CAD Recognition Using Three Mathematical Models

J. Martyniak, K.Stanisz-Wallis, and L.Walczycka

Deptartment of Biostatatistics & Medical Informatics,
Collegium Medicum Jagiellonian University, Cracow, Poland
mymartyn@cyf-kr.edu.pl, mywallis@kinga.cyf-kr.edu.pl

Abstract. In this paper we present the Bayesian diagnostic model and the logistic regression model with two different entrance strategies for determining factors predicting CAD risk. In the (R) methodological strategy the appropriate ratios lipids/apoproteins were considered as independent variables, in the (PC) the principal components analysis was used as a data reduction technique. The model based on Bayes' Theorem and the logistic regression model were good predictive models. It is the main objective of the project to try to design a tool for diagnosing.

1 The Aim

Coronary artery disease remains still the greatest cause of premature death for the Polish population. In the industrial countries it is responsible for 50% of all the death cases. The reasons of this are related to the adverse changes in the population's lifestyle resulting in the increase of the so called "risk factors" for development of arteriosclerosis, leading to coronary artery disease (CAD). The concept of "risk factor" comes from a long-term studies of Framingham population, in which the existence of several egzo- and endogenic phenomena has been shown, which are causally related to the CAD occurrence.

Taking into account the recent clinical studies, scientific achievements and more complete understanding of the peril and the action mechanism of the "risk factors" the International Commission for Prevention of Coronary Artery Disease, in cooperation with the International Society for Arteriosclerosis and Obesity, the European Society of Hypertension and the Diabetes Society, has prepared an accurate and comprehensive document concerning the primary and secondary prevention of coronary artery disease [2, 4].

The early detection and elimination of the risk factors of the CAD is particularly important, because in recent years the clinical form of the disease has been found in many young persons.

The earlier and greater is the awareness of the risk level for development of the disease in a given patient, the more appropriate and on-time decisions will be taken regarding the prevention of its development. The related changes in the patient's nourishment habits and lifestyle and the necessity of pharmacological treatment (choice of the appropriate medicine and its dose) depend on the global

P.M.A. Sloot et al. (Eds.): ICCS 2002, LNCS 2331, pp. 234–241, 2002.

risk assessment – the probability of coronary artery disease development in a given population, family and even particular patient.

Among more than 250 risk factors described up to date at least three have been recognized as independent i.e. individually responsible for development of the disease. These are: disorders in lipid level, uncured arterial hypertension and smoking. The lipid metabolism disorders have been found to be a primary risk factor for the CAD [8–10].

The aim of the study was to compare the anticipation accuracy of three different methods using a logistic regression analysis and the Bayesian diagnostic model.

2 The Data

The database used in the present study has been collected in the Dept. of Clinical Biochemistry, Chair of Clinical Biochemistry and Diagnosis in the Collegium Medicum of Jagiellonian University (CMUJ), Cracow, and it contained anonymous data of patients treated in the Clinic for Metabolic Diseases, CMUJ (including period 1985–1990), the Institute of Cardiology, CMUJ (1992–1998) and groups of patients from a selected population of healthy persons from the Tarnow Voiodship (the POL – MONICA project, realized in years 1985–1998).

Two hundred and twenty-eight subjects comprising 95 patients with arteriographically evidenced CAD (males, aged 47.99+11.12) and 133 healthy control (males, aged 47.11+7.29) without CAD were studied.

Blood samples were withdrawn after overnight fasting. Plasma levels of biochemical variables were assayed using automatic colorimetric methods, except insulin determined by radioimmunoassay.

Plasma lipid profile was determined as follows: total and free cholesterol and trigliceryde level in plasma (Ch, fCh, tg) and lipoproteins VLDL (VLDL-Ch, VLDL-fCh, VLDL-tg), LDL (LDL-Ch, LDL-fCh, LDL-tg) and HDL (HDL-Ch, HDL-fCh, HDL-tg).

Esterified cholesterol was calculated (eCh, VLDL-eCh, LDL-eCh, HLDL-eCh).

Apolipoprotein A1 level in plasma and HDL (apoA1, HDL-apoA1), alipoproteina B level in plasma and VLDL, LDL (apoB, VLDL-apoB, LDL-apoB); as well as protein B in VLDL, LDL and HDL (VB, LB, HB) were also determined.

Plasma glucose and insulin were determined in fasting state (Glu0, Ins0), and after 30, 60 and 120 min. during the oral glucose tolerance test (OGTT) (Glu30, Glu60, Glu120, Ins30, Ins60, Ins120 respectively).

Sum of glucose (SumGlu) and insulin (SumIns) were calculated as an area under glycemic and insulin kinetic curves estimated during OGTT respectively. Body mass index (BMI [kg/m^2]) was calculated.

3 Statistical Methods

A logistic regression model with two different entrance strategies for determining factors predicting CAD prevalence was used. In the first (R) methodological strategy the appropriate ratios lipids/apoproteins were considered as independent variables. Principal component analysis was used as data reduction techniques for the second strategy (PC). A model based on Bayes' theorem was used for calculating post-test probability.

The data set for the both methods was divided in two groups:

I – the learning set (121 patients – 72 persons in good health and 49 persons with CAD)

II – the test sets (107 patients – 57 persons in good health and 50 persons with CAD)

SAS and STATISTICA PL programmes were used for calculations [11, 12].

4 R Strategy

4.1 The Logistic Regression Model

For whole group of patients the relationship between five scores and CAD was explored by the logistic regression analysis. The probability of the appearance of the CAD (Y) conditional by the variable (x_p) can be calculated by means of the following form [3, 5].

$$P\,(Y = \text{the appearance of CAD}) = \frac{\exp\,(b_0 + b_1 x_1 + \ \ldots\ + b_p x_p)}{1 + \exp\,(b_0 + b_1 x_1 + \ \ldots\ + b_p x_p)} \qquad (1)$$

where: x_1, x_2, ..., x_p – vector of independent variables – ratios, calculated from biochemical variables: SumGlu/SumIns, LDL-Ch/LDL-apoB, HDL-Ch/HB, HDL-tg/HDL-apoA1 and BMI

b_0, b_1, ..., b_p – unknown coefficient of the regression

Y – dependent variable.

Table 1 shows the variables which significantly influence the appearance of the CAD.

Table 1. The significance of logistic regression coefficients assessed by the Wald chi-square statistic

Ratios of of variables	Coefficient b_i	Standard error	p - Value (Wald's test)
SumGlu/SumIns	–26,35	6,58	0,0001
HDL-Ch/HB	–6,42	3,23	0,049
HDL-tg/HDL-apoA1	14,51	6,16	0,02
b_0	4,50	2,39	0,062

5 PC Strategy

5.1 Principal Components Analysis

Principal components analysis is the multivariate data reduction technique [13].

This analysis selects the linear combination of the variables that best captures the variability of the data in a multidimensional space.

Given a data set with n numeric variables, m principal components can be computed. Each principal component is a linear combination of the original variables, with coefficients equal to the eigenvectors of the correlation or covariance matrix . The eigenvectors are customarily taken with unit length.

The principal components are sorted in descending order of the eigenvalues, which are equal to the variances of the components .

We have n variables, X_1, X_2, ..., X_n all of which are standardised, and we desire to generate m composite variables of the form:

$$Y_i = \omega_{i1}X_1 + \omega_{i2}X_2 + \ \ldots \ + \omega_{in}X_n \quad \text{for} \quad i = 1, \ \ldots, \ m \tag{2}$$

where $m < n$ and ω_{ij} are selected to "explain" the maximum possible variance. That is Y_1 will have the largest possible variance (λ_1) subject to the restriction:

$$\sum_{j=1}^{n} \omega_{1j}^2 = 1$$

Y_2 will be uncorrelated with Y_1 and have the next largest possible variances (λ_2 with $\lambda_1 > \lambda_2$) etc., until we obtain m such uncorrelated composite variables all of which have weights normalized and variances $\lambda_1 > \lambda_2 > \ldots > \lambda_m$.

The percentage of the variance of the original n variables explained by the m composite variables is

$$100(\ \lambda_1 + \ldots + \lambda_m\ )/(\ \lambda_1 + \ldots + \lambda_n\ )$$

We call the m composite scores in this context component scores .

5.2 Data Reduction

The resulting scores are linear combination of the 34 original variables which are selected to be uncorrelated with each other.

We can retain only scores with eigenvalues greater than 1. In essence this is like saying that, unless a score extracts at least as much as the equivalent of one original variable, we drop it. Using this, we would retain 8 principal components.

The selected principal components were:

fraction V : VLDL-tg, VLDL-Ch, VB, VLDL-eCh, VLDL-fCh, tg

fraction L : fCh, LDL-eCh, LDL-Ch, eCh, Ch, apoB, LDL-apoB

fraction LH: LDL-fCh, LDL-apo B, LB, HDL-tg, HDL-fCh, height, BMI

fraction Totalgluc: Glu60, Glu120, SumGlu

fraction Ins: Ins30, Ins60, SumIns

fraction H : HDL-Ch, HDL-eCh, HB, apoA1, HDL-apoA1
fraction Weight: weight
fraction Glu0: Glu0, Glu30

The resulting components were included in a logistic regression that selected a model with eight scores .

5.3 Logistic Regression Model

For the whole group of patients the relationship between five scores and CAD was explored by the logistic regression analysis. The probability of the appearance of the CAD (Y) conditional by the variable (x_p) can be calculated by means of the previous form (1);
where:

x_1, x_2, ..., x_p – vector of independent variables selected principal components:

- fraction V
- fraction L
- fraction LH
- fraction Totalgluc
- fraction Ins
- fraction H
- fraction Weight
- fraction Glu0.

The significance of logistic regression coefficients was assessed with the Wald chi-square statistic.

Table 2 shows the variables which significantly influence the appearance of the CAD.

Table 2. Significance levels of the Wald test

Principal Component	Coefficient b_i	Standard error	p - Value (Wald's test)
Fraction V	–1,45	0,45	0,0017
Fraction H	–1,80	0,45	0,0011
Fraction Ins	–3,24	0,68	0,000
Fraction Weight	0,99	0,4	0,016
Fraction Glu0	1,91	0,52	0,0004
b_0	0,13	0,38	0,743

6 The Bayes Theorem

Bayes' theorem is a quantitative method for calculating the posterior probability using the prior probability [6, 7].

Let $\{B_n\}$ be a countable measurable partition of Ω and $P(A) > 0$ and $B_1 \cup ... \cup B_n = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$ and $P(B_i) > 0$ for $i = 1, \ldots, n$. Then

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)} \tag{3}$$

where $P(A|B)$is called the conditional probability of event A given event B.

We propose to use the theorem of Bayes as a tool to predict CAD for the individual patient. This model computes the posterior probability of an appearance CAD given a set of patients symptoms, syndromes or laboratory values.

Let $X = \{x_1, x_1, \ldots x_6\}$ will be a set of patient's syndromes, symptoms or laboratory values and $x_i \in \{0, 1\}$ and x_i represent the presence or the absence of those variables. The Bayes probability we can calculate as

$$P(CAD|X) = \frac{P(X|CAD) \cdot P(CAD)}{P(X|CAD) \cdot P(CAD) + P(X|No\ CAD) \cdot P(No\ CAD)} \tag{4}$$

where

$P(CAD)$ is the prior probability of CAD

$P(No\ CAD) = 1 - P(CAD)$

$P(X|CAD) = P(x_1|CAD) \cdot P(x_2|CAD) \cdot \ldots \cdot P(x_6|CAD)$ for $x_i = 1, \ldots, 6$.

$P(X|No\ CAD) = P(x_1|No\ CAD) \cdot P(x_2|No\ CAD) \cdot \ldots \cdot P(x_6|No\ CAD)$

The following independent (Pearson's chi-square test) syndromes, symptoms and laboratory values were selected: tg, Ch, LDL-Ch, HDL-Ch, BMI, tg/HDL-Ch.

7 Results

The results of examination of model quality are only for the test sets [1].

Table 3. The results of PC model and R model

PC model				**R model**			
	expected				**expected**		
		CAD	No CAD			CAD	NoCAD
observed	CAD	40	6	**observed**	CAD	32	14
	No CAD	10	51		No CAD	46	15

Sensitivity = 86,96% specificity = 83,61% Sensitivity = 69,58% specificity = 75,41%

Preliminary tests indicate that the PC model is of greater diagnostic relevance, which can undoubtedly be attributed to a greater number of variables introduced to the model than in the case of the R model.

8 Predictive Value

The predictive value of a test is simply the post - test probability that a disease is present based on the results of a test. The predictive value of a positive test depends on the test's sensitivity, specificity, prevalence and can be calculated of the following form [6]:

$$PV^{+} = \frac{(sensitivity)(prevalence)}{(sensitivity)(prevalence) + (1 - specifity)(1 - prevalence)} \tag{5}$$

The PV^{+} index is also one of the form of Bayes' theorem.

Table 4 shows the predictive value (PV^{+} index) of a positive test in three methods.

Table 4. Results of comparison of different predictive models

Model	Number of parameters	PV^{+} index
Bayesian analysis	6	0.841
PC model	34	0.807
R model	3	0.679

It is evident that, the PV^{+} index has the highest value for Bayes'model and the lowest for the R model.

9 Conclusion

The principal component analysis followed by stepwise logistic regression analysis showed the better predictive power of prevalence in the large and complex CAD data base than methodological strategy of ratios lipids/apoproteins considered as independent variables. The Bayes' theorem model also shows a relatively high predictive power compared with other two models.

References

1. Marshall G., Grover F., Henderson W., Hammermeister K.: Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery. Statist. Med. **13** (1994) 1501–1511

2. Buring J.E., O'Conner G.T., Goldhaber S.Z.: Risk factors for coronary artery disease: a study comparing hypercholesteronemia and hypertriglyceridemia. Eur.J.Clin.Invest. **19** (1989) 419–423
3. Hosmer D.W., Lemeshow S.: Applied Logistic Regression . 1^{st} ed. New York, John Wiley & Sons (1989)
4. D'Agostino R.B., Belanger A.J., Markson E., Kelly-Hayes M. and Wolf P.A. : Development of health risk appraisal functions in the presence of multiple indicators: the Framingham Study nursing home institutionalization model. Statist. Med. **14** (1995) 1757–1770
5. Douglas G. Altman: Predictal statstics for Medical Research, CHAPMAN &HALL/CRC. rep.1999
6. Shortliffe E.H., Perreault L.E., Wiederhold G., Fagan L.M.: Medical Informatics. Springer-Verlag, New York (2001) 64–131
7. Grémy F., Salmon D. : Bases statistiques. Dunod, Paris (1969)
8. Wilhelmsen L., Wedel H., Tibblin G. : Multivariate Analysis of Risk Factors for Coronary Heart Disease. Circulation. **48** (1973) 950–958
9. Brand R.J., Rosenman R..H., Sholtz R.I., Friedman M. : Multivariate Prediction of Coronary Heart Disease in the Western Collaborative Study Compared to the Findings of the Framingham Study. Circulation **53** (1976) 348–355
10. Coronary Risk Handbook. Estimating Risk of Coronary Heart Disease in Daily Practice. American Heart Association (1973)
11. SAS Institute Inc., SAS Technical Report P-200, SAS/STAT Software: Logistic Procedure, Release 6.04, Cary. NC: SAS Institute Inc. (1990) 175–230
12. SAS Institute Inc. SAS User's Guide Statistics,Version 6.03 Edition, SAS Institute Inc., Cary North Carolina (1990)
13. Cureton E.E., D'Agostino R.B. : Factor Analysis, An Applied Approach. Erlbaum Publishers, New Jersey (1983)