Editorial:
Diet and Coronary Heart Disease:
The Statistical Analysis of Risk

DAVID W. SCOTT, PH.D., G. ANTHONY GORRY, PH.D., AND
ANTONIO M. GOTTO, M.D., D. PHIL.

DESPITE reports of more than a dozen authoritative bodies over the past 10 years, the relationship between diet and coronary heart disease (CHD) remains controversial. Comparisons of different populations have indicated a correlation between the consumption of saturated fat and plasma cholesterol levels, and similar studies have established the role of elevated plasma cholesterol as a risk factor for CHD. Within a given population, however, the correlation between dietary saturated fat and plasma cholesterol has not been established outside the controlled conditions of a metabolic ward. In Western societies, it is commonly recommended that consumption of saturated fat be reduced, but whether such a reduction in plasma cholesterol levels would result in a decrease in the incidence of CHD has not been established. In the World Health Organization's trial with clofibrate, the therapeutically induced reduction in cholesterol reduced nonfatal myocardial infarctions, but the overall mortality in the treated group after 4 years was 25% greater than that in the control groups. A recent report by the National Research Council of the National Academy of Sciences has heightened the argument concerning the value of dietary change in the general population by recommending that the "normal, healthy" American need not be concerned with dietary intake of fat or cholesterol.

Gordon et al. offered further evidence on the matter of diet and CHD through a study of three populations. They found that consumption of alcohol, starch and total calories in the diet each had an inverse relationship with CHD. These findings do not contradict the view that reduction of dietary saturated fat may decrease CHD in Western society. The substitution of starch for fat is illustrated in the Puerto Rico and Honolulu studies. The negative relationship between CHD and total caloric intake is consistent with beneficial effects of physical activity. The inverse relationship between CHD and alcohol consumption deserves some comment. Alcohol consumption is known to raise the level of HDL cholesterol. However, alcohol intake has several negative aspects, including the higher overall mortality in the present study, the relationship with cirrhosis of the liver, the association between high levels of alcohol consumption and hypertension, an alcohol-related cardiomyopathy and the acute effect of alcohol on reducing myocardial contractility.

Statistical studies of the relationship of diet to the development of CHD are tantalizing but not satisfying. Small but statistically significant differences in dietary intake may be suggestive, but the extent to which such statistical relationships reflect underlying causal relationships in the development of CHD is uncertain. This problem is compounded by uncertainty as to the exact nature of the statistical relationships themselves, for fallacious conclusions could be drawn about the nature of the dietary risk if an inappropriate or oversimplified statistical model for data analysis were used.

In this report we examine the problems of making sound statistical inferences about diet as a risk factor for CHD. The term "risk factor" denotes a characteristic that is disproportionately associated with people who develop a particular disease. Thus, for example, elevated plasma cholesterol is a risk factor for CHD because people who develop the disease, on the average, have higher cholesterol concentrations than those who do not. Similar associations with CHD have been found for smoking, hypertension, age and diabetes.

Suppose a particular characteristic is thought to be positively associated with an increased risk of developing CHD. Then, all else being equal, a person who has this characteristic should be more likely to develop the disease than one who does not. Alternatively, we can say that the odds of developing CHD in this patient should increase as the result of having this characteristic; for example, a plasma cholesterol level above 250 mg/dl. Using Bayes's rule, we can express the revised odds of CHD in this patient given the risk factor level "X" and the relative chance of "X" in disease and normal patients who are otherwise similar to the patient in question:

\[ O(CAD \mid X) = \frac{P(X \mid CHD)}{P(X \mid no \, CHD)} \times O(CHD) \]  (1)

Here, \( O(CHD \mid X) \) denotes the revised or "posterior" odds of developing CHD with the additional knowledge of the presence of a risk factor at the level "X"; \( O(CAD) \) is the former or "prior" odds; \( P(X \mid CHD) \) is the probability of finding level X in patients developing CHD; and \( P(X \mid CAD) \) is the probability of finding level X in patients not developing CHD. The functions P( ) are called probability density functions and essentially look like smooth
histograms. If the ratio of the two density functions in equation 1 for a particular level of X is greater than 1, then the posterior odds of developing CHD are increased and vice versa. A risk factor like plasma cholesterol varies greatly among diseased and nondiseased persons, so a more complete characterization of its role as a risk factor requires the determination of this ratio (the likelihood ratio) over the range of possible levels X of the risk factor.

A common statistical model for the density function of a risk factor X (such as plasma cholesterol) is the normal or Gaussian curve. Figure 1 is a sketch of normal curves for P(X | CHD) and P(X | no CHD) that have the same standard deviation, $\sigma$, but different means, $M_1$ and $M_2$, respectively, a common model that we call the shifted Gaussian model for the natural variations of a risk factor in two groups.

The likelihood ratio is also plotted in figure 1. For values of X above a certain cutoff level, a patient is more likely to develop CHD; below that cutoff level, less likely. An important characteristic of the risk factor is that its likelihood ratio is monotonic; that is, increasing levels of the risk factor imply increasing risk of developing CHD. In fact, the slope of the likelihood ratio curve is essentially $(M_1 - M_2)/\sigma$, the so-called coefficient of detection. One can show that when the probability densities of a risk factor follow the shifted Gaussian model, then the estimated risk is always monotonic. A t test statistic (which looks almost exactly like the coefficient of detection) may be used to decide if there really is a shift (or difference) in the means of the two density functions. In tables 2-4 of their paper, Gordon et al. make extensive use of t tests to detect differences in the diets of men in the three populations, so they apparently assume a normal distribution of their data.

Tables 2-4 of Gordon et al. illustrate a problem in statistical inference. The authors state that only risk factors with significant t tests will be entered into later multivariate regression analyses. Of the 171 t tests performed, we would expect by chance about 5%, or nine, of the differences to be significant, although the true differences are 0. Which nine of the 38 significant differences found are false is impossible to answer. Methods for appropriately determining statistical significance with so many individual tests have been discussed by Miller* and others.

However, suppose that the Gaussian density functions P( ) have equal means, but different standard deviations (fig. 2). The likelihood ratio function that results in figure 2 captures several unexpected features of the nature of the risk associated with X. First, the risk is initially decreasing as X increases, but then the risk increases for further increases in X. In fact, the risk of developing CHD is increased for either extremely reduced or elevated levels of the risk factor. A t test would show no shift in the means and would lead to an incorrect conclusion that this risk factor was unimportant.

Gordon et al. found that the characterization of alcohol consumption as a risk factor was a problem. While the authors accept the statistical significance of the mean-shift in alcohol consumption, they question the monotonicity of the risk factor, although monotonicity is built into their model. The authors constructed tables 5-7 so as to use a statistical model that does not depend on the monotonicity assumption, but the trends in these tables are not easily discernible. One hypothesis is that the situation for the risk factor alcohol is a combination of situations (figs. 1 and 2); that is, different means and variances.

Another situation in which the risk may not be monotonic may occur often. One or both of the probability density functions P( ) may have two or more peaks or modes (as opposed to Gaussian densities, which have exactly one mode). Only recently

![Figure 1. Shifted Gaussian model.](image)

![Figure 2. Gaussian models with same mean but different variables.](image)
have statistical tools been developed that can detect multiple modes in data, namely, nonparametric density estimation techniques. We have used those techniques to identify multiple modes in joint density functions of cholesterol and triglyceride in patients with CHD. For multiple modes, the risk as displayed in the likelihood ratio often resembles an S shape on its side, increasing, then decreasing, and then increasing again with increasing levels of the risk factor. Logistic regression as used by Gordon et al. is a powerful tool for analyzing risk factors, particularly in the multivariate case where one desires to consider the role of several risk factors simultaneously in predicting the development of CHD. When the shifted Gaussian statistical model provides an adequate model for the data, there is a fundamental relationship between the logistic regression coefficients and the means and standards deviations of the normal density. In particular, for the univariate logistic regression case, the regression coefficient is essentially the same as the coefficient of detection discussed in relation to $t$ tests. Thus, univariate logistic regression really provides no new information about risk, but only serves as a reference for the more important multivariate case. Once again, the monotonic increase of risk is a characteristic of the logistic regression model.

Our report has two purposes. The first is to emphasize the importance of the statistical model or technique chosen for analyzing medical data. Inappropriate conclusions can result from the model, not the data. For example, the monotonicity of a risk factor should be demonstrated by the statistical analysis of the data, not assumed a priori by a restricted choice of models. There will always be a compromise between the generality of the statistical model and the number of risk factors that may be considered. Logistic regression is the tool of choice when dealing with large numbers of risk factors, but the model may fail to capture important aspects of the data.

Our second purpose is to present a more general model for thinking about risk factors and for analyzing data. We have used a combination of this Bayesian model and nonparametric density estimation techniques to provide a useful statistical tool for analyzing risk factors with fewer assumptions. While it is true that such techniques generally require large amounts of data, the effort by Gordon et al. is an example of a developing trend toward the collection of large data sets for analysis. The conclusions of Gordon et al. from their data analysis are interesting and probably correct in the main. However, the use of other statistical models might illuminate interesting and important features of the relationships of diet to CHD.

References
Diet and coronary heart disease: the statistical analysis of risk.
D W Scott, G A Gorry and A M Gotto

Circulation. 1981;63:516-518
doi: 10.1161/01.CIR.63.3.516

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/63/3/516.citation

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org//subscriptions/