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, employ'ng 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 computing 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 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	$\frac{AD}{T}$ $\frac{(1)(5)}{(9)}$	$\frac{BC}{T}$ $\frac{(2)(4)}{(9)}$	E(A) $\frac{(3)(7)}{(9)}$	E(D) $\frac{(6)(8)}{(9)}$	V(A) $\frac{(12)(13)}{(9)-1.0}$	$\frac{N_1C}{N_2}$ $\frac{(3)(4)}{(6)}$	$\frac{N_1D}{N_2}$ $\frac{(3)(5)}{(6)}$	$\frac{N_2A}{N_1}$ $\frac{(1)(6)}{(3)}$	$\frac{N_2B}{N_1}$ $\frac{(2)(6)}{(3)}$
		A	B	N_1	C	D	N_2	M_1	M_2	T									
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(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: $\Sigma E(B) = \Sigma(2) + Y = 57.625$

$\Sigma E(C) = \Sigma(4) + Y = 24.625$

$\Sigma(AT/N_;) = \Sigma(1) + \Sigma(17) = 94.960$

$\Sigma(BT/N_1) = \Sigma(2) + \Sigma(18) = 218.040$

$\Sigma(CT/N_2) = \Sigma(4) + \Sigma(15) = 16.325$

$\Sigma(DT/N_2) = \Sigma(5) + \Sigma(16) = 296.675$

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

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

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

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

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

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

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

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 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.

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

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

If the combined distribution is taken as standard the estimator becomes

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

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: $R = H/G$; $R_1 = R_2 = R_3 = R_4 = AD/BC$.