THE CANADIAN MEDICAL ASSOCIATION JOURNAL

LE JOURNAL DE L'ASSOCIATION MÉDICALE CANADIENNE

Volume 97 • Number 20 • November 11, 1967

Some Difficulties in the Investigation of Genetic Factors in Coronary Artery Disease

EDMOND A. MURPHY, M.D., Sc.D.,* *Denver, Colorado, U.S.A.*

THE few facts available on the familial aggregation of coronary artery disease have been repeatedly reviewed.[1-5] Table I contains a brief summary of the findings in the major studies. So condensed a presentation does not do justice to the more elaborate analyses used in these papers, but suffices to show how uniform the findings are. The risk ratio of coronary artery disease in relatives of controls to that in relatives of probands with coronary artery disease is in all four studies[6-9] less than unity: in many cases considerably less. The ratios of affected offspring in the study of Thomas and Cohen[10] suggest that the liability is greatest if both parents are affected and least if neither is. Harvald and Hauge[11] find concordance for the disease somewhat (though not significantly) greater in identical than in non-identical twins. Two other case reports on coronary disease in identical twins have appeared in the last year.[12, 13] Despite their interest and the review they provide of the literature, they are too small to contribute much to the total picture.

These findings support the view that coronary disease is familial and suggest that there may be genetic factors at work. Nevertheless, no cogent proof of this is yet at hand. What is at issue is not whether aggregation occurs in families, but how it is to be interpreted. Familial aggregation may quite well be explained by dietary, smoking or other habits common to members of a family, communicated rather than inherited. Clearly an impasse has been reached, and it is unlikely that it will be resolved by the accumulation of further data unless it is accompanied by the development of means for extracting more pertinent information from them. In this paper little new evidence will be presented, but attention will be directed to defining the problems associated with investigation of the genetics of coronary disease. When the difficulties have been identified, an attempt will be made to deal with them, though many of the ideas discussed are as yet incompletely developed and cannot be put to the test because the right data are not at hand. Much of this has been recognized more or less explicitly but not always by persons most intimately concerned with data; and there is no organized account of the subject.

It will be apparent that these difficulties are for the most part not peculiar to coronary disease and recur repeatedly in a wide variety of chronic diseases, such as diabetes, gout, hypertension and schizophrenia.

We shall explore the genetics of coronary disease as if it were determined by the presence or absence of a single abnormal ("mutant") gene. The part of the chromosome occupied by the gene is called a "locus", so we shall refer to the conjectural pattern of inheritance as "unilocal". The reader may wonder whether in the light of experience this is wise. But there are two points to be made in justification of this point of departure which illustrate the difference (so often overlooked) between reason and motive. The reason is that I think we have dismissed too lightly and on too little evidence the notion that

From the Division of Biostatistics in the Department of Preventive Medicine and the Department of Medicine, University of Colorado Medical School, 4200 East Ninth Avenue, Denver, Colorado 80220.

Presented at the Sixth National Symposium of the Heart Association of Southeastern Pennsylvania, on the Metabolic Basis of Human Atherosclerosis, Philadelphia, February 23, 1967.

*Associate Professor, Medicine and Biostatistics, Department of Preventive Medicine and Comprehensive Health Care, School of Medicine, University of Colorado.

Reprint requests to: Dr. E. A. Murphy, Division of Biostatistics, Department of Preventive Medicine, University of Colorado Medical School, 4200 East Ninth Avenue, Denver, Colorado 80220, U.S.A.

Canad. Med. Ass. J.
Nov. 11, 1967, vol. 97

TABLE I.—SUMMARY OF THE RECENT LITERATURE ON THE FAMILIAL ASPECTS OF CORONARY ARTERY DISEASE

Authors	*Year*	*Centre*	*Probands*	*Controls*	*Relatives*	*Probands*	*Controls*	*Risk ratio*
Gertler and White[6]	1954	Boston, Mass.	Young males with coronary disease	"Unmatched"	Fathers	25/97 (25.8%	17/145(11.7%)	0.45
					Mothers	4/97 (4.1%)	5/145(3.4%)	0.84
					Brothers	7/201(3.5%)	0/244(0.0%)	0.00
					Sisters	3/174(1.7%)	0/236(0.0%)	0.00
Shanoff *et al.*[7]	1961	Toronto, Ont.	Male veterans	Male veterans matched for age	Fathers	39/102(38.2%)	24/100(24.0%)	0.63
					Mothers	27/102(26.5%)	16/100(16.0%)	0.60
					Brothers	25/205(12.2%)	2/164(1.2%)	0.10
					Sisters	6/199(3.0%)	1/147(0.7%)	0.23
Rose[8]	1964	Baltimore, Md.	White hospital patients	Hospital patients matched for age and sex		(a) Male Probands		
					Fathers	16/65 (24.6%)	12/65 (18.5%)	0.75
					Mothers	10/65 (15.4%)	4/65 (6.2%)	0.40
					Sibs	12/319(3.8%)	4/367(1.1%)	0.29
						(b) Female probands		
					Fathers	3/10 (30.0%)	2/10 (20.0%)	0.67
					Mothers	3/10 (30.0%)	1/10 (10.0%)	0.33
					Sibs	6/60 (10.0%)	2/45 (4.4%)	0.44
Slack and Evans[9]	1966	London, England	Hospital patients	Various		(a) Male probands)		
					Male	41/312(13.1%)	17/189(9.0%)	0.68
					Female	16/305(5.2%)	7/197(3.6%)	0.68
						(b) Female probands)		
					Male	43/255(16.9%)	13/214(6.1%)	0.36
					Female	21/235(8.9%)	8/245(3.3%)	0.37
Thomas and Cohen[10]	1955	Baltimore, Md.	Medical students	—	Grandparental phenotypes			
					GF GM	Sons	Daughters	
					+ +	11/63 (21.2%)	2/49 (4.1%)	1.00
					+ −	11/91 (12.1%)	5/120(4.2%)	0.59
					− +	4/93 (4.3%)	0/87 (0.0%)	0.17
					− −	17/414(4.1%)	10/380(3.4%)	0.26
Harvald and Hauge[11]	1963	Denmark	Twins			Concordance rates		
					Monozygotic	20/102	(19.6%)	
					Dizygotic	24/155	(15.5%)	

a common disorder which has a high mortality may be unilocal. This standpoint seems to rest on a view of the theory of population genetics which is probably grossly oversimplified. For example, the classical argument against the single locus hypothesis supposes that the selection against the putative gene would eventually lead to extinction of the mutant line. In the genetic sense, selection means that the affected organisms produce offspring who live to maturity, on a smaller scale than normal. But Winkelstein and Rekate[14] have found that despite a higher fetal wastage, the average size of completed family is, if anything, slightly larger in women with coronary disease than in controls. Furthermore, the argument supposes that selection pressure has remained constant for many generations so that equilibration has had the opportunity to occur, and this assumption has been called into question in this instance.[15] Again, the argument supposes that there is no selection against the "wild type" (i.e. normal) gene in the homozygous state, but this too might be disputed. It can be argued that the disorder is complex, so that on prior grounds we know that dozens of separate functions must be involved and hence numerous loci. This is true about almost any disorder we care to think of: hemostasis, the characteristic by which we become aware of the existence of hemophilia, is very complex, and yet the disorder was recognized as a clearly Mendelian characteristic long before the refinements of modern hematology were developed. That is the reason. The motive for adhering to the single gene hypothesis is that, in man at least, it is difficult, perhaps impossible, to devise cogent tests of heritability of a character, even a measurable character, once we abandon simple Mendelian patterns. We do know of a number of rare disorders under the control of a single locus which cause coronary artery disease (e.g. pseudoxanthoma elasticum, homocystinuria and perhaps some lipid disorders) and the recent work of Vallance-Owen[16] on the relationship of insulin antagonists to myocardial infarction raises new hopes that even in the common kind of coronary disease it may be possible to identify a single defect. Such a view implies that the so-called risk factors influence the rate at which the disease progresses but are not in themselves sufficient to produce it.

What is it that makes the inheritance of this disorder hard to study? We can identify three major problems: penetrance, misclassification and segregation analysis. These are all intimately interlocking, but I shall try to tease them apart and analyze them separately.

I. PENETRANCE

The first problem we shall examine is penetrance. Many orthodox geneticists believe that this term serves no useful purpose. I would not contend that there is any basic mechanism to which it could be said to correspond and which could be the subject of experimental investigation. However, this is also true of "dominance", which modern biochemical genetics has shown to be an artifact of our tools of study and not an entity in its own right; nevertheless, it would be foolhardy to deny that it has had a contribution to make in the study of human genetics. If we wished to study the genetics of aortic stenosis, we have elegant and exact methods, such as cardiac catheterization, which will allow those with the disorder to be distinguished from those without it. But this is too expensive, too elaborate and too dangerous for use other than in a selected group of subjects. And we must first select that group. Feeling for thrills and listening for diamond-shaped murmurs may be by comparison crude; but they can be applied on an epidemiological scale, which cardiac catheterization cannot. By means of the simpler methods, however, we may be able to identify perhaps 80% of the cases; and in this sense, we can say that this disorder (one form of which apparently depends on a mutant at one locus) is 80% penetrant by these criteria. It might be that what we should be looking at is one specific but as yet unidentified factor which need not necessarily reside in the coronary vessels. The factor might be the coronary angiogram or myocardial oxygen consumption, but these could hardly be studied on a large scale in healthy people. Thus we must in general await some complication of the disease before we can recognize it. How are we to adjust for incomplete penetrance? It seems reasonable in many, perhaps most, instances to regard the recognition of the disorder as a chance event and we might identify two kinds of situation:

1. Where the recognition of the disorder is accidental and thus diagnosis in the several members of a family are uncorrelated events. A well-established example of this kind of penetrance would be the manifestation of G6PD deficiency by exposure to certain drugs. This is reasonable where a challenge is more or less uniformly distributed and where there is no process whereby genetic factors at other loci prevent the expression of the putative gene (epistasis). With the greater homogeneity of living conditions which has resulted from urbanization, this representation of the state of affairs is in many cases becoming progressively more realistic. Thus in regard to coronary disease, the more we eat out, the less our coronaries will reflect the peculiarities of the domestic cooking pot. If stress, or infection, or smoking is important in atherogenesis, then the effects of the general environment will progressively swamp those of the home. Evidence on the extent of the familial component on several factors incriminated in atherogenesis has recently been presented by Deutscher, Epstein and Kjelsberg.[17] Denborough, Clarke and Paterson[18] believe that familial similarities in blood lipids are due to sharing a common environment.

2. The second situation is where there are modifier genes at work, and there the manifestation or otherwise of the gene by two members of a family will be correlated. This process of epistatic interaction can readily be demonstrated in bacterial or drosophila genetics, and though it is hard to illustrate in man, there is little doubt that it is a realistic model. Only if one gene is of major importance is there any point in calling the system "unilocal". If a gene is so unimportant that its path cannot be traced with moderate confidence, then the system is best considered "multilocal". But the distinction is clearly not a sharp one.

In some disorders, however, penetrance, i.e. manifestation of the abnormal gene, may not be accidental but may, like Huntington's chorea, be relentlessly age-dependent. It is said that if they live long enough, all white horses will die of melanoma. This does not appear to be a very useful statement, since the only alternative is to suppose that the horses might be immortal. But conceptually at least we could suppose that it is merely a matter of time before any person with the right genetic constitution would develop manifest occlusive coronary disease.

The question at once arises as to why the onset of the disease should be delayed and why it should not occur at a uniform rate. Two mechanisms, both of which perhaps operate, suggest themselves. First, the speed of development of the disease varies and this might account for the importance of the "risk factors". For example, it might be that the rate of diffusion of lipid, or the accretion of deposit, or the activity of scavenger mechanisms shows variation from patient to patient. Alternatively, the crucial factor may be the size of the coronary

Canad. Med. Ass. J.
Nov. 11, 1967, vol. 97

vessels. A recent study by Wilens, Plair and Henderson[19] reports that the total external area of the epicardial vessels varies by a factor of three, from 12 to 36 sq. cm. There was no clear age trend. These are autopsy data and must be taken with reservation. Nonetheless, there seems little doubt that this is a considerable source of variation. If we assumed a constant environment, these factors would depend on regulation by genes at other loci—in effect we are again arguing in terms of partial and complete epistasis. One consequence of this mechanism and one which could easily be tested would be that not only should there be familial aggregation of coronary disease, but also that the age of onset should show a correlation within families. To my knowledge, no one has so far paid any attention to this point, except that Douglas[13] comments on the similarities in ages of onset of the disease in identical twins. The second mechanism to explain the delayed and variable age of onset is to suppose that atherogenesis is not a continuous or quasi-continuous process but proceeds by steps, each episode of deterioration being precipitated by some insult in the environment such as a thrombus. These episodes may be important because they make the underlying disease manifest; but they may be important because thrombi may subsequently become organized and actually produce atheromatous lesions.[20-23] If the insults were of much the same size and occurred independently and with constant risk, the waiting time until a threshold number of hits was received and the disease became manifest would follow a pattern known as the gamma distribution (Fig. 1). The diagram shows cumulative risk functions, and the incidence of new cases at any age is represented by the slope of the curve at that point. A gamma of order 1 or the exponential curve, that is, a model which supposed that a single insult would lead to overt coronary disease, would mean that the greatest incidence of new cases occurs at birth, which is manifestly not the case. The cumulative curve based on the supposition that half the cases had occurred before the age of 50 is shown in the diagram (Fig. 1). If the number of hits required were two, the median age being the same, we get the second curve, the steepest segment of which is at 29.6 years of age. The six-hit curve, also shown, would predict a peak incidence at 47.2 years. This is a little more realistic, but we can do better. With the 25-hit curve the age of onset has approximately a normal distribution and the cumulative curve has the familiar sigmoid shape. The median—50—now virtually coincides with the mean and the age of maximum incidence; and

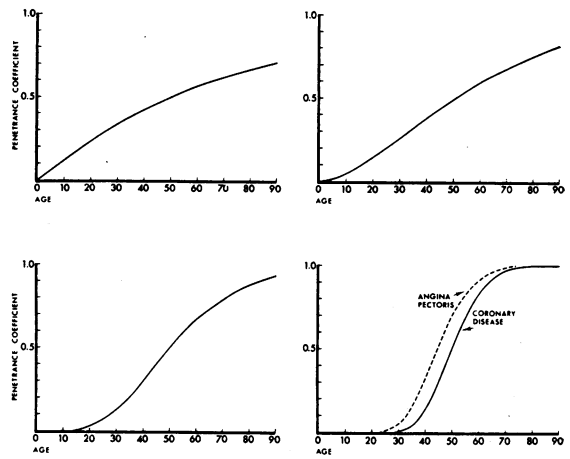

Fig. 1.—A model of coronary atherogenesis as a multiple hit phenomenon. Top left: the cumulative distribution of the age of onset if one "hit" is enough to cause clinically overt coronary disease; top right, two hits; bottom left, six hits; bottom right, 20 hits for angina pectoris, 25 for coronary occlusion. For details see the text.

the standard deviation is 10. About 2½% of those at risk would develop their disease before 30 and 2½% would develop it after 70, which corresponds roughly to reality. We might suppose that when 20 episodes had occurred, angina pectoris develops: the age of onset of angina is shown by the interrupted line. It would be pointless to push this argument further or make it more precise, because it is not immediately germane to our problem; but at least it provides one fairly satisfying notion (though certainly not the only one) of why the age of onset follows the pattern it does. In analyzing family patterns we would have to make due allowance for the effects of age and this might be done by fitting an appropriate curve through survivorship data. This is a major statistical problem so far unsolved. We may note that there is no good reason why we should not combine the two models and suppose that the development of occlusive disease depends on two kinds of chance factors—on the habitual risk of insult controlled by random genetic and environmental factors and on the accumulation of insults.

II. MISCLASSIFICATION

We now move on to the second problem, that of misclassification and the related problem of discrimination. Hitherto, we have been discussing the subject as if there were no difficulty in deciding who has coronary disease. It is plainly desirable to have uniform standards of diagnosis; there are real temptations to be avoided, and analysis of such conditions as the hyperlipemias has suffered at the hands of shifting diagnostic standards. These changes have been

necessary to reconcile discrepancies between the predictions of the model and the facts which develop when enough data have accumulated. This situation may in part result because the investigator has attempted by one means or another to fit the data to some preconceived model rather than to fit the model to the data; but even if this is not the case, it may be that he is trying to be too exact—that he is attempting by some test to attain a degree of resolution which is greater than that by which the test is validated. If, for example, the diagnostic criteria for the diagnosis of coronary disease by electrocardiogram are established by the study of cases classified by postmortem changes, the electrocardiogram can have no better power of discrimination than postmortem findings: indeed it is very much less sensitive. A derived test cannot at best agree with the defining criterion more than 100% of the time. If as a result of future developments the electrocardiogram were to become a superior discriminant, it would necessarily mean that, implicitly at least, another criterion of coronary disease was (perhaps correctly) being used and hence that the validation of the changes was no longer dependent on postmortem evidence. It could, of course, be argued that when we talk of the imperfections of, let us say, the ballistocardiogram as a detector of the coronary disease diathesis, we have inverted the problem; that the test is highly reliable and that it is the clinical manifestations which are defective.

Before we develop the question of optimizing discrimination, we might consider for a moment the useful kinds of evidence. These fall into two classes—measurable characters (e.g. the white cell count, the blood pressure and transaminase levels) and attributes (e.g. pain, gallop rhythms and pericardial rubs). For each of these it is possible, at least in principle, to find how often they occur, or what values they assume, in those with and those without coronary disease. Thus we can establish how reliable they are as diagnostic criteria, and the risks of error of classification. There is little use to be made of evidence not admitting of this minimum description. We would specifically exclude such evidence as transcendental instincts and incommunicable judgments based on experience which cannot be precisely defined, validated and put in the hopper of analytical reduction. Such judgments, perhaps indispensable in clinical practice, would likely prove on analysis to represent compounds of the simpler facts, and an optimal compounding of such characters should ideally be handled by mathematical means. If they were to be admitted in genetic analysis, there is some danger of the same fact being used twice in the analysis, once as a basic fact and once, implicitly, in the clinical judgment, and this may lead to too much weight being given to it.

The other kind of evidence which must be excluded is family history. It has been pointed out elsewhere[3] that if we admit this as evidence, bearing however little weight, for the diagnosis of coronary disease in the individual, then naturally when we come to analyze our data, they will show familial aggregation for coronary disease—that statistical analysis is merely rearranging our prior convictions.

Granted, then, that we have some measurement, how can we best use it to establish a diagnosis? It will help if we begin with the case where we have a single measurement from which we have to make a diagnosis and consider later how this is to be generalized to the case where there are multiple measurable criteria.

The single measurement problem is a familiar one to the chemist. On rare occasions, results of a test fall into groups which are clearly and unambiguously separable. This is the real state of affairs when the investigator claims to have discovered an attribute such as arachnodactyly or an aminoaciduria. In such instances we may be rarely in doubt, and yet it is desirable that we should not lose sight of the fact that the difference between the groups is but one of degree, not of kind, and there is considerable variation within groups. In the majority of cases, however, we are dealing with much less clear separation and this may lead to many difficulties. To preserve a sense of proportion we may note that the distribution of galactose-1-phosphate uridyl transferase in the homozygous and heterozygous states of galactosemia, though bimodal, does not segregate cleanly; it shows one curve evidently representing a mixture of two distributions corresponding to the two genetic groups (genotypes);[24] yet this is fairly well established as a single locus disorder. Thus a very high degree of discrimination is not necessary for a character to be suitable for genetic study from a Mendelian standpoint. However, there is an important aside here; while such a character may be studied with profit, much more than bimodality, even clear and incontrovertible bimodality, is required to establish that in fact it represents the operation of a single locus. The problem is discussed in some detail and with numerous illustrations elsewhere.[24] To show a single locus effect, it is necessary to produce genetic evidence, i.e. evidence of transmissibility. And even that is formally insufficient; education and wealth are transmissible,

Canad. Med. Ass. J.
Nov. 11, 1967, vol. 97

TABLE II—POLYDACTYLY IN GUINEA PIGS*

Mating	*Offspring phenotype* *Polydactylous*	*Wild type*	χ^2
PP × ++	0	76	
$F_1 \times F_1$	45	188	3.72
	(58.25)	(174.75)	
$F_1 \times PP$	160	129	3.11
	(144.5)	(144.5)	

Numbers in parentheses are those expected under the hypothesis that polydactyly is an autosomal recessive character.

*Data of Wright (1934)[25]

though not as such genetic; and some classical studies by Sewell Wright on the inheritance of polydactyly in guinea pigs showed that for two generations after the crossing of two inbred strains, this condition can mimic to a nicety the operation of a single locus (Table II); but more extensive studies led to rejection of this view.[25] There is an illuminating paper by Edwards[26] on the simulation of Mendelism. We have assumed, however, that coronary disease is a character under the control of a single locus and that imperfect segregation is due to "noise" from other loci and from environment.

The problem of drawing dividing lines between the two or three genotypes now arises. It is unrealistic to suppose the distribution of the character known. For the case where we cannot assume a distribution, what is a reasonable procedure? It seems to me that in classifying we aim to minimize the variance within groups and to maximize that between them. However, we shall discuss this problem later.

Difficulties are presented when there are several characteristics to be considered. Ultimately a single composite criterion is required and in the absence of any clear genetic theory, a weighted linear combination of the observations seems reasonable. The solution to the problem of finding the weights provided by the technique of classical discriminant analysis is, strictly speaking, predicated on the assumptions that the members of the two groups follow multivariate normal distributions and have identical variance matrices and, what is much more important, that they can be sorted out by some independent criterion. A recent paper provides evidence that, even for quite small samples, discriminant analysis works quite well where there is a moderate amount of misclassification.[27] However, in the present situation the problem is to find an optimal *definition* of coronary disease, and other means must be sought. One fact in our favour is that under certain plausible assumptions[28] an index composed of a weighted linear combination of observations will tend to be distributed normally, so that if the optimal

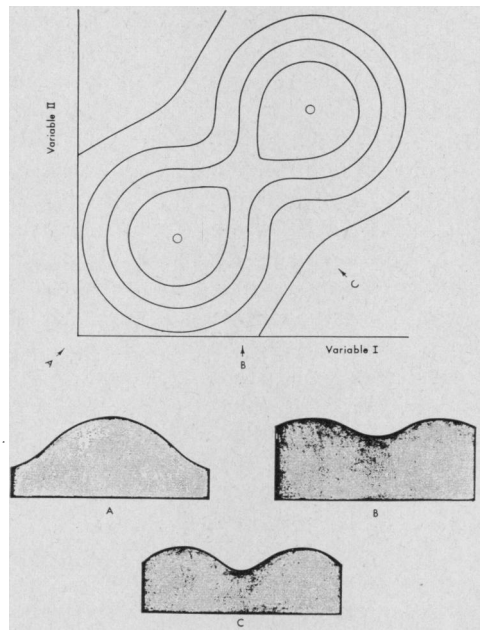

Fig. 2.—Maximizing the discriminating power of a linear combination of results from two tests. The upper diagram is a contour map of the probability density function of the possible pairs of values. The values tend to cluster around two poles (the small circles) corresponding to the coronary and control populations. The hatched diagrams below show the profile of the distribution seen from A, B and C. It will be evident that the trough is deepest and the separation of the two populations maximized when the distribution is seen from C.

combination can be found, in drawing a dividing line, we will no longer have to restrict ourselves to distribution-free methods.

So far, little successful work has been done to find how we may best combine the measurements. The problem may be thought of geometrically as rotating the co-ordinate axis in such a way as to maximize the amount of "daylight" between the groups, that is, to minimize the overlap. The idea can be conveniently illustrated in the two-dimensional case (Fig. 2). Here we have two measurements—for definiteness let us suppose the serum glutamic oxaloacetic transaminase (SGOT) level is on the horizontal axis and the amount of depression of the ST segment in lead V6 is on the vertical. The probability density, i.e. the height of the distribution surface corresponding to any pair of values (the co-ordinates of the point), stands up off the page; and points on the surface of equal height lie on the same contour. Thus we have two peaks representing the "rallying points" of the control and coronary populations. The apices of the two peaks are enclosed within the small circles. If we could look at a three-dimensional model of this arrangement from the positions indicated by arrows, A, B and C, we would get the silhouettes correspondingly marked below, each with a different depth of

trough and degree of overlap. When the trough is deepest (C), the discriminating power is maximized and we may then proceed as before to minimize the misclassification. How the problem is to be handled if we have three or more measurements to deal with is, it is to be supposed, not really very different, but nobody has so far been able to see how it is done even for the multivariate normal case. The problem has recently been discussed by Moran[29] in relation to psychiatric disorders but without any very encouraging conclusions. However, we are in a somewhat better situation here. In the first place the criteria used in the diagnosis of coronary disease are much better defined than those used in psychoses, and they are metrical (measurable) rather than categorical in character. In the second, provided we are prepared as a starting point to beg the question and assume that coronary disease is hereditary, we have some outside information which will help us to sort out our groups by means other than the traditional diagnostic criteria.

III. Segregation Analysis

We come finally to the questions which to many human geneticists are the most important of all, the analysis of segregation ratios and the testing of genetic hypotheses. What sort of segregation ratios are we to expect where there is incomplete and perhaps variable penetrance and where there is misclassification? The latter aspect we may deal with first because it is the easiest. We have a number of choices.

First we may in certain cases assume the form of the component distributions and by maximum likelihood or some such technique decompose the mixture into its component parts, then by suitable minimum chi-square procedures we can, at least for reasonably large samples, do a statistical test of the goodness of fit of the observed values to those expected under the genetic hypotheses. For this to work without family data, we need to be able to recognize all three phenotypes which will leave one degree of freedom over for the test of goodness of fit to a Hardy-Weinberg equilibrium. This method leads to difficulties if we do not find some means of ensuring that none of the variances of the underlying distributions is estimated as zero. The whole matter is discussed with illustrations elsewhere.[30] Recently Cohen[31] has proposed an alternative method of tackling the problem, suitable for large samples.

The second kind of procedure is, instead of attempting to separate out the constituent groups, to draw a dividing line and classify individuals as falling above or below it. The test then consists of comparing expected with observed numbers in each class. This requires only two classes of phenotypes provided there are family data. Sufficient degrees of freedom for a test are furnished by grouping the offspring according to the parental phenotypes. The dividing line can be drawn in various ways.

1. We may divide it so that misclassification is minimized. It can readily be shown that under circumstances which we may confidently assume in almost any real situation, this is done by taking a dividing line through the intersection point of the two curves, and Rao[32] has shown that this generalizes to multivariate distributions. Inasmuch as misclassification means noise, intuition suggests that the less misclassification, the more powerful the tests should be. The difficulty is to find the dividing point—there is no method that I know other than to decompose the curve into components and find the intersection point, and this presupposes that we know the form of the distribution functions.

2. We may use the method of Smith,[33] developed by Penrose[34] and by Kalmus and Sheila Maynard-Smith,[35] which consists of dividing the combined distribution curve in such a way as to make misclassification in the two directions equal. As a result, the genetic analysis is much simplified. However, the method also assumes that the distributions are Gaussian, that the character is known to be under the control of a single locus, and that Hardy-Weinberg equilibrium exists.

3. We may do what the biochemists do when they quantitate plasma proteins—take the lowest points (antimodes) between the peaks as the dividing points. This has the advantages that it is very simply done and that it is distribution-free, i.e. we do not need to know what the underlying distribution functions are. But it may be that the component curves "fuse" and that there is no antimode. This does not necessarily mean that there are no possibilities for discrimination at all.

4. We may take the common mean for the whole sample as the dividing point. This too is simple and distribution-free.

As an illustration of these methods, we present some data for normal distributions. Supposing that the control and coronary populations have the same variance and that the former population is 10 times as large as the latter, we get the misclassifications (expressed as a proportion) plotted against the distance between the means for the two populations (expressed in multiples of the standard deviations) shown in

Canad. Med. Ass. J.
Nov. 11, 1967, vol. 97

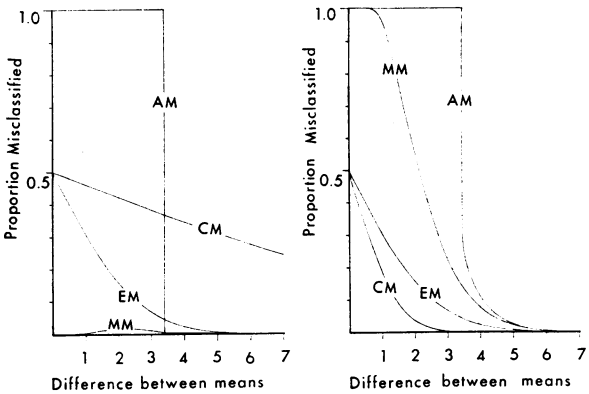

Fig. 3.—The expected proportions of subjects misclassified by the various methods of dividing a mixture of two Gaussian distributions with the same variance. We suppose that the normal population is 10 times as large as the coronary population. On the left are shown the proportions of the normal population misdiagnosed as having coronary disease, on the right the proportions of subjects with coronary disease mistakenly considered normal. The horizontal scale represents how far apart the means are, expressed as multiples of the standard deviation. The results are those obtained when the criteria of subdivision are respectively the antimode or lowest point between the peaks (AM), the mean for the pooled values (CM) the point which equalizes the percentage misclassification of the two groups (EM) and that which minimizes the total misclassification (MM).

Fig. 3. An antimode does not appear until the means are separated by about 3.36 standard deviations; thereafter the misclassification shown by the line marked "AM" is reasonably small especially for the normal group. Minimum misclassification (MM) is achieved very much at the expense of the smaller population. The common mean (CM) reverses this effect: the larger part of the misclassifications being of the normal group. Equal misclassification (EM) by definition treats both populations in the same way. In Fig. 4 are shown actual results when the means are separated by four standard deviations. With

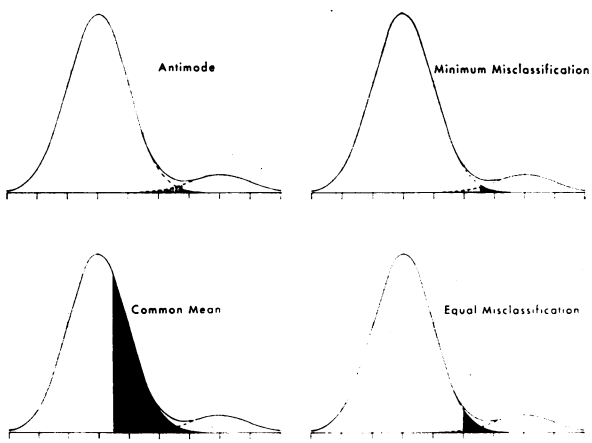

Fig. 4.—Misclassification where the means are separated by four standard deviations. In each case the proportion of coronary subjects misclassified is indicated by the hatched area, the proportion of normal subjects misclassified in black. The interrupted lines refer to the distributions of the component populations, the continuous lines to their combined distribution which is what would be observed in practice.

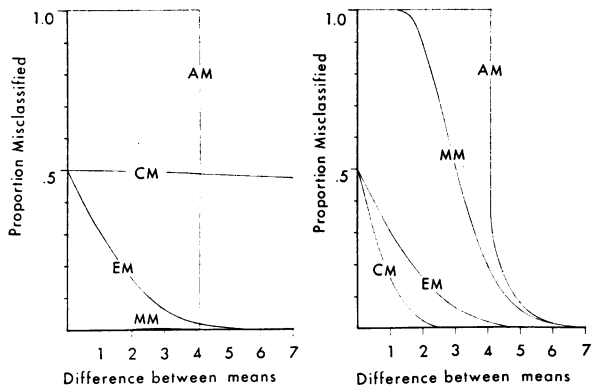

Fig. 5.—This is similar to Fig. 3 except that the normal population outnumbers the coronary disease population by 100 to 1.

a ratio of healthy to diseased subjects of 100:1 these effects are magnified (Fig. 5); conversely, if the two populations are of equal size, all four methods give the same dividing point.

The four methods will be applied to data on serum cholesterol levels in Johns Hopkins medical students.[36] Decomposition of the distribution of 1018 readings into two Gaussian components leads to the following results. The larger (normocholesterolemic) population comprises 88.54% of the whole with a mean of 218.2 mg. per 100 ml. and a standard deviation of 33.6. The "hypercholesterolemic" population comprises 11.46% with a mean of 281.1 mg. per 100 ml. and a standard deviation of 50.9. The percentage misclassifications in the two groups for each method are given in Table III. In addition the results are given of dividing the observations at the 99.9 percentile for the control group. This ensures that virtually all the misclassification will be in one direction only. In practice, if the forms of the underlying distributions were not known, this might be done in a number of ways, assuming that the distributions were not too pathological. Thus supposing that they are roughly symmetrical one could reconstruct the left-hand curve by supposing that its ascending

TABLE III.—SUBDIVISION OF BLOOD CHOLESTEROL LEVELS IN WHITE MALE MEDICAL STUDENTS*

Criterion of division	*Dividing point (mg./100 ml.)*	*Percentage misclassification Of left-hand population*	*Of right-hand population*
Minimum misclassification	293.1	1.29	59.32
Equal misclassification	243.2	22.84	22.84
Antimode	—	—	—
Common mean	225.4	41.52	13.70
99.9 percentile of the left-hand population	322.1	0.10	78.90

*Data of Thomas, Murphy and Bolling (1964)[36]

limb is mirrored in the descending limb. The actual cut-off point is not very critical. We now treat the problem as one of incomplete penetrance of the right-hand population, and part of the problem of analysis consists of estimating the corresponding penetrance coefficient.

This last procedure provides a link with the other facet of our problem. Where, for example, the age-penetrance relationship is well established, segregation ratios under any given hypothesis can readily be predicted and tested. The relationship is rarely known, however, and we have first to estimate the penetrance function from the data and then do a test of goodness to fit to some genetic hypothesis on the same data. This is a complicated and messy problem for which only partial solutions have been worked out. Batschelet[37] has explored this area for autosomal recessive characters with age-dependent penetrance, but ignoring the problem of ascertainment bias. Murphy worked the problem out for autosomal recessives for both single and complete ascertainment but supposing a fixed penetrance coefficient which is not age-dependent.[38] The dominant and the X-linked cases have so far been totally ignored.

The phenomenon of ascertainment bias is well known. In many situations the families are discovered through affected members of a sibship. If, as may happen by chance, no member of the sibship is affected, then such a family will be left out of the collection of families ascertained, which leads to an inflation in the proportion of affected children in those families which are ascertained. There are various ways of dealing with this according to the precise assumptions made; the problem has been fairly well explored in the fully penetrant case (see for example reference 39) but because of uncertainty about parental genotypes, the matter is considerably more complicated in the incompletely penetrant cases and there is much work to be done on this subject. Viewing this in the context of coronary disease, we might for argument's sake (and without in any sense committing ourselves to the view that it is true) consider the angiocardiographic findings as the best clinical criterion available. A recent massive study[40] has shown that in patients with typical clinical findings there is an excellent correspondence between their severity and the extent of the angiocardiographic changes. However, since subjects are selected for angiocardiography because they have symptoms, we are in no position to say how often significant disease is missed because it produces no symptoms.

There is one further problem of special interest to be brought forward for consideration. It

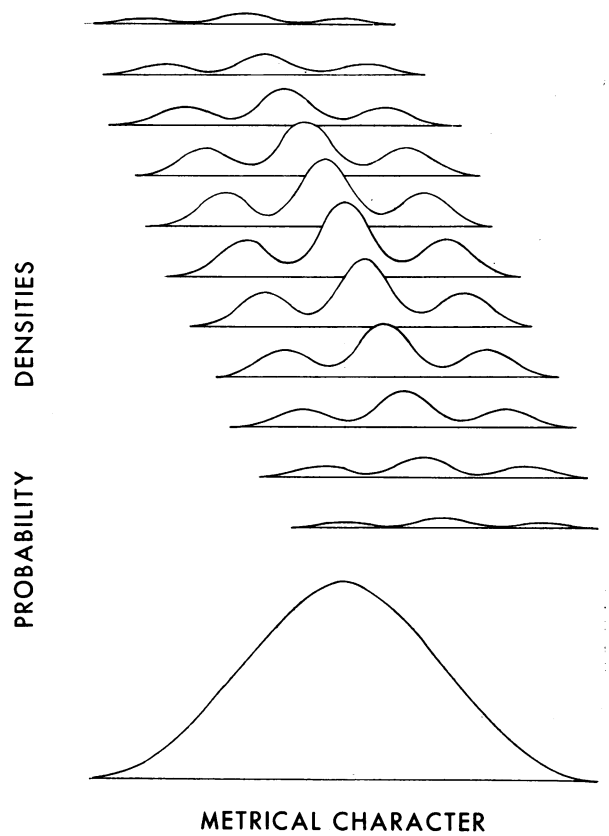

Fig. 6.—A theoretical model of the distribution of the hypercholesterolemic gene. For any given "background level" there is a fairly clean separation into three phenotypes. When all the values are put together regardless of background, the distribution at the bottom, in which all but one mode have been obliterated, results. (It is arranged that the individual phenotypes are all normally distributed; also the different "background values" have approximately a normal distribution.)

arose in connection with Khachadurian's data on hypercholesterolemia.[41] Suppose that there is a large variation in cholesterol levels between one family and another, and that there is nonetheless a disorder, hypercholesterolemia, which segregates within families in a Mendelian fashion, giving two or three different phenotypes. Lumping the readings together may be of little help because the points about which the different phenotypes cluster may be so "out of phase" in the several families that the net result is a hopelessly confusing and perhaps unimodal curve (Fig. 6). Testing conformity of the observed proportions to classical segregation ratios within a family would provide a series of very weak tests which, taken severally, provide little useful information since the null hypothesis would be almost impossible to reject even for unusually large sibships. However, there are techniques for combining non-significant p values to give an overall test of goodness of fit. One of the oldest and perhaps the simplest is due to Fisher.[42]

Canad. Med. Ass. J.
Nov. 11, 1967, vol. 97

TABLE IV.—THE PROBABILITY OF HETEROGENEITY AMONG SIBS

Parental genotypes	*Probability of mating*	*The conditional probability of the sibship*		
		One kind of sib	*Two kinds of sib*	*Three kinds of sib*
HH × HH	q^4	1	0	0
HH × H+	$4q^3p$	$(\frac{1}{2})^{S-1}$	$\binom{S}{R}(\frac{1}{2})^{S-1}$	0
H+ × H+	$4q^2p^2$	$(\frac{1}{2})^{S}+(\frac{1}{2})^{2S-1}$	$\binom{S}{R}\left[(\frac{1}{2})^{S+R}+(\frac{1}{2})^{2S-R}+(\frac{1}{2})^{2S}\right]$	$\binom{S}{R:T}(\frac{1}{2})^{S+R+T}$
H+ × ++	$4qp^3$	$(\frac{1}{2})^{S-1}$	$\binom{S}{R}(\frac{1}{2})^{S-1}$	0
++ × ++	p^4	1	0	0
HH × ++	$2q^2p^2$	1	0	0

S = Size of family. q = Frequency of the mutant gene. p = (1 − q).

R = size of left-hand group. $\binom{S}{R}=\frac{S!}{R!\,(S-R)!}$

T = Size of right-hand group. $\binom{S}{R:T}=\frac{S!}{R!T!\,(S-R-T)!}$

But there is one difficulty right at the start, and that is the uncertainty about the genotypes. If in a sibship of, for example, 12, the subjects segregate into three cleanly separated groups, there is no difficulty. But if there are only two groups out of a possible three, how are we to decide which two they are? We may note first that there is reason to believe that for cholesterol levels both in man[36] and animals[43] the standard deviation is proportional to the mean, and it is not implausible to suppose that the effects of a hypercholesterolemic gene would be proportional also. Thus we might either deal with the ratios of the readings or use their logarithms. One might reasonably suppose that if the two groups represented in the family were the two homozygous states, the gap between their means should be double that expected if one of them were the heterozygous state. But clearly if there is a single gap, we cannot on internal evidence decide which is the heterozygous state. In this situation, there seems to be no choice but to suppose each of the two possible interpretations in turn and weight by the prior probabilities of the corresponding matings occurring.

In Table IV are listed the probabilities associated with the possible outcomes. These probabilities are of two kinds:

1. The prior probabilities. In each case the probability that the mating of two particular genotypes, assuming random mating, is given. The prior probabilities are all functions of q, the frequency of the hypercholesterolemic gene.

2. Conditional probabilities. Conditional on a particular mating, the probability of a particular family occurring may be written down. These probabilities are not functions of q.

The total likelihood for the observed family is the sum of the products of the prior and conditional probabilities. For example, suppose that a couple has five children whose cholesterol levels group themselves naturally into two classes: two of them higher and three of them lower. This may be interpreted as two homozygous mutant, and three heterozygous, phenotypes, or as two heterozygous, and three homozygous wild-type, phenotypes. Since there are two kinds of children, we take the conditional probabilities from the fourth column. We get the following results:

Mating	*Prior probability*	*Conditional probability*
HH × H+	$4q^3p$	$\binom{5}{3}(\frac{1}{2})^4$
H+ × H+	$4q^2p^2$	$\binom{5}{3}[(\frac{1}{2})^8+(\frac{1}{2})^7+(\frac{1}{2})^{10}]$
H+ × ++	$4qp^3$	$\binom{5}{3}(\frac{1}{2})^4$

Multiplying each prior probability by the corresponding conditional probability and adding them we get as the total likelihood of the sibship:

$$\frac{5}{2}(q^3p+qp^3)+\frac{65}{128}q^2p^2$$

Fig. 7 shows a plot of the readings in the sibships F, G, H, I and J in Khachadurian's paper. Sibs considered to be of the same genotype (with respect to the putative gene) are

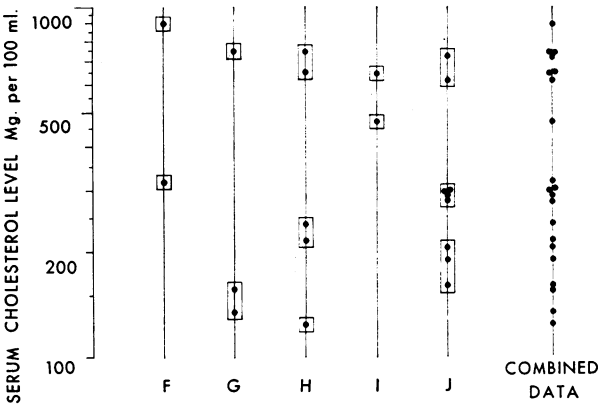

Fig. 7.—A plot on a logarithmic scale of serum cholesterol values in five sibships from Khachadurian's data,[41] families F, G, H, I and J. Sibs judged to be of the same genotype are enclosed within a common box. At the right are shown the pooled values in which the grouping observed within families is considerably blurred.

"boxed in". The vertical scale is logarithmic. The likelihood function of all five sibships is shown for various frequency values of the mutant gene in Fig. 8. It will be noted that, as is always the case, it is symmetrical about $q = \frac{1}{2}$. Thus in every case, either the maximum likelihood of the gene frequency will always be given by $q = \frac{1}{2}$, or there will be two maxima, symmetrically placed about this point. In the nature of the problem, we cannot estimate q from outside sources. Our maximum likelihood estimate is estimating not q for the total population but the gene frequency in the subpopulation which we are sampling, and this is not clearly defined

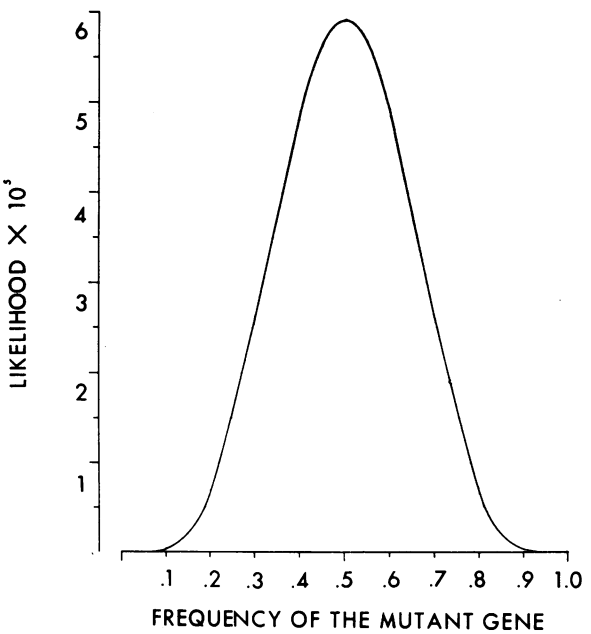

Fig. 8.—The combined likelihood of the five sibships (see Fig. 7) as a function of the frequency of the hypercholesterolemic gene.

since the ascertainment procedure is probably very complicated. However, we are not concerned mainly with the estimation of q, but with testing a hypothesis, and our maximum likelihood estimate may be regarded as suitable for our purposes. Where there are two estimates, it is perhaps reasonable to take the lower.

How the single focus hypothesis is to be tested is not at once clear. The method usually used in genetics is to find the probability of the observed outcome or any less likely, and reject the hypothesis if this is less than some preassigned value such as 0.05. In the present situation we can calculate this quantity for each family and combine them by Fisher's method. However, the calculation of the quantity in each family will be tedious because of the ambiguity of the outcomes. The symmetry of the likelihood function will be our main asset here.

An alternative approach is to compare the likelihood of the outcome with the likelihood under some alternative model. Except in the case where one model is a special case of the other, there is considerable difficulty in carrying through this procedure. Two papers by Cox[44, 45] throw light on the subject, but they do not treat the competing models symmetrically: one model is "in possession" and the test consists of finding whether the other may displace it. The whole problem will be taken up in greater depth elsewhere.

CONCLUSION

The main objective of this paper has been to define certain areas worthy of fuller exploration and in some of these an attempt has been made to sketch what may be useful approaches. It is too soon to abandon hope that at least some cases of coronary disease may be traced to a mutant gene at one locus. Should this prove not to be the case, our time will not have been wasted, since these problems occur repeatedly in the study of chronic disease.

The ideal, set by modern biochemical genetics, is the isolation of a protein polymorphism or a specific enzyme as a suitable subject for genetic analysis. But success in attaining this end, though considerable, has not always been in the areas of greatest clinical importance: coronary disease is vastly more common than oroticaciduria. There seems to be place for more refined analysis of less refined data. It is noteworthy that the three most firmly established examples of genetic linkage in man involve hemophilia with colour blindness, the nail-patella syndrome with the ABO blood group, and elliptocytosis with the Rh blood group sys-

tem. In none of these characters has an enzyme been isolated or protein structure analyzed.

Analyses of family history in coronary disease so far published have virtually ignored the defects and fallacies listed above, and have still yielded results suggesting a considerable hereditary component. There is reason to believe that a more detailed analysis will reduce the question to a form in which cogent genetic analysis is rendered possible.

Summary To demonstrate that coronary artery disease in man is hereditary, it will be necessary to show that its occurrence conforms to some Mendelian pattern. Three major problems require to be solved: age-dependence in the clinical manifestations of the disease and the interaction of ascertainment bias with it; misclassification and the means by which it is to be minimized, especially where diagnosis depends on measurements on several different variables; and testing of genetic hypotheses when due allowance has been made for the foregoing effects. These problems are explored in some detail with a review of pertinent literature. In particular, the problem of testing genetic hypotheses about quantitative data where intrafamilial variance is large enough to obscure in pooled data the natural groupings within families is expounded at some length.

Résumé Le rôle des facteurs génétiques dans la pathogénèse de la maladie coronarienne chez l'homme ne peut être prouvé que si son apparition se conforme à un des modes mendéliens classiques de transmission héréditaire. Pour ce faire, trois difficultés majeures doivent être d'abord résolues: la relation entre l'âge et les manifestations cliniques de la pathologie et l'interaction avec cet élément des prédispositions corroborées; les erreurs de classification et les moyens d'y remédier dans les cas où le diagnostic dépend de la mesure de plusieurs variables différentes; l'évaluation du bienfondé des hypothèses génétiques compte-tenu des effets qu'on vient de citer. L'auteur discute ces trois difficultés majeures à la lumière des travaux qui ont été inspirés par ces problèmes. Il insiste en particulier sur le problème consistant à traduire les hypothèses génétiques dans les statistiques quantitatives dans les cas où les variations intrafamiliales sont assez considérables pour fausser, dans les statistiques générales, les groupements naturels au sein des familles.

REFERENCES

1. McKUSICK, V. A.: *Mod. Conc. Cardiov. Dis.*, **28**: 535, 1959.
2. SCHWEITZER, M. D. *et al.*: *J. Chronic Dis.*, **15**: 1093, 1962.
3. MURPHY, E. A.: Genetics and atherosclerosis. *In*: Coronary heart disease, edited by W. Likoff and J. H. Moyer, Grune & Stratton Inc., New York, 1963, p. 89.
4. EPSTEIN, F. H.: *Amer. Heart J.*, **67**: 445, 1964.
5. McKUSICK, V. A.: Coronary artery disease. *In*: Genetics and the epidemiology of chronic disease, edited by J. V. Neel, M. W. Shaw and W. J. Schull, U.S. Department of Health, Education and Welfare, Division of Chronic Diseases, Public Health Service Publication No. 1163, Superintendent of Documents, U.S. Government Printing Office, Washington, 1965, p. 133.
6. GERTLER, M. M. AND WHITE, P. D.: Coronary heart disease in young adults: a multidisciplinary study, Harvard University Press, Cambridge, 1954.
7. SHANOFF, H. M. *et al.*: *Canad. Med. Ass. J.*, **84**: 519, 1961.
8. ROSE, G.: *Brit. J. Prev. Soc. Med.*, **18**: 75, 1964.
9. SLACK, J. AND EVANS, K. A.: *J. Med. Genet.*, **3**: 239, 1966.
10. THOMAS, C. B. AND COHEN, B. H.: *Ann. Intern. Med.*, **42**: 90, 1955.
11. HARVALD, B. AND HAUGE, M.: Hereditary factors elucidated by twin studies. *In*: Genetics and the epidemiology of chronic disease, edited by J. V. Neel, M. W. Shaw and W. J. Schull, U.S. Department of Health, Education, and Welfare, Division of Chronic Diseases, Public Health Service Publication No. 1163, Superintendent of Documents, U.S. Government Printing Office, Washington, 1965, p. 61.
12. SIDD, J. J., SASAHARA, A. A. AND LITTMANN, D.: *New Eng. J. Med.*, **274**: 55, 1966.
13. DOUGLAS, A. H.: *Dis. Chest*, **49**: 522, 1966.
14. WINKELSTEIN, W., JR. AND REKATE, A. C.: *Amer. Heart J.*, **67**: 481, 1964.
15. McKUSICK, V. A. AND MURPHY, E. A.: Genetic factors in the etiology of myocardial infarction. *In*: The etiology of myocardial infarction, edited by T. N. James and J. W. Keyes, Little, Brown & Co. Inc., Boston, 1963, p. 13.
16. VALLANCE-OWEN, J.: *Diabetes*, **13**: 241, 1964.
17. DEUTSCHER, S., EPSTEIN, F. H. AND KJELSBERG, M. O.: *Circulation*, **33**: 911, 1966.
18. DENBOROUGH, M. A., CLARKE, P. AND PATERSON, B.: *Aust. Ann. Med.*, **13**: 328, 1964.
19. WILENS, S. L., PLAIR, C. M. AND HENDERSON, D.: *J. A. M. A.*, **198**: 1325, 1966.
20. DUGUID, J. B.: *J. Path. Bact.*, **60**: 57, 1948.
21. MOVAT, H. Z., HAUST, M. D. AND MORE, R. H.: *Amer. J. Path.*, **35**: 93, 1959.
22. MORGAN, A. D.: The pathogenesis of coronary occlusion, Blackwell Scientific Publications, Oxford, 1956.
23. MUSTARD, J. F. *et al.*: *J. Atheroscler. Res.*, **4**: 1, 1964.
24. MURPHY, E. A.: *J. Chronic Dis.*, **17**: 301, 1964.
25. WRIGHT, S.: *Genetics*, **19**: 537, 1934.
26. EDWARDS, J. H.: *Acta Genet. (Basel)*, **10**: 63, 1960.
27. LACHENBRUCH, P. A.: *Technometrics*, **8**: 657, 1966.
28. WILKS, S. S.: Mathematical statistics, 2nd ed., John Wiley & Sons Inc., New York City, 1962, p. 257.
29. MORAN, P. A. P.: *Brit. J. Psychiat.*, **112**: 1165, 1966.
30. MURPHY, E. A. AND BOLLING, D. R.: *Amer. J. Hum. Genet.*, **19**: 322, 1967.
31. COHEN, A. C.: *Technometrics*, **9**: 15, 1967.
32. RAO, C. R.: Advanced statistical methods in biometric research, John Wiley & Sons Inc., New York City, 1952.
33. SMITH, C. A. B.: *Ann. Eugen. (Lond.)*, **13**: 272, 1947.
34. PENROSE, L. S.: *Ibid.*, **16**: 134, 1951.
35. KALMUS, H. AND SMITH, S. M.: *Ann. Hum. Genet.*, **29**: 127, 1965.
36. THOMAS, C. B., MURPHY, E. A. AND BOLLING, D. R.: *Bull. Hopkins Hosp.*, **114**: 290, 1964.
37. BATSCHELET, E.: *Biometrika*, **50**: 265, 1963.
38. MURPHY, E. A.: An exploration of some effects of incomplete penetrance on the ascertainment of recessive characteristics. Doctoral thesis, Johns Hopkins University, 1964.
39. CROW, J. F.: Problems of ascertainment in the analysis of family data. *In*: Genetics and the epidemiology of chronic disease, edited by J. V. Neel, M. W. Shaw and W. J. Schull, U.S. Department of Health, Education, and Welfare, Division of Chronic Diseases, Public Health Service Publication No. 1163, Superintendent of Documents, U.S. Government Printing Office, Washington, 1965, p. 23.
40. PROUDFIT, W. L., SHIREY, E. K. AND SONES, F. M., JR.: *Circulation*, **33**: 901, 1966.
41. KHACHADURIAN, A. K.: *Amer. J. Med.*, **37**: 402, 1964.
42. FISHER, R. A.: Statistical methods for research workers, 12th ed., revised, Oliver & Boyd Ltd., Edinburgh, 1954.
43. ROWSELL, H. C. *et al.*: Unpublished observations.
44. COX, D. R.: Tests of separate families of hypotheses. *In*: Proceedings of the 4th Berkeley symposium on mathematical statistics and probability, vol. 1, University of California Press, Berkeley, 1961, p. 105.
45. *Idem*: *Journal of the Royal Statistical Society*, Series B, 24: 406, 1962.