Statistical Consultation

CANTAB

CANTAB: Computerized Neuropsychiatric Test in Dementia

Group 3 members:

- 流預所 博一 黃煜鈞
- 流預所 碩二 葉憲周
- 流預所 碩二 張宏卿
- 流預所 碩二 余奇祐

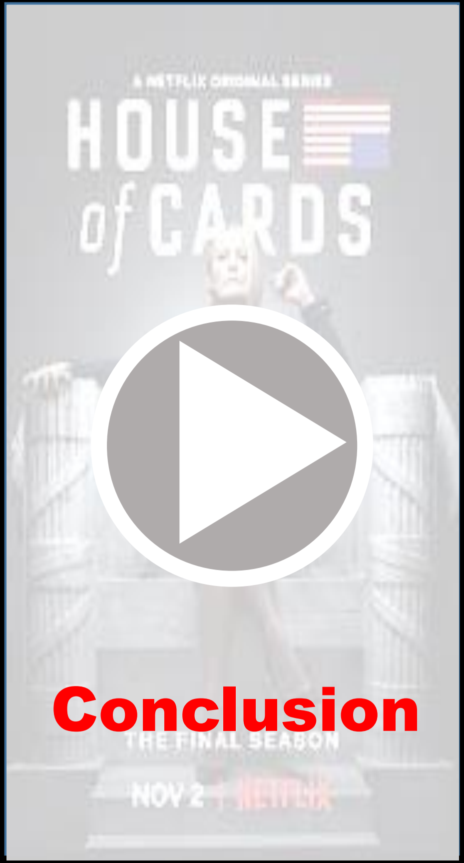
Statistical Data Analysis

Outline


Introduction


Methods


Results

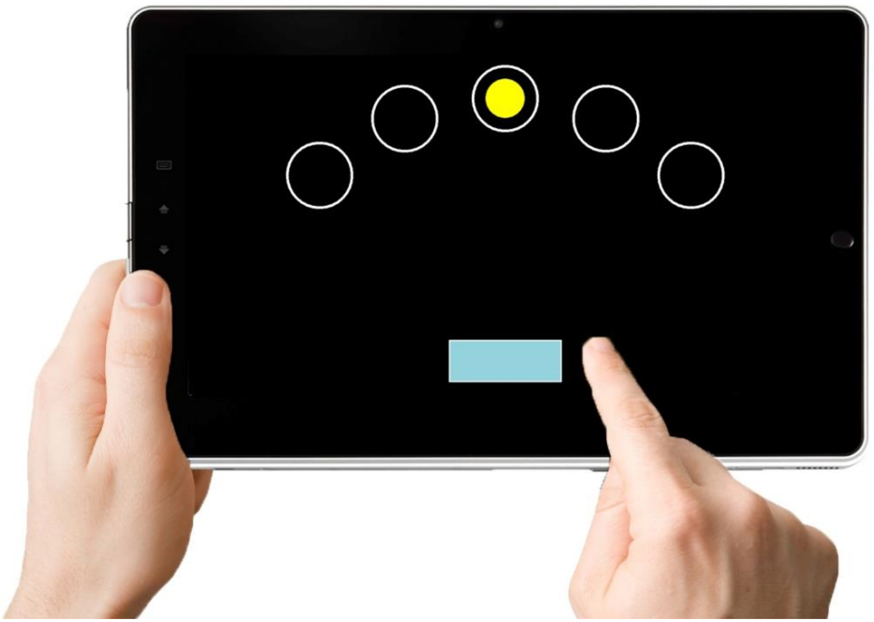


Conclusion


Supplementary

Introduction

What is CANTAB?

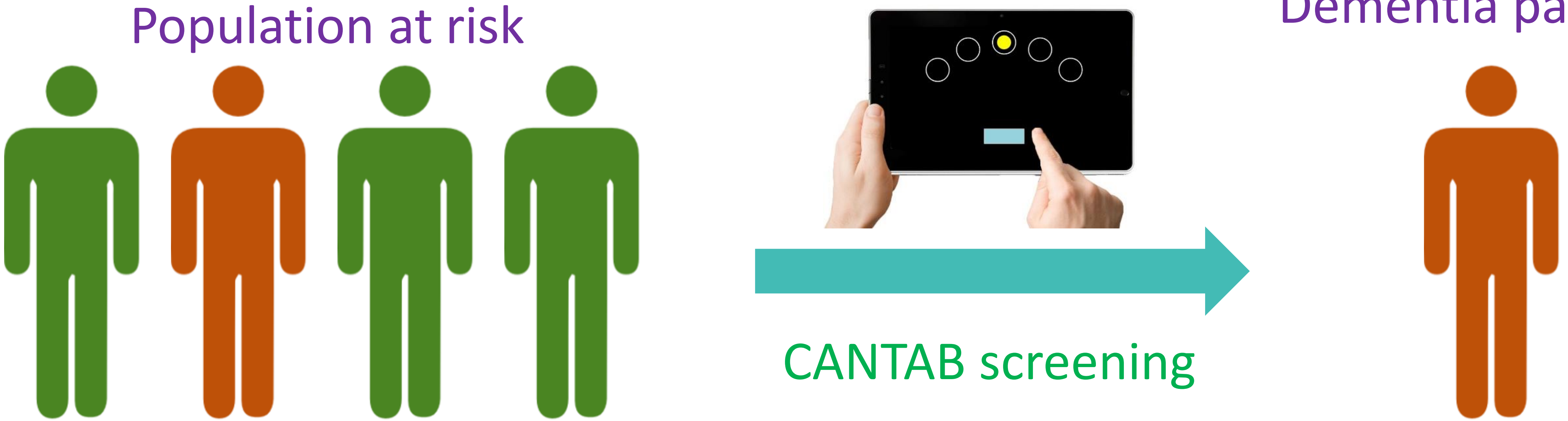


- *Cambridge Neuropsychological Test Automated Battery* (CANTAB)
- Measures of cognitive function
- 8 tests:

Memory				Psychomotor speed/ Attention	Executive function/ Decision making		Social emotional function
DMS	PRM	PAL	VRM	RTI	SSP	MTT	ERT

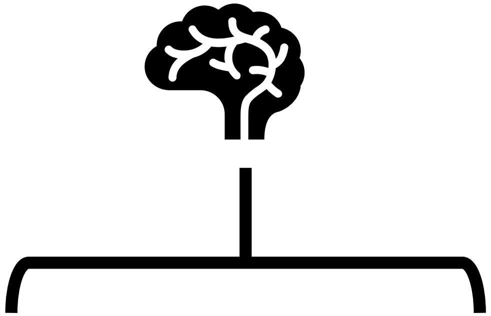
Research Aim

- Early detection of dementia
  - Inefficiency in traditional methods (e.g. MMSE, CDR, NPT)
- Which information from CANTAB is required?
  - Which variables can be used in prediction?



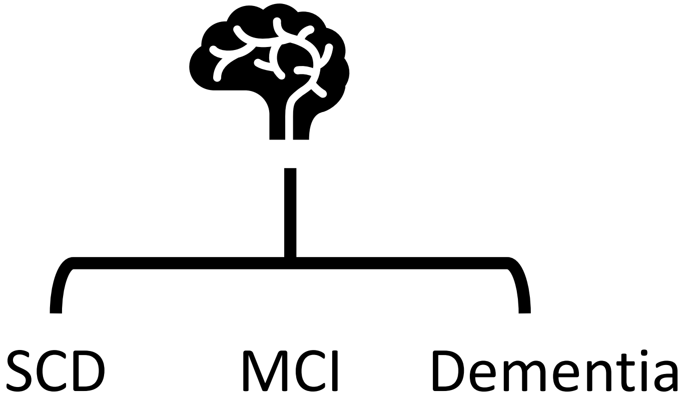
Introduction to data

Response

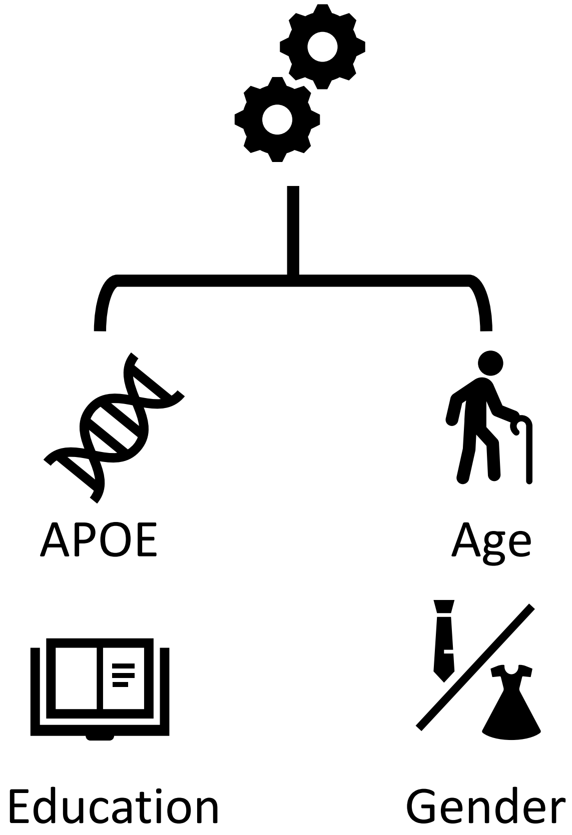


Introduction to data

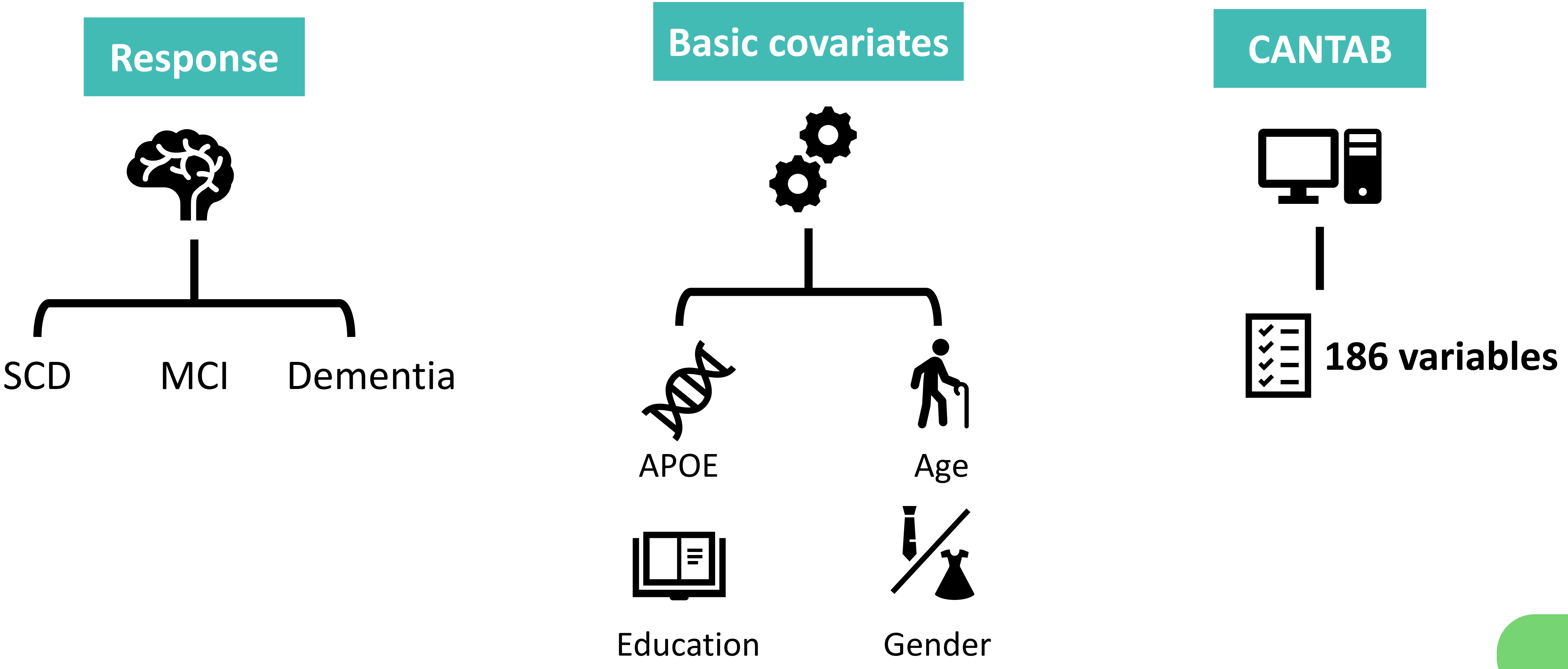
Response



Basic covariates



Introduction to data



Basic covariates

79 samples → 78 samples
- Remove one sample (chart no. 3324878)

- Gender
- APOE
  - Not used in the following analysis
- Age
- Education

Gender

	Female	Male
SCD	23	19
MCI	12	9
Dementia	6	10

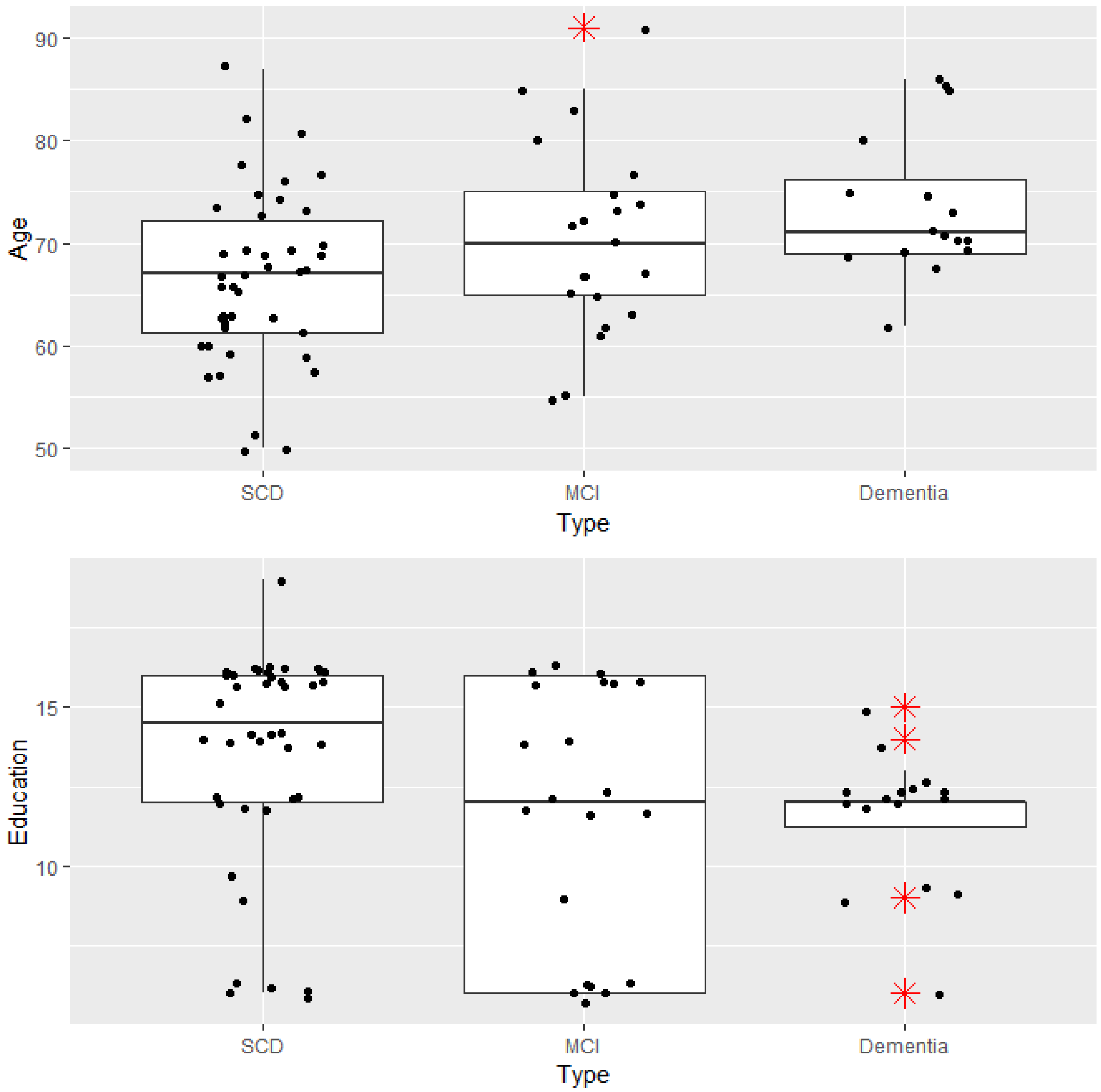
APOE

	E2/E3	E2/E4	E3/E3	E3/E4	E4/E4
SCD	6	1	27	7	1
MCI	2	0	12	7	0
Dementia	3	0	10	3	0

Basic covariates

78 samples

- Age
- Education



Re-define response variable

SCD	MCI	Dementia
	Mild/moderate	Severe
（主觀認知下降）	（輕度知能障礙）	（失智）

Re-define response variable

	Mild/moderate	Severe
SCD	MCI	Dementia
（主觀認知下降）	（輕度知能障礙）	（失智）
42 people	20 people	16 people

Re-define response variable

Mild/moderate

Severe

SCD
（主觀認知下降）

42 people

Normal
(coding=0)

MCI
（輕度知能障礙）

20 people

Dementia
（失智）

16 people

Disease
(coding=1)

Data processing flow chart

Remove one sample with missing values (79 → 78 samples)

Re-define response variable (NL=0; DZ=1)

Remove CANTAB variables with missing values (186 → 153 variables)

First part

Find out important variables

Second part

Focus on PAL, MTT and ERT (153 → 81 variables)

Remove CANTAB variables with high correlation (>0.4 or <-0.4) with Age or Education (81 → 66 variables)

Methods

First Part: Single Variable Selection

Using all available variables as inputs **(153 variables)**

Goal: find the first 5 predictor variables which provide the best prediction

4-fold

Cross-validation

Split into training set and testing set

One variable at a time

single logistic regression

Prediction

Performance measurement: Brier score

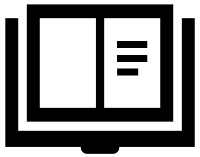
$$\sum (prob_{test} - Obs_{test})^2$$

Methods

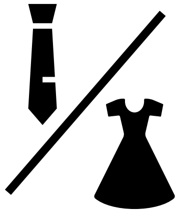
The rule of thumb: $n > 10 \times p$
(sample size 至少要是變數個數的10倍)

我們的 training set 大約有60筆資料(78*3/4)，
建議最多選6個變數放入模型

Age



Education



Gender

+

PALTEA
ERTMDRTH
MTTICE

Model 1

PALNPR
ERTMDRTH
MTTICE

Model 2

PALFAMS
ERTMDRTH
MTTICE

Model 3

Full model 擺放在附錄(詳見P.33)

模型的預測表現擺放在附錄(詳見P.34)

How to use our model

Take model 1 for example

從資料中隨機挑選一個個案來舉例說明 (Chart No. 2446901)

- A new subject: (Age, Gender, Education, PALTEA, MTTICE, ERTMDRTH)
  = (50, Male, 15, 19, 5, 877)

- $log\left(\widehat{\frac{P(Y=1)}{1-P(Y=1)}}\right) = -1.085 - 0.058 \times \mathbf{50} + 0.321 \times \mathbf{1} - 0.033 \times \mathbf{15} + 0.075 \times \mathbf{19}$
  $+0.081 \times \mathbf{5} + 0.0004 \times \mathbf{877} = -1.9782$

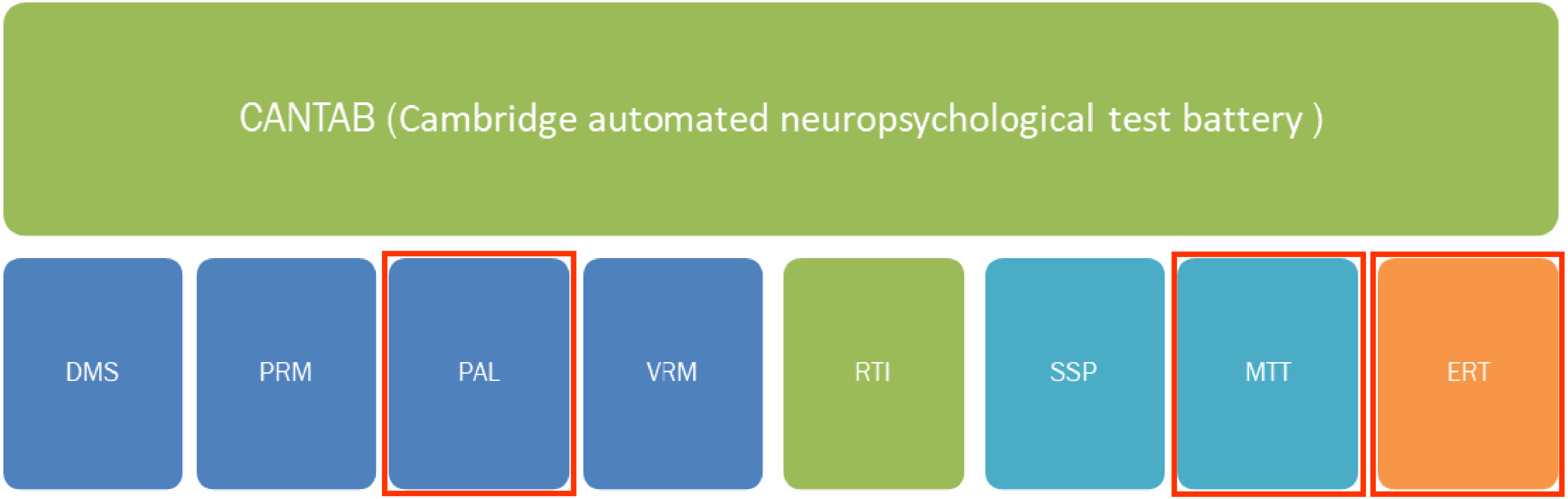
- The probability of the subject to be clinically diagnosed with dementia :

$$\widehat{P(Y=1)} = \frac{e^{-1.9782}}{1+e^{-1.9782}} = 0.1215$$

$\boldsymbol{e \approx 2.718}$

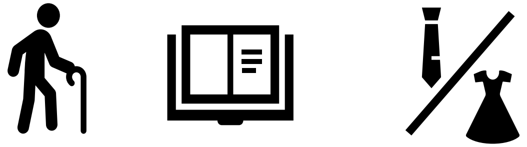
真實資料為normal

More focus on these three tests

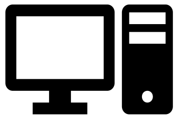


Second Part: Ensemble tree

Covariates + CANTAB variables



Age Education Gender



PAL = 13 variables
MTT = 33 variables
ERT = 20 variables

Logistic regression
(3 covariates as adjustment and one CANTAB variable at a time)

↓ 4-fold cross validation
Replicate 10 times

Brier Score

$$\sum (prob_{test} - Obs_{test})^2$$

↓ Select variables with the smallest 3 Brier scores
Within each category

PALTA8, PALTEA4, PALTEA8,
MTTDBE, MTTMTCM, MTTDE,
ERTOCRTSD, ERTTHD, ERTOMDCRT

Cut-off value investigation

PALTA8, PALTEA4, PALTEA8, MTTDBE, MTTMTCM, MTTDE, ERTOCRTSD, ERTTHD, ERTOMDCRT

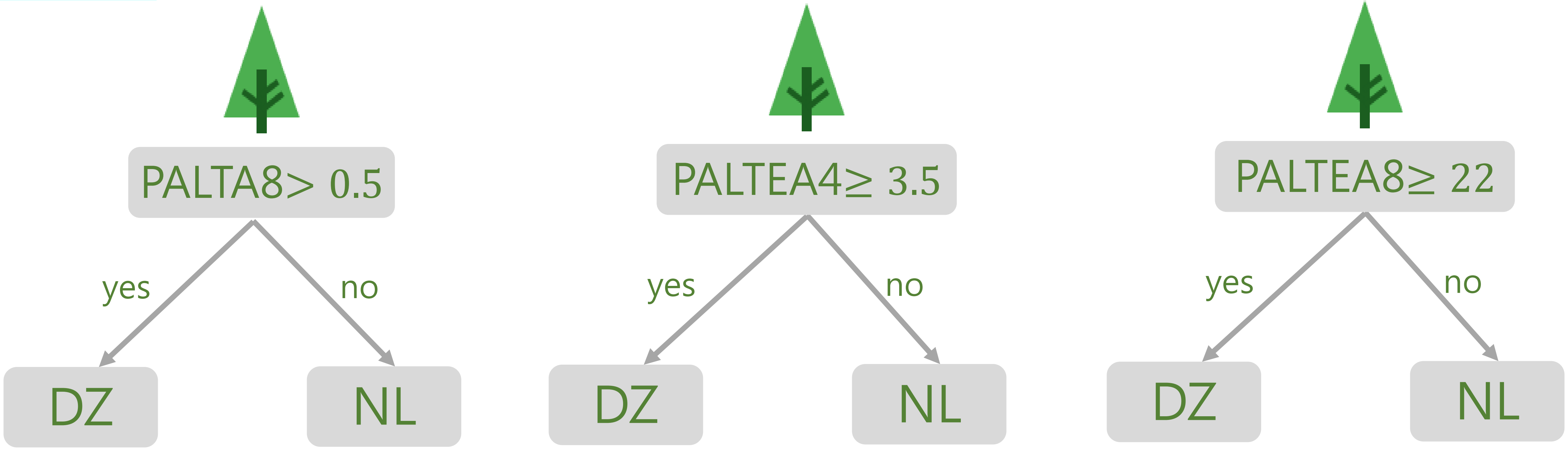
Classification tree
(find out cut-off value for each)

CANTAB variable	DZ	NL
PALTA8	>0.5	<=0.5
PALTEA4	>=3.5	<3.5
PALTEA8	>=22	<22
MTTDBE	>=12	<12
MTTMTCM	<39	>=39
MTTDE	>=21	<21
ERTOCRTSD	>=2427	<2427
ERTTHD	<0.5	>=0.5
ERTOMDCRT	>=1650	<1650

Results

Results

PAL



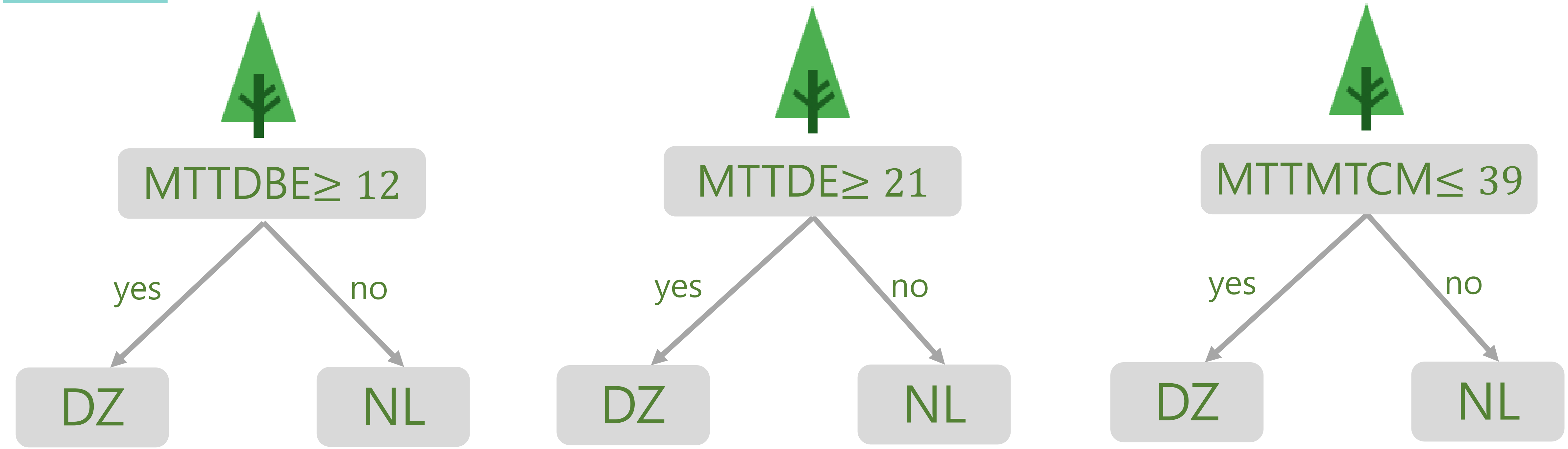
Example: (Chart No. 4160135) --- Dementia

PALTA8 = 0	PALTEA4 = 12	PALTEA8 = 28
NL	DZ	DZ

→ DZ

Results

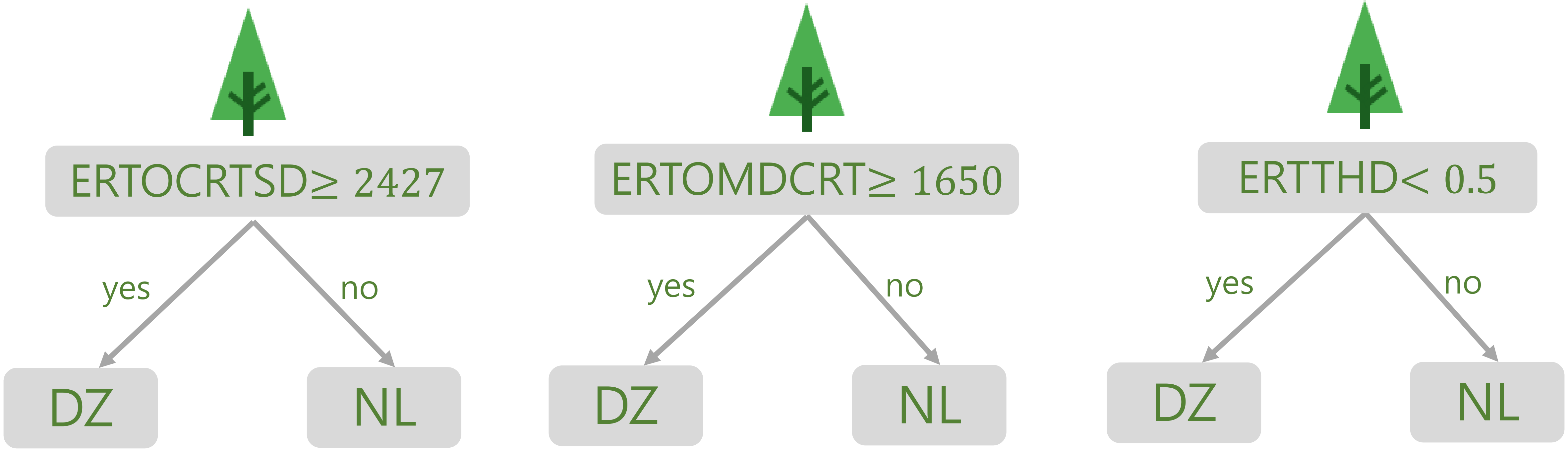
MTT



MTTDBE≥ 12
yes
no
DZ
NL
MTTDE≥ 21
yes
no
DZ
NL
MTTMTCM≤ 39
yes
no
DZ
NL

Results

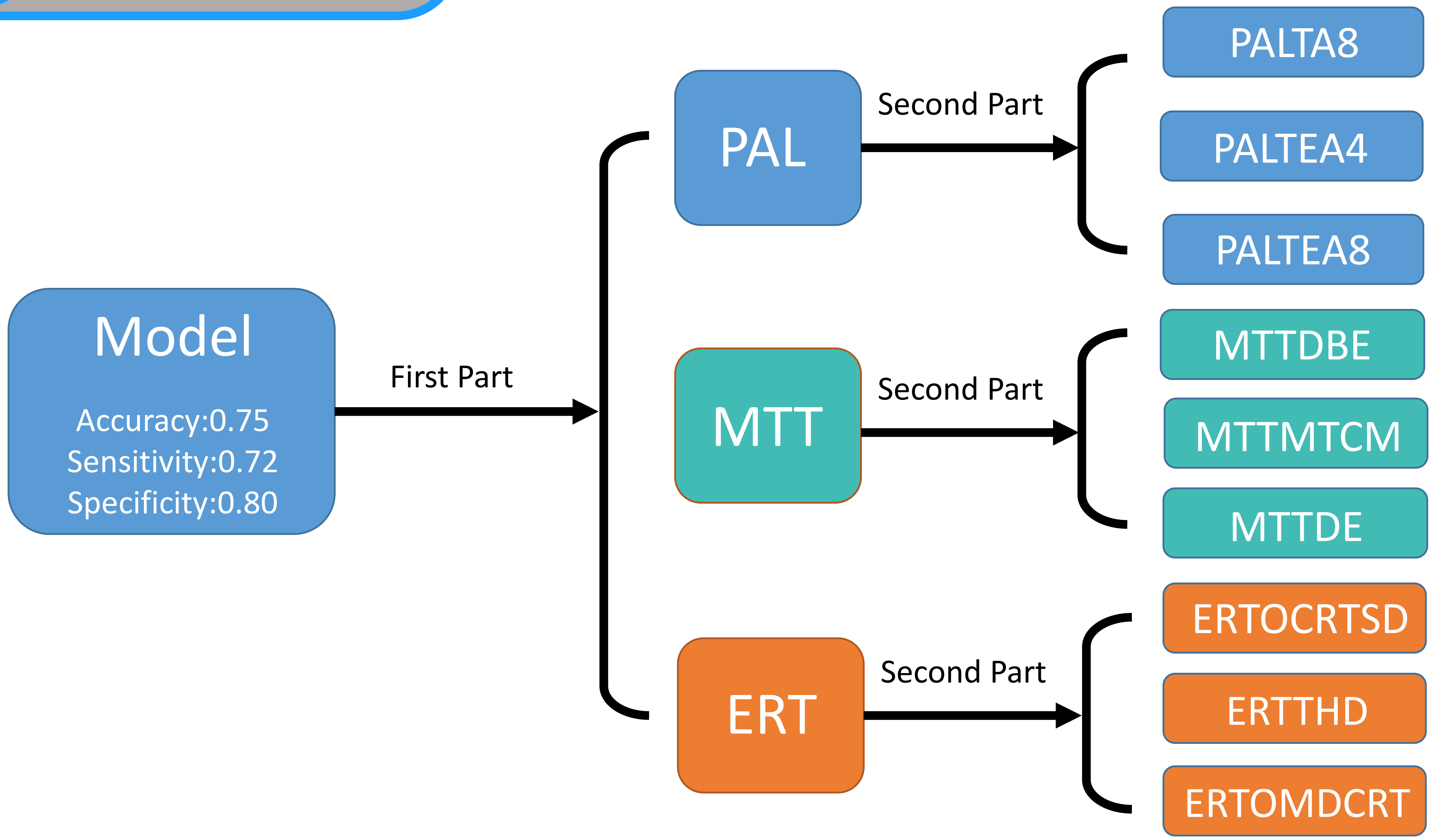
ERT



ERTOCRTSD≥ 2427
yes
no
DZ
NL
ERTOMDCRT≥ 1650
yes
no
DZ
NL
ERTTHD< 0.5
yes
no
DZ
NL

Conclusion

Conclusion


Model
Accuracy:0.75
Sensitivity:0.72
Specificity:0.80
First Part
PAL
Second Part
PALTA8
PALTEA4
PALTEA8
MTT
Second Part
MTTDBE
MTTMTCM
MTTDE
ERT
Second Part
ERTOCRTSD
ERTTHD
ERTOMDCRT

THE END

Supplementary

Supplementary

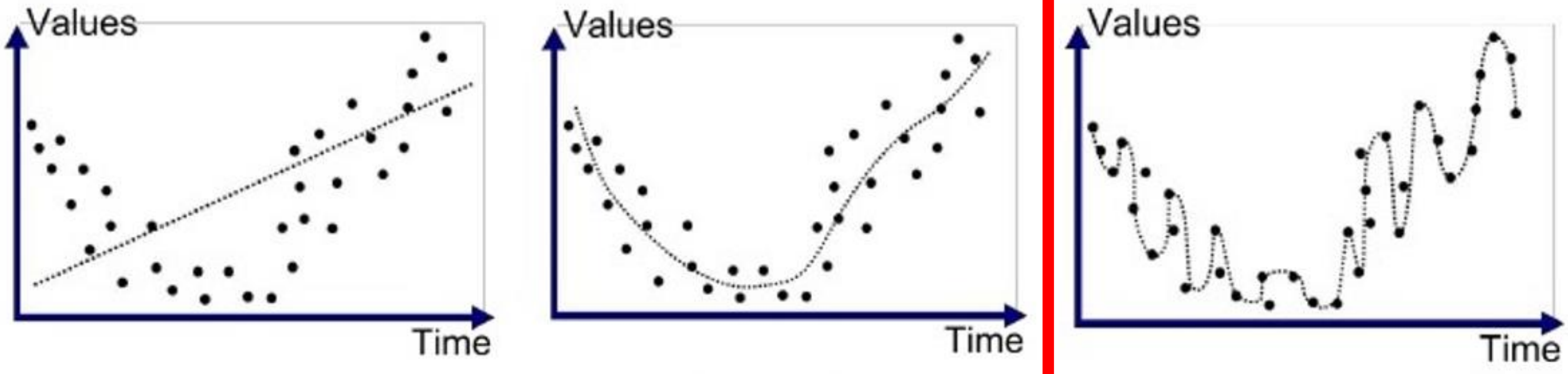
Model Evaluation

- **如何衡量一個模型(model)的好壞?**
  - 用各種衡量的指標，例如正確率(accuracy rate)、敏感度、特異度
  - 選擇一個你覺得最重要的指標就好
  - 不可能每個指標都好，在統計的世界，魚與熊掌不可兼得，其中一個好，通常另外的指標就會有些相對較差
- **正確率越高越好?**
  - 只答對一半而已!為什麼?
  - 如果只有考慮這小小的資料，可以達到非常高的正確率，就會發生過度配適(overfitting)的狀況

What is overfitting?

- 什麼是過度配適?

針對手中現有的資料，
為了達到最高的正確率
而選擇沒有彈性的模型

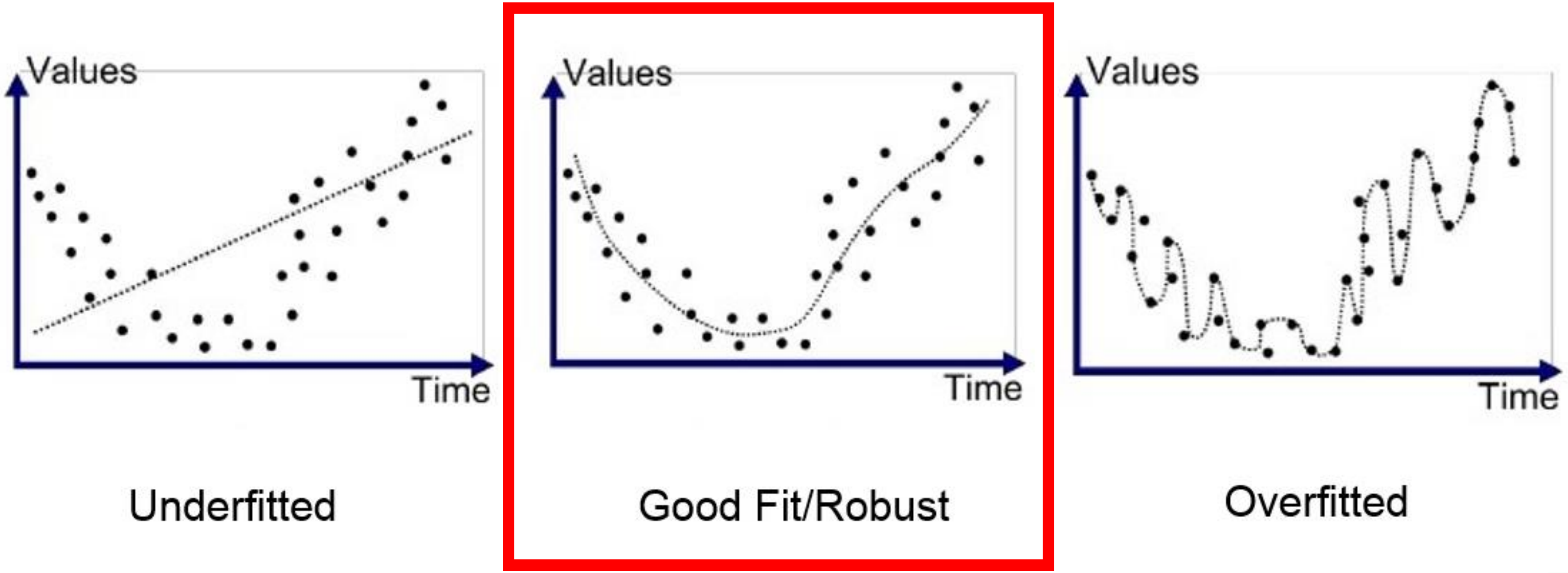


Underfitted | Good Fit/Robust | Overfitted

What is overfitting?

- 什麼是過度配適?

具有彈性的模型，可以幫助外推作更好的預測，所以黑色虛線，並沒有追求和每個點連線，這個就是**容忍誤差的彈性**



Model Evaluation

- **所以要怎麼樣才能衡量模型好壞，而且又確保模型有彈性可外推執行更好的預測呢?**
  - 利用外部的資料，也就是請醫生再額外蒐集樣本，讓模型預測看看
    (這個通常很困難，因為沒有那麼多時間和金錢)
  - 資料切割(data splitting)
    - Resample
    - Holdout sets
    - K-fold cross validation (K折交叉驗證)
    - Leave-one-out cross validation (留一驗證) (為K折交互驗證的特例)
- **選擇使用K折交叉驗證**
  - 相對穩健且安全的做法

Supplementary

K-fold Cross Validation

- 例如四折交叉驗證

Part 1 Training | Part 2 Training | Part 3 Training | Part 4 Testing

其中三等份作為訓練資料集(training set)，先拿來建立模型

Model

再拿剩下的一等份作為測試資料集(testing set)，丟入模型預測，並得到一個正確率P_1

Supplementary

K-fold Cross Validation

- 例如四折交叉驗證
  - 重複同樣的方式，但是是針對**Part 3作為測試資料集**

Part 1 Training | Part 2 Training | Part 4 Training | Part 3 Testing

其中三等份作為訓練資料集(training set)，先拿來建立模型

Model

再拿剩下的一等份作為測試資料集(testing set)，丟入模型預測，並得到一個正確率P_2

K-fold Cross Validation

- 完成四次的交叉驗證後，得到了四個正確率
  - 通常取平均作為這個模型的交叉驗證正確率

$$Accuracy\ rate = \frac{(P_1 + P_2 + P_3 + P_4)}{4}$$

- 比較不同模型的預測力好壞

	Model A	Model B	Model C
交叉驗證正確率:	60%	80%	75%

Full Model

$$log(\frac{\widehat{P(Y=1)}}{1-P(Y=1)})$$
$$= -1.085 - 0.058\,age + 0.321\,Gender - 0.033\,Edu + 0.075\,PALTEA + 0.081\,MTTICE + 0.0004\,ERTMDRTH$$

$$log(\frac{\widehat{P(Y=1)}}{1-P(Y=1)})$$
$$= 7.4 - 0.062\,age + 0.393\,Gender - 0.057\,Edu - 0.825\,PALNPR + 0.086\,MTTICE + 0.0006\,ERTMDRTH$$

$$log(\frac{\widehat{P(Y=1)}}{1-P(Y=1)})$$
$$= 2.168 - 0.044\,age + 0.213\,Gender - 0.011\,Edu - 0.272\,PALFAMS + 0.076\,MTTICE + 0.0007\,ERTMDRTH$$

Predictive performance

	Model 1	Model 2	Model 3
Accuracy	0.764	0.753	0.753
Specificity	0.81	0.802	0.791
Sensitivity	0.717	0.706	0.717

Results

PAL

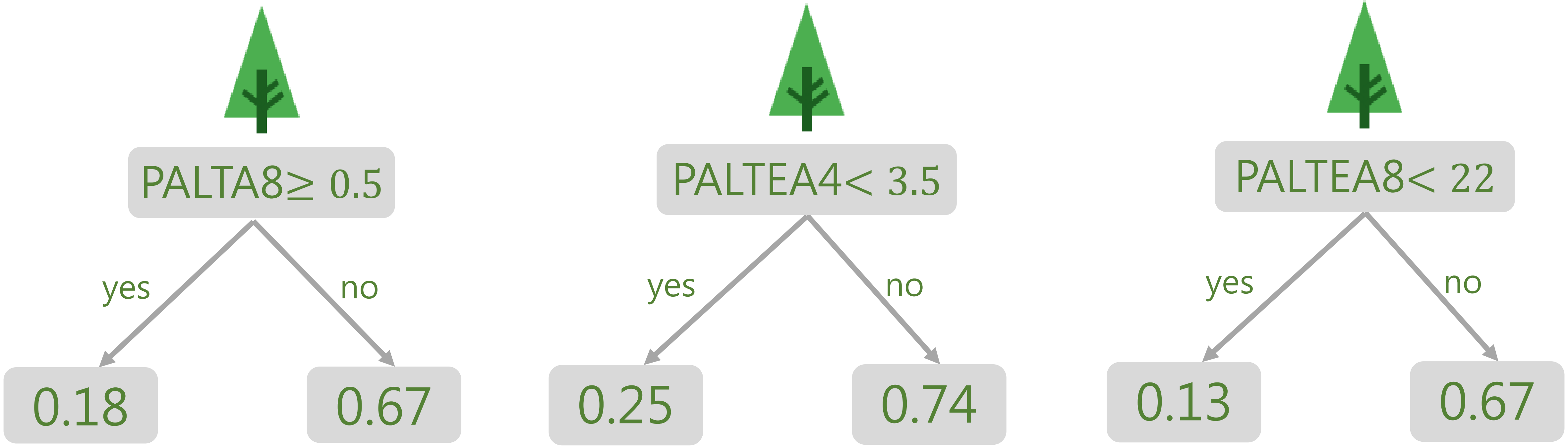
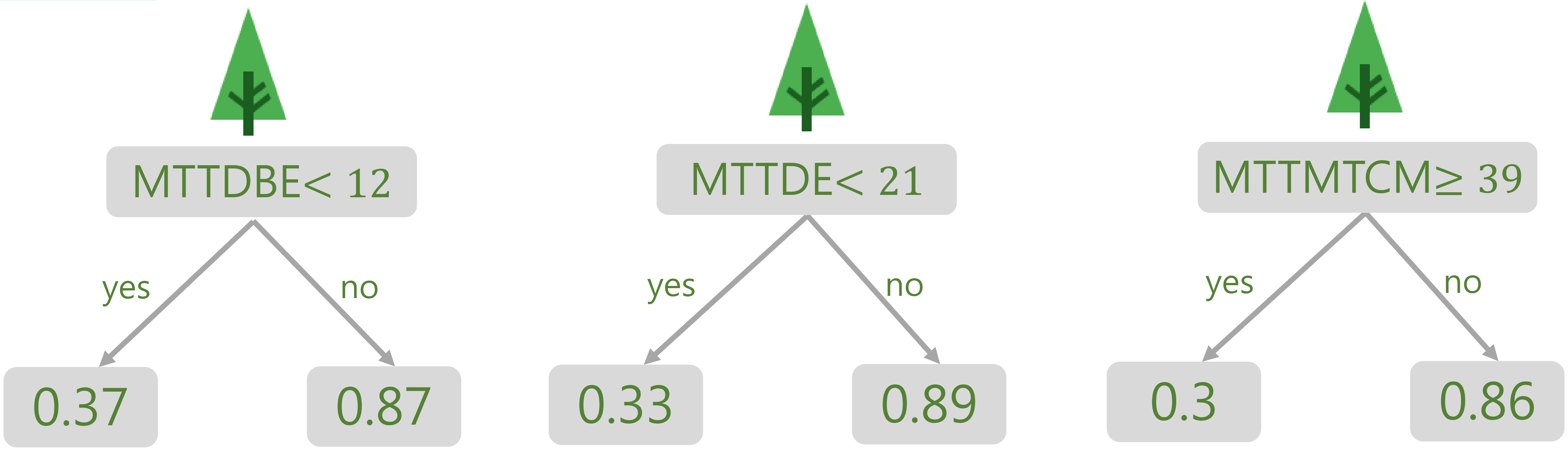


Chart No. 8044822

PALTA8 = 0   PALTEA4 = 4   PALTEA8 = 28

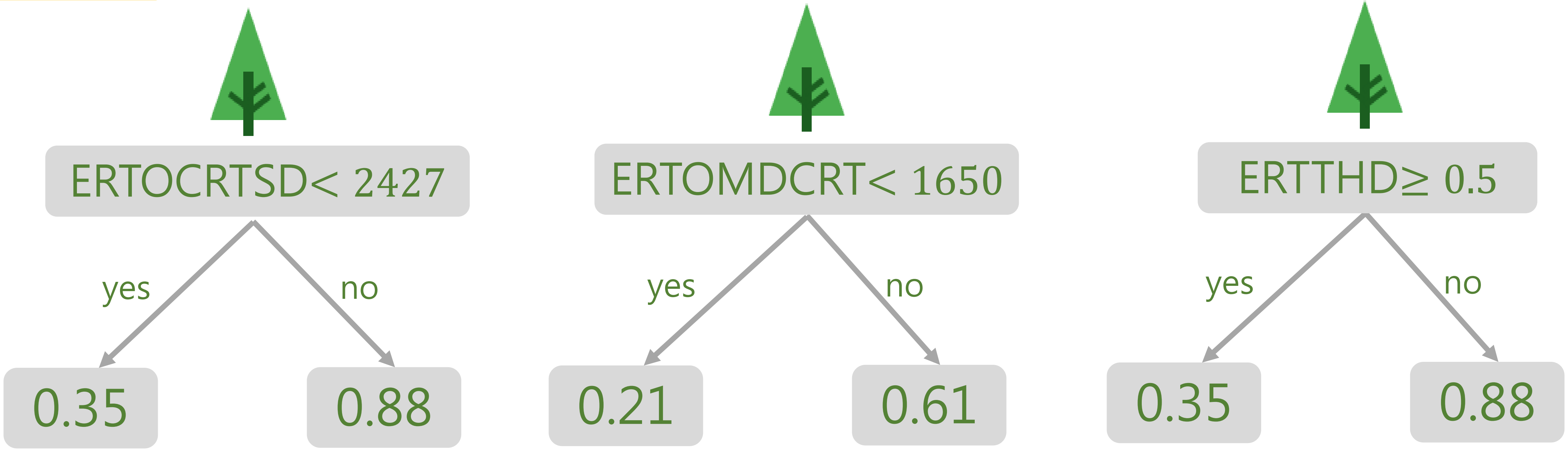
Results

MTT

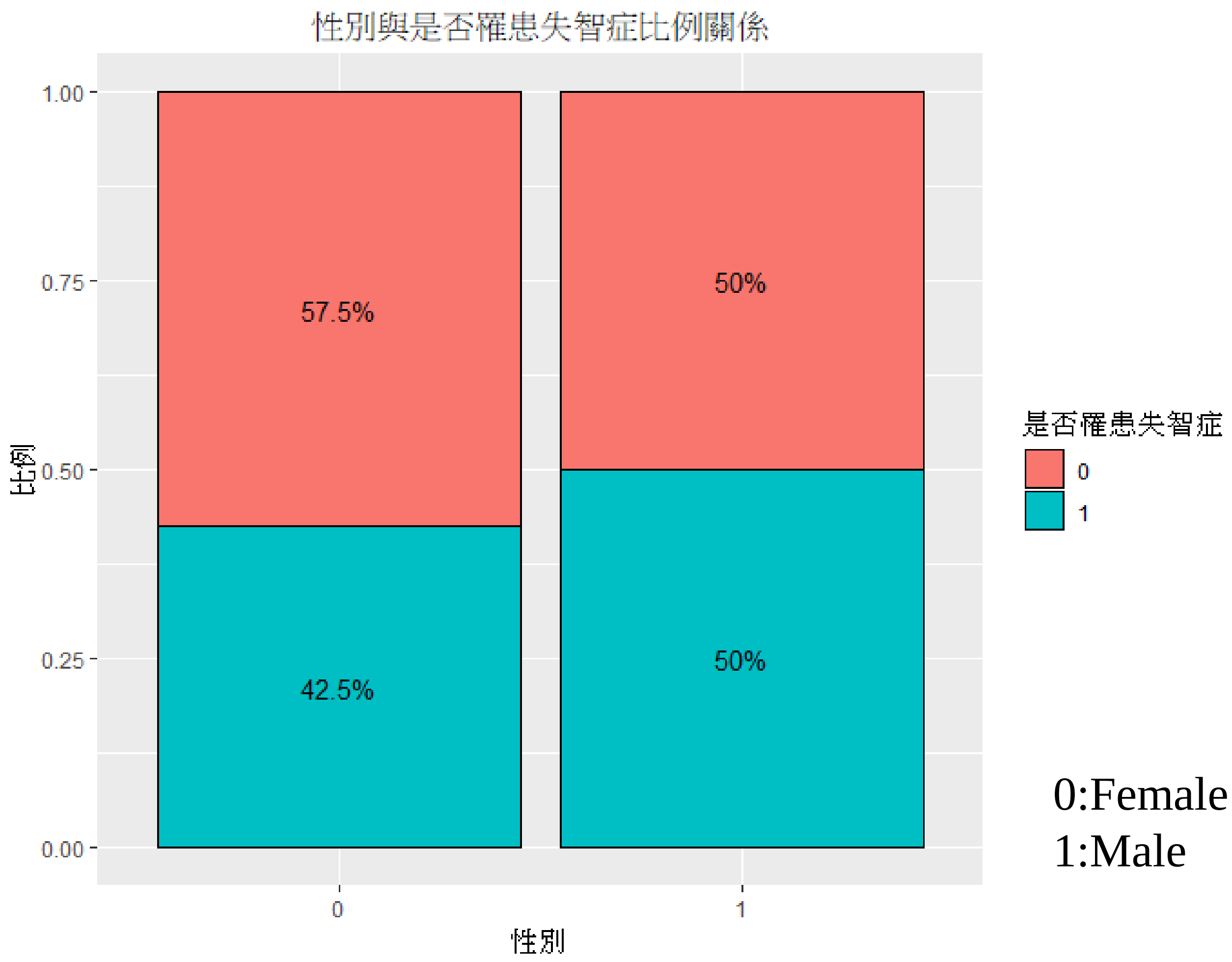

MTTDBE< 12
yes
no
0.37
0.87
MTTDE< 21
yes
no
0.33
0.89
MTTMTCM≥ 39
yes
no
0.3
0.86

Results

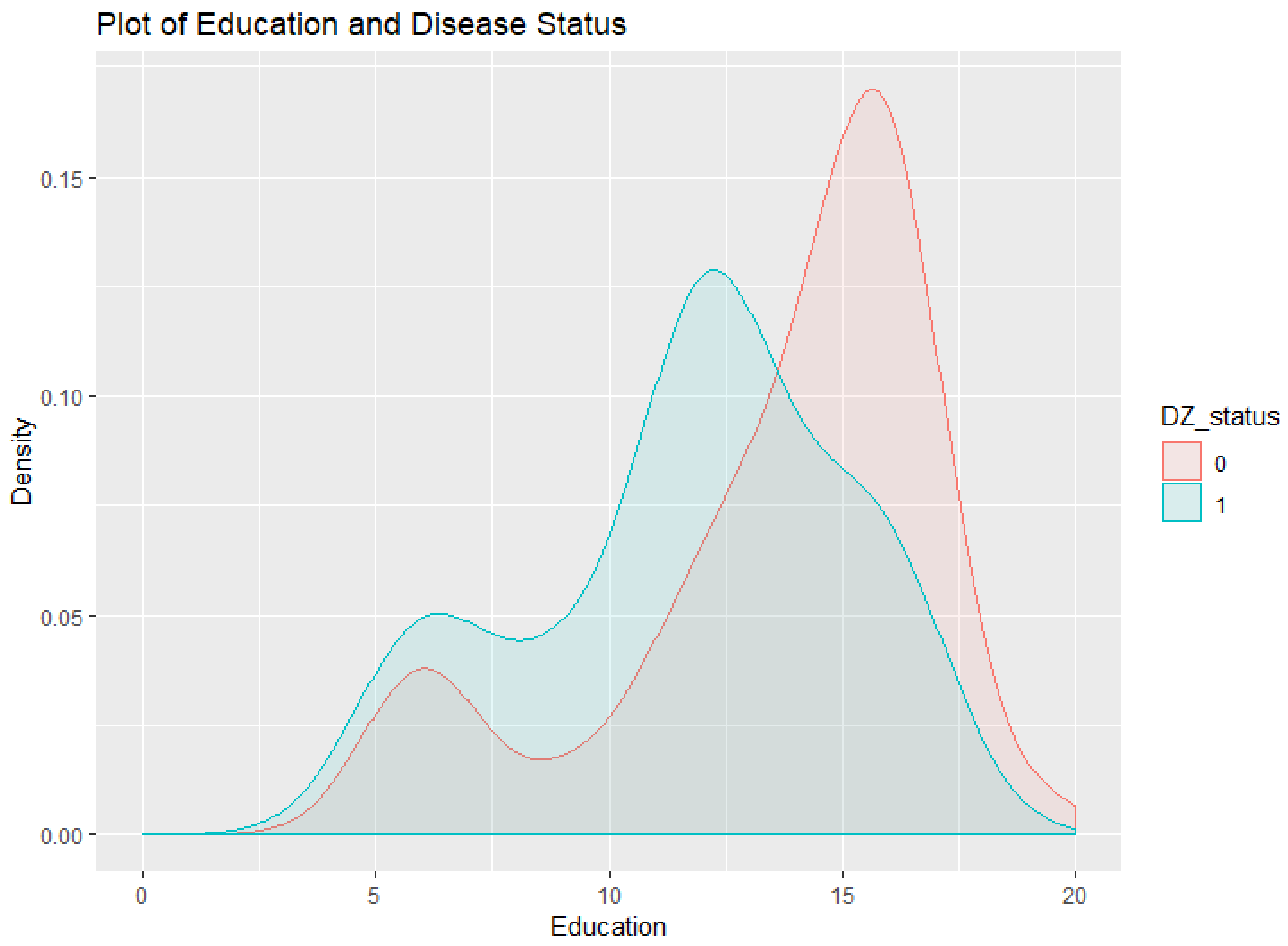
ERT


ERTOCRTSD< 2427
yes
no
0.35
0.88
ERTOMDCRT< 1650
yes
no
0.21
0.61
ERTTHD≥ 0.5
yes
no
0.35
0.88

Descriptive Statistics



0:Female
1:Male


Plot of Education and Disease Status
0.15
0.10
0.05
0.00
Density
0
5
10
15
20
Education
DZ_status
0
1

Descriptive Statistics