# A METHOD OF ESTIMATING COMPARATIVE RATES FROM CLINICAL DATA. APPLICATIONS TO CANCER OF THE LUNG, BREAST, AND CERVIX[1]

JEROME CORNFIELD, *National Cancer Institute, National Institutes of Health, U. S. Public Health Service, Bethesda, Md.*

A frequent problem in epidemiological research is the attempt to determine whether the probability of having or incurring a stated disease, such as cancer of the lung, during a specified interval of time is related to the possession of a certain characteristic, such as smoking. In principle, such a question offers no difficulty. One selects representative groups of persons having and not having the characteristic and determines the percentage in each group who have or develop the disease during this time period. This yields a true rate. The difference in the magnitudes of the rates for those possessing and lacking the characteristic indicates the strength of the association. If it were true, for example, that a very large percentage of cigarette smokers eventually contracted lung cancer, this would suggest the possibility that tobacco is a strong carcinogen.

An investigation that involves selecting representative groups of those having and not having a characteristic is expensive and time consuming, however, and is rarely if ever used. Actual practice in the field is to take two groups presumed to be representative of persons who do and do not have the disease and determine the percentage in each group who have the characteristic. Thus rather than determine the percentage of smokers and nonsmokers who have cancer of the lung, one determines the percentage of persons with and without cancer of the lung who are smokers. This yields, not a true rate, but rather what is usually referred to as a relative frequency. Relative frequencies can be computed with comparative ease from hospital or other clinical records, and in consequence most investigations based on clinical records yield nothing but relative frequencies. The difference in the magnitudes of the relative frequencies does not indicate the strength of the association, however. Even if it were true that there were many more smokers among those with lung cancer than among those without it, this would not by itself suggest whether tobacco was a weak or a strong carcinogen. We are consequently interested in whether it is possible to deduce the rates from knowledge of the relative frequencies.

[1] Received for publication February 23, 1951.

## A GENERAL METHOD

To fix our ideas we may illustrate how the general problem can be attacked with some data recently published by Schrek, Baker, Ballard, and Dolgoff (*1*). They report that 77 percent of the white males studied, aged 40–49, with cancer of the lung, smoked 10 or more cigarettes per day, while only 58 percent of a group of white males, aged 40–49, presumed to be representative of the non-lung-cancer population, smoked that much. Can we estimate from these data the frequency with which cancer of the lung occurs among smokers and nonsmokers?

Denote by $p_1$ (=0.77) the proportion of smokers among those with cancer of the lung, by $p_2$ (=0.58) the proportion of smokers among those without cancer of the lung, and by $X$ the proportion of the general population that has cancer of the lung during a specified period of time. We may then summarize the relevant information for the general population in a two-by-two table showing the proportion of the population falling in each of the four possible categories.

| Characteristic | Having cancer of the lung | Not having cancer of the lung |
|---|---|---|
| Smokers | $p_1X$ | $p_2(1-X)$ |
| Nonsmokers | $(1-p_1)X$ | $(1-p_2)(1-X)$ |
| Total | $X$ | $1-X$ |

One can now compute that the percentage of the general population that smokes is $p_2+X(p_1-p_2)$, that the proportion of smokers having cancer of the lung is:

$$(1)\quad p_1X/[(p_2+X(p_1-p_2)].$$

Similarly, the proportion of nonsmokers having cancer of the lung is

$$(2)\quad (1-p_1)X/[(1-p_2)-X(p_1-p_2)].$$

Formulas (1) and (2) yield the true rates we seek.

Given the appropriate data, formulas (1) and (2) are easy to compute. They are somewhat cumbersome algebraically, however. The following approximation to the true rates, therefore, seems useful. If the proportion of the general population having cancer of the lung, $X$, is small relative to both the proportion of the control group smoking and not smoking, $p_2$ and $1-p_2$, the contribution of the term $X(p_1-p_2)$ to the denominator of formulas (1) and (2) is trivial and may be neglected. In that case the approximate rate of cancer of the lung among smokers becomes $\frac{p_1X}{p_2}$ and the corresponding rate for nonsmokers $\frac{(1-p_1)X}{1-p_2}$. Whenever $p_1-p_2$ is greater than zero, $p_1/p_2$ is greater than unity. We may conclude from the approximation, therefore, that whenever a greater proportion of the diseased than of the control group possess a characteristic, the incidence of the disease is always higher among those possessing the characteristic. This is the intuition on which the procedures used in such clinical studies

are based. Although it has frequently been questioned, it can now be easily seen to be correct.

It also follows from this analysis, however, that if one knows $X$, the prevalence of cancer of the lung in the general population, one can compute its prevalence among the smoking and nonsmoking population. Hospital or clinical records usually cannot furnish an estimate of $X$, however, since one seldom knows the size of the population exposed to risk from which the actual cases are drawn. Its value is frequently known, at least approximately, from other sources. Thus, we have estimated from Dorn's data (*2*) that the annual prevalence of cancer of the lung among all white males aged 40–49 is 15.5 per 100,000.[2] $X$ consequently is equal to $0.155\times10^{-3}$. We may now construct a table showing the proportion of the population in each of the four categories from the data of Schrek *et al.*

| | Having cancer of the lung | Not having cancer of the lung | Total |
|---|---|---|---|
| Smokers | $0.119\times10^{-3}$ | 0.579910 | 0.580029 |
| Nonsmokers | $.036\times10^{-3}$ | .419935 | .419971 |
| Total | $.155\times10^{-3}$ | .999845 | 1.000000 |

The proportion of smokers who have cancer of the lung using formulas (1) and (2) is thus $0.205\times10^{-3}$ as contrasted with $0.086\times10^{-3}$ for nonsmokers. The corresponding rates are 20.5 and 8.6 per 100,000 per year. These rates clearly provide a sounder basis for appraising the effect of cigarette smoking than does the knowledge that 77 percent of those with cancer of the lung and 58 percent without it smoke.

If one is interested only in knowing the relative amount by which the prevalence of the disease is augmented by the possession of the attribute, one may calculate this without knowledge of $X$, since the ratio of the two rates is $\frac{p_1(1-p_2)}{p_2(1-p_1)}$ when $X$ is small. One can thus conclude from the Schrek data alone that the prevalence of cancer of the lung among white males aged 40–49 is 2.4 times as high among those who smoke 10 or more cigarettes a day as among those who do not.

The more extensive, but age-standardized, data of Levin, Goldstein, and Gerhardt (*3*) on the same subject may be used to illustrate the same calculation. They show that 66.1 percent of all (presumably white) males at all age groups who had cancer of the lung smoked some cigarettes as compared with 44.1 percent smoking among the control group. Setting $.661=p_1$ and $.441=p_2$, we have $\frac{p_1(1-p_2)}{p_2(1-p_1)}=2.5$. The prevalence of lung cancer, according to these data is 2.5 times as high among cigarette

[2] Dorn's published data show an annual prevalence rate in the period 1937–1939 of 29.7 per 100,000 for cancer of all respiratory organs among white and colored males, aged 40–49. In the North 52.1 percent of the respiratory cases in all age groups for both males and females was accounted for by lung cancer. The estimate of 15.5 (=29.7×0.521), is consequently somewhat rough.

amokers as among nonsmokers. (The agreement with the Schrek data is closer than would be expected in view of differences in the population covered, definitions used, and number of cases studied. The application of the present method to other studies of lung cancer and tobacco yields much more divergent results.)

The calculations may also be applied to multiple classifications such as the data on cancer of the cervix in Cardiff, Wales, recently published by Maliphant (*4*). In table 1, the first column gives the percent distribution of women who develop cancer of the cervix by marital status and number of children borne, while the second column shows the same distribution for all women. Women under 40 have been excluded. From other data given by Maliphant we have estimated that the incidence rate of cervical cancer for women over 40 in Cardiff was 79.7 per 100,000 (somewhat below the corresponding rate in this country.) This yields $X$ and we accordingly have been able to calculate the incidence rates by marital status and number of children shown in the third column. The relation between cervical cancer and number of children born is obviously shown more clearly and usefully by the rates in the third column than by the relative frequencies in the first two.

TABLE 1.—*Distribution of women with and without cervical cancer by marital status and number of children*

| | Women contracting cancer of the cervix, $100\,p_1$ | All women, $100\,p_2$ | Incidence rate per 100,000, $\frac{p_1 X}{p_2}$ |
|---|---|---|---|
| Unmarried | 1.3 | 10.5 | 9.9 |
| Married: | | | |
| No children | 5.0 | 13.0 | 30.7 |
| 1 or more children, total | 93.7 | 76.5 | 97.6 |
| 1 child | 13.3 | 15.3 | 69.3 |
| 2 children | 18.3 | 17.0 | 85.8 |
| 3 children | 15.0 | 13.0 | 92.0 |
| 4 children | 11.0 | 9.6 | 91.3 |
| 5 children | 9.2 | 6.4 | 114.6 |
| 6 or more | 26.9 | 15.2 | 141.6 |
| Total | 100.0 | 100.0 | |

## TESTS OF SIGNIFICANCE ON THE COMPUTED RATE

Since most clinical studies are based on limited numbers of cases, it is of some importance to be able to estimate the limits of error of rates calculated according to this procedure. The approximate formula for the variance of a ratio sometimes used is inappropriate for this purpose, since it will sometimes show $\frac{p_2 X}{p_1}$ differing significantly from $X$ when a test on the difference $p_2$-$p_1$ shows that it does not differ significantly from zero. To avoid this we employ a test of Fieller's (*5*). Thus, writing the computed prevalence rate as $\frac{p_1 X}{p_2} = r$ and denoting by

$n_1$ = the number of disease cases

$n_2$ = the number of control cases
$t_\alpha$ = the value of $t$ in the normal curve corresponding to the $100\alpha$-percent probability level
$pq$ = the unbiased estimate of the unknown population value $PQ$

$$\left(=\frac{n_1p_1+n_2p_2}{n_1+n_2-1}\left[1-\frac{n_1p_1+n_2p_2}{n_1+n_2}\right]\right)$$

the upper and lower confidence limits for the $100\alpha$-percent probability level of the estimate $r$ are given by

$$\frac{r\pm\frac{t_\alpha X}{p_2}\sqrt{\frac{pq}{n_1}\left[1+\frac{1}{n_2p_2^2}\left(n_1p_1^2-t_\alpha^2pq\right)\right]^{\frac{1}{2}}}}{1-\frac{t_\alpha^2pq}{n_2p_2^2}},$$

when $X$ is considered free from sampling error. We may use the Schrek data to illustrate the use of this formula. Thus, letting $n_1=35$, $n_2=171$, setting $t_\alpha=2$, and using $p$, $X$, and $r$ as previously calculated, we compute the upper limit to the rate as 25.6 per 100,000 and the lower limit as 16.1 per 100,000. Since the value of $X$ used, 15.5 per 100,000, falls outside these limits, we conclude that the rates for smokers and nonsmokers differ significantly at the 5-percent probability level. Whenever $p_1$ and $p_2$ differ significantly at the $100\alpha$-percent level, the limits computed in this fashion will not include $X$, and *vice versa*. Thus, if one simply wishes to test significance, it is sufficient to test the difference between $p_1$ and $p_2$. If one wishes to express error limits in the same units that the prevalence rate is expressed, however, one must use the formula given.[3]

## PITFALLS

Our major purpose in preparing this note has been to show that any set of data that furnishes estimates of relative frequencies can be used to obtain estimates of rates. The procedure suggested, however, has assumed that the diseased and control groups used are representative of these same groups in the general population. If this assumption is not satisfied, then neither the rates, the relative frequencies, nor any other statistics calculated from the data will have applicability beyond the particular group studied.

We may illustrate the difficulties that can arise on this score with 2 examples. The first relates to Lane-Claypon's study of cancer of the breast (*6*). In this study a detailed questionnaire was filled in for 508

[3] The procedure discussed in the text yields a two-sided test of significance; *i. e.*, it tests the hypothesis that the rate for smokers is significantly different from that for nonsmokers. It would be more realistic to use a one-sided test; *i. e.*, test the hypothesis that the rate for smokers is significantly *higher* than that for nonsmokers. To do this one uses the same formula but calculates only a lower limit, using a value of $t_\alpha$ appropriate to the one-sided test. Thus, for $\alpha=0.05$, $t_\alpha=1.645$.

In testing whether $p_1$ and $p_2$ are drawn from the same population it is appropriate to compute a pooled variance as has been done. When the results of such a test of significance suggest that $p_1$ and $p_2$ could not have been drawn from the same population, however, the use of a pooled variance to compute error limits is no longer correct. In fact, exact confidence limits can no longer be calculated for this case. The results yielded by the formula will nevertheless be sufficiently accurate for most practical purposes.

women with breast cancer and 509 control women, who were being treated by the cooperating hospitals for "some trouble, other than cancer." We reproduce in table 2 the percent distribution by number of children ever borne for each group. Only women having passed the menopause are included. We do not know $X$, the prevalence rate of breast cancer in the United Kingdom at the time the data were collected, and have therefore confined ourselves to computing relative prevalence.

TABLE 2.—*Distribution of women with and without breast cancer by marital status and number of children*

| Characteristic | Cancer group, $p_1$ | Control group, $p_2$ | $\frac{p_1}{p_2}$ | Relative prevalence |
|---|---|---|---|---|
| Unmarried | 20. 91 | 16. 42 | 1. 273 | 100 |
| Married: | | | | |
| No children | 14. 55 | 10. 45 | 1. 392 | 109 |
| 1 to 3 children | 29. 09 | 24. 78 | 1. 174 | 92 |
| 4 to 6 children | 21. 21 | 22. 39 | . 947 | 74 |
| 7 or more | 14. 24 | 25. 97 | . 548 | 43 |
| Total | 100. 00 | 100. 00 | | |

If the data are to be taken at their face value, one must conclude that lowered prevalence of breast cancer is associated with increasing numbers of children. Greenwood in an analysis of Lane-Claypon's data (*6*) in fact concludes, "we think then that an etiological factor of importance has now been fully demonstrated." At the very beginning of his analysis, however, he points out, without attaching any significance to it, that the control group had borne an average of about 25 percent more children than had all women in England and Wales with the same duration of marriage. This would appear to provide definite evidence for the unrepresentative character of the control group and to cast doubt on the adequacy of the evidence.

The basic difficulty in this example is the unrepresentative nature of the control group. Since there is always some doubt whether or not a control group selected from among hospital patients can provide an accurate estimate of the frequency of a characteristic in the population at large, the difficulty may be quite general. The possibility that the diseased group is not representative either, cannot be entirely disregarded, however. We reproduce in table 3 the distribution by age of 413 patients with adenocarcinoma of the breast admitted to the Ellis Fischel State Cancer Hospital in the years 1940–46 as given by Ackerman and Regato (*7*). For comparison we give the expected distribution on the basis of known incidence rates by age.

It is obvious from inspection that an excess number in the older age groups were encountered, and that to some extent the hospital was functioning as a home for the aged. An epidemiological investigation the results of which would be sensitive to the age distribution of the persons studied might consequently be adversely affected.

Any set of hospital or clinical data that is worth analyzing at all is

TABLE 3.—*Actual and expected distribution of breast cancer cases by age—Ellis Fischel State Cancer Hospital*

| Age | Number of breast cancer cases [1] | | (3) (4) × (5) | Incidence of breast cancer per 100,000 [2] (4) | Percent distribution Missouri population 1940 [3] (5) |
|---|---|---|---|---|---|
| | Reported (1) | Expected (2) (3) $\frac{\text{Tot. (1)}}{\text{Tot. (3)}}$ | | | |
| Less than 30 | 3 | 7 | 1.1 | 2.2 | 0.4825 |
| 30–34 | 10 | 13 | 1.9 | 25.1 | .0795 |
| 35–39 | 24 | 25 | 3.7 | 50.7 | .0774 |
| 40–44 | 26 | 41 | 6.1 | 91.1 | .0734 |
| 45–49 | 40 | 54 | 8.0 | 122.9 | .0674 |
| 50–54 | 51 | 51 | 7.5 | 129.0 | .0648 |
| 55–59 | 54 | 57 | 8.4 | 169.6 | .0580 |
| 60–64 | 54 | 53 | 7.8 | 190.4 | .0495 |
| 65–69 | 59 | 45 | 6.6 | 193.3 | .0410 |
| 70–74 | 48 | 34 | 5.1 | 205.5 | .0344 |
| 75 and over | 44 | 33 | 4.9 | 184.5 | .0516 |
| Total | 413 | 413 | 61.1 | ---------- | 1.0000 |

[1] Chi square for difference=24.5, $P<0.01$.
[2] As estimated by Dorn (*2*).
[3] U. S. Bureau of the Census, Population, vol. II, pt. 4, table 7.

worth analyzing properly. It is from this point of view that the technique proposed seems useful. The preceding two examples suggest, however, that the results of even the most carefully analyzed set of such data may be open to question, and that these doubts can be resolved only by methods of data collection that provide representative samples of diseased and nondiseased persons.

## REFERENCES

(*1*) SCHREK, R., BAKER, L. A., BALLARD, G. P., and DOLGOFF, S.: Tobacco smoking as an etiologic factor in disease. I. Cancer. Cancer Research 10: 49–58, 1950.
(*2*) DORN, H. F.: Illness from cancer in the United States. Pub. Health Rep. 59, Nos. 2, 3 and 4, 1944.
(*3*) LEVIN, M. L., GOLDSTEIN, H., and GERHARDT, P. R.: Cancer and tobacco smoking. J. A. M. A. 143: 336–338, 1950.
(*4*) MALIPHANT, R. G.: The incidence of cancer of the uterine cervix. Brit. M. J., I: 978–982, 1949.
(*5*) FIELLER, E. C.: The biological standardization of insulin. Supp. J. Roy. Stat. Soc. 7: 1–64, 1940.
(*6*) LANE-CLAYPON, J. E.: A further report on cancer of the breast. Reports on Public Health and Medical Subjects, No. 32, Ministry of Health, London, 1926.
(*7*) ACKERMAN, L. V., and del REGATO, J. A.: Cancer, Diagnosis, Treatment and Prognosis. C. V. Mosby Co., St. Louis, 1947, p. 927.

947977—51——23