

Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis

JEROME CORNFELD

Biometrics Research Branch, National Heart Institute, Bethesda, Maryland

THE ASSOCIATION BETWEEN INCREASED RISK OF CORONARY heart disease and elevated levels of serum cholesterol and systolic blood pressure is well known. Several questions about the magnitude of this association remain unanswered, however. Thus, are there critical values of serum cholesterol (or blood pressure) below which no relation between disease risk and cholesterol value exists, but which when exceeded are accompanied by abrupt elevations in risks? If no critical value exists is it nevertheless true that a given elevation is associated with a larger "effect" when added to a high rather than a low level? Does the apparent effect of serum cholesterol persist when levels of systolic blood pressure are held constant, and does the apparent blood pressure effect persist when serum cholesterol is held constant? If they do, are the effects independent of each other or do they, in the sense of the pharmacologists, potentiate each other? Answers to these and related questions offer no difficulty in principle. They require continued observation of the frequency with which coronary heart disease develops in a well-defined population, for each member of which measurements of the magnitudes of possible risk factors have been made. Several such longitudinal studies are now in progress. The present analysis is based on the long-term follow-up study of heart disease in Framingham, Massachusetts (2). The results discussed were obtained from clinical examination of a sample of 1,329 of the male population (aged 40-59) of the town carried out from September 1948 through August 1952 and from their subsequent follow-up over a 6-year period. During this period 92 of the original sample developed clinically manifest coronary heart disease (myocardial infarction or angina pectoris). The combination of these two conditions into a single category is not, of course, intended to imply that they necessarily have the same etiology.

Table 1 shows the ratio of number of new events to number of men exposed for different combinations of serum cholesterol and systolic blood pressure after 6 years of follow-up. Certain qualitative conclusions emerge from an examination of this table. Thus, for

those with serum cholesterol below 200 mg/100 cc the associated disease risk increases from 2 new cases in 6 years out of 119 exposed at systolic blood pressures below 127 mm Hg to 4 out of 26 at pressures above 167 mm Hg. A similar increase in risk with increasing blood pressure is suggested for the other levels of serum cholesterol shown. In the same way for those with blood pressures below 127 mm Hg the 6-year risk increases from 2/119 for those with cholesterol levels below 200 to 7/74 for those with cholesterol levels above 260. A similar increase in risk with increasing cholesterol level is suggested for each of the other blood pressure groups shown.

Simple inspection of the results is thus sufficient to suggest certain qualitative conclusions—the absence of a critical value for either cholesterol or blood pressure and the persistence of at least some association with each variable when the value of the other is held constant within broad limits. The thinness of the data nevertheless imposes a clear limit to the kinds of conclusions that this form of analysis will support. Inspection of the coarse groupings of Table 1 is, for example, hardly sufficient to indicate the way in which the effects of cholesterol level and blood pressure combine to influence the risk of the disease. Because of these limitations and because additional information will accumulate only slowly, one is led to seek a more searching form of analysis than simple inspection. The use of a mathematical model which summarizes the observations in a small number of disposable parameters seems to offer the only present hope of obtaining quantitative answers to questions of interest.

The model that we use is that of discriminant functions (1, 4), but because our objective is not that of discriminating between two populations, it is useful to start with a brief description of the method, emphasizing those aspects most important for the present application. We begin with the case of one variable and consider as two separate populations those who did (CHD) and did not (NCHD) experience a new coronary event during the study period. Each of these populations is

characterized by a frequency distribution with respect to the variable being considered. Thus, the CHD frequency distribution is such that the estimated proportion with serum cholesterols below 200 mg is 12/92, whereas NCHD frequency distribution is such that the estimated proportion with serum cholesterols below 200 mg is 307/1237. The essential characteristic of the method of discriminant functions is that for the observed frequencies one substitutes a mathematical formula describing a theoretical frequency distribution. The particular theoretical distribution used should, of course, be compatible with the observations and should be completely characterized by the values of a small number of disposable parameters, such as the mean or standard deviation.

Deferring the choice of distribution momentarily, suppose that the proportion of CHD individuals with serum cholesterols between Y and $Y + h$ can be adequately described by some simple function which we denote by $f_1(Y)h$, when h is reasonably small, while that for the NCHD population is $f_0(Y)h$. Denote the proportion of the combined populations who belong to the CHD population by p ($= 92/1329$). Finally, introduce the risk function, $P(Y)$, which denotes the proportion of individuals with serum cholesterol of Y who belong to the CHD population. Then elementary algebraic manipulation, or an application of Bayes' formula (3), is sufficient to show that

$$P(Y) = \frac{1}{1 + (1 - p)f_0(Y)/pf_1(Y)} \quad (1)$$

Thus, if theoretical frequency distributions can be found which adequately characterize the CHD and NCHD populations, then the risk function can be deduced from equation 1. An excellent description of both Framingham CHD and NCHD distributions with respect to cholesterol is provided by the log normal distribution. That is to say, in both populations log cholesterol has a normal distribution (Fig. 1 of ref. 1).

If f_0 and f_1 in equation 1 are replaced by appropriate expressions for normal frequency distributions with different means but the same standard deviation, it is easy to show that

$$P = 1/[1 + e^{-(\alpha + \beta X)}] \quad (2)$$

where

- $X = \log_{10}$ cholesterol
- $\beta = (\mu_1 - \mu_0)/\sigma^2$
- $\mu_1 =$ mean \log_{10} cholesterol for the CHD population
- $\mu_0 =$ mean \log_{10} cholesterol for the NCHD population
- $\sigma^2 =$ the common variance of \log_{10} cholesterol
- $\alpha = -\log_e (1 - p)/p - \beta[(\mu_0 + \mu_1)/2]$

The curve (2) is S shaped, starting at $P = 0$ and increasing up to a level of $P = 1$. In the range of observations made up to now, e.g., with P less than one-fourth, the full sigmoid appearance is not apparent and the curve appears exponential. For many purposes it is convenient to express equation 2 in the linear form:

$$\log_e P/(1 - P) = \alpha + \beta X \quad (3)$$

The effect of inequality of the variances in the two populations, an inequality not suggested by present

TABLE 1. Ratio of Number of New Events in 6 Years to Number Exposed to Risk by Initial Systolic Blood Pressure and Serum Cholesterol

Serum Cholesterol, mg/100 cc	Blood Pressure, mm Hg				
	Total	<127	127-146	147-166	167+
Total	92/1329	20/408	28/555	20/224	24/142
<200	12/319	2/119	3/124	3/50	4/26
200-219	8/254	3/88	2/100	0/43	3/23
220-259	31/470	8/127	11/220	6/74	6/49
260+	41/286	7/74	12/111	11/57	11/44

Framingham results, is to add a third term to the right hand side of equation 3 of the form γX^2 , where

$$\gamma = \frac{1}{2} \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \quad (4)$$

(and to change the value of α).

The same analysis applies to the single variable, systolic blood pressure, except that the frequency distribution of log blood pressure is not quite normal. This can be seen by plotting on probability paper the cumulative frequencies against log blood pressure. The curve is not linear, as it would be if normality applied. The plot against the logarithm of (blood pressure - 75) is linear, however, so that if Y denotes blood pressure, $\log_{10}(Y - 75)$ has a normal frequency distribution for both CHD and NCHD populations.

To consider both the cholesterol and systolic blood pressure variables simultaneously, it is sufficient to find frequency distributions in the two variables which characterize both CHD and NCHD populations. If a frequency distribution in a single variable is thought of as a curve in which frequency is plotted against X , then a bivariate frequency distribution can be thought of as a surface in which the frequency is plotted against the values of both variables, say X_1 and X_2 . Empirical analysis of the Framingham results indicates that both CHD and NCHD populations are described by bivariate normal frequency functions in \log_{10} cholesterol and \log_{10} (blood pressure - 75) with differing means but having equal variances and correlation coefficient (section 9 of ref. 1).

The equivalent of equation 2 is then:

$$P = 1/[1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2)}] \quad (5)$$

where

- $X_1 = \log_{10}$ cholesterol
- $X_2 = \log_{10}$ (blood pressure - 75)

and α , β_1 , and β_2 are easily computed functions of the means of log cholesterol and log (blood pressure - 75) of the CHD and NCHD population, and of the variances of the two variables and correlation coefficient between them. The linear version of equation 5 is

$$\log_e P/(1 - P) = \alpha + \beta_1 X_1 + \beta_2 X_2 \quad (6)$$

Formulas for approximate confidence limits on the coefficients β_1 and β_2 are available (formula 7.8 of ref.

1). The right-hand side of equation 6 is usually referred to as the discriminant function (α is often omitted), since its numerical value may discriminate between those likely and unlikely to become members of the CHD

TABLE 2. New CHD in 6 Years Follow-Up: Actual and Expected Number of Cases Among Men 40-59

Serum Cholesterol, mg/100 cc	New CHD		Population at Risk
	Actual	Expected*	
200	12	10.5	319
<117	0.8	0.8	53
117-126	1.4	1.4	66
127-136	1.8	1.8	59
137-146	2.3	2.3	65
147-156	1.6	1.6	37
157-166	0.7	0.7	13
167-186	1.4	1.4	21
187+ over	0.5	0.5	5
Total	5.9	5.9	133
<117	0.5	0.5	21
117-126	0.8	0.8	27
127-136	1.4	1.4	34
137-146	1.0	1.0	19
147-156	1.0	1.0	16
157-166	0.7	0.7	10
167-186	0.4	0.4	5
187+ over	0.1	0.1	1
Total	6.9	6.9	121
<117	0.4	0.4	15
117-126	0.9	0.9	25
127-136	1.0	1.0	21
137-146	1.5	1.5	26
147-156	0.4	0.4	6
157-166	0.9	0.9	11
167-186	1.1	1.1	11
187+ over	0.8	0.8	6
Total	22.3	22.3	334
<117	0.6	0.6	20
117-126	2.9	2.9	69
127-136	4.7	4.7	83
137-146	6	6	81
147-156	2.4	2.4	29
157-166	1.5	1.5	13
167-186	3.1	3.1	27
187+ over	1.6	1.6	10
Total	11.6	11.6	196
<117	0.5	0.5	14
117-126	1.3	1.3	24
127-136	2.3	2.3	33
137-146	2.0	2.0	23
147-156	1.9	1.9	19
157-166	1.3	1.3	11
167-186	0.7	0.7	5
187+ over	1.5	1.5	7
Total	16.0	16.0	156
<117	1.0	1.0	22
117-126	1.4	1.4	22
127-136	2.1	2.1	26
137-146	3.5	3.5	34
147-156	1.9	1.9	16
157-166	1.7	1.7	13
167-186	2.7	2.7	16
187+ over	1.5	1.5	7

Serum Cholesterol, mg/100 cc	New CHD		Population at Risk
	Actual	Expected*	
285+ over	18	18.9	130
<117	0.7	0.7	11
117-126	1.7	1.7	19
127-136	3.1	3.1	28
137-146	4	4	23
147-156	2.8	2.8	16
157-166	4	4	12
167-186	3.2	3.2	14
187+ over	1.9	1.9	7
Total	18.9	18.9	130

* Expected number of cases = (population at risk)/(1 + $\frac{1}{2}(\alpha_1 X_1 + \alpha_2 X_2)$)

population. Our present interest is not in discrimination in this sense, but in using equation 6 to study the quantitative nature of the dependence of risk on serum cholesterol and systolic blood pressure levels. If the variances and correlation coefficients of the CHD and NCHD populations had been unequal three terms would have been added to the right-hand side of equation 6 of the form $\gamma_1 X_1^2 + \gamma_2 X_2^2 + \gamma_3 X_1 X_2$.

Granting the bivariate normal model, all the data on which Table 1 is based can then be summarized in the six constants $\alpha, \beta, \beta_1, \beta_2, \gamma_1, \gamma_2$, and γ_3 , and the answers to any questions about the joint dependence of risk on cholesterol and blood pressure levels are contained in them. Because of the near equality of the observed variances and correlation coefficient in the two populations the last three constants can be treated as zero. The estimates for the other three are: $\beta_1 = 6.14$ (3.35-9.00), $\beta_2 = 3.29$ (1.75-4.88), and $\alpha = 23.13$, where the numbers in parentheses are 95% confidence limits. Table 2 compares the actual and expected number of new CHD cases obtained on this basis. The agreement appears satisfactory.

The following approximation simplifies the subsequent discussion. When the value of P is small (say $> 1/3$), $\log P/(1 - P)$ can be written as $\log P$,* in which case it is easy to verify (remembering that logs of cholesterol and blood pressure are to base 10) that

$$P = .0091 \left(\frac{100}{X_1} \right)^{2.68} \left(\frac{100}{X_2 - 75} \right)^{1.41} \quad (7)$$

where X_1 and X_2 are serum cholesterol (in mg/100 cc) and systolic blood pressure (in mm Hg). The confidence limits on the exponents are from 1.45 to 3.90 and 0.76 to 2.12.

We may now return to the questions with which we started. It is clear first of all that the description provided by equation 7 is incompatible with the idea of a critical value for either cholesterol or blood pressure. Although the notion of decision values, e.g., of 260 mg for cholesterol or 160 mm Hg for systolic blood pressure, may be used for some of the values of X_1 and X_2 observed, P considered as the 6-year risk is too large for this approximation to be entirely accurate. For the 1-year risk, which is one-sixth of the 6-year risk, the approximation holds for all observed X_1 and X_2 .

CORONARY HEART DISEASE RISK

convenient clinically, these values are not associated with abrupt changes in risks.

Rather than ask about critical values, however, we may ask how the change in risk associated with changes in either variable depends on the level from which the change takes place. This includes the question of critical values as a special case. It is important here to distinguish between absolute and relative changes. Because both exponents in equation 7 exceed unity, a given absolute increase in either variable will be associated with a larger absolute increase in risk when the change takes place from a high rather than a low level. A given difference in milligrams per 100 cc of cholesterol will thus be associated with a larger absolute difference in risk when added to a large rather than a small initial level. A similar conclusion applies to systolic blood pressure, except that since the lower 95% confidence limit is below unity, it is less certain.

The conclusion is different, however, if one considers percentage changes. Because of the form of equation 7 a given percentage difference in cholesterol is associated with the same percentage difference in risk no matter what the level to which the difference is added. A 1% difference in cholesterol is associated with a 2.66% difference in risk throughout the range of cholesterol values. If one would lower cholesterol levels by a given amount, say 15%, and if equation 7 describes not only the association between risk and cholesterol level in Framingham, but also the change in risk that would accompany an experimental alteration in serum cholesterol, then the relative risk would be $(.85)^{2.66}$, or a reduction of about 35%. This would be true no matter what the level of cholesterol from which the 15% reduction occurred.

Because of the subtractive 75 in systolic blood pressure, given percentage changes in blood pressure are associated with larger percentage changes in risk when starting from low than from high blood pressure. A 1% difference in systolic blood pressure will be associated with a percentage difference in risk of $1.47 Y/(Y - 75)$, where Y is the starting level. Thus, at a starting blood pressure of 110 mm Hg a 1% difference in blood pressure is associated with a 4.62% difference in risk. This contrasts with a starting level of 175 mm Hg, where a 1% difference is associated with a 2.57% difference in risk.

For both cholesterol and blood pressure the form of equation 7 implies that the percentage effect on risk of a given percentage change in either variables is independent of the level of the other variable. No matter what

the blood pressure a 1% difference in cholesterol is associated with a 2.66% difference in risk. A 1% increase in systolic blood pressure from 110 mm Hg is associated with a 4.62% difference in risk no matter what the cholesterol level. Finally, from equation 7 the effect of simultaneous differences in blood pressure and cholesterol on percentage differences in risk is multiplicative. A simultaneous 1% change in cholesterol and in blood pressure from a starting level of 110 mm Hg will result in a relative risk of $(1 + .0266)(1 + .0462)$ %. A useful way of summarizing these relations is by considering the percentage difference in risk associated with cholesterol values at the 5th and 95th percentiles, with blood pressures at the 5th and 95th percentiles, and with both cholesterol and blood pressure at their 5th as compared with their 95th percentiles. These are summarized below.

	Risk 95th Percentile Relative to 5th Percentile
Cholesterol alone (166-301 mg)	4.8
Blood pressure alone (110-177 mm)	4.8
Both	23.4

The population in the upper 5% of either the cholesterol or blood pressure distribution has a risk more than 4.8 times as large as that in the lower 5%, while individuals who are in the upper 5% with respect to both (about $\frac{1}{4}$ of 1% of the population) have a risk more than 23.4 times as large as those who are in the lower 5% with respect to both.

SUMMARY

1) Discriminant functions have been used to describe the relationships in the Framingham population between risk of developing coronary heart disease, serum cholesterol level, and systolic blood pressure.

2) Critical values of either variable, at which sharp increases in risk occur, were not found.

3) A 1% difference in serum cholesterol is associated with a 2.66% difference in risk at all levels of serum cholesterol.

4) A 1% difference in systolic blood pressure is associated with a 4.62% difference in risk at blood pressures of 110 mm Hg and of 2.57% at 175 mm Hg.

5) There is an almost fivefold difference in risk between the upper and lower 5% with respect to serum cholesterol or systolic blood pressure alone, and an almost 25-fold difference between those who are in the upper 5% of the population with respect to both variables and those who are in the lower 5% with respect to both.

REFERENCES

- CORNFIELD, J., T. GORDON, AND W. W. SMITH. *Bull. Intern. Statistical Inst.* 38: 97, 1961.
- KANNEL, W. B., T. R. DAWBER, A. KAGAN, N. REVOTSKIE, AND J. STOKES III. *Ann. Internal Med.* 55: 33, 1961.
- PARZEN, E. *Modern Probability Theory and Its Applications*. New York: Wiley, 1960, p. 119.
- RAO, C. R. *Advanced Statistical Methods in Biometric Research*. New York: Wiley, 1952.