Haenszel

Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease [1]

NATHAN MANTEL *and* WILLIAM HAENSZEL, *Biometry Branch, National Cancer Institute,* [2] *Bethesda, Maryland*

Summary

The role and limitations of retrospective investigations of factors possibly associated with the occurrence of a disease are discussed and their relationship to forward-type studies emphasized. Examples of situations in which misleading associations could arise through the use of inappropriate control groups are presented. The possibility of misleading associations may be minimized by controlling or matching on factors which could produce such associations; the statistical analysis will then be modified. Statistical methodology is presented for analyzing retrospective study data, including chi-square measures of statistical significance of the observed association between the disease and the factor under study, and measures for interpreting the association in terms of an increased relative risk of disease. An extension of the chi-square test to the situation where data are subclassified by factors controlled in the analysis is given. A summary relative risk formula, R, is presented and discussed in connection with the problem of weighting the individual subcategory relative risks according to their importance or their precision. Alternative relative-risk formulas, R_1, R_2, R_3, and R_4, which require the calculation of subcategory-adjusted proportions of the study factor among diseased persons and controls for the computation of relative risks, are discussed. While these latter formulas may be useful in many instances, they may be biased or inconsistent and are not, in fact, averages of the relative risks observed in the separate subcategories. Only the relative-risk formula, R, of those presented, can be viewed as such an average. The relationship of the matched-sample method to the subclassification approach is indicated. The statistical methodology presented is illustrated with examples from a study of women with epidermoid and undifferentiated pulmonary carcinoma.—J. Nat. Cancer Inst. 22: 719-748, 1959.

Introduction

A retrospective study of disease occurrence may be defined as one in which the determination of association of a disease with some factor is based on an unusually high or low frequency of that factor among diseased persons. This contrasts with a forward study in which one looks instead

[1] Received for publication November 6, 1958.

[2] National Institutes of Health, Public Health Service, U.S. Department of Health, Education, and Welfare.

Reprinted by the

for an unusually high or low occurrence of the disease among individuals possessing the factor in question. Each approach has its advantages. Among the desirable attributes of the retrospective study is the ability to yield results from presently collectible data, whereas the forward study usually requires future observation of individuals over an extended period (this is not always true; if the status of individuals can be determined as of some past date, the data for a forward study may already be at hand). The retrospective approach is also adapted to the limited resources of an individual investigator and places a premium on the formulation of hypotheses for testing, rather than on facilities for data collection. For especially rare diseases a retrospective study may be the only feasible approach, since the forward study may prove too expensive to consider and the study size required to obtain a respectable number of cases completely unmanageable.

In the absence of important biases in the study setting, the retrospective method could be regarded, according to sound statistical theory, as the study method of choice. This follows from the much reduced sample sizes required by this approach and may be illustrated by the following extreme example. If a disease attack rate of 10 per 100,000 among 50 percent of the population free of some factor were increased tenfold among the other half of the population subject to the factor, a retrospective study of 100 cases and 100 controls would, with high probability, reveal this significantly increased risk. On the other hand, a forward study covering 2,000 persons, half with and half without the factor, would almost certainly fail to detect a significant difference. For comparable ability to find the type of increased risk just indicated, a forward study would need to cover about 500 times as many individuals as the corresponding retrospective study. The disparity in the required number of persons to be studied could, of course, be reduced by lengthening the follow-up period for forward studies to increase the experience in terms of person-years observed. The larger sample size required for the forward study reflects principally the infrequent occurrence of the disease entity under investigation. In the example illustrated, uncovering 100 cases of disease in a forward study would require either 100,000 individuals with the factor or 1,000,000 without. For diseases with a higher probability of occurrence the disparity in required size between retrospective and forward studies would be progressively reduced.

The retrospective study might be looked upon as a natural extension of the practice of physicians since the time of Hippocrates, to take case histories as an aid to diagnosis. Its guise has varied with respect to the means of measuring the prevalence of the suspect factor among diseased persons and the criteria for determining unusual departures from normal experience. When an association is so marked, as in Percival Pott's observations on the representation of chimney sweeps among cases of scrotal cancer, no further quantitative data are required to perceive its significance.

The retrospective approach has often been employed in studies of com-

municable diseases, one illustration being Snow's observations (*1*) on a common water supply for cholera cases in an area served by several sources (there would have been no element of unusualness had there been but one water supply). When a disease is epidemic in a circumscribed locality, the disease-free population in the same area offers a natural contrast. The method may be used successfully for endemic diseases as well. Holmes, in reaching his conclusions on the communicable nature of puerperal fever (*2*), noted particularly that a large number of women with puerperal fever had been attended by the same physicians. In this context it should be emphasized that communicable disease investigations have often combined retrospective and forward study methods. For example, Snow supplemented his retrospective observations on water supply by a contrast of cholera rates among subscribers of the Southwark and Vauxhall water company with the experience of persons served by the Lambeth water company within the same area.

When a disease occurs sporadically, or its occurrence is not confined to a well-defined group (such as women at childbirth), a choice of controls is not immediately evident. For cancer and other diseases characterized by high fatality rates, a study restricted to decedents might use persons dying from other causes as controls. Rigoni Stern adopted this technique in deducing the relationship of cancer of the breast and of the uterus to pregnancy history (*3*). Some contemporary studies have also used deaths from other causes as controls (*4*, *5*).

The present-day controlled retrospective studies of cancer date from the Lane-Claypon paper on breast cancer published in 1926 (*6*). This report is significant in setting forth procedures for selecting matched hospital controls and relating them to a consideration of study objectives. Retrospective techniques have since been applied in several investigations of cancer, including the following partial list of current references for a few primary sites: bladder (*7–10*), breast (*11–13*), cervix (*13–16*), larynx (*17*, *18*), leukemia (*19*), lung (*18*, *20–27*), and stomach (*13*, *28–30*).

Statisticians have been somewhat reluctant to discuss the analysis of data gathered by retrospective techniques, possibly because their training emphasizes the importance of defining a universe and specifying rules for counting events or drawing samples possessing certain properties. To them, proceeding from "effect to cause," with its consequent lack of specificity of a study population at risk, seems an unnatural approach. Certainly, the retrospective study raises some questions concerning the representative nature of the cases and controls in a given situation which cannot be completely satisfied by internal examination of any single set of data.

Only a few published papers have treated the statistical aspects of retrospective studies. Cornfield discussed the problem in terms of estimated measures of relative and absolute risks arising from contrasts of persons with and without specified characteristics (*31*). His paper was concerned with the simple situation of a homogeneous population of cases and controls, presumably alike in all characteristics except the one under

investigation, which could be represented by a single contingency table. In a later contribution he handled the problem of controlling for other variables by adjusting the distribution of controls to the observed distribution of cases (*16*). Dorn briefly mentions retrospective studies with emphasis on such topics as sources of data, choice of controls, and validity of inferences (*32*).

This paper presents a method for computing relative risks for retrospective study contrasts, which controls for the effects of other variables by use of the basic statistical principle of subclassification of data. The related problem of significance testing is also considered. Since details of statistical treatment are conditioned by study objectives, data collection methods, choice of a control series, and the use of matched or unmatched controls, these topics are also discussed briefly.

Objectives

Retrospective studies are relatively inexpensive and can play a valuable role as scouting forays to uncover leads on hitherto unknown effects, which can then be explored further by other techniques. The effects may be novel and not suggested by existing data, as in the pioneer work on the association of smoking and lung cancer or the association of blood type and gastric cancer, or they may represent refinements of current knowledge. The latter category might include collection of lifetime residence and/or work histories to elaborate differences in incidence and mortality which appear when some diseases are classified by last place of residence or last occupation of the newly diagnosed case or decedent.

With diseases of low incidence the controlled retrospective study may be the only feasible approach. Here emphasis should be placed on assembling results from several studies. Before accepting a finding and offering an interpretation, scientific caution calls for ascertaining whether it can be reproduced by others and in other administrative settings having their own peculiar biases.

A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done. Even when observations for a forward study have been collected, a supplementary retrospective approach to the same body of material may prove useful in collecting more data on points not covered in the original study design or in amplifying suggestive associations appearing in the initial forward-study results.

The findings of a retrospective study are necessarily in the form of statements about associations between diseases and factors, rather than about cause and effect relationships. This is due to the inability of the retrospective study to distinguish among the possible forms of association—cause and effect, association due to common causes, etc. Similar difficulties of interpretation arise in forward studies as well. A forward study, to avoid these difficulties, would need to be performed with the preciseness of a laboratory experiment. For example, such a study of associations with cigarette smoking would require that an investigator

randomly assign his subjects in advance to the various smoking categories, rather than simply note the categories to which they belong. The inherent practical difficulties of such an enterprise are evident.

In addition to the failings shared with the forward study, the retrospective study is further exposed to misleading associations arising from the circumstances under which test and control subjects are obtained. The retrospective study picks up factors associated with becoming a diseased or a disease-free *subject*, rather than simply factors associated with presence or absence of the disease. The difficulties in this regard may be most pronounced when the study group represents a cross section of patients alive at any time (prevalence), including some who have been ill for a long period. Inclusion of the latter may lead to identification of items associated with the course of the illness, unrelated to increased or decreased risk of developing the disease. The theoretical point has been raised that factors conducive to longer survival of patients may be found in "prevalence" samples and interpreted erroneously as being associated with excess liability to the disease (*33*). Loopholes of this type are minimized when investigations are restricted to samples of newly diagnosed patients (incidence).

A partial remedy for these uncertainties lies in employing a conservative approach to interpretation of the associations observed. Recognizing the ease with which associations may be influenced by extraneous factors, the investigator may require not only that the measure of relative risk be significantly different from unity but also that it be importantly different. He may, for instance, require that the data indicate an increased relative risk for a characteristic of at least 50 percent, on the assumption that an excess of this magnitude would not arise from extraneous factors alone. However, the use of such conservative procedures emphasizes a corresponding need to pinpoint the disease entity under study. A strong relationship between a factor and a disease entity might fail to be revealed, if the entity was included in a larger, less well-defined, disease category. After the event from data now at hand, we know that a study of the association of cigarette smoking with epidermoid and undifferentiated pulmonary carcinoma is more revealing than an inquiry covering all histologic types of lung cancer.

Multiple Comparison Problem

The present-day retrospective study is usually concerned with investigating a variety of associations with a disease, little effort being involved in acquiring, within limits, added information from respondents. The results may be analyzed in a number of ways: the various factors may be investigated separately, without regard to the other factors; they may be investigated in conjunction with each other, a particular conjunction being considered a factor in its own right; or, more commonly, a factor may be tested with control for the presence or absence of other factors. Thus, if the role of cigarette smoking and coffee drinking in a given disease are under study, the possible comparisons include the relative

risk of disease for individuals who both smoke and drink as opposed to all other persons, or as opposed to those who neither smoke, nor drink coffee. In addition, the relative risk associated with smoking might be obtained separately for drinkers and nondrinkers of coffee, with a weighted average of these two relative risks constituting still another item. Conversely, risks associated with coffee drinking, with adjustments for cigarette smoking, could be computed.

The potential comparisons arising from a comprehensive retrospective study can be large. Almost any reasonable level of statistical significance used to test a single contrast, when applied to a long series of contrasts, will, with a high degree of probability, result in some contrasts testing significant, even in the absence of any real associations. The usual prescription for coping with this multiple comparison problem—requiring individual comparisons to test significant at an extreme probability level to reduce the number of associations incorrectly asserted to be true—would result only in making real associations difficult to detect.

However, the multiple comparison problem exists only when inferences are to be drawn from a single set of data. If the purpose of the retrospective study is to uncover leads for fuller investigation, it becomes clear there is no real multiple significance testing problem—a single retrospective study does not yield conclusions, only leads. Also, the problem does not exist when several retrospective and other type studies are at hand, since the inferences will be based on a collation of evidence, the degree of agreement and reproducibility among studies, and their consistency with other types of available evidence, and not on the findings of a single study.

Nevertheless, it would be wise to employ testing procedures which do not lead to a superabundance of potential clues from any one study. This may be achieved by employing nominal significance levels in testing factors of primary interest incorporated into the design of an investigation and applying more stringent significance tests to comparisons of secondary interest or to comparisons suggested by the data. For the usual problem of multiple significance testing, this would be equivalent to allocating a large part of the desired risk of erroneous acceptance of an association as real to a small group of comparisons where fruitful results were anticipated, and parceling out the remainder of the available risk to the large bulk of comparisons of a more secondary nature. This minimizes the risk of diluting, through inclusion of many secondary comparisons, the chances for detecting an important primary effect.

Representative Nature of Data

The fundamental assumption underlying the analysis of retrospective data is that the assembled cases and controls are representative of the universe defined for investigation. This obligates the investigator not only to examine the data which are the end product but also to go behind the scenes and evaluate the forces which have channeled the material to his attention, including such items as local practices of referral to special-

ists and hospitals and the patient's condition and the effect of these items on the probability of diagnosis or hospital admission. We re-emphasize that this requires the exercise of judgment on the potential magnitude of biases and as to whether they could result in factors seeming to be related to a disease, in the absence of a real association of the factor with presence or absence of the disease. The danger of bias may be greatest in working with material from a single diagnostic source or institution.

Among the more important practical considerations affecting retrospective studies is that they are ordinarily designed to follow the line of least resistance in obtaining case and control histories. This means that cases and controls will often be hospital patients rather than persons in the general population outside hospitals. As a result, any factor which increases the probability that a diseased individual will be hospitalized for the disease may mistakenly be found to be associated with the disease. For example, Berkson (*34*) and White (*35*) have pointed out that positive association between two diseases, not present in the general population, may be produced when hospital admissions alone are studied, because persons with a combination of complaints are more likely to require hospital treatment. In theory, bias might also be produced in reverse manner, if the suspect factor diminished the probability of hospitalization for other diagnoses used as controls. The difficulties are not unique for hospital patients. Similar loopholes in interpretation may be advanced for any special groups used as sources of cases and controls.

However, a mere catalogue of biases arising from the possibly unrepresentative nature of a sample of cases and controls should not *ipso facto* invalidate any study findings. This is a substantive issue to be resolved on its merits for a specific investigation. Collateral evidence may provide information on the potential magnitude of bias and the size of spurious associations which could result. In some situations the difference between cases and controls may be so great that postulation of an unreasonably large bias would be required. Whether he consciously recognizes it or not, the investigator must always balance the risks confronting him and decide whether it is more important to detect an effect, when present, or to reject findings, when they may not reflect the true situation. If opportunities for further testing exist, one should not be too hasty in rejecting an association as an artifact arising from the method of data collection, and in foreclosing exploration of a potentially fruitful lead.

Because of the important role retrospective studies play in studies of human genetics, mention may be made of a bias frequently encountered in studies dealing with the familial distribution of diseases. A frequently used procedure takes a group of diagnosed cases for a disease in question and a group of controls and compares the prevalence of this disease among relatives of the probands and controls. The bias arises from the unrepresentative nature of the probands with respect to familial distribution and is known in other fields as "the problem of the index case" or "the effect of method of ascertainment." It has long been recognized that the

characteristics for a random sample of families will differ from those for families to whom the investigator's attention has been directed because the family rosters include individuals selected for study on the basis of a specified attribute. For example, data on family size (number of children) obtained from siblings, rather than parents, are biased, since two or three potential index cases are present in the population for two- and three-child families as opposed to one for one-child families and none for childless couples. The analogy for disease occurrence is apparent. Families with two or three cases of the disease under study may have double or triple the probability of being represented by individuals in source material and having a representative selected as a proband than families with only one case. An appropriate analysis for this situation in studies of family size and birth order has been discussed by Greenwood and Yule (*36*), which takes account of the probability of family representation in proband data. Haenszel (*37*) has applied their correction to gastric-cancer data reported by Videbaek and Mosbech (*38*) and found the correction to reduce the originally reported fourfold excess of gastric cancer among relatives of probands, as compared to relatives of controls, to one of about 60 percent.

One remedy for the weakness of the retrospective approach to problems involving association of diseases and familial distribution would be to place greater reliance on forward observations of defined cohorts for data on these topics.

Controls

While easier accessibility to and lesser expense of hospital controls are important considerations, they should not deter one from collecting control data for a sample representing a more general population, if the latter are demonstrably superior. Some of the uncertainties about the superiority of hospital or general population controls arise from the need to maintain comparability in responses. The dependence of retrospective studies on comparability of responses from cases and controls cannot be overemphasized. When more accurate answers can be obtained from controls in a medical-care environment, the gain in comparability of responses for these controls could outweigh the other advantages to be derived from the more representative nature of general population controls. The difficulties may be illustrated by the experience with smoking histories. Hospital controls invariably yield a higher proportion of smokers for each sex than controls of comparable age drawn from the general population (*27*). Does this mean more complete smoking histories are collected in hospitals or does it imply that smokers have higher hospital admission rates? If the first alternative is correct, hospital controls are the appropriate choice for measuring the association of smoking history with a given disease. The second alternative calls for general population controls and in this situation the use of hospital controls yields underestimates of the degree of association.

Dual hospital and general population controls would have some merit. If control data from the two sources were in agreement, this would rule

out some alternative interpretations of the findings. In the event of disagreement, its extent could be measured and alternate calculations made on the degree of association between an event and a suspect antecedent characteristic. Where the two sets of controls lead to substantially different results, a cautious and conservative interpretation is indicated.

Some topics, such as those bearing on sex practices and use of alcohol, may be amenable to study only within a clinical setting, and the collection of general population data on these items may prove impractical. The limitations of general population controls in this regard may have been overstressed, and empirical trials to test what information can be collected in household surveys should be encouraged instead of dismissing the possibility with no investigation whatsoever. Whelpton and Freedman, for example, have reported some success in collecting histories of contraceptive practices in interviews of a random sample of housewives (*39*).

When hospital controls are chosen, some precautions may be built into the study. Within limitations on the nature of controls imposed by a study hypothesis, controls drawn from a wide variety of diseases or admission diagnoses should be preferred. This permits examination of the distribution of the study characteristics among subgroups to check on internal consistency or variation among controls. This affords protection against two sources of error: *a*) attributing an association to the disease under investigation, when the effect is really linked to the diagnosis from which controls were drawn, and *b*) failure to detect an effect because both the study and control diseases are associated with the suspect factor. The latter is far from impossible. Both tuberculosis and bronchitis have exhibited association with smoking history and the use of one disease or the other as a control could easily lead to missing the association with smoking history. Similarly, patients with coronary artery disease would not constitute suitable controls for a study of the relationship of smoking and bladder cancer and *vice versa*, since the investigator would probably conclude that smoking was not related to either disease, when in truth it appears related to both. When there is definite evidence that two diseases are associated, for example, pernicious anemia and stomach cancer, the use of one as a control for the other is contraindicated, unless the study is specially designed to elucidate some aspects of the relationship.

It is always advantageous to include several items in a questionnaire for which general population data are available. This could be considered a partial substitute for dual hospital and general population controls. Disparity among cases, hospital controls, and general population controls on several general characteristics unrelated to the study hypothesis may be regarded as warning signals of the unrepresentative nature of the hospital cases and controls.

Where possible, interviews should be conducted without knowledge of the identity of cases and controls to guard against interviewer bias, although administrative reasons will often prevent attainment of "blind" interviews. In cooperative studies employing several interviewers, the

magnitude of interviewer bias may be diminished, since it is unlikely that all interviewers will share the same bias in concert. In special circumstances, such as those prevailing at Roswell Park Memorial Institute, admissions may be interviewed before diagnosis, and hence before the identity of cases and controls is established. This feature requires a comprehensive, general purpose interview routinely administered to all admissions, which may restrict its use to publicly supported institutions diagnosing and treating neoplastic diseases or other specialized disease entities. Several epidemiological contributions for specific cancer sites have been based on the unique control data available from Roswell Park Memorial Institute (*9*, *11*, *12*, *30*, *40–43*), which are particularly valuable for collation with studies depending on more conventional sources of controls to evaluate interviewer bias and related issues.

Some patients interviewed as diagnosed cases will subsequently have their diagnoses changed. This may be turned to advantage. If scrutiny of the data for the erroneously diagnosed group reveals they had histories resembling those for the control rather than the case series, as Doll and Hill found in their study of smoking and lung cancer (*21*), this would constitute evidence against interviewer bias.

In investigations of a cancer site the association of a factor may often be restricted to a specific histologic type or a well-defined portion of an organ. The finding that epidermoid and undifferentiated pulmonary carcinoma is more strongly related to smoking history than adenocarcinoma of the lung is now well established. The range of explanations for the observed deficit of epidermoid carcinoma of the cervix in Jewish women as compared to other white women is greatly circumscribed by the presence of about equal numbers of adenocarcinoma of the corpus in both groups. When these finer diagnostic details or their significance are unknown to the interviewer, another check on interviewer bias is provided. Furthermore, the confirmation in repeated studies of an association limited to a specific histologic type or a detailed site will lend credence to an etiological interpretation of the association. Repeated confirmation is an essential element. Otherwise, a very specific association may be a reflection of the multiple comparison problem; if enough contrasts are created by fractionation of a single set of data, some apparently significant result is likely to appear. For this reason it would be desirable to reproduce such provocative results as Wynder's finding that use of alcohol was more strongly associated with cancer of the extrinsic larynx than of the intrinsic larynx (*18*), and Billington's report that prepyloric and cardiac neoplasms of the stomach were associated with blood group A and those located in the fundus with blood group O (*44*).

Discussion of matched controls in relation to the analysis and the computation of relative risks is deferred to a later section. One consideration on matched controls arising in the planning and development of a study should be mentioned here. Obviously, if the risk of disease changes with age an apparent association of the disease with other age-related factors may result. Other apparent associations with race, sex,

nativity, etc., may arise in a similar manner. In devising rules for selecting controls, those factors known or strongly suspected to be related to disease occurrence should be taken into account if unbiased and more precise tests of the significance of the factors under investigation are desired. A sensible rule is to match those factors, such as age and sex, the effect of which may be conceded in advance and for which strong evidence is available from other sources, such as mortality data and morbidity surveys. When a factor is matched, however, it is eliminated as an independent study variable; it can be used only as a control on other factors. This suggests caution in the amount of matching attempted. If the effect of a factor is in doubt, the preferable strategy will be not to match but to control it in the statistical analysis. While the logical absurdity of attempting to measure an effect for a factor controlled by matching must be obvious, it is surprising how often investigators must be restrained from attempting this.

When a minimum of matching is involved, the importance of establishing, precisely and in advance, the method by which controls are selected for study increases. The rule should be rigid and unambiguous to avoid creating effects by subconscious selection and manipulation of controls. The problem is similar to that encountered in therapeutic trials where a protocol spelling out all the contingencies and actions to be taken in advance is, along with random assignment of cases and controls, the major bulwark against bias.

To reduce interview time and expense there are advantages in procedures for selecting controls which permit a case and the corresponding controls to be interviewed in a single session, particularly if travel to several institutions is involved. In practice, this favors selecting controls from a hospital patient census rather than from hospital admission lists. The difficulty with hospital admissions is that there is no guarantee that the controls will be available in the hospital at the time the diagnosed case is interviewed. This point seems more important than the fact that patients with diagnoses requiring long-term stays are overrepresented in a current hospital census (*45*). If the latter is an important issue, it may be handled in analysis through subclassification of controls by diagnosis.

Normally there will be little difficulty in reconciling these considerations into a harmonious set of rules. The items to be matched often lend themselves to a procedure for specifying controls. In a recent study on female lung cancer we found that the definition of two controls as the next older and the next younger women in the same hospital service, present on the day the case was interviewed, met the requirements just outlined (*27*). The controls were uniquely defined, the records establishing their identity were readily available on the service floor, interviews could be completed in one day, and a provision for balancing ages of cases and controls was incorporated. Simultaneous interviews of cases and controls may be more than an administrative convenience. If the prevalence of the associated factor is rapidly shifting over time,

failure to control time of interview could obscure or exaggerate an association.

Some Statistical Tools

To progress further, questions on the representative nature of the case and control series must have been resolved affirmatively. With this condition in mind, let us suppose that a controlled retrospective study has been conducted and that the number of diseased cases, N_1, consists of A individuals with the factor being investigated and B free of the factor, while the number of controls, N_2, consists of C individuals with, and D individuals without the factor. Let $M_1 = A + C$, $M_2 = B + D$, $T = N_1 + N_2 = M_1 + M_2 = A + B + C + D$. What statistical evidence is there for the presence of an association and what is an appropriate measure of the strength of the association?

A commonly employed statistical test of association is the chi-square test on the difference between the cases and controls in the proportion of individuals having the factor under test. A corrected chi square may be calculated routinely as

$$(|AD-BC|-\tfrac{1}{2}T)^2T/N_1M_1N_2M_2$$

and tested as a chi square with 1 degree of freedom in the usual manner.

A suggested measure of the strength of the association of the disease with the factor is the apparent risk of the disease for those with the factor, relative to the risk for those without the factor. Consider that a population falls into the four possible categories and in the proportions indicated by the following table:

	With factor	Free of factor	Total
With disease	P_1	P_2	$P_1 + P_2$
Free of disease	P_3	P_4	$P_3 + P_4$
Total	$P_1 + P_3$	$P_2 + P_4$	1

The proportion of persons with the factor having the disease is $P_1/(P_1 + P_3)$, while the corresponding proportion for those free of the factor is $P_2/(P_2 + P_4)$. Relatively then, the risk of the disease for those with the factor is $P_1(P_2 + P_4)/P_2(P_1 + P_3)$. On a sampling basis this quantity may be estimated either by drawing a sample of the general population and estimating P_1, P_2, P_3, and P_4 therefrom or estimating $P_1/(P_1 + P_3)$ and $P_2/(P_2 + P_4)$ separately from samples of persons with, and persons free of, the factor.

It may be noted, however, that if the relative risk as defined equals unity, then the quantity P_1P_4/P_2P_3 will also equal unity. Further, for diseases of low incidence where the values for P_1 and P_2 are small in comparison with P_3 and P_4 it follows, as has been pointed out by Cornfield (*31*), that P_1P_4/P_2P_3 is also a close approximation to the relative risk. This latter approximate relative risk can properly be estimated from the two sample approaches described or from samples drawn on a retrospective basis; that is, separate samples of persons with, and persons free of, the disease. The sample proportions of persons with, and free

of, the factor in the retrospective approach provide estimates of $P_1/(P_1 + P_2)$ and of $P_2/(P_1 + P_2)$ from the sample having the disease and of $P_3/(P_3 + P_4)$ and of $P_4/(P_3 + P_4)$ from the disease-free sample. The estimate of P_1P_4/P_2P_3 is obtained by appropriate multiplication and division of these four quantities.

Whichever of the three methods of sampling is employed, the estimate of the approximate relative risk, P_1P_4/P_2P_3, reduces simply to AD/BC, where A, B, C, and D are defined in the manner stated in the first paragraph of this section. Also, the chi-square test of association given, which is essentially a test of whether or not the relative risk is unity, is equally applicable to all three sampling methods.

In the foregoing the two basic statistical tools of the epidemiologist for retrospective studies, the chi-square significance test and the measure of a relative risk, have been described for a relatively simple situation, one in which to all intents there is a single homogeneous population. The more complex situations confronting the epidemiologist in actual practice and the corresponding modifications in the statistical procedures will be presented.

Two other statistical problems may be noted here. One is the determination of how large a retrospective study to conduct. This depends on how sure we wish to be that the study will yield clear evidence that the relative risk is not unity, when it in fact differs from unity to some important degree. Application of this statistical technique requires reinterpreting a relative risk greater than unity into the corresponding difference between the diseased and the disease-free groups in the proportion of persons with the factor. For example, suppose an attack rate of 20 percent, given a normal rate of 10 percent, is worth uncovering. Suppose further that the factor associated with the increased disease rate affects 20 percent of the population. The population would then be distributed as follows:

	With factor	Free of factor	Total
With disease	P_1=4%	P_2=8%	12%
Free of disease	P_3=16%	P_4=72%	88%
Total	20%	80%	100%

The required retrospective study should be large enough to differentiate between a 33.3 percent $[P_1/(P_1 + P_2)]$ relative frequency of the factor among diseased individuals and an 18.2 percent $[P_3/(P_3+P_4)]$ relative frequency among disease-free individuals. The usual procedures for determining required sample sizes to differentiate between two binomial proportions are applicable in this situation.

While rigorous extension of this procedure to the more complex situations to be considered is not too simple, it can readily be adapted to secure approximations of the necessary study size. One might, for example, start by estimating the over-all required sample size following the procedure just indicated for differentiating between two sample proportions, assuming that cases and controls are homogeneous with

respect to factors other than the one under investigation. Suppose on an over-all basis it is determined that the study should include $N_1 = 200$ disease cases and $N_2 = 200$ controls, but that the study data will be subclassified for purposes of analysis. Ignoring mathematical complications resulting from variations in binomial parameter values within individual subclassifications, we may interpret the above values of N_1 and N_2 as roughly meaning that the total information required for the study is $N_1N_2/(N_1 + N_2) = 100$. The objective should then be to assign values to N_{1i} and N_{2i} to obtain a total score of 100 for the cumulated information over all the subclassifications, $\Sigma N_{1i}N_{2i}/(N_{1i} + N_{2i})$, where N_{1i} and N_{2i} are the number of cases and controls in the *ith* subclassification.

This formulation of required total information brings out some aspects of retrospective study planning which are considered later in this paper. For instance, if any N_{1i} or N_{2i} is zero, no information is available from that particular category. Much of the benefit of a large N_{1i} (or N_{2i}) in any particular category is lost if the corresponding N_{2i} (or N_{1i}) is small. It is normally desirable to have N_{1i} and N_{2i} values commensurate with each other; for fixed totals, ΣN_{1i} and ΣN_{2i}, the total information in an investigation will be at a maximum if the degree of crossmatching is equal in all subclassifications with a constant case-control ratio of $\Sigma N_{1i}/\Sigma N_{2i}$. Maintaining a fixed case-control ratio among categories need not preclude assigning more cases and controls to specific categories. Larger numbers may be desired for categories of crucial interest to the study or for categories which represent greater segments of the population.

The information formula also reveals the limits for adjusting the relative numbers of diseased and control cases. It shows that if the number of controls (N_2) becomes indefinitely large, the required N_1 value can at most be reduced only by a factor of 2. Furthermore, this reduction in required diseased cases may be inappropriate if one wishes to obtain clear results for the separate subcategories.

The study size requirements suggested by the information formula may be seriously in error if the binomial parameters show excessive variation among subcategories. Ordinary precautions, however, should serve to keep the formula useful. In some situations it may be desirable to modify the information formula indicated above to reflect the contribution due to variation in the binomial parameters involved.

The second statistical procedure involves setting reasonable limits on the relative risk when it is in fact different from unity. For the homogeneous case considered, formulas for such limits have been published in (*46*). The chi-square test as stated is essentially a test of whether or not the confidence limits include unity. Extension of this procedure to more complex cases is fairly involved and depends primarily on the measure of relative risk adopted. In the absence of a clear justification for any single measure of over-all relative risk, the burden of extremely involved computation of confidence limits in such cases would not seem warranted. Instead, we feel that emphasis should be directed to obtaining an over-all measure of risk, coupled with an over-all test of statistical significance.

Statistical Procedures for Factor Control

A major problem in any epidemiological study is the avoidance of spurious associations. It has been remarked that where the risk of disease changes with age, apparent association of the disease with other age-related factors can result. However, there are appropriate statistical procedures for controlling those factors known or suspected to be related to disease occurrence. They serve not only to remove bias from the investigation but, in addition, can add to its precision.

Two simple procedures for obtaining factor control may first be mentioned. One is simply to restrict the investigation to individuals homogeneous on the factors to be controlled. For this situation the statistical procedures already outlined would be appropriate. The potential number of individuals available for such a study would, of course, be sharply restricted.

There is also the matching case method. A sample of N diseased individuals is drawn and the characteristics of each individual noted with respect to the control factors. Subsequently, a sample of N well individuals is drawn, with each individual matched on the control factors to one of the diseased individuals. The statistical procedures to be presented can be shown to cover the matched-sample approach as a special case, and a discussion of the analysis of such data will be given in that context. Some difficulties of the matched-sample study may be mentioned here. One is that when matching is made on a large number of factors, not even the fiction of a random sampling of control individuals can be maintained. Instead, one must be grateful for each matching control available. Another difficulty is that the method cannot be applied to factors under control, since diseased and control individuals are identical with respect to these factors. Conversely, factors under study in matched samples cannot themselves be controlled statistically. They can be analyzed separately or in particular conjunctions but cannot be employed as control factors.

An alternative to case matching is to draw independent samples of cases and controls, and adjust for other factors in the analysis. This approach requires simply the classification of individuals according to the various control and study factors desired, and an analysis for each separate subclassification as well as an appropriate summary analysis. Its success will depend on a reasonable degree of cross-matching between observations on diseased and control persons. In a small study various devices for reducing the number of subclassifications and for increasing the chances of cross-matching may be necessary, including a limit on the number of factors on which individuals are classified in any one analysis and the use of broad categories for any particular classification. Thus, a 10-year interval for age classification might permit a reasonable degree of cross-matching, whereas a 1-month interval would not.

The need for some degree of deliberate matching, even when the classification approach is employed, can be seen. If the disease under consideration occurs at advanced ages, little cross-matching would result

if controls were selected from the general population. The remedy lies in deliberately selecting controls from the same age groups anticipated for persons with the disease, perhaps even matching one or more controls on age for each diseased person. This principle can be extended to matching on several control factors, *solely for the purpose of increasing the extent of cross-matching in the analysis.*

One of the subtle effects which can occur in a retrospective study, even with careful planning, may be pointed out. It can be shown, for instance, that within a given age interval the average age of individuals with cancer of certain sites will be greater than the average age of individuals from the general population in the same age interval. This can arise when incidence increases rapidly with age and may pose a serious problem with broad age intervals. This effect can be offset by close matching of cases and controls on age in drawing samples, even though they are classified by a broad age category in the analysis.

When a random sample of diseased and disease-free individuals is classified according to various control factors the distribution of the factor under study within the *ith* classification may be represented as follows:

	With factor	Free of factor	Total
With disease	A_i	B_i	N_{1i}
Free of disease	C_i	D_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

Within this subgroup the approximate relative risk associated with the disease may be written as A_iD_i/B_iC_i. One may compare the observed number of diseased persons having the factor, A_i, with its expectation under the hypothesis of a relative risk of unity, $E(A_i)=N_{1i}M_{1i}/T_i$. The discrepancy between A_i and $E(A_i)$ (which is also the discrepancy for any other cell within a 2×2 table) can be tested relative to its variance which, subject to the fixed marginal totals—N_{1i}, N_{2i}, M_{1i}, and M_{2i}—is given by $V(A_i) = N_{1i}N_{2i}M_{1i}M_{2i}/T_i^2(T_i-1)$. The corrected chi square with 1 degree of freedom $(|A_i-E(A_i)|-\frac{1}{2})^2/V(A_i)$ reduces in this case to $(|A_iD_i-B_iC_i|-\frac{1}{2}T_i)^2(T_i-1)/N_{1i}N_{2i}M_{1i}M_{2i}$. This formula for the variance of A_i is obtained as the variance of the binomial variable $N_1PQ(P = M_1/T, Q = M_2/T)$, multiplied by a finite population correction factor $(T-N_1)/(T-1) = N_2/(T-1)$. The earlier chi-square formula, which is ordinarily used, essentially employs a finite population correction factor of N_2/T.

There is thus a difference between the two chi-square formulas of a factor of $(T-1)/T$ which, though trivial for any single significance test with respectably large T, can become important in the over-all significance test. It is with the latter formula, just presented, that chi square is computed as the ratio of the square of a deviation from its expected value to its variance.

The adjustment for control factors is at this point resolved for the resulting separate subclassifications. The problem of over-all measures of relative risk and statistical significance still remains. A reasonable over-all

significance test which has power for alternative hypotheses, where there is a consistent association in the same direction over the various subclassifications between the disease and a study factor, is provided by relating the summation of the discrepancy between observation and expectation to its variance. The corrected chi square with 1 degree of freedom then becomes $(|\Sigma A_i - \Sigma E(A_i)| - \frac{1}{2})^2/\Sigma V(A_i)$ where $E(A_i)$ and $V(A_i)$ are defined as above.

The specification of a summary estimate of the relative risk associated with a factor is not so readily resolved as that for an over-all significance test, and involves consideration of alternate approaches to a weighted average of the approximate relative risks for each subclassification (A_iD_i/B_iC_i). If one could assume that the increased relative risk associated with a factor was constant over all subclassifications, the estimation problem would reduce to weighting the several subclassification estimates according to their respective precisions. The complex maximum likelihood iterative procedure necessary for obtaining such a weighted estimate would seem to be unjustified, since the assumption of a constant relative risk can be discarded as usually untenable.

Another possible criterion for obtaining a summary estimate of relative risk would involve weighting the risks for subclassification by "importance." A twofold increase of a large risk is more important than a twofold increase of a small risk. An increased risk for a large group is more important than one for a small group. An increased risk for young individuals may be more important than for older individuals with a shorter life expectation. Difficulties arise in attempts to weight relative risk by measures of importance. For one, the necessary information on importance, in terms of the size of the populations affected or in terms of the absolute level of rates prevailing in the subgroups, is generally not contained within the scope of the investigation. A problem in definition of the precise terms of the weighted comparison also appears. Does one want to adjust the risks of disease among persons with the factor to the distribution of the population without the factor, or *vice versa*, or adjust the risks for the populations with and without the factor to a combined standard population? These procedures, and the different phrasing of the comparisons which they entail, could yield different answers. If only a small proportion of the population with the factor was in a subcategory with a high relative risk, while most of the factor-free population fell into this subcategory, and in other categories the relative risk associated with the factor was less than unity, the factor would appear to exert a protective influence under one set of weights but a harmful effect under the other.

Published instances of summary relative risks do not fall clearly into either of the two categories—weighting by precision or weighting by importance. They do follow an approach usually employed in age-adjusting mortality data. Since the relative risk for a single 2×2 table can be obtained from the incidence of the factor among diseased and well individuals, the problem would appear translatable into terms of obtaining

over-all, category-adjusted incidence figures. Direct or indirect methods of adjustment can be used, employing as a standard of reference the frequency distribution or rates corresponding to the sample of diseased persons, of controls, or the diseased persons and controls combined.

While such adjustment procedures provide weighting by importance in their customary application to mortality rates, this is not so in the relative risk situation. This may be illustrated in the following extreme example. Suppose that in each of two subcategories the approximate relative risk for a contrast between the presence and absence of a factor is about 5, which arises in the first subcategory from contrasting percentages of 1 and 5, and in the second subcategory from contrasting percentages of 95 and 99. If these percentages were based on equal numbers of individuals, all methods of category adjusting would yield contrasting adjusted summary percentages of 46 and 52, and a resultant relative risk of slightly less than 1.3. Some other approach for obtaining category-adjusted relative risks would seem desirable. However, to the extent that such extreme situations are not encountered in actual practice, results based on these more conventional adjustment procedures will not be grossly in error.

A suggested compromise formula for over-all relative risk is given by $R = \Sigma(A_iD_i/T_i)/\Sigma(B_iC_i/T_i)$. As a weighted average of relative risks this formula would, in the illustration given, yield the over-all relative risk of 5 found in each of the two subcategories. The weights are of the order $N_{1i}N_{2i}/(N_{1i} + N_{2i})$ and as such can be considered to weight approximately according to the precision of the relative risks for each subcategory. The weights can also be regarded as providing a reasonable weighting by importance.

An interesting property of this summary relative risk formula is that it equals unity only when $\Sigma A_i = \Sigma E(A_i)$ and hence the corresponding chi square is zero. From the fact that $A_i - E(A_i) = (A_iD_i - B_iC_i)/T_i$, it follows that when $\Sigma A_i = \Sigma E(A_i)$, $\Sigma A_iD_i/T_i$ will equal $\Sigma B_iC_i/T_i$, chi square will be zero, and R will be unity. The chi-square significance test can thus be construed as a significance test of the departure of R from unity.

Of some other procedures for measuring over-all relative risks, the one following also has the interesting property of being equal to unity when $\Sigma(A_i) = \Sigma E(A_i)$ and therefore subject to the chi-square test:

$$R_1 = \frac{\Sigma A_i \Sigma D_i}{\Sigma B_i \Sigma C_i} \Big/ \frac{\Sigma E(A_i) \Sigma E(D_i)}{\Sigma E(B_i) \Sigma E(C_i)} \text{ where } E(A_i) = N_{1i}M_{1i}/T_i,\ E(B_i)$$

$$= N_{1i}M_{2i}/T_i,\ E(C_i) = N_{2i}M_{1i}/T_i, \text{ and } E(D_i) = N_{2i}M_{2i}/T_i.$$

In this formula the numerator represents the crude value for the relative risk, which would result from pooling the data into one table and ignoring all subclassification on other factors. The denominator represents the crude value for relative risk, which would have resulted from pooling in the situation where all relative risks within each subclassification were exactly unity. Readers familiar with the "indirect" method of com-

puting standardized mortality ratios will recognize an analogy between the "indirect" method and the above procedure.

The estimator R_1 can be seen to have a bias toward unity. One reason is covered by the illustration which indicated that adjusted percentages (or frequencies) do not yield an appropriate adjusted relative risk. In addition, when either cases or controls have little representation in a subcategory, there will be lack of cross-matching and little information about relative risk, and the observed cell frequencies and their expectations will be numerically close. Such results will, in the process of summation used by the estimator, tend to force its value toward unity. This weakness will not be too important if the degree of cross-matching is roughly equal in the various subclassifications—an optimum goal one would normally attempt to achieve. The bias will become more pronounced as the number of control factors increases and as the prospects for good cross-matching become poorer.

We used the estimator R_1 in a recent paper (*27*), knowing its potential weaknesses. This was done to present results more nearly comparable with those reported by other investigators using similarly biased estimators. One set of results from this paper on lung cancer among women illustrates the conservative behavior of estimator R_1 compared with R, as additional factors are controlled. The relative risk (R_1) for epidermoid and undifferentiated pulmonary carcinoma associated with smoking more than one pack of cigarettes daily as compared to nonsmokers decreased from 7.1 (controlled for age) to 5.6 (controlled for age and coffee consumption). The corresponding figures, with R as a measure of relative risk, were 9.7 and 9.9.

Computational procedures for R and R_1 are presented in table 1, drawing on material comparing smoking histories of women diagnosed as cases of epidermoid and undifferentiated pulmonary carcinoma with those of female controls. For simplicity in presentation only two smoking levels are considered—nonsmokers and smokers of more than one pack of cigarettes daily. An extension of the significance testing procedures to the case of study factors at more than two levels is discussed later. The control factors are age and occupation. The basic data are given in the first 9 columns. Columns 10 and 11 carry the derivative calculations required for R. Columns 12 and 13 are used in the computation for R_1 and for the variance estimate in column 14—the latter being needed for the chi-square test. Only columns 1 to 10, 12, and 14 would be necessary to compute chi square, R and R_1. Column 13 is not essential for the computation of $E(D)$ but simplifies computation of $V(A)$, while providing a check on $E(A)$. Column 11 serves as a check on 10 and 12. A system of checks and computations is outlined at the bottom of table 1. Not all the computations shown would ordinarily be necessary for an analysis.

The corrected chi-square value of 30.66 (1 degree of freedom) would indicate a highly significant association between epidermoid and undifferentiated pulmonary carcinoma and cigarette smoking in women, after adjusting for possible effects connected with age or occupation. The

TABLE 1.—*Illustrative computations for chi square and for summary measures of undifferentiated pulmonary carcinoma*

Group		Epidermoid-undifferentiated pulmonary carcinoma			Controls			Cases and controls		
		1 + Pack cigarettes daily	Nonsmokers	Total	1 + Pack cigarettes daily	Nonsmokers	Total	1 + Pack cigarettes daily	Nonsmokers	Total
		A	B	N_1	C	D	N_2	M_1	M_2	T
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Housewives	under age 45	0	2	2	0	7	7	0	9	9
	45–54	2	5	7	1	24	25	3	29	32
	55–64	3	6	9	0	49	49	3	55	58
	65 and over	0	11	11	0	42	42	0	53	53
White-collar workers	under age 45	3	0	3	2	6	8	5	6	11
	45–54	2	2	4	2	18	20	4	20	24
	55–64	2	4	6	2	23	25	4	27	31
	65 and over	0	6	6	1	11	12	1	17	18
Other occupations	under age 45	1	0	1	3	10	13	4	10	14
	45–54	4	1	5	1	12	13	5	13	18
	55–64	0	6	6	1	19	20	1	25	26
	65 and over	1	3	4	0	15	15	1	18	19
Total		18	46	64	13	236	249	31	282	313

Checks: Total discrepancy, Y, $= \Sigma A - \Sigma E(A) = \Sigma(1) - \Sigma(12) = 11.625$
$= \Sigma D - \Sigma E(D) = \Sigma(5) - \Sigma(13) = 11.625$
$= \Sigma(AD/T) - \Sigma(BC/T) = \Sigma(10) - \Sigma(11) = 11.625$

$\Sigma(15) + \Sigma(16) = 64.000$; $\Sigma(3) = 64$
$\Sigma(17) + \Sigma(18) = 249.000$; $\Sigma(6) = 249$

Derivative computations: $\Sigma E(B) = \Sigma(2) + Y = 57.625$
$\Sigma E(C) = \Sigma(4) + Y = 24.625$
$\Sigma(AT/N_1) = \Sigma(1) + \Sigma(17) = 94.960$
$\Sigma(BT/N_1) = \Sigma(2) + \Sigma(18) = 218.040$
$\Sigma(CT/N_2) = \Sigma(4) + \Sigma(15) = 16.325$
$\Sigma(DT/N_2) = \Sigma(5) + \Sigma(16) = 296.675$

value of R implies that the risk of these cancers is 10.7 times as great for women currently smoking in excess of 1 pack a day than for women who never used cigarettes. The value of R_1, 7.05, is almost identical with the crude relative risk, 7.10, which results from pooling the data with no attention to the control factors. The difference from the published R_1 value of 6.3 in (*27*) arises from the exclusion in the illustrative example, of data for women currently smoking 1 pack a day or less and for occasional or discontinued smokers.

The computation of three other summary estimates of relative risk is also outlined in table 1. The additional derivative computations required for this purpose appear in columns 15 to 18. All three estimates are based on a direct method of category adjustment, that is, the use of a standard distribution to which both the case and control distributions are

relative risk (R, R_1, R_2, R_3, *and* R_4) *relating to the association of epidermoid and in women with smoking history*

Derivative computations								
$\frac{AD}{T}$ $\frac{(1)(5)}{(9)}$	$\frac{BC}{T}$ $\frac{(2)(4)}{(9)}$	$\frac{E(A)}{}$ $\frac{(3)(7)}{(9)}$	$\frac{E(D)}{}$ $\frac{(6)(8)}{(9)}$	$\frac{V(A)}{}$ $\frac{(12)(13)}{(9)-1.0}$	$\frac{N_1C}{N_2}$ $\frac{(3)(4)}{(6)}$	$\frac{N_1D}{N_2}$ $\frac{(3)(5)}{(6)}$	$\frac{N_2A}{N_1}$ $\frac{(1)(6)}{(3)}$	$\frac{N_2B}{N_1}$ $\frac{(2)(6)}{(3)}$
(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
0	0	0	7. 000	0	0	2. 000	0	7. 000
1. 500	0. 156	0. 656	22. 656	0. 480	0. 280	6. 720	7. 143	17. 857
2. 534	0	0. 466	46. 466	0. 380	0	9. 000	16. 333	32. 667
0	0	0	42. 000	0	0	11. 000	0	42. 000
1. 636	0	1. 364	4. 364	0. 595	0. 750	2. 250	8. 000	0
1. 500	0. 167	0. 667	16. 667	0. 483	0. 400	3. 600	10. 000	10. 000
1. 484	0. 258	0. 774	21. 774	0. 562	0. 480	5. 520	8. 333	16. 667
0	0. 333	0. 333	11. 333	0. 222	0. 500	5. 500	0	12. 000
0. 714	0	0. 286	9. 286	0. 204	0. 231	. 769	13. 000	0
2. 667	0. 056	1. 389	9. 389	0. 767	0. 385	4. 615	10. 400	2. 600
0	0. 231	0. 231	19. 231	0. 178	0. 300	5. 700	0	20. 000
0. 790	0	0. 211	14. 211	0. 166	0	4. 000	3. 750	11. 250
12. 825	1. 201	6. 375	224. 375	4. 036	3. 325	60. 675	76. 960	172. 040

Chi-square: $X^2 = (|\text{discrepancy}| - 0.5)^2/\Sigma V(A) = (|Y| - 0.5)^2/\Sigma(14) = 30.66$
Relative risk: $R = \Sigma(AD/T)/\Sigma(BC/T) = \Sigma(10)/\Sigma(11) = 10.68$
R_1 { crude relative risk, $r = \Sigma A\Sigma D/\Sigma B\Sigma C = \Sigma(1)\Sigma(5)/\Sigma(2)\Sigma(4) = 7.10$
adjustment factor, $f = \Sigma E(A)\Sigma E(D)/\Sigma E(B)\Sigma E(C) = \Sigma(12\Sigma(13)/\Sigma E(B)\Sigma E(C) = 1.0081$
$R_1 = r/f = 7.05$
$R_2 = \Sigma A\Sigma(N_1D/N_2)/\Sigma B\Sigma(N_1C/N_2) = \Sigma(1)\Sigma(16)/\Sigma(2)\Sigma(15) = 7.14$
$R_3 = \Sigma(N_2A/N_1)\Sigma D/\Sigma(N_2B/N_1)\Sigma C = \Sigma(5)\Sigma(17)/\Sigma(4)\Sigma(18) = 8.12$
$R_4 = \Sigma(AT/N_1)\Sigma(DT/N_2)/\Sigma(BT/N_1)\Sigma(CT/N_2) = 7.91$

Note: Figures shown are rounded from those actually calculated and consequently are not fully consistent. Column totals and figures shown do not necessarily agree.

adjusted. If the distribution of diseased cases is taken as the standard distribution to which the controls are adjusted, the estimator becomes

$$R_2 = \frac{\Sigma A_i \Sigma\left(D_i \times \frac{N_{1i}}{N_{2i}}\right)}{\Sigma B_i \Sigma\left(C_i \times \frac{N_{1i}}{N_{2i}}\right)}.$$

Estimator R_2 was used by Wynder *et al.* in a study of the association of cervical cancer in women with circumcision status of sex partners (*16*). The merit of employing the cervical cancer case-distribution as the standard presumably rests on the fact that this distribution at least would be well defined by the study.

TABLE 1.—*Illustrative computations for chi square and for summary measures of relative risk (R, R_1, R_2, R_3, and R_4) relating to the association of epidermoid and undifferentiated pulmonary carcinoma in women with smoking history*

Group		Epidermoid-undifferentiated pulmonary carcinoma			Controls			Cases and controls			Derivative computations								
		1+ Pack cigarettes daily	Nonsmokers	Total	1+ Pack cigarettes daily	Nonsmokers	Total	1+ Pack cigarettes daily	Nonsmokers	Total	$\frac{AD}{T}$ $\frac{(1)(5)}{(9)}$	$\frac{BC}{T}$ $\frac{(2)(4)}{(9)}$	$\frac{E(A)}{(3)(7)}$ $\frac{}{(9)}$	$\frac{E(D)}{(6)(8)}$ $\frac{}{(9)}$	$\frac{V(A)}{(12)(13)}$ $\frac{}{(9)-1.0}$	$\frac{N_1C}{N_2}$ $\frac{(3)(4)}{(6)}$	$\frac{N_1D}{N_2}$ $\frac{(3)(5)}{(6)}$	$\frac{N_2A}{N_1}$ $\frac{(1)(6)}{(3)}$	$\frac{N_2B}{N_1}$ $\frac{(2)(6)}{(3)}$
		A (1)	B (2)	N_1 (3)	C (4)	D (5)	N_2 (6)	M_1 (7)	M_2 (8)	T (9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Housewives	under age 45	0	2	2	0	7	7	0	9	9	0	0	0	7.000	0	0	2.000	0	7.000
	45–54	2	5	7	1	24	25	3	29	32	1.500	0.156	0.656	22.656	0.480	0.280	6.720	7.143	17.857
	55–64	3	6	9	0	49	49	3	55	58	2.534	0	0.466	46.466	0.380	0	9.000	16.333	32.667
	65 and over	0	11	11	0	42	42	0	53	53	0	0	0	42.000	0	0	11.000	0	42.000
White-collar workers	under age 45	3	0	3	2	6	8	5	6	11	1.636	0	1.364	4.364	0.595	0.750	2.250	8.000	0
	45–54	2	2	4	2	18	20	4	20	24	1.500	0.167	0.667	16.667	0.483	0.400	3.600	10.000	10.000
	55–64	2	4	6	2	23	25	4	27	31	1.484	0.258	0.774	21.774	0.562	0.480	5.520	8.333	16.667
	65 and over	0	6	6	1	11	12	1	17	18	0	0.333	0.333	11.333	0.222	0.500	5.500	0	12.000
Other occupations	under age 45	1	0	1	3	10	13	4	10	14	0.714	0	0.286	9.286	0.204	0.231	.769	13.000	0
	45–54	4	1	5	1	12	13	5	13	18	2.667	0.056	1.389	9.389	0.767	0.385	4.615	10.400	2.600
	55–64	0	6	6	1	19	20	1	25	26	0	0.231	0.231	19.231	0.178	0.300	5.700	0	20.000
	65 and over	1	3	4	0	15	15	1	18	19	0.790	0	0.211	14.211	0.166	0	4.000	3.750	11.250
Total		18	46	64	13	236	249	31	282	313	12.825	1.201	6.375	224.375	4.036	3.325	60.675	76.960	172.040

Checks: Total discrepancy, Y, $= \Sigma A - \Sigma E(A) = \Sigma(1) - \Sigma(12) = 11.625$
$= \Sigma D - \Sigma E(D) = \Sigma(5) - \Sigma(13) = 11.625$
$= \Sigma(AD/T) - \Sigma(BC/T) = \Sigma(10) - \Sigma(11) = 11.625$

$\Sigma(15) + \Sigma(16) = 64.000$; $\Sigma(3) = 64$
$\Sigma(17) + \Sigma(18) = 249.000$; $\Sigma(6) = 249$

Derivative computations: $\Sigma E(B) = \Sigma(2) + Y = 57.625$
$\Sigma E(C) = \Sigma(4) + Y = 24.625$
$\Sigma(AT/N_1) = \Sigma(1) + \Sigma(17) = 94.960$
$\Sigma(BT/N_1) = \Sigma(2) + \Sigma(18) = 218.040$
$\Sigma(CT/N_2) = \Sigma(4) + \Sigma(15) = 16.325$
$\Sigma(DT/N_2) = \Sigma(5) + \Sigma(16) = 296.675$

Chi-square: $X^2 = (|\text{discrepancy}| - 0.5)^2/\Sigma V(A) = (|Y| - 0.5)^2/\Sigma(14) = 30.66$

Relative risk: $R = \Sigma(AD/T)/\Sigma(BC/T) = \Sigma(10)/\Sigma(11) = 10.68$

R_1 { crude relative risk, $r = \Sigma A\Sigma D/\Sigma B\Sigma C = \Sigma(1)\Sigma(5)/\Sigma(2)\Sigma(4) = 7.10$
adjustment factor, $f = \Sigma E(A)\Sigma E(D)/\Sigma E(B)\Sigma E(C) = \Sigma(12\Sigma(13)/\Sigma E(B)\Sigma E(C = 1.0081$
$R_1 = rf = 7.05$

$R_2 = \Sigma A\Sigma(N_1D/N_2)/\Sigma B\Sigma(N_1C/N_2) = \Sigma(1)\Sigma(16)/\Sigma(2)\Sigma(15) = 7.14$

$R_3 = \Sigma(N_2A/N_1)\Sigma D/\Sigma(N_2B/N_1)\Sigma C = \Sigma(5)\Sigma(17)/\Sigma(4)\Sigma(18) = 8.12$

$R_4 = \Sigma(AT/N_1)\Sigma(DT/N_2)/\Sigma(BT/N_1)\Sigma(CT/N_2) = 7.91$

Note: Figures shown are rounded from those actually calculated and consequently are not fully consistent. Column totals and figures shown do not necessarily agree.

If the distribution of control cases is taken as standard the estimator becomes

$$R_3 = \frac{\Sigma\left(A_i \times \frac{N_{2i}}{N_{1i}}\right)\Sigma D_i}{\Sigma\left(B_i \times \frac{N_{2i}}{N_{1i}}\right)\Sigma C_i}.$$

If the combined distribution is taken as standard the estimator becomes

$$R_4 = \frac{\Sigma\left(A_i \times \frac{T_i}{N_{1i}}\right)\Sigma\left(D_i \times \frac{T_i}{N_{2i}}\right)}{\Sigma\left(B_i \times \frac{T_i}{N_{1i}}\right)\Sigma\left(C_i \times \frac{T_i}{N_{2i}}\right)}.$$

If any N_{1i} or N_{2i} should equal zero, the estimator R_4 would not be defined. R_2 is not defined for any zero-valued N_{2i}, and R_3 is not defined for any zero-valued N_{1i}. In these instances it would be necessary to exclude the zero-frequency categories to define the estimators. The estimator R_1 retains these categories at the expense of greater bias toward unity. The estimator R gives such categories zero weight, since they contain no information about relative risk. The chi-square significance test gives no weight to these categories.

While R_4 is clearly a direct adjusted estimate of relative risk employing the combined distribution as standard, R_2 and R_3 may be viewed alternatively as either direct or indirect adjusted estimates. The same estimates will result if a direct adjustment is made using the distribution of cases as standard, or an indirect adjustment is made using the factor incidence rates for controls as the standard rates.

It may be noted that in the example used, the values for R_2, R_3, and R_4 (7.14, 8.12, and 7.91, respectively) were roughly comparable to R_1, and all were smaller than R. The example was selected because all the N_{1i} and N_{2i} values were non-zero, so that the values of R_2, R_3, and R_4 were all defined.

The over-all relative risk estimates are averages and as averages may conceal substantial variation in the magnitudes of the relative risk among subgroups. Ordinarily, the individual subcategory data should be examined, paying special attention to relative risks based on reasonably large sample sizes. This will provide protection against the potential deficiencies of any particular summary relative risk formula employed. The over-all chi-square significance test in any case will remain appropriate for detecting any strong general tendency for the risk of disease to be associated with the presence or absence of the test factor.

The Matched-Sample Study

The matched-sample study previously described can be considered a special case of the classification procedure with the number of classifications equal to the number of pairs of individuals. The status of pairs of well and diseased individuals classified with respect to the presence or absence of the suspect factor in each individual will be represented as

F, G, H, or J in the following fourfold table. The meanings attached to the marginal totals A, B, C, and D are the same as those in the first schematic representation.

	Diseased individuals		
Well individuals	With factor	Free of factor	Total
With factor	F	G	C
Free of factor	H	J	D
Total	A	B	N

In the absence of association between the disease and the factor, we expect the same number of individuals with the factor to appear among both diseased and well individuals; that is, we expect $A(=F+H)$ to equal $C(=F+G)$. This can occur only when $G=H$ and the statistical test is simply whether or not G differs significantly from 50 percent of $G+H$. G is tested as a binomial variable with parameter $\frac{1}{2}$, $G+H$ being the number of cases. G thus has expectation $\frac{1}{2}(G+H)$, variance $\frac{1}{4}(G+H)$ and the corrected chi square with 1 degree of freedom can readily be shown to reduce to $(|G-H|-1)^2/(G+H)$.

Treating the data as consisting of N classifications each with $N_{1i}=N_{2i}=1$, $T_i=2$ and applying the previously described procedures will lead to the same value of chi square. For F of the N classifications, $A_i=1$, $M_{1i}=2$, $M_{2i}=0$, $E(A_i)=1$, $V(A_i)=0$; for G classifications $A_i=0$, $M_{1i}=M_{2i}=1$, $E(A_i)=\frac{1}{2}$, $V(A_i)=\frac{1}{4}$; for H classifications $A_i=1$, $M_{1i}=M_{2i}=1$, $E(A_i)=\frac{1}{2}$, $V(A_i)=\frac{1}{4}$; and for J classifications, $A_i=0$, $M_{1i}=0$, $M_{2i}=2$, $E(A_i)=0$, $V(A_i)=0$. Thus, $\Sigma A_i=F+H$, $\Sigma E(A_i)=F+\frac{1}{2}(G+H)$, $\Sigma V(A_i)=\frac{1}{4}(G+H)$, and the resultant corrected chi square can again be seen to be $(|G-H|-1)^2/(G+H)$.

It is of interest to observe that the summary chi-square formula is appropriate in the matched-sample case, even though the frequencies for each of the separate subclassifications are small. Its appropriateness, despite the small frequencies, stems from the fact that it is a test on a summation of random variables, A_i, and thus tends to approach normality rapidly, making the chi-square test valid, even though the individual A_i's are not normally distributed. This property of the chi-square formula applies in the general classification as well as the matched-sample situation. Only substantial lack of cross-matching in the general case would tend to make the chi-square test invalid. It is also essential, of course, that there be some appreciable variation in the presence or absence of the factor under study.

It should be noted that in the matched-sample study with $T_i=2$ for each of the N pairs of individuals, the variances of the A_i's would have been understated by a factor of 2, had $T-1$ been replaced by T in the variance formulas. The usual formula for chi square does essentially make this replacement, but it is usually of little consequence if T is of any reasonable magnitude. The formulas for relative risk in the matched-sample study reduce simply to the following: $R=H/G$; $R_1=R_2=R_3=R_4=AD/BC$.

Study Factors at More Than Two Levels

The preceding discussion on the analysis of retrospective data has been in terms of the test factor under study taking only two values. This framework has sufficed for discussion of the underlying statistical ideas and issues. In practice, the study factor will frequently take on more than two, perhaps many, potential values. When the number of study factor values is large, grouping can reduce them to manageable proportions.

The need to consider only a limited number of classes for the study factor stems from the fact that, when an association is anticipated, most of the significant information about the association will come from the results for the more extreme values of the study factor. While it is efficient to concentrate attention on the test factor classes expected to show the greatest differences in association with the disease, it is also profitable to consider intermediate values for the test factor to seek evidence for a consistent pattern of association. For example, in table 1, a highly significant difference between nonsmokers and women currently smoking more than 1 pack of cigarettes daily was illustrated. Inclusion of data for smokers of 1 pack or less a day showing results intermediate between the other classes would have added little, if anything, to the statistical significance of the results, and might actually lower it, if one made an over-all test of the differences among the three smoking classes. However, the observation that the intermediate smoking class does, in fact, show an intermediate relative risk contributes to an orderly pattern and increases our confidence in the conclusions suggested by the data for the remaining two classes.

For any two particular test-factor levels, the relative risk for one over the other may be calculated using only the data pertaining to those two levels or by using the results for all test levels. In the formulas previously given for R, R_1, R_2, R_3, and R_4, the difference between the two calculating procedures is simply one of setting the values of N_{1i}, N_{2i}, and $T_i = N_{1i} + N_{2i}$ in terms of number of cases and controls occurring at the two study-factor levels only, or defining them in terms of total number of cases and controls in the entire study. When total cases and controls are used in defining N_{1i}, N_{2i}, and T_i, it can be shown that for R_1, R_2, R_3, and R_4 the various relative risks will be internally consistent with each other. If the relative risk for the first level is twice that for the second level, which in turn is twice that for the third level, then the relative risk for the first level will be four times that of the third. These exact relationships do not hold for R as an estimator of relative risk, and a somewhat sophisticated extension of the formula for R would be required to secure this property.

The problem of obtaining a summary chi square when the study factor is at more than two levels is complicated by the fact that the deviations from expectation at the various study-factor levels are intercorrelated. When there are but two levels, the two deviations will have perfect negative correlation, and attention need be directed to only one of the devia-

tions. Irrespective of the number of levels, at any one level the deviation from expectation among diseased persons will be equal, but opposite in sign, to the deviation from expectation among controls, so that attention can be confined to the deviations for diseased persons.

The problem can be stated as one of reducing a set of correlated deviations into a summary chi square. Table 2 applies this process for obtaining a summary chi square to the study of the association of epidermoid and undifferentiated pulmonary carcinoma in women and maximum cigarette-smoking rate, classified into three levels, after adjustment for age and occupation.

The general expressions for the expectations and variances of the number of cases at a particular test-factor level are given in the lower right section of table 2. Also shown is the expression for the covariance between the number of cases at two different test-factor levels. Since the total of all the deviations is zero, one would in general need the variances of, and covariances between, the number of cases at all but one of the levels. The number of covariance terms will rise sharply as the number of test levels are increased. At 3 test levels, there are 2 variance terms and 1 covariance term, while at 10 test levels, there would be 9 variances and 36 covariance terms of interest.

For the general case the burden of computation could be heavy. After all the necessary computation for the deviations, their variances and covariances, there would still remain the problem of converting these, presumably by matrix methods, into a summary chi square. Since the retrospective problem will normally involve only a limited number of test-factor levels, precise procedures will be given only for the three-level situation, and approximate procedures outlined for the general case.

The exact computation procedure for the three-level case is detailed in table 2. Lines (1), (2), and (4) show the total observed and expected frequencies and variances of the number of cases (and controls) at each of the three smoking-rate levels, after adjusting for age and occupation. These are the summary totals over each subclassification obtained by application of the formulas appearing in table 2.

Lines (5) and (6) give the chi squares corresponding to the total deviation from expectation at each of the smoking-rate levels. The chi squares in line (5) are corrected for continuity. They relate to the difference of the particular level to which they apply, from the two other levels combined. Following the usual practice of making no continuity corrections when chi squares with more than 1 degree of freedom are under consideration, line (6) shows the uncorrected chi squares.

The computing procedure of table 2 takes advantage of the fact that, since the sum of the deviations from expectation is zero, the variance of the third deviation must equal the sum of the other two variances plus twice the covariance for the first two deviations. The covariance of the first two deviations is readily obtained as illustrated and is used in calculating the summary chi square. The summary chi square is obtained as the sum of squares of two orthogonal deviates, with each

TABLE 2.—*Illustrative computation of summary chi square, when there are 3 levels for study factor. The data relate to the association of epidermoid and undifferentiated pulmonary carcinoma in women with smoking history*

	1+ Pack cigarettes daily			1 Pack or less of cigarettes daily			Occasional or nonsmokers			Total		
	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_1)	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_2)	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_3)	Epidermoid-undifferentiated pulmonary carcinoma (ΣN_1)	Controls (ΣN_2)	Total ($\sim T$)
(1) Total observed frequencies	19	17	36	32	71	103	51	251	302	102	339	441
(2) Total expected frequencies, adjusted for age and occupation	9. 09	26. 91	36	23. 76	79. 24	103	69. 15	232. 85	302	102	339	441
(3) Total deviation from expectation (1) − (2)	$+9.91 = Y_1$			$+8.24 = Y_2$			$-18.15 = Y_3$					
(4) Variance of total observed frequencies, subject to fixed marginal totals in each age and occupation group	$5.9163 = V_1$			$12.2900 = V_2$			$14.0723 = V_3$					
(5) Individual corrected chi squares $(\|Y\|-0.5)^2/V$	$14.97 = X^2_{1c}$			$4.88 = X^2_{2c}$			$22.15 = X^2_{3c}$					
(6) Individual uncorrected chi squares Y^2/V	$16.60 = X^2_1$			$5.53 = X^2_2$			$23.42 = X^2_3$					
(7) Covariance (Y_1, Y_2) $(V_3 - V_1 - V_2)/2$				-2.0670								
(8) Adjusted Y_2 $Y_2 - (7)\,Y_1/V_1$				11.70								
(9) Adjusted V_2 $V_2 - (7)^2/V_1$				11.5678								
(10) Adjusted X^2_2 $(8)^2/(9)$				$11.83 = X^2_2$(ad.)								
(11) Summary chi square (2 degrees of freedom) $X^2_1 + X^2_2$(ad.)				$16.60 + 11.83 = 28.43$								

For the general situation the total expected case frequency at the *jth* level of a test factor is

$$\sum_i N_{1i} M_{ji} / T_i$$

The variance of the total case frequency is

$$V_j = \sum_i \frac{N_{1i} N_{2i} M_{ji} (T_i - M_{ji})}{T_i^2 (T_i - 1)}$$

The covariance of the total case frequencies at test levels j and k is

$$-\sum_i \frac{N_{1i} N_{2i} M_{ji} M_{ki}}{T_i^2 (T_i - 1)}$$

The index of summation, i, represents the various subclassifications into which the results are divided

For 3 test levels only, since $Y_3 = -(Y_1 + Y_2)$, it follows that $V_3 = V_1 + V_2 + 2$ Covariance (Y_1, Y_2)

square adjusted for its own variance. The first deviate squared is simply the uncorrected chi square at the first level in line (6)—the variance of the deviate remaining as initially calculated. The second deviate is the deviation at the second level adjusted for its correlation with the first deviation [adjusted $Y_2 = Y_2 - b_{21}Y_1$; b_{21} = covariance (Y_1, Y_2)/variance Y_1)]. The variance of the adjusted second deviate is the initial value reduced by that portion of the variation accounted for by the first deviation [Var. (adjusted Y_2) = variance Y_2 − covariance$^2(Y_1, Y_2)$/variance Y_1)].

In the present instance the summary chi square with 2 degrees of freedom is 28.43 [line (11)]. This presumably is close to the chi square with 1 degree of freedom which would have obtained had only the two most extreme smoking classes been compared. If one examines the individual uncorrected chi squares [line (6)], their total is found to be 45.55, the maximum individual figure being 23.42. *It will necessarily be true that the summary chi-square value will lie between the largest of the three chi squares and their total. At almost any reasonable probability level these limits would be sufficient to establish statistical significance without further calculation.* In our companion paper (*27*) this rule sufficed in almost all instances to separate the significant from the nonsignificant results.

Comments on Extensions to More Than Three Factors

Two procedures can be suggested for getting approximate summary chi squares, when there are a large number of levels for the test factors, without the burden of computation that the exact method would entail. Both methods calculate the approximate summary chi square as a sum of squares of approximately orthogonal standardized deviates.

In the first method one computes an uncorrected chi square with 1 degree of freedom for the difference of the first level from all the remaining levels combined (the same first step as in the illustration for the three-level case). Discarding the data from the first level, a second chi square is computed for the difference between the second test-factor level and the remaining levels combined. This is done successively up to and including the last two remaining levels. The approximate summary chi square is then the sum of the separate chi squares with the number of degrees of freedom being one less than the number of test levels.

Exactly orthogonal standardized deviates would be obtained if, in the summary analysis, as each successive total deviation from expectation were evaluated, it was adjusted for its multiple regression on the preceding deviations, and then standardized by the adjusted variance. This, of course, would no longer be a simplified approximate procedure. However, it can be shown that for a single classification, in the multiple regression of any deviation from expectation on any subset of deviations, the regression coefficients will all be equal; the multiple regression on the set of deviations will be the same as the simple regression on their sum. The equality of regression coefficients, while holding true exactly for deviations in the separate subclassifications, will hold only approximately for the total

deviations from expectation (it would hold exactly if equal numbers of individuals were observed from level to level at each subclassification). Nevertheless, this result suggests that approximately orthogonal deviates would be obtained if, in evaluating each successive total deviation, it were adjusted for the cumulative total of deviations already evaluated. Computing procedures to accomplish this can readily be devised.

Both approximate chi-square procedures just outlined, which may have merit when more than three groups are being compared simultaneously, should, in theory, yield linear combinations of independent chi squares. While testing the chi-square values obtained as though they were exact is not likely to be too inappropriate, it may be more correct to obtain a modified number of degrees of freedom, along the lines suggested by Satterthwaite (*47*) for problems involving such linear combinations. What the modified number of degrees of freedom would be has not been investigated by us, and it may prove as easy to apply the exact chi-square procedure, indicated later, as to determine the appropriate degrees of freedom for the approximate chi square.

It is of interest that a somewhat similar task of obtaining an appropriate summary chi square appears in the birth-order problems described by Halperin (*48*). There, it was necessary to compare a set of total observations (across family sizes) with a set of total expectations, one for each birth order. Halperin described a matrix-inversion procedure for reducing the set of correlated deviations into a summary chi square. In that problem it can be shown that all the regression coefficients are equal in the multiple regression of the deviation at a particular birth order on the set of deviations at all succeeding birth orders. The second approximate method described previously for the present problem could thus be used exactly for the birth-order problem, permitting simplified computation of chi square. The procedure indicated by Halperin has the advantage of generality and could be applied to the current and related problems, if one obtained all the necessary variances and covariances and inverted the resulting matrix.

References

(*1*) Snow, J.: On the mode of communication of cholera. *In* Snow on Cholera. New York, The Commonwealth Fund, 1936, pp. 1–139.

(*2*) Holmes, O. W.: The contagiousness of puerperal fever. *In* Medical Classics. Baltimore, Williams & Wilkins Co., vol. 1, 1936, pp. 211–243.

(*3*) Stern, R.: Nota sulle ricerche del dottore Tanchon intorno la frequenza del cancro. Annali Universali di Medicina 110: 484–503, 1844.

(*4*) Stocks, P., and Campbell, J. M.: Lung cancer death rates among non-smokers and pipe and cigarette smokers. Brit. M. J. 2: 923–929, 1955.

(*5*) Wynder, E. L., and Cornfield, J.: Cancer of the lung in physicians. New England J. Med. 248: 441–444, 1953.

(*6*) Lane-Claypon, J. E.: A further report on cancer of the breast, with special reference to its associated antecedent conditions. Rept. Publ. Health & M. Subj., No. 32, 1926, pp. 1–189.

(*7*) Clemmesen, J., Lockwood, K., and Nielsen, A.: Smoking habits of patients with papilloma of urinary bladder. Danish M. Bull. 5: 123–128, 1958.

(*8*) Denoix, P. R., and Schwartz, D.: Tobacco and cancer of the bladder. (Bulletin de L'Association francaise pour l'étude du Cancer.) Cancer 43: 387–393, 1956.

(9) LILIENFELD, A. M., LEVIN, M. L., and MOORE, G. E.: The association of smoking with cancer of the urinary bladder in humans. A.M.A. Arch. Int. Med., 1956.

(10) MUSTACCHI, P., and SHIMKIN, M. B.: Cancer of the bladder and infestation with *Schistosoma hematobium*. J. Nat. Cancer Inst. 20: 825–842, 1958.

(11) LILIENFELD, A. M.: The relationship of cancer of the female breast to artificial menopause and marital status. Cancer 9: 927–934, 1956.

(12) LILIENFELD, A. M., and LEVIN, M. L.: Some factors involved in the incidence of breast cancer. *In* Proc. Third National Cancer Conference. Philadelphia, J. B. Lippincott Co., 1957, pp. 105–112.

(13) SEGI, M., FUKUSHIMA, I., FUJISAKU, S., KURIHARA, M., SAITO S., ASANO, K., and KAMOI, M.: An epidemiological study on cancer in Japan. Gann Supp. 48, 1957.

(14) DUNHAM, L. J., THOMAS, L. B., EDGCOMB, J. H., and STEWART, H. L.: Some environmental factors and the development of uterine cancers in Israel and New York City. To be published in Acta Unio internat. contra cancrum.

(15) STOCKS, P.: Cancer of the uterine cervix and social conditions. Brit. J. Cancer 9: 487–494, 1955.

(16) WYNDER, E. L., CORNFIELD, J., SCHROFF, P. D., and DORAISWAMI, K. R.: A study of environmental factors in carcinoma of the cervix. Am. J. Obst. & Gynec. 68: 1016–1052, 1954.

(17) MILLS, C. A., and PORTER, M. M.: Tobacco smoking habits and cancer of the mouth and respiratory system. Cancer Res. 10: 539–542, 1950.

(18) WYNDER, E. L., BROSS, I. J., and DAY, E.: A study of environmental factors in cancer of the larynx. Cancer 9: 86–110, 1956.

(19) MANNING, M. D., and CARROLL, B. E.: Some epidemiological aspects of leukemia in children. J. Nat. Cancer Inst. 19: 1087–1094, 1957.

(20) BRESLOW, L., HOAGLIN, L., RASMUSSEN, G., and ABRAMS, H. K.: Occupations and cigarette smoking as factors in lung cancer. Am. J. Pub. Health 44: 171–181, 1954.

(21) DOLL, R., and HILL, A. B.: A study of the aetiology of carcinoma of the lung. Brit. M. J. 2: 1271–1286, 1952.

(22) LEVIN, M. L.: Etiology of lung cancer; present status. New York J. Med. 54: 769–777, 1954.

(23) SADOWSKY, D. A., GILLIAM, A. G., and CORNFIELD, J.: The statistical association between smoking and carcinoma of the lung. J. Nat. Cancer Inst. 13: 1237–1258, 1953.

(24) WATSON, W. L., and CONTE, A. J.: Lung cancer and smoking. Am. J. Surg. 89: 447–456, 1955.

(25) WYNDER, E. L., and GRAHAM, E. A.: Tobacco smoking as possible etiologic factor in bronchiogenic carcinoma. J.A.M.A. 143: 329–336, 1950.

(26) WYNDER, E. L., BROSS, I. J., CORNFIELD, J., and O'DONNELL, W. E.: Lung cancer in women. New England J. Med. 255: 1111–1121, 1956.

(27) HAENSZEL, W., SHIMKIN, M. B., and MANTEL, N.: A retrospective study of lung cancer in women. J. Nat. Cancer Inst. 21: 825–842, 1958.

(28) AIRD, I., BENTALL, H. H., and ROBERTS, J. A. F.: A relationship between cancer of stomach and the ABO blood groups. Brit. M. J. 1: 799–801, 1953.

(29) BUCKWALTER, J. A., WOHLWEND, C. B., COLTER, D. C., TIDRICK, R. T., and KNOWLER, L. A.: The association of the ABO blood groups to gastric carcinoma. Surg. Gynec. & Obst. 104: 176–179, 1957.

(30) KRAUS, A. S., LEVIN, M. L., and GERHARDT, P. R.: A study of occupational associations with gastric cancer. Am. J. Pub. Health 47: 961–970, 1957.

(31) CORNFIELD, J.: A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J. Nat. Cancer Inst. 11: 1269–1275, 1951.

(32) DORN, H. F.: Some applications of biometry in the collection and evaluation of medical data. J. Chron. Dis. 1: 638–664, 1955.

(33) NEYMAN, J.: Statistics—servants of all sciences. Science 122: 3166, 1955.
(34) BERKSON, J.: Limitations of the application of fourfo'd table analysis to hospital data. biometrics Bull. 2: 47–53, 1946.
(35) WHITE, C.: Sampling in medical research. Brit. M. J. 2: 1284–1288, 1953.
(36) GREENWOOD, M., and YULE, G. U.: On the determination of size of family and of the distribution of characters in order of birth from samples taken through members of the sibships. Roy. Stat. Soc. J. 77: 179–197, 1914.
(37) HAENSZEL, W.: Variation in incidence of and mortality from stomach cancer with particular reference to the United States. J. Nat. Cancer Inst. 21: 213–262, 1958.
(38) VIDEBAEK, A., and MOSBECH, J.: The aetiology of gastric carcinoma elucidated by a study of 302 pedigrees. Acta med. scandinav. 149: 137–159, 1954.
(39) WHELPTON, P. K., and FREEDMAN, R.: A study of the growth of American families. Am. J. Sociol. 61: 595–601, 1956.
(40) LEVIN, M. L., GOLDSTEIN, H., and GERHARDT, P. R.: Cancer and tobacco smoking. J.A.M.A. 143: 336–338, 1950.
(41) LEVIN, M. L., KRAUS, A. S., GOLDBERG, I. D., and GERHARDT, P. R.: Problems in the study of occupation and smoking in relation to lung cancer. Cancer 8: 932–936, 1955.
(42) LILIENFELD, A. M.: Possible existence of predisposing factors in the etiology of selected cancers of nonsexual sites in females. A preliminary inquiry. Cancer 9: 111–122, 1956.
(43) WINKELSTEIN, W., JR., STENCHEVER, M. A., and LILIENFELD, A. M.: Occurrence of pregnancy, abortion and artificial menopause among women with coronary artery disease: a preliminary study. J. Chron. Dis. 7: 273–286, 1958.
(44) BILLINGTON, B. P.: Gastric cancer—relationships between ABO blood-groups, site, and epidemiology. Lancet 2: 859–862, 1956.
(45) SCHWARTZ, D., and ANGUERA, G.: Une cause de biais dans certaines enquêtes médicales: le temps de séjour des malades a l'hôpital. Communication à l'Institut International de Statistique, 30ème Session. Stockholm, 1957.
(46) CORNFIELD, J.: A statistical problem arising from retrospective studies. Proc. Third Berkeley Symposium on Mathematical Statistics and Probability 4: 135–148, 1956.
(47) SATTERTHWAITE, F. E.: Synthesis of variance. Psychometrika 6: 309–316, 1941.
(48) HALPERIN, M.: The use of X^2 in testing effect of birth order. Ann. Eugenics 18: 99–106, 1953.