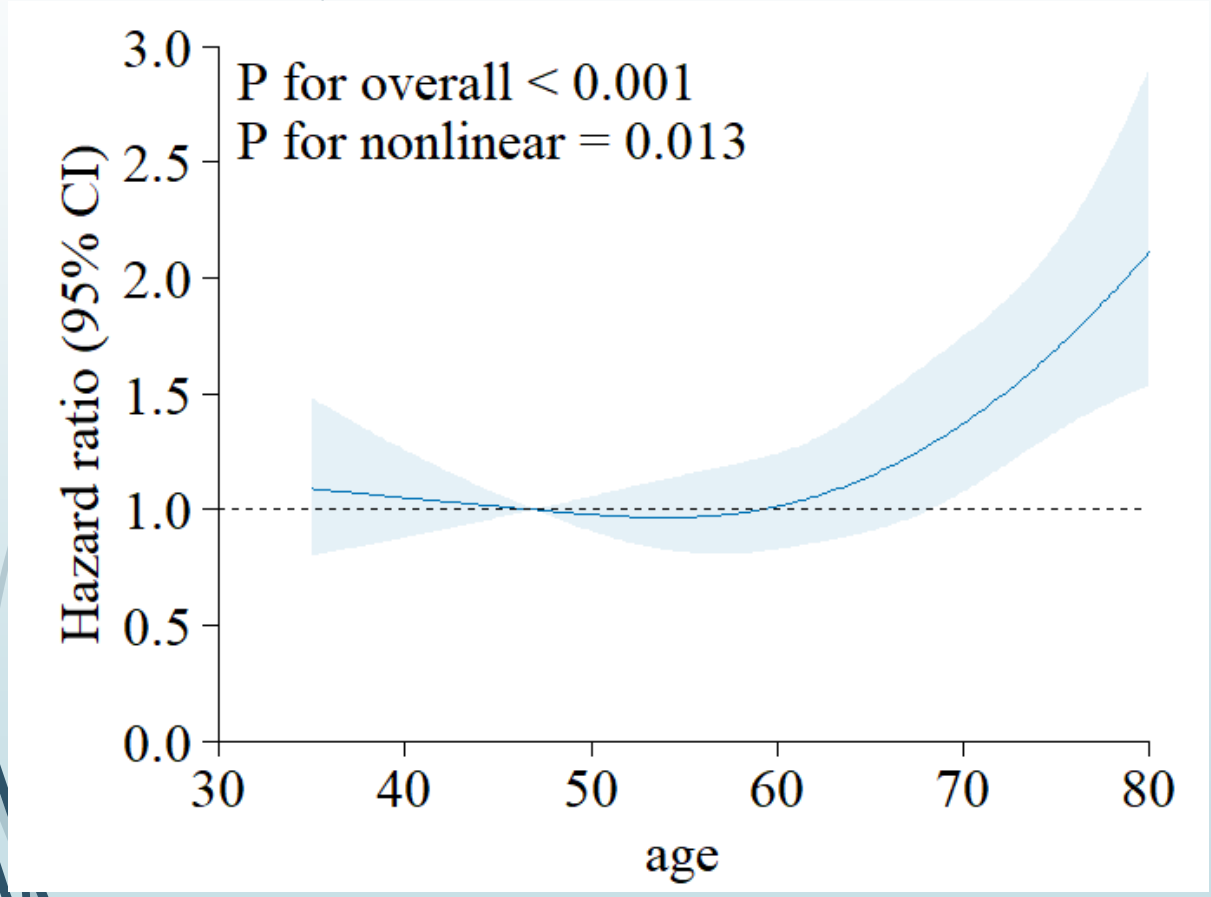
RCS Curves for A Logistic Regression Model




```
library(readxl)

Cancer1 <- read_excel("D:/副研究員/中榮醫研部-生統小組/全院教育課程規劃-2022oct/113年生統課程規劃/20240618-探討非線性關係-Amelia/cancer.xlsx")
```

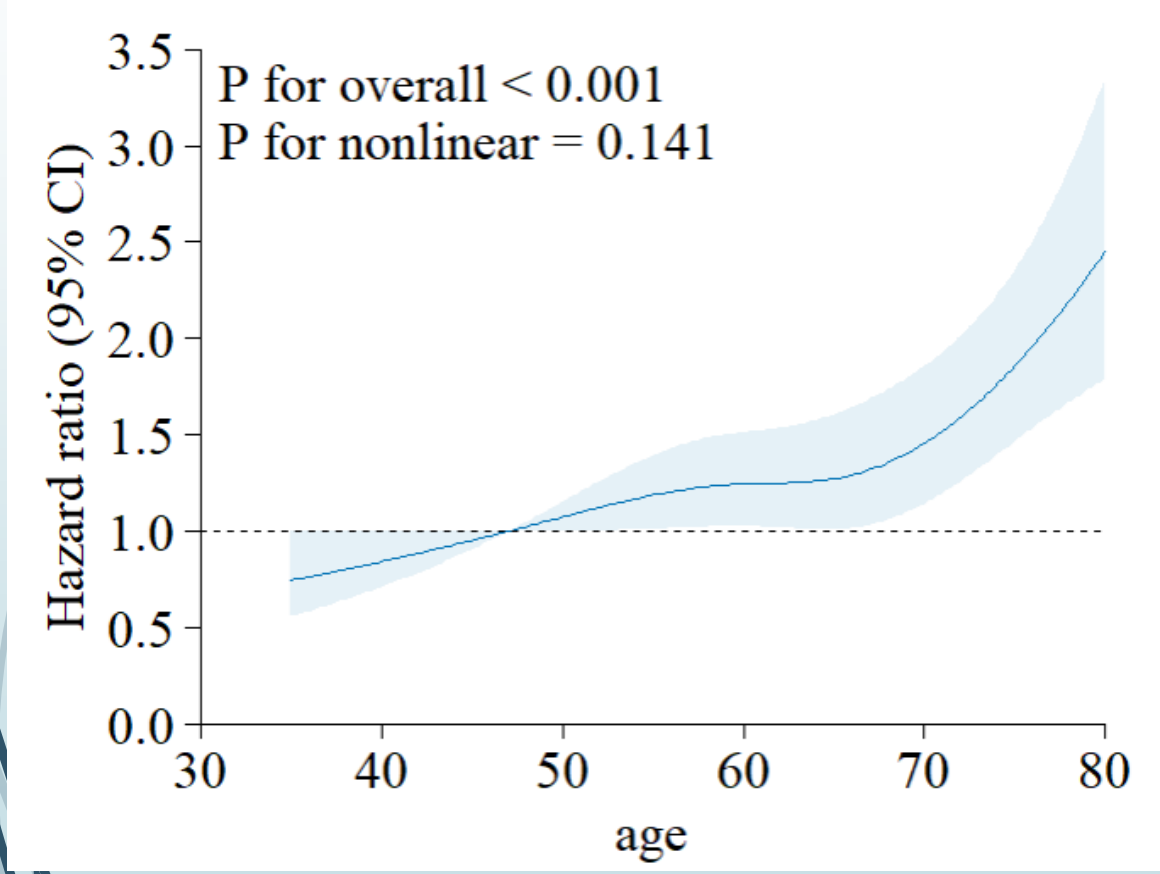
```
library(plotRCS)
library(rms)
library(ggplot2)

data(cancer)
head(cancer)

# RCS curves for a logistic regression model
rcsplot(data = cancer,
outcome = "status",
exposure = "age",
covariates = c("sex", "race", "size", "metastasis"))

# RCS curves for a logistic regression model
rcsplot(data = Cancer1,
outcome = "status",
exposure = "age",
covariates = c("sex", "race", "size", "metastasis"),
positive = "Dead"))
```

Cox Regression: Crude Effect



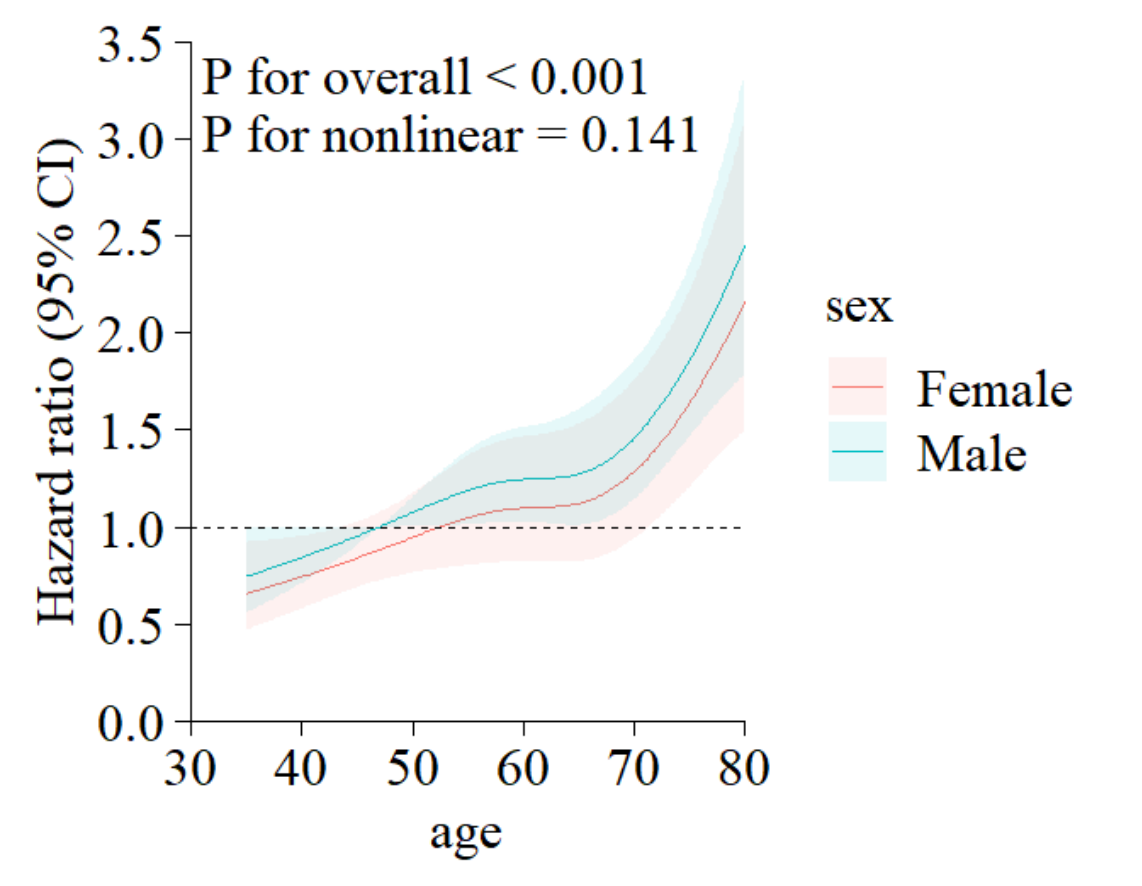

```
library(plotRCS)
library(rms)
library(ggplot2)

data(cancer)
head(cancer)

# Cox regression: Unadjusted covariates
rcsplot(data = cancer,
        outcome = "status",
        time = "time",
        exposure = "age")
```

Cox Regression: Adjusted Effect



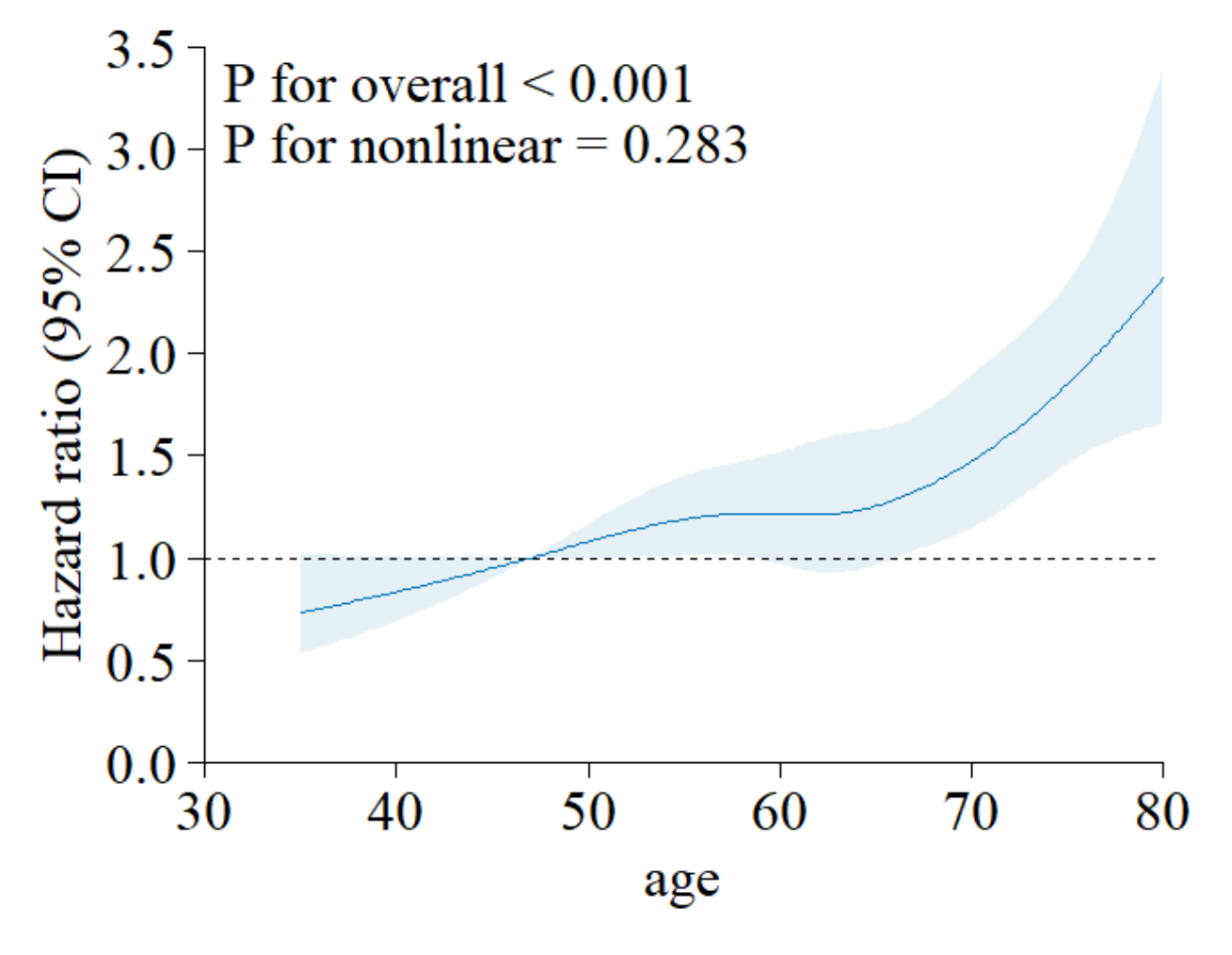

```
library(plotRCS)
library(rms)
library(ggplot2)

data(cancer)
head(cancer)

# RCS curves for adjusted Cox regression model
rcsplot(data = cancer,
outcome = "status",
time = "time",
exposure = "age",
covariates = c("sex", "race", "size", "metastasis"))
```

Cox Regression: Adjusted Effect (by group)



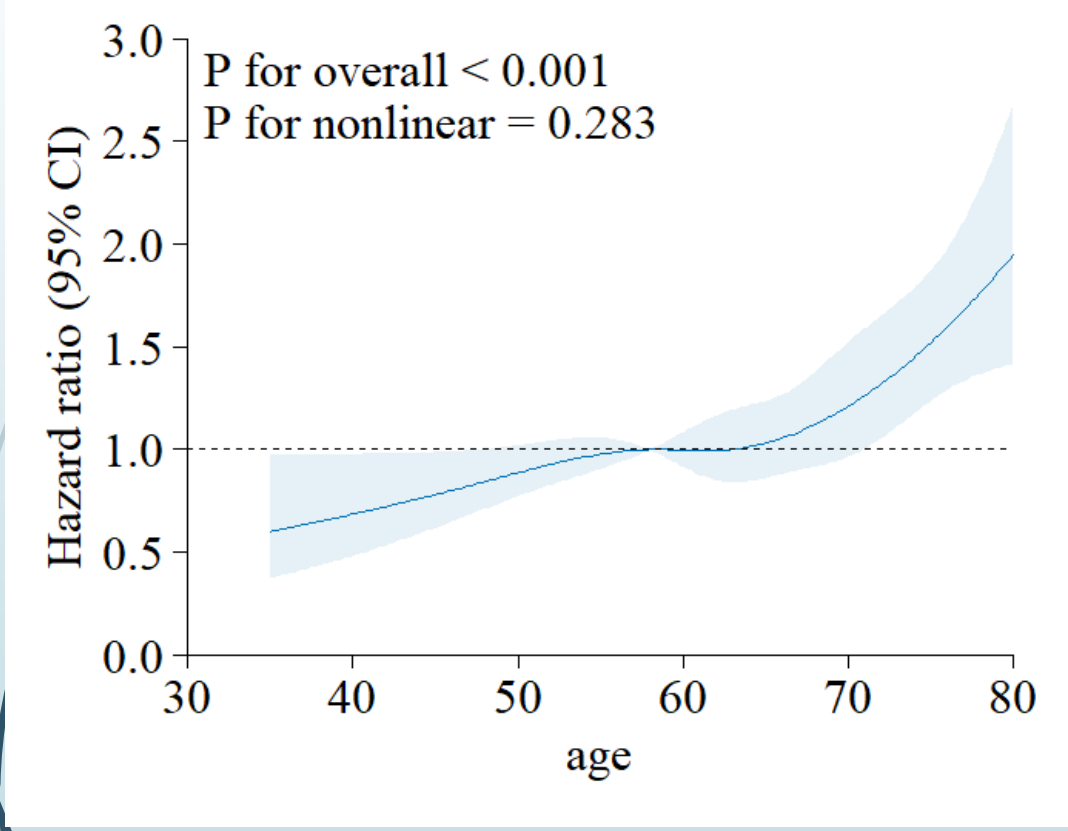

```
library(plotRCS)
library(rms)
library(ggplot2)


data(cancer)
head(cancer)

# By group
rcsplot(data = cancer,
outcome = "status",
time = "time",
exposure = "age",
covariates = c("sex", "race", "size",
"metastasis"),
group = "sex")
```

Cox Regression: Adjusted Effect (by K nots)




```
library(plotRCS)
library(rms)
library(ggplot2)

data(cancer)
head(cancer)

# Set 5 knots from 'knot' function
rcsplot(data = cancer,
outcome = "status",
time = "time",
exposure = "age",
covariates = c("sex", "race", "size",
"metastasis"),
knots = knot(5))
```

Cox Regression: Adjusted Effect (by K nots)




```
library(plotRCS)
library(rms)
library(ggplot2)

data(cancer)
head(cancer)
```

```
# Set the second knot as the reference value
rcsplot(data = cancer,
outcome = "status",
time = "time",
exposure = "age",
covariates = c("sex", "race", "size", "metastasis"),
knots = knot(5), ref.value = "k2")
```

當代主流的非線性關係

- Restricted cubic spline (RCS, R有完整套組)
  - Cubic spline 是指在不同的節點之間，函數可以是三次方的關係
  - 使用 RCS 時，使用「k - 1」個自由度即可（不包含截距項），k 為節點的數量
  - 每個 RCS 只需要估計線性效果 X 以及 k-2個分段 piecewise cubic變項
  - 需要決定兩件事情：第一是節點的數量，第二是節點的位置
  - 節點數量的決定會直接跟樣本數有關係，通常樣本數越小（例如 <100）則建議採取三點節點就好，若樣本數很大（例如上千、上萬）或許最多可以考慮到五個節
  - 大多數的資料使用四個節點應該是合適
  - 節點的位置，若有事先具有臨床意義的切點則直接採用，例如像是血壓的控制會就有國際學會制訂的標準（<90, 91-120, 121-140, 141-160, >160 mmHg），然而大多數的臨床研究都是探索性的，因此極少一開始就已經知道節點要切在哪裡
  - 「節點的位置」對 RCS 結果的影響比較小，反而是「節點的數量」的選擇比較重要

Number of knots K	Knot locations expressed in quantiles of the *x* variable						
3	0.1	0.5	0.9				
4	0.05	0.35	0.65	0.95			
5	0.05	0.275	0.5	0.725	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975

Table 2. Location of knots. From Harrell (2001), Regression Modeling Strategies.

（資料來源：https://support.sas.com/resources/papers/proceedings16/5621-2016.pdf）

使用 R 軟體的 rms 套件的預設節點所在分位數位置

Cox proportional hazards models were used to estimate hazard ratios for cardiovascular and mortality events.

Restricted cubic splines (RCS) under a Cox regression were applied when analyzing time to event data with continuous exposures to "smooth" continuous measurements with a specific outcome.

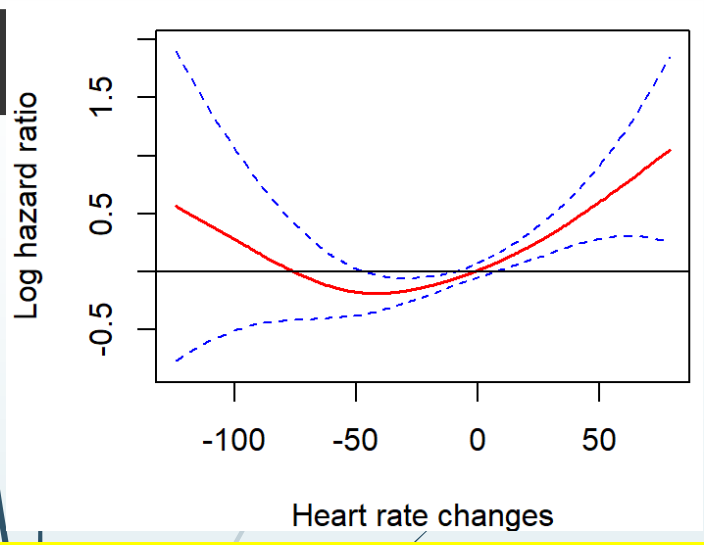
All analyses were performed using R (version 3.6.3)

其他Smooth Model繪圖參考

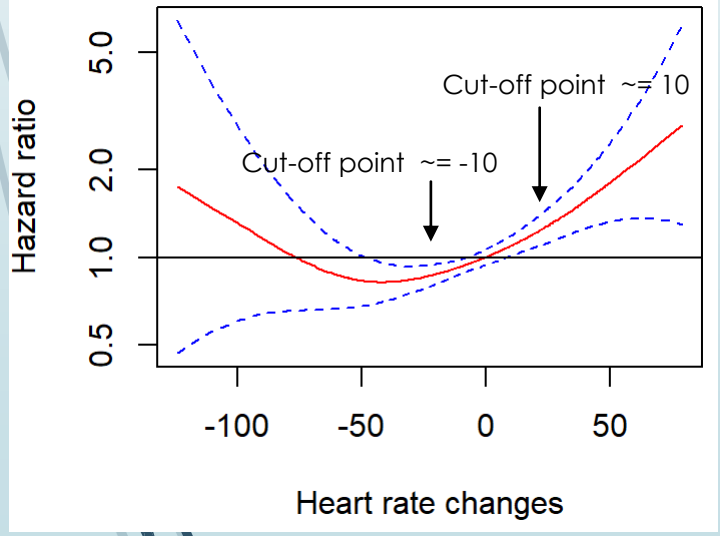
```
#安裝讀檔軟體
install.packages("survival")
install.packages("haven")
install.packages("readxl")

library(survival)
library(haven)
library(readxl)
```

Non-AF: **Model 0**

All deaths



上排是 Log (hazard ratio): ref=0
下排是轉換後製的 Hazard ratio: ref=1



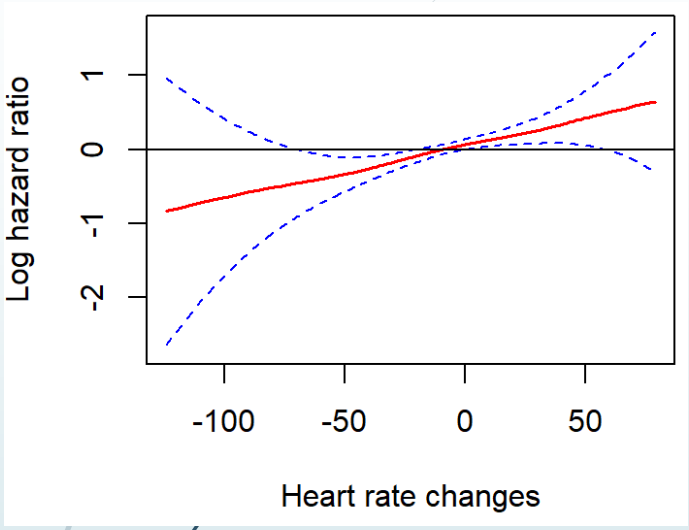

```
HR_HF_nonAF <- read_excel("D:/work/Project 60/HR_vs_HF_nonAF.xlsx")

# 做 Cox regression的 restricted cube smooth model
library(survival)

mfit0 <- coxph(Surv(yrs_all_death, death_all) ~
pspline( hr_change , df=3) , data=HR_HF_nonAF)

termplot(mfit0, term=1, se=TRUE, col.term=2, col.se=4, xlab="Heart rate changes",
ylab="Log hazard ratio")

abline(h=0)

ptemp <- termplot(mfit0, se=TRUE, plot=FALSE)
attributes(ptemp)
HR_base <- ptemp$hr_change # this will be a data fram
center <- with(HR_base, y[x==0])    #log scale with HR 100 as the reference
ytemp <- HR_base$y + outer(HR_base$se, c(0, -1.96, 1.96), '*')
matplot(HR_base$x, exp(ytemp - center), log='y',
type='l', lty=c(1,2,2), col=c(2,4,4),
xlab="Heart rate changes", ylab="Hazard ratio")
abline(h=1)
```

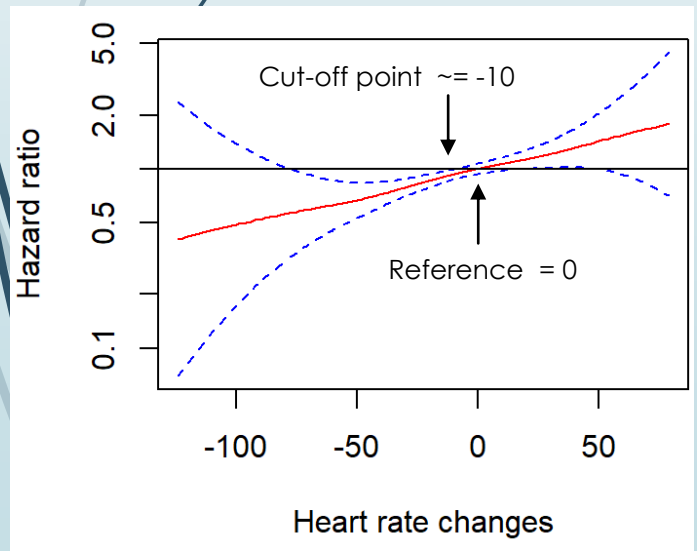
#term是指第幾個變數的圖
#col.se =2 信賴區間的顏色是紅色

HR_HF_nonAF <- read_excel("D:/work/Project 60/HR_vs_HF_nonAF.xlsx")

Non-AF: Model 0

Urgent HF visit






```
# 做 Cox regression的 restricted cube smooth model
library(survival)

mfit0 <- coxph(Surv(yrs_urgent_hf_visit, urgent_hf_er_visit) ~
pspline( hr_change , df=3) , data=HR_HF_nonAF)

termplot(mfit0, term=1, se=TRUE, col.term=2, col.se=4, xlab="Heart
rate changes", ylab="Log hazard ratio")

abline(h=0)

ptemp <- termplot(mfit0, se=TRUE, plot=FALSE)
attributes(ptemp)
HR_base <- ptemp$hr_change # this will be a data fram
center <- with(HR_base, y[x==0])    #log scale with HR 100 as the
reference
ytemp <- HR_base$y + outer(HR_base$se, c(0, -1.96, 1.96), '*')
matplot(HR_base$x, exp(ytemp - center), log='y',
type='l', lty=c(1,2,2), col=c(2,4,4),
xlab="Heart rate changes", ylab="Hazard ratio")
abline(h=1)
```

#term是指第幾個變數的圖
#col.se =2 信賴區間的顏色是紅色

生統小組：統計方法教育訓練

滿意度問卷QR code

探討非線性關係

臺中榮民總醫院 Taichung Veterans General Hospital

Thank you for listening