/ Predicting the risk of Coronary Heart Disease (Framingham Heart Study)

Tool : Logistic regression

The Analytics Edge :

The risk of having CHD ten years from now can be predicted from data available today on patients such as the blood pressure, cholestrol level, Smoking habits, age. Using a simple logistic regression model it is possible to do predictions and this has spawned prediction (clinical decision rules) and new markets for drugs and intervention programs.

Overview

www.framinghamheartstudy.

Framingham Heart Study (FHS)

In 1948, the FHS under the direction of the National Heart Institute embarked on an ambitious and landmark project in health research. At that time little was known about the causes of heart disease and stroke and the death rates for cardiovascular diseases (CVD) were increasing.

Objective of FHS study: To identify the common factors on characteristics that contribute to CVD by following its development over a long period of time in a large group of patients who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.

TOTAL DEATHS BY BROAD CAUSES

Year	Tuberculosis	Other Communicable Diseases	Neoplasms	Cardiovascular Diseases	External Causes of deaths	Disease of Early Infancy	Other Causes of Deaths	Total
1950	12.0%	32.5%	2.8%	6.3%	4.0%	7.2%	35.3%	100%
1955	9.1%	21.7%	6.5%	8.5%	4.9%	11.1%	38.3%	100%
1960	6.3%	18.7%	10.4%	10.6%	5.0%	11.2%	37.8%	100%
1965	6.2%	12.6%	13.9%	13.7%	6.5%	9.9%	37.2%	100%
1970	4.2%	12.7%	15.1%	27.0%	7.9%	5.9%	27.1%	100%
1975	3.7%	11.7%	18.5%	29.4%	4.1%	3.5%	29.0%	100%
1980	1.8%	11.4%	21.0%	34.4%	7.2%	3.3%	20.9%	100%
1985	1.3%	11.6%	22.0%	34.8%	8.1%	2.5%	19.7%	100%
1990	0.8%	10.3%	23.9%	37.1%	7.3%	2.2%	18.4%	100%
1995	0.8%	14.3%	25.2%	35.7%	7.1%	1.3%	15.6%	100%
2000	0.6%	13.9%	27.0%	36.6%	7.2%	0.8%	13.8%	100%
2005	0.4%	17.9%	26.5%	33.3%	6.3%	0.7%	15.1%	100%
2006	0.4%	15.3%	28.8%	33.2%	6.3%	0.7%	15.4%	100%
2007	0.5%	16.1%	27.7%	34.0%	6.0%	0.5%	15.2%	100%
2008	0.5%	15.3%	29.3%	33.6%	5.8%	0.6%	14.9%	100%
2009	0.4%	16.8%	29.3%	32.8%	5.7%	0.6%	14.3%	100%
2010	0.4%	17.2%	28.8%	33.0%	5.5%	0.5%	14.5%	100%
2011	0.4%	17.7%	30.4%	31.7%	4.6%	0.5%	14.6%	100%
2012	0.4%	18.0%	30.6%	31.1%	5.6%	0.5%	13.9%	100%
2013	0.3%	19.7%	30.9%	30.4%	4.9%	0.5%	13.3%	100%

Source : MoH, Singapore

HDL (ie 'good cholesterol').
- Replace food rich in saturated fats (pork, beef, mutton, cheese, coconut milk) and cholesterol (egg yolk, liver, kidney, brain) with skinless poultry, fish and low fat milk. Avoid oily, fatty and fried food as well as sugary food and starches.
- Increase intake of fruits and vegetables.
- Quit cigarette smoking, if you are a smoker.
- Repeat your cholesterol test in 3-6 months.

Glucose

Your blood glucose level is normal. There is no evidence of diabetes noted.

CORONARY HEART DISEASE (CHD) RISK ASSESSMENT

Based on the results of the health questionnaire and screening parameters, your risk of developing myocardial infraction or coronary death in the next 10 years is ***Low (3%)***.

Lipid goal for Low risk group	
Total Cholesterol	<200
LDL	<160
HDL	≥40

ADVICE:

- Do take special effort to minimize your modifiable risk factors with healthy dietary, exercise and lifestyle habits. Find out more at *http://www.dayspring.sg/results#cdra*.

Note: The 10-Year CHD Risk Score for Chinese, Malay and Indian males and females in Singapore is derived from the Framingham-based NCEP ATP III 10-Year Risk Score Tables which have been modified taking into account the Singapore cardiovascular epidemiological data. This modification was carried out as part of a collaboration between investigators at the Singapore Ministry of Health, Singapore General Hospital, National University of Singapore and Prof. Ralph B D'Agostino from the Framingham Heart Study, USA. Since there are insufficient data for other ethnic minorities, it is recommended that the 10-Year CHD Risk Score for the lowest risk group (i.e. Chinese) be used for these individuals.

CONCLUSION

Once again, thank you for screening with Dayspring.For more information on the screening results, please refer to *http://www.dayspring.sg/results*.

No definite diagnoses may be made from the test results alone. If there are abnormal findings, please consult your doctor for follow-up. In addition, normal test results may not necessarily mean the absence of a medical condition.

In accordance with a Ministry of Health (MOH) advisory, any person undergoing general health screening is to be attended by a medical practitioner when the screening results are available. A referral letter is attached for your convenience.

Origins of FHS

The origins of the Framingham Heart Study is closely linked to the cardiovascular health of the US President Franklin D Roosevelt and his premature death from hypertensive heart disease and stroke in 1945.

1932 : Roosevelt blood pressure 140/100

1935 to 1941 : Gradual rise in his blood pressure from 136/78 to 188/105

Despite the rising BP, his personal physician insisted that the Presidents health was fine and his blood pressure "no more normal than a man of his age".

Leading up to his death at the age of 63 : 186/108 → 240/130 → $\underbrace{300/190}_{\text{Time of death pressure}}$

Cerebral haemorrhage.

Harry Truman, who was Vice President under Roosevelt became President and signed into law the National Heart Act. The law allocated a US $500000 seed grant for a 20 year epidemiological study of heart diseases (FHS).

Today blood pressure $\frac{\text{Systolic}}{\text{diastolic}}$ $\frac{90\text{to}120}{60 \text{ to } 80}$

1948 : 5209 men and women between the ages of 30 and 62 were enrolled in the FHS from the town of Framingham, Massachusetts, United States of America

First round of extensive physical examination & lifestyle interviews

↓

Participants return every two years for examinations and tests

1971 : Second generation of 5124 original participant's children and spouses enrolled in the program

1994 : Omni cohort consisting of 507 men and women of African-American, Hispanic, Asian,... origins who were residents of Framingham enrolled to reflect increasing diversity of residents in the area

2002 : Third generation of participants

2003 : Second group of Omni cohort

Most of the knowledge concerning heart disease such as the effect of diet, exercise, BP is based on this longitudinal study.

The dataset and model that we will study is inspired from an influential paper in 1998 "Prediction of Coronary Heart Disease Using Risk Factor Categories" by Wilson, Agostino, Levy, Belanger, Silbershatz, Kannel in the journal Circulation published by the American Heart Association.

(Number of Google citations for this paper as of 14/5/2015 : 7033)

Main contribution of the paper :

Develop a predictive algorithm (clinical decision rule) to predict the 10-year CHD (coronary heart disease) risk using risk factors of blood pressure, total cholestrol & LDL cholestrol in a white middle aged population sample from the Framingham heart study. This allows physicians to predict CHD risk in patients without overt CHD.

In our analysis, we will use a dataset from the website https://biolincc.nhlbi.nih.gov/home/ which is anonymized.

Data on 4434 participants with data collected during 3 examination periods, approximately 6 years apart from 1956 to 1968. Participant followed for 24 years.

Intervention strategies & impact

The results from the Framingham study has been validated with external validation by generalizing to different populations such as black male, Asians

The advantage of such an approach is that it can be used to develop intervention strategies for example drugs to lower blood pressure, drugs for lower cholestrol. The effects of these on reduced chances of coronary heart disease can be tested by doing clinical trials. The markets for diuretics (to reduce blood pressure) statins (to reduce cholestrol levels) are now in billions of dollars.

The FHS also led to the increase in clinical decision rules in many areas of medicine that predict clinical outcomes using patient data & test results. These models are unbiased, unemotional & can assist new physicians with little experience to make decisions

Analytics on FHS data : R

Read and basic analysis of data with preprocessing

```
framing ← read.csv("framingham.csv")

Str(framing)          11627 obs of 39 variables,
```

We need to preprocess this data to do the prediction

Identify the subset of data frame such that each individual has only one observation (corresponding to PERIOD = 1) and is free of prevalent CHD at this time (PREVCHD = 0)

```
framing1 ← subset(framing, PERIOD==1 & PREVCHD==0)

Str(framing1)          4240 obs of 39 variables

length(unique(framing1$RANDID))
```

4240 helps verify that each individual is represented only once

To model the event data, we need to identify for the patients if they had CHD in 10 years from their first visit.

```
framing1$TENCHD = as.integer((framing1$TIMECHD/365) ≤ 10)
```

Converts time for CHD to years (roughly) & checks if CHD occurs in 10 years
Note that the maximum range is 24 years in data

```
Colnames(framing1)
```

Provides a list of all the Column names of the dataframe

```
which (colnames (framing1) == "LDLC")
```

Find column number with name of LDL

```
framing1 <- framing1 [, c(1:21,40)]

Str (framing 1)
```

Keeps only the risk factor and output.
The variables HDLC & LDLC are dropped as the data is not available

Variables

1) RANDID - Identification number

2) SEX - 1 = M, 2 = F

3) TOTCHOL - Total cholestrol (mg/dL)

4) AGE - age at exam (years)

5) SYSBP - Systolic blood pressure

6) DIABP - diastolic blood pressure

7) CURSMOKE - Current cigarette smoking 0 = NO, 1 = YES

8) CIGPDAY - Cigarettes per day

9) BMI - Body mass index

10) DIABETES - 0 = Not diabetic, 1 = Diabetic

11) BPMEDS - Use of antihypertensive medication = 1, 0 otherwise

12) HEARTRTE - Heart rate (beats/min)

13) GLUCOSE - mg/dL

14) educ - 1 = 0-11 years, 2 = High school, 3 = Some College Vocational, 4 = College or more

15) PREVCHD - 0 = FREE of disease , 1 = Prevalent
(In this data due to our preprocessing all individuals have PREVCHD = 0)

16) PREVAP - Prevalent Angina Pectoris = 1 , else 0

17) PREVMI - Prevalent myocardial infection = 1, else = 0

18) PREVSTRK - Prevalent stroke = 1, 0 otherwise

19) PREVHYP - Prevalent hypertensive = 1, 0 otherwise

20) TIME - No of days since examination (all 0 since first exam date)

21) PERIOD - Period = 1

To predict

$$\text{TEN CHD} = \begin{cases} 1 & \text{if individual develops CHD within 10 years} \\ 0 & \text{Otherwise} \end{cases}$$

Note that when you collect data from other sources, often you need to spend some time preprocessing it before applying analytics methods

Split dataset into traning & test sets

To develop a predictiie model it is important to be able to split the datasets into two parts - one used for training the dataset and the second to test the dataset.

This should be done while preserving ratios of the outcome variable in the two sets.

install.packages("caTools")	Installs a package contribt in R from CRAN
library (caTools)	Loads the package
set.seed (1)	Sets a seed so that results can be replicated by users

Split ← Sample.split(framing1$TENCHD, SplitRatio = 0.65)

Uses sample.split fncton to split data into 65% - 35%. by maintaing ratio of TENCHD variable into the two sets

training ← subset (framing1, split == TRUE)

test ← subset (framing1, split == FALSE)

mean (framing1 $TENCHD)	mean (traing$TENCHD)	mean (test$TENCHD)
0.1518	0.1520	0.1516

Maintains a similar balance of the fraction of patients with CHD

Perform logistic regression

```
help(glm)
```

Help on generalised linear model

```
model1 <- glm(TENCHD ~ . , data = training, )
                  family = binomial
```

Performs logistic regression with dependent variable as TENCHD and all other variables as predictors

```
Summary (model 1)
```

Provides summary

- 5 of the coefficients of PREVCHD, PREVAP, PREVMI, TIME, PERIOD not defined because of singularities
- 372 observations deleted due to missing values

Note that we should not expecte the variables such as RANDID or EDUC to play a role here possibly though one might argue that a person who is educated more gives greater importance to health. So we will leave the EDUC variable in.

```
model2 <- glm(TENCHD ~ SEX+TOTCHOL+AGE+SYSBP
              +DIABP+CIGPDAY+CURSMOKE+BMI+
              DIABETES+ BPMEDS+ HEARTRTE + GLUCOSE
              + educ+ PREVSTRK+ PREVHYP,
              data=training , family = binomial)
```

From the fit, significant variables at 0.001 level are intercept, SEX of individual, AGE of individual SYSBP (systolic blood pressure), CIGPDAY (number of cigarettes per day), GLUCOSE (glucose).

Summary (model 2)

AIC = 1866.1

Suppose, we solve a smaller model only using these variables

```
model3 ← glm(TENCHD ~ SEX + AGE + SYSBP
             + CIGPDAY + GLUCOSE, data = training,
             family = "binomial")
```

Summary (model 3)

AIC = 1941.3

All variables are extremely significant

While the AIC increases, we decide to stick with this model since it appears to be more interpretable due to fewer variables.

However one could also work with the earlier model ~~if~~ so desired, due to a better fit at the cost of more variables.

$$\text{Logit (CHD)} = -7.16 - 0.54\ \text{SEX} + 0.059\ \text{AGE} + 0.016\ \text{SYSBP} + 0.018\ \text{CIGPDAY} + 0.0099\ \text{GLUCOSE}$$

$$\text{Probability of CHD} = \frac{1}{1 + e^{-\text{Logit (CHD)}}}$$

Consider a patient who is 60 years old, male, has systolic blood pressure of 145, smokes two cigarettes per day with a glucose level of 80.

For this patient, we can predict

$$\text{Logit (CHD)} = -1.012$$

$$\text{Probability of CHD} = 0.266$$

Prediction

```
predict_test ← predict (model 3, type = "response",
                        newdate = test )
```

Performs a prediction on the test set with logistic regression where type = "response" gies predicted probabilities

```
table ( predict_test > 0.5, test $ TENCHD )
```

	0	1
FALSE	1113	187
TRUE	9	21

Model
FALSE = Predict no CHD
TRUE = Predict CHD

Actual
0 = No disease
1 = TENCHD

$$\text{Accuray in test set} = \frac{1113 + 21}{1113 + 21 + 9 + 187} = \underline{0.852}$$

```
table ( training $ TENCHD )
```

0	1
2337	419

In training set, majority of patients do not have CHD. So a baseline model is to predict no one has CHD in test set table (predict_test > 1, test $TENCHD)

	0	1
FALSE	1122	208

Predict

Actual

$$\text{Accuracy} = \frac{1122}{1122+208} = \underline{0.8436}$$

Using a 0.5 threshold, the model beats the baseline model by a small amount.

Suppose we use a lower threshold (0.25). In this case we are more prone to make false positives but this seems more important than false negatives in this application.

Note that long term effects of false negatives are often much more than short term costs of false positive.

table (predict_test > 0.25, test $TENCHD)

	0	1
FALSE	987	$\boxed{121}$
TRUE	$\boxed{135}$	87

Predicted

Actual

If clinicians used this model then, 135 + 87 = 222 patients need treatment (preventive), out of which 135 would be unnecessary but 87 would be necessary.

By choosing threshold = 0.5, the observation is classified into a class for which the probability is highest.

However in some instances one type of error is more preferred to another.

For example in predicting diseases as in the FHS (Framingham Heart Study), say

1 = Patient develops CHD
0 = Patient does not develop CHD

High threshold (t) implies we will make more false negative errors (predict person does not have CHD when they actually do).

Lower threshold (t) implies we will make more false positive errors (predict person develops CHD when they do not)

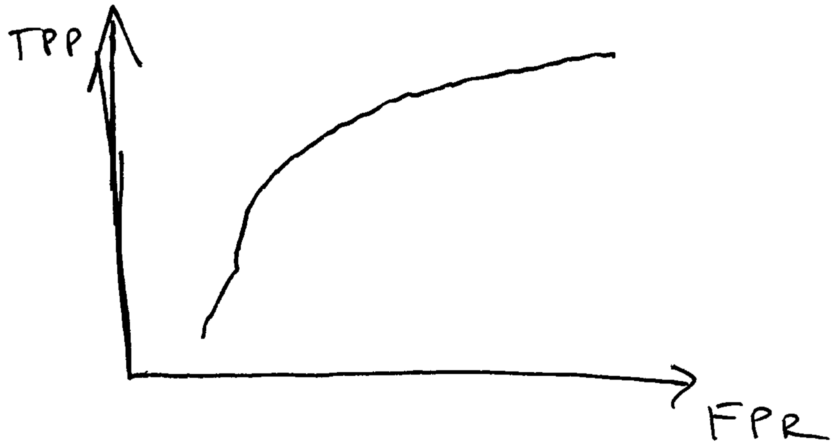
FPR might be more preferred here though more resources are spent on unnecessary patients who do not need it.

More detailed test

```
Install.packages("ROCR")
library(ROCR)

predict <- prediction(predict_test, test$TENCHD)
perf <- performance(predict, mesure="tpr",
                    X.mesure="fpr")

plot(perf)          Plots ROC curve
```



You can add arguments colorize=TRUE to visualize ROC curve better

```
performance(predict, "auc")
performance(predict, "auc")@y.values
```

AUC= 0.7574 } Model can distinguish between low & high risk patients better than random guess

Points systems

While the analytics approach predicts risk of getting a coronary heart disease for patients, it is not particularly easy for physicians & patients to use.

Points systems are often implemented to make the results more usable.

From the logistic regression, we can develop a base case as follows.

Lowest risk is for the person who is in the low age category, is female, has a systolic blood pressure of lower than 120 (from clinically meaningful status), smokes zero cigarettes per day, and has lower glucose level.

We consider how such ratings can be developed next.

Such points systems are particularly useful since it can tell patients on how improving on a certain aspect can decrease CHD risk.

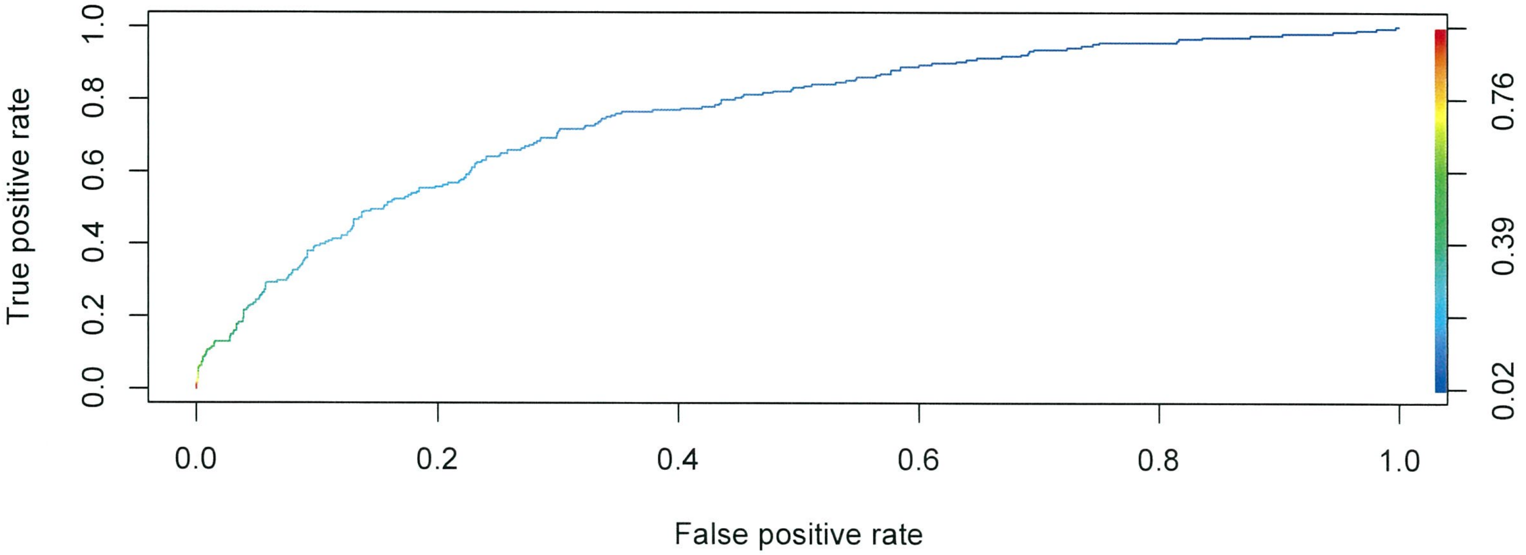
For example consider,

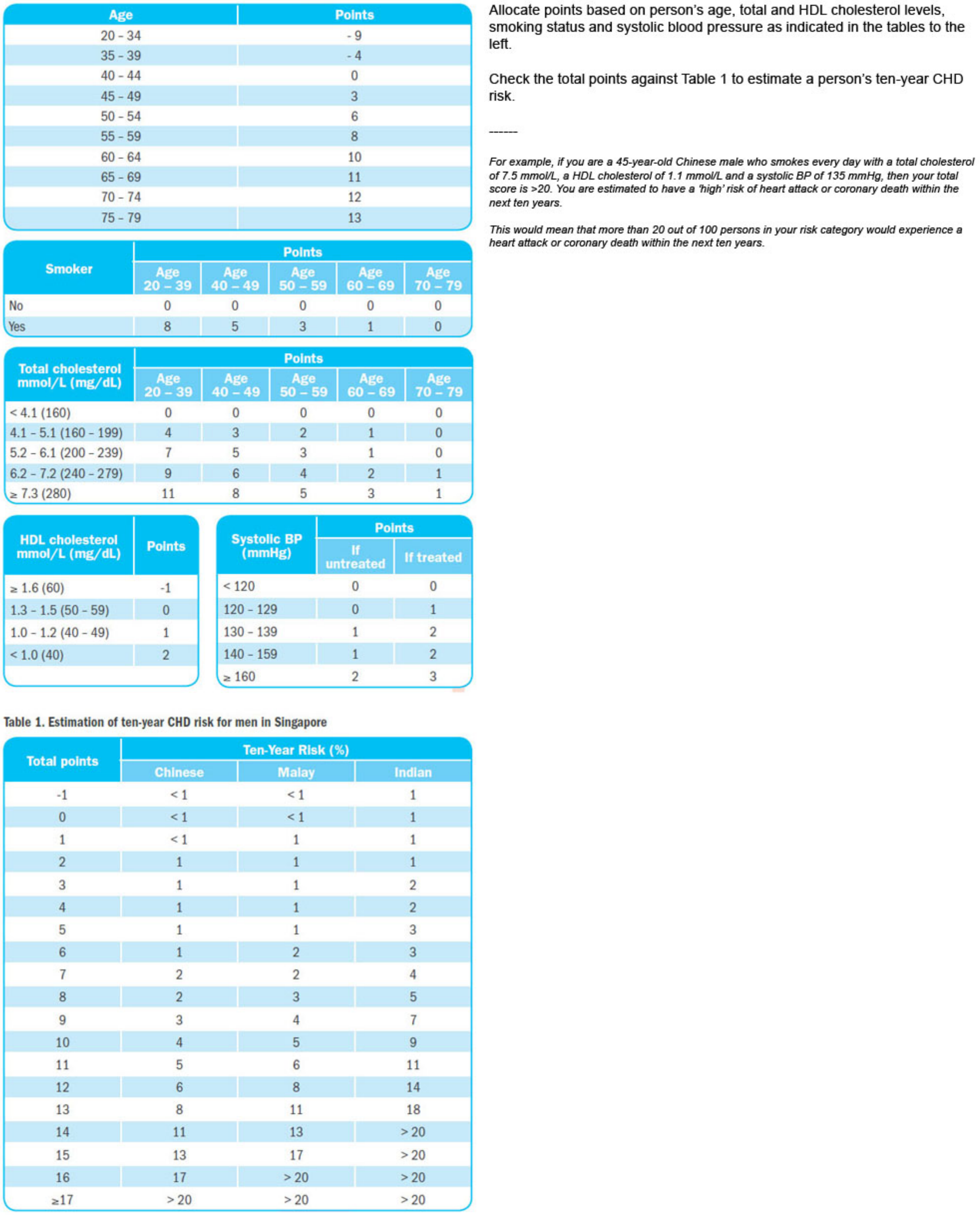
Variables	Category	Reference	Base difference	Logit units	Points
Age	30-39	34.5	0	0	0
	40-49	44.5	10	0.59	2
	50-59	54.5	20	1.18	4
	60-69	64.5	30	1.77	6
	70-79	74.5	40	2.36	8
Sex	Male	1	-1	0.54	2
	Female	2	0	0	0
Systolic blood pressure	<120	106.5	-13	-0.208	-1
	120-139	129.5	0	0	0
	140-159	149.5	20	0.32	1
	≥160	174.5	45	0.72	2

Similarly it can be done for cigarettes smoking and glucose.

To get points, we divide here by 0.059×5 & then round to nearest integers.

We given -ve points (bonus). in this example too.


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.02
0.39
0.76



Age	Points
20 - 34	- 9
35 - 39	- 4
40 - 44	0
45 - 49	3
50 - 54	6
55 - 59	8
60 - 64	10
65 - 69	11
70 - 74	12
75 - 79	13

Smoker	Points				
	Age 20 – 39	Age 40 – 49	Age 50 – 59	Age 60 – 69	Age 70 – 79
No	0	0	0	0	0
Yes	8	5	3	1	0

Total cholesterol mmol/L (mg/dL)	Points				
	Age 20 – 39	Age 40 – 49	Age 50 – 59	Age 60 – 69	Age 70 – 79
< 4.1 (160)	0	0	0	0	0
4.1 - 5.1 (160 - 199)	4	3	2	1	0
5.2 - 6.1 (200 - 239)	7	5	3	1	0
6.2 - 7.2 (240 - 279)	9	6	4	2	1
≥ 7.3 (280)	11	8	5	3	1

HDL cholesterol mmol/L (mg/dL)	Points
≥ 1.6 (60)	-1
1.3 - 1.5 (50 - 59)	0
1.0 - 1.2 (40 - 49)	1
< 1.0 (40)	2

Systolic BP (mmHg)	Points	
	If untreated	If treated
< 120	0	0
120 - 129	0	1
130 - 139	1	2
140 - 159	1	2
≥ 160	2	3

Table 1. Estimation of ten-year CHD risk for men in Singapore

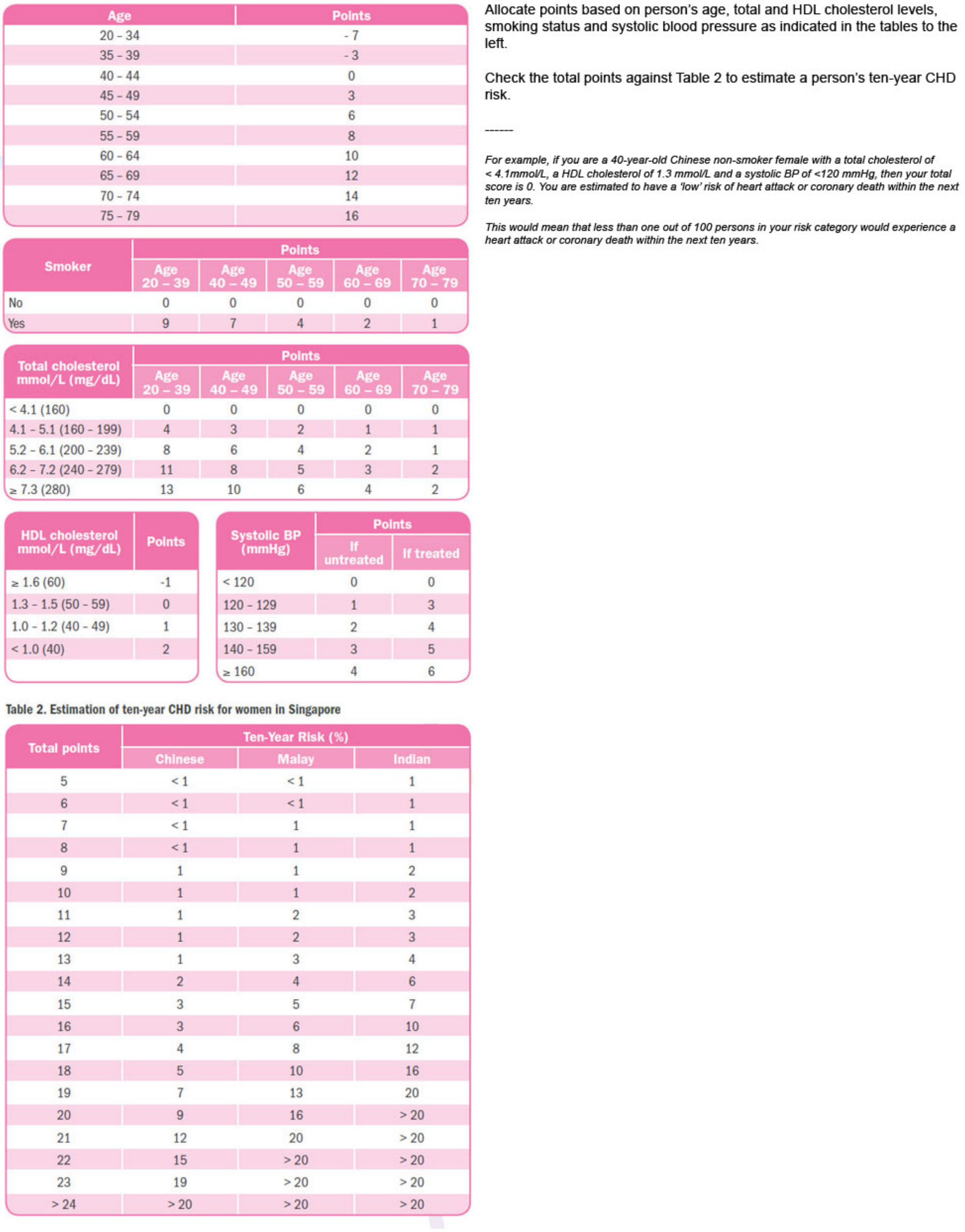
Total points	Ten-Year Risk (%)		
	Chinese	Malay	Indian
-1	< 1	< 1	1
0	< 1	< 1	1
1	< 1	1	1
2	1	1	1
3	1	1	2
4	1	1	2
5	1	1	3
6	1	2	3
7	2	2	4
8	2	3	5
9	3	4	7
10	4	5	9
11	5	6	11
12	6	8	14
13	8	11	18
14	11	13	> 20
15	13	17	> 20
16	17	> 20	> 20
≥17	> 20	> 20	> 20

Allocate points based on person's age, total and HDL cholesterol levels, smoking status and systolic blood pressure as indicated in the tables to the left.

Check the total points against Table 1 to estimate a person's ten-year CHD risk.

For example, if you are a 45-year-old Chinese male who smokes every day with a total cholesterol of 7.5 mmol/L, a HDL cholesterol of 1.1 mmol/L and a systolic BP of 135 mmHg, then your total score is >20. You are estimated to have a 'high' risk of heart attack or coronary death within the next ten years.

This would mean that more than 20 out of 100 persons in your risk category would experience a heart attack or coronary death within the next ten years.



Age	Points
20 - 34	- 7
35 - 39	- 3
40 - 44	0
45 - 49	3
50 - 54	6
55 - 59	8
60 - 64	10
65 - 69	12
70 - 74	14
75 - 79	16

Smoker	Points				
	Age 20 – 39	Age 40 – 49	Age 50 – 59	Age 60 – 69	Age 70 – 79
No	0	0	0	0	0
Yes	9	7	4	2	1

Total cholesterol mmol/L (mg/dL)	Points				
	Age 20 – 39	Age 40 – 49	Age 50 – 59	Age 60 – 69	Age 70 – 79
< 4.1 (160)	0	0	0	0	0
4.1 - 5.1 (160 - 199)	4	3	2	1	1
5.2 - 6.1 (200 - 239)	8	6	4	2	1
6.2 - 7.2 (240 - 279)	11	8	5	3	2
≥ 7.3 (280)	13	10	6	4	2

HDL cholesterol mmol/L (mg/dL)	Points
≥ 1.6 (60)	-1
1.3 - 1.5 (50 - 59)	0
1.0 - 1.2 (40 - 49)	1
< 1.0 (40)	2

Systolic BP (mmHg)	Points	
	If untreated	If treated
< 120	0	0
120 - 129	1	3
130 - 139	2	4
140 - 159	3	5
≥ 160	4	6

Table 2. Estimation of ten-year CHD risk for women in Singapore

Total points	Ten-Year Risk (%)		
	Chinese	Malay	Indian
5	< 1	< 1	1
6	< 1	< 1	1
7	< 1	1	1
8	< 1	1	1
9	1	1	2
10	1	1	2
11	1	2	3
12	1	2	3
13	1	3	4
14	2	4	6
15	3	5	7
16	3	6	10
17	4	8	12
18	5	10	16
19	7	13	20
20	9	16	> 20
21	12	20	> 20
22	15	> 20	> 20
23	19	> 20	> 20
> 24	> 20	> 20	> 20

Allocate points based on person's age, total and HDL cholesterol levels, smoking status and systolic blood pressure as indicated in the tables to the left.

Check the total points against Table 2 to estimate a person's ten-year CHD risk.

For example, if you are a 40-year-old Chinese non-smoker female with a total cholesterol of < 4.1mmol/L, a HDL cholesterol of 1.3 mmol/L and a systolic BP of <120 mmHg, then your total score is 0. You are estimated to have a 'low' risk of heart attack or coronary death within the next ten years.

This would mean that less than one out of 100 persons in your risk category would experience a heart attack or coronary death within the next ten years.