AN IMPORTANT SOURCE of current misunderstanding and controversy about diet-disease relationships is the inappropriate extrapolation of evidence from group data to the individual and vice versa. In this editorial piece the difference is illustrated between correlations of risk factors and disease found in individuals and those found between populations. An attempt is made to clarify the causal inference possible when these correlations are concordant and discordant. An analytic approach is encouraged in which both sources of information are considered, i.e., relationships among individuals and between populations, as well as their consistency with other data, before arriving at inference about the meaning of diet-heart, diet-cancer, or, for that matter, any statistical association. It is further suggested here that factors that determine the average and usual population levels of risk characteristics are likely to be the major determinants of mass disease. These influences on population risk may be the same as, or different from, those factors that predominantly influence risk factor level and the risk factor–disease relationship among individuals. We demonstrate that the causes of individual and population correlations are mathematically and conceptually independent. From an understanding of these ideas, the fallacy may be avoided of rejecting important causal influences in populations when relationships appear weak or absent in an individual patient—as well as the inappropriate application of group data to the individual, which is called "the ecological fallacy."1, 2, *

From this point of view a rational preventive approach is derived in which risk assessment and therapy are appropriately tailored to the individual in preventive practice but in which a complementary strategy attempts to shift downward the entire population distribution of risk. The goal of this strategy is the most effective control and prevention of disease and eventually prevention of elevated risk in the first place.3

Case I. Concordant positive correlations for populations and individuals

Figure 1, A, illustrates schematically the concordant condition between population and individual correlations. A significant, positive, and continuous relationship is found between population rates of disease and mean population levels of a risk factor. Similarly, within a given population, a strong correlation is found between individual values for the risk factor and future personal risk of disease.

Example. An example of this situation is the positive correlation of plasma total cholesterol (TC) levels or low-density lipoprotein (LDL) levels with the risk of coronary heart disease (CHD). In the population or ecologic correlation, the rates of CHD are highly correlated with mean population TC values.4 More dramatically, average TC level actually predicts future population rates of CHD deaths and the precision of this prediction increases over time.5 In addition, the risk of CHD for an individual living in populations having moderate or high CHD incidence is significantly related to the individual level of TC.6 Thus, the positive, graded relationship is concordant for populations and for individuals. Individual variations are averaged out by computation of the population correlation and a significant population effect emerges. On the other hand, when the environment is held constant and individual correlations are computed within a population, an individual TC-CHD relationship is demonstrable across the range of values encountered.

Inference. The fact that the TC-CHD relationship holds both for individuals and for populations suggests a single source of the relationship or a single cause. It suggests that TC is on the pathogenetic pathway for CHD. Moreover, the major lipid "pathogen," which

Vol. 70, No. 5, November 1984

775

*For those interested, we have filed in the National Auxiliary Publication Service a short appendix providing a unified mathematical model from which both individual and population relationships may be determined. It contains terms to represent levels for populations and for individuals; individual and population correlations are shown to be distinct and separate mathematical entities. See NAPS document No. 04230 for 2 pages of supplementary material. Order from NAPS c/o Microfiche Publications, P.O. Box 3513, Grand Central Station, New York, NY 10163. Remit in advance in U.S. funds only. $7.75 for photocopies or $4.00 for microfiche. Outside the U.S. and Canada add postage of $4.50 for the first 20 pages and $1.00 for each additional page; $1.50 for microfiche postage.
A SCHEMATIC OF
POPULATION (P) VERSUS INDIVIDUAL (I) CORRELATIONS

FIGURE 1. Schematic representation of population (P) and individual (I) correlations between risk characteristics and disease, showing numerous possible combinations of their concordance and discordance.

we assume to be an excess of LDL, is prevalent enough to produce a strong population effect and a significant disease burden. These concordant correlations, along with experimentally demonstrated TC (and LDL) responses to dietary change, suggest that habitual diet is a powerful environmental influence on population values and that many individuals are intrinsically susceptible to its effect. In this condition, because the significant concordant associations are accompanied by plausible and demonstrated mechanisms, causal inference is strong and the potential for prevention appears great.

Case II. Concordant negative correlations for populations and individuals

In the concordant condition illustrated schematically in figure 1, B, there is a significant, continuous, and negative association between population rates of disease and mean values for the risk factor. Similarly, individuals with higher levels of the factor have lower disease risk.

Example. There is no well-established example of this condition in cardiovascular epidemiology, although diets high in carbohydrate and fiber, the almost necessary inverse of high-fat diets, may be suggested.7

Inference. As with concordant positive correlations, it may be inferred that in this hypothetical condition a single factor (or its absence) is likely to operate, affecting importantly both individual and population risk. That determinant is likely to be an environmental-cultural factor.

Case III. Discordant correlations: positive population and absent individual correlations

In the discordant correlations illustrated schematically in figure 1, C, a positive association is found between a population phenomenon and mean levels of a risk factor, whereas, within populations, significant individual correlations are not found.

Example. This condition is a frequent, major, and we believe, inappropriate source of diet-heart controversy. Two prominent examples exist from cardiovasc-
lar population studies: (1) the substantial correlations found between the composition in saturated fat and cholesterol of the habitual diet of populations, their average TC (LDL) levels, and their incidence of CHD, and (2) the correlation between average salt (sodium) intake of a population and its average level of blood pressure or frequency of adult hypertension. In contrast, correlations between ordinary measures of diet, TC, and CHD, or of sodium intake, blood pressure, and hypertension, made in individuals within a population are usually weak or absent.

Inference. This condition is often encountered in high-exposure, high-incidence populations. However, it is important to avoid the fallacy of rejecting causation out of hand. Rather, this condition requires that the search go further, particularly if the ecologic (population) correlation is strong and if plausible mechanisms for a causal relationship exist from clinical or laboratory evidence. Commonly, the methods of risk measurement to characterize the individual (as in diet) are inadequate as single measures. Finer methods and repeat measurements, along with more careful study design, tend to reduce variability. Real and significant individual associations may then emerge.

Case IV. Discordant correlations: positive population and negative individual correlations

In the unusual discordant condition illustrated in the schema of figure 1, D, a positive correlation is found between mean risk factor levels and population events, while a significant but negative correlation exists for the same characteristic of individuals within a population.

Example. An example of this less common condition may be the relationship between TC levels and cancer of the colon. Populations having lower average TC values generally have lower colon (and other) cancer death rates. In contrast, half the systematic studies carried out within populations show a greater risk of cancer of the colon among individuals with low TC levels.

Inference. In this discordant condition, when individual correlations are found inconsistently, and in this case only in men, it may be reasonable to infer that the characteristic, if causal, is intrinsic rather than environmental, affecting some individuals but not the population overall. If it were an environmental influence, it may be inferred that it is not a direct cause. Again, in this discordant condition, further search is needed for explanations and underlying mechanisms. For this specific instance, TC and cancer of the colon, there are, so far, no established mechanisms.

In computing the population correlation in this condition, individual variation in the intrinsic factor or its effects that lead to a low TC level is averaged out. It would be fallacious, however, to extrapolate the association between low TC levels and excess cancer risk in individuals to an argument, for populations, that low TC, or TC-lowering diets would increase the population risk of cancer. A public health strategy of encouraging eating patterns to shift downward the mean population level of TC in a preventive approach to CHD would not necessarily affect unfavorably cancer rates in the population — to the contrary. On the other hand, it would be equally fallacious to infer from the finding of low cancer risk in populations having low average TC levels that an individual would necessarily reduce his risk of cancer by lowering his blood cholesterol level.

Case V. Discordant correlations: absent population and significant positive individual correlations

In figure 1, E, the discordant situation is illustrated schematically in which the population correlation is absent while a significant positive correlation is found for individuals within a population.

Example. An example of this condition may be the finding of a significant, graded, and independent relationship of work and leisure-time physical activity habits to individual CHD risk, and the contrasting finding that activity and fitness measured in populations are unrelated to their CHD disease rates.

Inference. Here again, it would be fallacious to reject automatically a causal influence of physical activity on CHD because of the absence of concordant correlations. In fact, logical mechanisms exist for a protective effect from atherosclerosis and CHD in the metabolic and physiologic effects of regular physical activity and fitness, especially among individuals who live in high-CHD incidence populations. However, risk factor effects and interactions appear to be culture-dependent, e.g., a protective effect of one factor may be overwhelmed by the excess risk from a pathogenic factor. A good example may be the mass hypercholesterolemia and high cardiovascular disease rates found among the physically active and fit rural Finns.

Case VI. Discordant correlations: absent population and significant negative individual correlations

In figure 1, F, is illustrated the last permutation considered here of concordant and discordant correlations. There is no significant population correlation, while the individual correlation is significant but negative.

Example. A very good example of this condition is
the significant negative association found between individual levels of plasma high-density lipoprotein (HDL) and future CHD risk in several Western societies. This phenomenon among individuals is apparently found only in populations in which there is relative hypercholesterolemia (due to high LDL) attributable, we believe, to mass responses to a ubiquitous and habitual diet high in saturated fat.\textsuperscript{13, 14} In contrast, it has been shown that mean HDL levels in populations correlate poorly with population TC levels and presumably correlate little with population diet or CHD rates.\textsuperscript{6}

**Inference.** Despite this discordant condition for HDL, the individual negative relationship within some populations is quite strong. Furthermore, logical mechanisms can account for a protective effect of HDL in individuals, especially within high-risk cultures. It is conceivable that this effect is not operative in the absence of mass phenomena leading to high LDL levels, the major diet-influenced pathogen for mass atherosclerosis.\textsuperscript{15} HDL level helps account, therefore, for considerable individual variation in CHD risk within a high-incidence population with mass hypercholesterolemia. In contrast, HDL apparently has little bearing on CHD incidence, or on the changes in CHD death rates observed in the United States and other countries in recent years. Changes in HDL levels may be less dynamic and less culturally influenced than those in LDL levels.

Figure 2 illustrates a final example of discordant correlations that has contributed to diet-heart controversy, i.e., the differing relationships found between habitual dietary fat intake and TC levels in populations and individuals. The lower slope represents the average response of individual TC levels to change in dietary fatty acid composition found in a classic series of metabolic ward experiments by Keys et al. in Minnesota.\textsuperscript{16} The regression is adjusted so that the experimentally induced individual change of 1 mg/dl in TC level corresponds to 1 unit of change in an individual’s dietary score. Over the years, this useful estimate of TC response to diet has provided confirmation of the effect of specific fatty acids and diets. Departures from expectation and good fit have also led to new knowledge about the effect of diet.

Here, in figure 2, the upper slope applies the Minnesota diet score derived from the correlation in individuals to the population condition, i.e., to the average fatty acid composition of the diets of the populations of the Seven Countries. It computes the regression on mean population TC levels.\textsuperscript{4} The figure shows that in these long-term “natural experiments,” population TC levels are about 2 mg/dl different for every 1 unit difference in the population dietary score. Thus, the dietary–total cholesterol association in populations is twice as strong as it is in individuals. Which is the biologic truth? What are the determinants of different degrees of correlation for individuals and for populations?

This discordance of population and individual diet effects illustrates further the fallacy of direct inference for individuals from a set of population data and vice versa. It illustrates further the different force of a causal influence on populations than on individuals. It is surmised for this case that the two predictions actually reflect the influence of the same biologic phenomenon, which we assume to be the effect of diet composition, but that the apparent inconsistency results from the lack of control over confounding factors in the population regression (including duration of exposure). In contrast, in the metabolic unit from which the individual correlation was derived, all dietary and other influences except the one under investigation were controlled.

**Discussion and conclusions**

The ecological fallacy is more broadly defined here than elsewhere as the inappropriate extrapolation of evidence from one set of relationships to a different set. Traditionally and properly, science gives more weight and greater inferential credence to individual than to population correlations. This is because population correlations may exaggerate associations because of confounding and of the way that individuals
are grouped to obtain aggregate measures. It is shown here, however, that the important determinants of individual correlations may differ in the force of their effect or actually be different from those resulting in population correlations. Thus, discordance between the correlations is insufficient grounds for rejection of causal effect. A characteristic may, in fact, operate strongly to cause population differences, despite weak individual correlations within populations, under several conditions: when population exposure to a risk factor is heavy (above a population threshold), when exposure is ubiquitous and homogeneous (such as the general U.S. exposure to high-salt or high saturated fat–cholesterol diets), and when variability of a factor measured within the individual approaches the variability found between individuals (whatever the reason: genetic variation in response, technically unreliable measurement, age, season, etc.). These conditions virtually guarantee that individual correlations will approach zero, as we have mathematically elaborated elsewhere. However, in this rather common condition, care to reduce individual variability by repeated, standardized measurements and by study of the relationship across the widest possible range available within a population will often elicit significant individual correlations. For example, multiple standard dietary assessments, which more accurately represent an individual’s diet than single measures, have caused to emerge significant correlations between individual dietary fat and cholesterol intake and individual TC levels and future CHD risk, even within the high-exposure, high-risk U.S. population. Similarly, individual dietary sodium intake (or the sodium/potassium ratio in urine) correlates with blood pressure level under conditions that reduce variation in measurements of individual sodium intake and blood pressure. Consequently, the absence of significant individual correlations should not lead automatically to the rejection of inference about causal factors of mass disease in populations. This is particularly the case when the population correlations are strong, when logical mechanisms exist to explain the relationships between populations, and, obviously, when the relationship can be established experimentally.

The criteria for drawing causal inference from statistical associations are well established. They include the strength and graded nature of the association, the power to predict future risk independently of other factors, a clear temporal relationship, consistency between studies, and congruence between scientific methodologies. To these well-established characteristics might now be added concordance between population and individual correlations. However, the main point made here is that discordance in these relationships does not necessarily mitigate causality. We need to pay attention to ecologic (population) correlations. Although concordant individual and population correlations are likely to emerge from some element in common, here it is illustrated that serious errors in interpretation can be made by extrapolation from a part to the whole, or from the whole to parts of a system—the ecological fallacy. It is also illustrated that discordant correlations can be informative.

These observations, including the mathematical demonstration of the independence of population and individual correlations, lead us to the further inferential extrapolation that important determinants of mass disease may be different, at least quantitatively, from those predominantly determining disease in individuals. We propose a strategy for simultaneous interpretation of individual and population correlations.

In our hierarchical mathematical model, the level of a given risk factor is algebraically separated into population and individual components. For comparison between individuals within a population, the population component is absorbed into the constant. For comparison between populations, across individuals, the individual components, which are deviations from the population level, are summed, becoming zero within each population. Thus, determinants of the individual and population components do not necessarily coincide and the differences may be subtle. For instance, sodium intake in Japan may be determined predominantly by traditional diets, including soy sauce and miso, a cultural characteristic that varies little among individuals. In contrast, in a Western country, primary determinants of sodium intake may be salt added to foods in processing, in cooking, or at the table. These sources of sodium are less ubiquitous and culture-bound than use of soy sauce or miso in Japan and are subject to more individual control and thus may vary greatly between individuals.

Under this model, individual and population regression coefficients are distinct and separate entities. The magnitude of each reflects the possibly different determinants of the individual and population components. With this model, the investigation of discordant population and individual correlations reduces to a search for differing determinants of the population and individual levels, i.e., influences in the mass and for the individual.

We postulate that generalized and powerful environmental factors, acting on widespread, inherent susceptibility, are probably responsible for the concordant
positive correlations illustrated here in the case of TC and CHD risk. Although intrinsic (genetic) factors account greatly for individual variation in levels and risk within a population, they may have relatively little to do with large differences in disease rates between entire populations. Computation of the population or ecologic correlation averages out these individual variations and lets emerge (or not) the major determinants of population differences. Thus, information from population and individual sources together is important to arrive at inference of cause, and, eventually, to develop rational preventive advice and public health policy. Contributions from the three major disciplines of medical science, i.e., those from clinical, laboratory, and population studies, are needed to examine disease-risk factor relationships completely. Information from one set of observations complements that from the other. Understanding is most complete with evidence from all.

It is likely that correlations between individual behavior affecting health, risk factor levels, and disease risk will remain elusive under the common condition in affluent societies today in which there is ubiquitous, high-level exposure to a risk (such as habitual diet) and in which there is substantial genetic variation in response. Moreover, significant individual correlations will remain difficult to find when individual behaviors and risk are weakly characterized (such as current dietary measures). Finally, causation will be difficult to "prove" when randomized controlled trials, such as the experimental study of long-term changes in life patterns in free-living communities, are not feasible. In fact, this discordant and uncertain condition is the real and usual one in which decisions on preventive practice and public policy must be formulated for many health and disease prevention issues in modern society. This is particularly the case for diet-disease relationships and diet policy. Greater awareness of the two types of correlations discussed here, population and individual, and their different determinants and meanings should help avoid the ecological fallacy and reduce unnecessary controversy.

References
Sources of the diet-heart controversy: confusion over population versus individual correlations.
H Blackburn and D Jacobs

Circulation. 1984;70:775-780
doi: 10.1161/01.CIR.70.5.775

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/70/5/775.citation