



---

Dynamic Risk Analysis in Retrospective Matched Pair Studies of Disease

Author(s): Paul R. Sheehe

Source: *Biometrics*, Vol. 18, No. 3 (Sep., 1962), pp. 323-341

Published by: International Biometric Society

Stable URL: <http://www.jstor.org/stable/2527475>

Accessed: 16/03/2010 11:06

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=ibs>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).



*International Biometric Society* is collaborating with JSTOR to digitize, preserve and extend access to *Biometrics*.

<http://www.jstor.org>

# DYNAMIC RISK ANALYSIS IN RETROSPECTIVE MATCHED PAIR STUDIES OF DISEASE

PAUL R. SHEEHE

*Roswell Park Memorial Institute,  
Buffalo, New York, U. S. A.*

## SUMMARY

It has been shown how a general population model of an exponentially changing risk of disease may be tested against retrospective matched pair data. In the illustrative analysis of breast cancer data it was found that the risk of breast cancer increased exponentially at a greater rate during menstrual years than during other phases of life.

Analysis of proportional risk functions and linear risk functions, as well as the extension of exponential risk functions to more than one variable, were discussed.

## INTRODUCTION

Retrospective studies of disease are so called because the investigator observes an effect, that being a number of individuals with and without the disease, and "looks back" into the histories of the diseased cases and non-diseased controls for an explanation of the effect. If the relative frequency of some characteristic is sufficiently greater among the cases than among the controls, then the characteristic may be an indicator of disease.

The principal reason for looking backward at the problem is evident when the alternative prospective approach is considered. In a prospective, or forward looking, study a chosen number of individuals with the characteristic is compared with a chosen number without the characteristic to see if the relative frequency of disease in the two classes is different. But if disease is infrequent, as is generally true, either very large samples or a very long follow-up period, or both, would be required in order to obtain reliable frequencies of disease. The retrospective study short circuits this problem, but in accomplishing this it presents problems of its own. There is of course the very practical problem of obtaining histories of individuals—memories are faulty and records may be missing or inaccurate. But assuming the historical data to be adequate, there is the added problem of interpretation.

TABLE 1  
POPULATION DISTRIBUTION

	With Characteristic	Without Characteristic	Total
With disease	$N$	$N'$	$N + N'$
Without disease	$M$	$M'$	$M + M'$
Total	$N + M$	$N' + M'$	$N + N' + M + M'$

While one may “look back” by practical necessity, one prefers to “think forward” for theoretical purposes. It is preferable to ask whether the risk of disease is greater for individuals with the characteristic, to reason from causes or antecedent conditions to effects, rather than the reverse.

Cornfield [1951] showed that the retrospective study is amenable to a prospective interpretation when he considered a situation similar to that schematized in Table 1. Table 1 shows a population consisting of four classes of individuals. There are  $N$  individuals who have both a certain characteristic and the disease,  $M$  with the characteristic but not the disease. Among those without the characteristic there are  $N'$  and  $M'$  respectively with and without disease. Among those with the characteristic, the relative frequency, or prevalence, of disease is

$$P_1 = N/(N + M).$$

This is assumed to be very low. Similarly, for those without the characteristic, the prevalence is

$$P_2 = N'/(N' + M'),$$

also assumed to be very low. Thus, the prevalence of disease among those with the characteristic, relative to those without, is

$$R = \frac{N}{N + M} \bigg/ \frac{N'}{N' + M'} = \frac{N(N' + M')}{N'(N + M)}$$

There is a second expression which closely approximates the relative prevalence. This expression is the relative odds. Among those with the characteristic the odds for disease are  $N/M$ . Among those without, the odds are  $N'/M'$ . The relative odds are therefore

$$(N/M)/(N'/M') = NM'/N'M.$$

Since the disease is rare,  $N$  and  $N'$  are relatively small, so that

$$(N' + M')/(N + M) \simeq M'/M.$$

Hence, the expression for relative prevalence reduces approximately to the relative odds,

$$R \simeq \frac{NM'}{N'M}$$

Under the retrospective method of study, a relatively large fraction,  $f$ , of diseased cases are sampled, and a much smaller fraction,  $g$ , of non-diseased controls are sampled. Ignoring any sampling error for the moment, the number of sampled cases with and without the characteristic is

$$n = fN \quad \text{and} \quad n' = fN',$$

and

$$m = gM \quad \text{and} \quad m' = gM',$$

respectively. Then the relative odds computed from the retrospective data are

$$\frac{nm'}{n'm} = \frac{NM'}{N'M} \simeq R.$$

That is to say, the retrospective method has not disturbed the relative odds, and the relative odds, in turn, closely approximate the relative prevalence in the population.

Relative incidence can be dealt with in a retrospective study in similar fashion by defining the population differently. Instead of a population of individuals with and without disease, a population wholly free of the disease is initially defined. As before, this population is divided into those individuals with and without a certain characteristic. In either of these two classes of the population, incidence is defined as the proportion of individuals who come down with the disease within a specified interval of time. The situation is again represented by Table 1, except that now the population is considered to be initially free of the disease, and now the class "With disease" should be taken to mean "With disease at some time during the specified interval", also, "Without disease", should be taken to mean "Without disease throughout the specified interval". The subsequent manipulations of cell frequencies apply equally well to this situation. Ignoring sampling error, as before, and assuming that the incidence of disease is low, the relative odds this time approximate relative incidence instead of relative prevalence.

The problems of sampling error and the combination of relative odds have been dealt with recently by Woolf [1954] and by Haldane

[1955]. In addition, an interesting method of estimation in retrospective studies of matched cases and controls has been demonstrated by Kraus [1960]. While these methods all pertain to dichotomous characteristics, this paper will deal with certain risk functions of *measured* variables which can be tested in retrospective matched pair studies. The method to be developed will be illustrated with data from a study of breast cancer cases and matched controls.

### 1. RISK FUNCTIONS OF AGE

As shown in the introduction, either a relative prevalence or a relative incidence interpretation can be obtained from retrospective data when populations appropriate to each interpretation are defined and sampled. This paper will deal more directly with relative incidence than with relative prevalence. However, it will deal with incidence in a multiplicity of classes, rather than just two classes. These classes will be defined according to age, among other variables.

In order to obtain some notion of the nature of these age-specific classes, consider some large population of females. Each of the individuals will live free of breast cancer, for example, for a certain number of years. Therefore, we can visualize classes of all the individuals alive and free of past or current history of breast cancer at age 0, 1 year, 2 years, and so on. (Note that these classes are not mutually exclusive with respect to individuals. The same individual may appear in many classes.) The one-year incidence of breast cancer among 0 year olds is, by definition, the number of individuals coming down with breast cancer in the one year interval following birth, divided by the number of 0 year olds. Similarly, the 1 year incidence is defined for every age class. An individual who survives one year without breast cancer enters (as in a life-table) a new age class at the end of the year. Thus, we can visualize the passage of an individual from one age class to the next throughout her lifetime until either breast cancer or death occurs. But age classes need not be limited to annual increments. Half year incidences could be considered for classes specified at every half year of age. Indeed, the age interval, and correspondingly the interval over which incidence is observed, may be made very small. For theoretical purposes, we postulate, finally, an instantaneous incidence, which we call risk, for every instant of age. Thus, at any defined age instant in the breast cancer-free lifetime of an individual, that individual belongs to a class with a certain postulated risk (instantaneous incidence) of breast cancer. This postulated risk is introduced here as a convenience in overcoming the difficulty of expressing a large number of hypothesized short-interval incidences.

As will be done later in the study of breast cancer, it may be postulated that risk is some integrable function of age. This postulated risk function, when weighted by the number of individuals at each instant of age can be summated over a specified interval to yield an hypothesized number of new cases of breast cancer in that interval\*. When this number is divided by the size of the class at the beginning of the age interval, an hypothesized incidence is obtained. Thus, there is a direct connection between risk functions and hypothesized age-specific incidence in the population. This connection establishes the meaning of risk functions of age. In the following section, a risk function will be constructed as a model of the development of breast cancer in the lifetimes of individuals in the general female population. Subsequently, it will be shown how such a risk function can be brought to bear on retrospective data. Since risk functions of age have an immediate prospective connotation in terms of age-specific incidences in the population, the establishment of a connection between risk functions and retrospective data will provide a prospective interpretation of certain patterns of change observed in retrospective studies of disease.

## 2. CONSTRUCTION OF AN HYPOTHETICAL RISK FUNCTION FOR BREAST CANCER

Now, with some misgivings, we shall propose a model for the risk of breast cancer as it develops in the general female population. This model is something more than a shot in the dark because it does not seem to conflict in any obvious way with previous epidemiological findings. Yet we are almost certain that this model will sooner or later be found to err in one or many respects. Why, then, attempt any model at all? There are, we think, some good reasons. First by setting up an explicit mathematical model, its shortcomings become quite obvious to readers familiar with the complexity of the problem. This, we hope, minimizes any tendency for this study to be taken as "final" in any sense. But, second, setting up a model which is something more sophisticated than an ill-considered null hypothesis helps to provide a structure for subsequent study. And third, it is a fact that this model furnished the stimulus to develop an analytical procedure which may well have applications in the epidemiological study of many

---

\*Let  $r_t$  be an integrable risk function of age,  $t$ . To find the hypothesized number of cases in a given time interval, denote the small component intervals between successive occurrences of death or breast cancer by  $k = 0, 1, \dots, D$ , such that  $\sum_{k=0}^D \Delta t_k$  equals the duration of the defined interval. Let  $l_k$  be the class size during each component interval. Then

$$S = \sum_{k=0}^D l_k \int r_t dt)_k$$

is the hypothesized number of new cases occurring during the interval.

other diseases. We therefore ask the reader to make allowances for our naivete in attempting a bald mathematical model in the face of so complex a problem.

With what is known now of the epidemiology of breast cancer, many hypotheses might be formed, but the most promising line seems to be some hypothesis connecting breast cancer with ovarian function. Lilienfeld [1955, 1956] found that the age-specific incidence of breast cancer rose exponentially with age for both single and married women up to approximately age 40 or 45, corresponding roughly with age at menopause. From that age on, the incidence continued to rise, but at a definitely lower rate of increase. Moreover, the break in the rise was somewhat postponed for single women. Further study uncovered the fact that the average age at menopause for single women was later than for married women. This was not because of later natural menopause for single women but rather because artificial menopause was much more frequent among married women.

In a combined review and international study of the epidemiology of breast cancer, Wynder et al., [1960] discussed these and other consistent factors which have been observed in studies of the disease, such as: the more frequent occurrence of breast cancer among single than among married women; the later age of marriage of breast cancer patients; the reduced incidence of breast cancer brought about by pregnancy, or events subsequent to pregnancy, such as nursing; the reduced frequency of breast cancer among castrated women; and an increased incidence presumably due to prolonged ovarian activity, i.e. late menopause. In addition to these factors related to marriage and hormonal functioning, cognizance was taken of "background" variables such as heredity, familiarity, race, religion, and socio-economic status, each of which may play its own etiological role. But the basic conception which the authors developed was that "any factor that can reduce endocrine function (through its effect on menstruation) tends to reduce the risk of developing breast cancer".

In relation to endocrine function, five phases of a woman's life can be identified: (1) pre-menarchal; (2) menstrual; (3) pregnancy; (4) lactation; (5) post-menopausal. In general, the conception expressed by the authors reviewed above could be interpreted to mean that a long menstrual history would be indicative of high risk of breast cancer, while long histories of the other four phases lead to a relatively lower risk of breast cancer. In view of Lilienfeld's results, it would seem plausible to relate the risk of breast cancer to an exponential function of years spent in each phase of life. Thus, it might be hypothesized that risk of breast cancer increases exponentially as men-

strual years increase, while the increase is not so rapid or the risk even diminishes exponentially during the other phases. The rate of increase or decline in risk might be different for each of the five phases, and indeed, if we were to specify phases in more detail, the risk function might be exceedingly complex. But in the interest of simplicity and feasibility of testing the hypothesis, it seems desirable to confine our hypothesis within the limits of these five phases.

The preceding paragraph contains the most general form of the breast cancer model which we shall adopt. That model is, simply, that the risk of breast cancer at any moment, under certain conditions which will be specified, is an exponential function of the number of pre-menarchal, menstrual, pregnancy, lactation, and post-menopausal years lived by a woman up to that moment. When we refer to a "woman's risk" at a given moment, we mean that at a given moment the woman belongs to the general class of all women free of breast cancer after a certain number of years spent in each type of phase, and we refer to the risk for that class. If the woman is in the menstruating phase, then "her risk", so to speak, changes with each passage of time to that of the class of women with a longer history of menstruation. If the woman is in another phase, then "her risk" changes with each passage of time to that of the class of women with a longer history in that phase.

According to the model, risk may be increasing or decreasing at a constant exponential rate during any given phase of life. The rate of change may be different in each of the five phases of life. But both for practical and expository purposes, we now make a simplifying assumption. We shall assume that the rate of change is the same in pre-menarchal, lactating and post-menopausal phases. This is not to say that we prefer this over letting the pre-menarchal and lactating years "ride free". There seems to be some evidence in the literature that lactation furnishes some added retardation of the risk of breast cancer, and little is known about how risk may change during the pre-menarchal years. But in the United States at least, there is little variation in the age of menarche, this being usually somewhere between 12 and 14 years, and there is even less variation in lactating years, since relatively few women have nursed more than 10 months during their lives. Therefore, the error, if any, of assuming the same rate of change in risk during these phases as during the post-menopausal phase will be of about the same relative magnitude for most women. Subsequently, when we come to compare the hypothesized risk of a breast cancer patient (i.e., the risk of breast cancer in the class of women to which the breast cancer patient belonged just prior to diagnosis)

with an age-matched control, the two errors will cancel out in the ratio of the two risks, if age of menarche and history of lactation are the same. If the menarche and lactation histories differ, they will usually differ only slightly, as indicated above, and unless the rates of change in risk during these two phases are radically different from that in post-menopausal years, the *relative* risk of case versus control will be affected only slightly. The simplifying assumption reduces the number of identified phases of life to three: (1) menstrual, (2) pregnancy, (3) other. For post-menopausal women, the principal source of variation in non-menstrual and non-pregnant years is, of course, the number of post-menopausal years.

In addition to phase of life, the growth or decay of risk may depend on certain "background" conditions for a defined class of women. Among the "background" conditions are such factors as age, race, country of birth, religion and marital status. No specific hypothesis is made here as to how risk of breast cancer may be affected by these background variables. It is considered that risk of breast cancer may or may not be affected by these factors. It is only hypothesized that risk of breast cancer is an exponential function of phase history for classes of women within specified age, racial, country of birth, religious and marital status groups.

In the illustrative analysis which will be presented, all the background variables mentioned above will be employed. In addition, to simplify the analysis, years of pregnancy will also be relegated to the background, so that we shall be directly concerned only with testing the exponential risk hypothesis with respect to variations in menstruating years and "other" years (excluding years of pregnancy). The exponential risk hypothesis is expressed precisely as follows:

$$r_{ix'x''} = C_i a_1^{x'} a_2^{x''}, \quad i = 1, 2, \dots, s,$$

where

- $x'$  denotes number of menstrual years,
- $x''$  denotes number of other years (excluding pregnancy),
- $a_1$  and  $a_2$  are constants relating the number of menstrual and "other" years, respectively, to risk,
- $i$  denotes a class of women with a specified age, race, country of birth, religion, marital status, and pregnancy history,
- $C_i$  is an unknown  $i$ th class constant,

and

$r_{ix'x''}$  is the risk of breast cancer in the  $i$ th class of women in the sub-class denoted by variables  $x'$  and  $x''$ .

Thus, the hypothesized risk for the sub-class of women denoted by  $(ix'_1x''_1)$ , relative to the risk for another sub-class,  $(ix'_2x''_2)$ , is given by

$$R_{i12} = r_{ix'_1x''_1} / r_{ix'_2x''_2} = a_1^{x'_1 - x'_2} a_2^{x''_1 - x''_2}.$$

One final simplification results from the fact that in any relative risk comparison of two sub-classes in the  $i$ th class of women, age and years of pregnancy are constant. Hence the sum of menstrual and 'other' (excluding pregnancy years) is constant:

$$x'_1 + x''_1 = x'_2 + x''_2,$$

so that

$$(x'_1 - x'_2) = -(x''_1 - x''_2).$$

Consequently, the expression of the relative risk for two sub-classes reduces to

$$\begin{aligned} R_{i12} &= a_1^{x'_1 - x'_2} \cdot a_2^{-(x''_1 - x''_2)} = (a_1/a_2)^{x'_1 - x'_2} \\ &= b^{x'_1 - x'_2}. \end{aligned}$$

Since the relative risk for the two sub-classes of  $i$  has been reduced to an expression involving only menstrual years, we shall drop the prime notation. Through the remainder of this paper, the quantity  $x$  means menstrual years. Thus,

$$R_{i12} = b^{x_1 - x_2},$$

is the hypothetical relative risk for two sub-classes of  $i$ , where  $b$  is some constant to be determined. The latter expression can be restated in the following form,

$$R_{i12} = e^{(x_1 - x_2) \ln b},$$

or in a form which will be used later on,

$$\ln R_{i12} = (x_1 - x_2) \ln b.$$

Referring back to Lilienfeld's studies, if  $\ln b$  were positive, this would be in general agreement with the greater slope in the plot of log incidence versus age during the generally pre-menopausal years than during the generally post-menopausal years.

Under the exponential hypothesis, or for that matter any other risk hypothesis expressed as a function of  $x$ , we are obviously dealing with a measured variable. Moreover, the chosen risk function is a dynamic hypothesis in the sense that hypothesized risk changes with age. The problem which lies before us is to see how such an exponential risk function can be brought to bear on retrospective matched pair

data. This problem is taken up in the next section, following which an illustrative analysis of the breast cancer hypothesis will be presented.

#### 4. THE CONNECTION BETWEEN RISK FUNCTIONS AND RETROSPECTIVE MATCHED CASE-CONTROL PAIRS

Consider all non-diseased individuals at all points of time in the  $i$ th class. Within this class it is hypothesized that risk of disease is some function of  $x$ ,

$$r_{ix} = f_i(x).$$

Thus, in the sub-class designated by  $ix_1$ ,

$$r_{ix_1} = f_i(x_1),$$

and in  $ix_2$ ,

$$r_{ix_2} = f_i(x_2).$$

Consequently, the relative risk for sub-class  $ix_1$  versus  $ix_2$  is

$$R_{i12} = f_i(x_1)/f_i(x_2),$$

For example, under the exponential hypothesis for breast cancer,

$$R_{i12} = b^{x_1 - x_2}.$$

Now consider in detail the process of obtaining cases and matched controls. Let the number of individuals in the  $ix$  sub-class be denoted  $H_{ix}$  and the number of individuals in the  $i$ th class by  $K_i$ .

Also, let

$$H_{ix} = K_i L_{ix},$$

where  $L_{ix}$  is the relative distribution of  $x$  for individuals in the  $i$ th class.

Now, *as part of the hypothesis*, assume that the *conditional probability* that the next case which comes from class  $i$  will be from the  $ix$  sub-class is proportional to the number of individuals times the risk of disease in that sub-class:

$$P_{ix} = cH_{ix}r_{ix},$$

where  $c$  is some proportionality constant. Assume also that the match-control for the next case is obtained at random from the same class ( $i$ ) as the case. Thus, the conditional probability (given  $i$ ) of a control from an  $ix$  sub-class is

$$P_{x:i} = L_{ix}.$$

Then the probability of an  $x_1$  case and  $x_2$  control given that they are both in the  $i$ th class is

$$P_{i12} = cH_{ix_1}r_{ix_1}L_{ix_2} .$$

Similarly, the probability of an  $x_2$  case and  $x_1$  control (inverse pair) is given by,

$$P_{i21} = cH_{ix_2}r_{ix_2}L_{ix_1} .$$

The relative probability (odds) of an  $x_1x_2$  pair versus the inverse pair  $x_2x_1$  in the  $i$ th class is therefore

$$\begin{aligned} P_{i12}/P_{i21} &= cH_{ix_1}r_{ix_1}L_{ix_2}/cH_{ix_2}r_{ix_2}L_{ix_1} \\ &= cK_iL_{ix_1}L_{ix_2}r_{ix_1}/cK_iL_{ix_2}L_{ix_1}r_{ix_2} \\ &= r_{ix_1}/r_{ix_2} \\ &= R_{i12} . \end{aligned}$$

That is to say, the relative probability (odds) of a given pair ( $ix_1$ ,  $ix_2$ ) versus its inverse is equal to the relative risk for individuals in ( $ix_1$ ) versus individuals in ( $ix_2$ ). Finally, given that one individual in a matched pair is  $x_1$  and the other is  $x_2$ , and that they both come from class  $i$ , the conditional probability of an  $x_1$  case and  $x_2$  control is given by

$$R_{i12}/(R_{i12} + 1),$$

while the complementary probability of the inverted pair is given by

$$1/(R_{i12} + 1)$$

or, what is equivalent, by

$$R_{i21}/(R_{i21} + 1).$$

This connection between an hypothesized risk function of  $x$  and the retrospective matched pair method of collecting data is the key to the analysis of matched pair studies presented in the next section.

#### 5. ANALYSIS OF THE EXPONENTIAL RISK HYPOTHESIS

Suppose that  $h$  matched pairs,  $j = 1, 2, \dots, h$ , have been obtained in the manner already described. Denote by  $x_{j1}$  the higher measurement of  $x$  in the  $j$ th pair and by  $x_{j2}$  the lower measurement of  $x$ . The odds for the case having the higher measurement are given by  $R_{j12}$ , and the odds for an inverse pair (case lower than control) are given by  $R_{j21}$ , as discussed in the previous section. That is, the conditional probability of an  $x_1$  case and  $x_2$  control (obverse pair) is  $R_{j12}/(R_{j12} + 1)$ , while the conditional probability of an inverse pair is the complement,  $1/(R_{j12} + 1)$ . For example under the exponential risk hypothesis in

the study of breast cancer, the relative risk of breast cancer in subclasses  $jx_1$  versus  $jx_2$  is

$$R_{j12} = b^{x_1 - x_2}.$$

Thus the odds for an obverse pair are

$$R_{j12} = b^{x_{j1} - x_{j2}}, \quad \text{where } x_{j1} > x_{j2}.$$

Now the  $h$  pairs can be listed in rank order of the hypothesized odds for an obverse pair, or what is equivalent, in rank order of the *absolute* difference between  $x_{j1}$  and  $x_{j2}$ . Categories containing 15 or more neighboring pairs can be formed and the frequencies of obverse and inverse pairs in these categories can be counted.

Now we may obtain an estimate of the natural log relative risk in each of the categories. Haldane has shown [1955] that for a binomially distributed variable the natural logarithm of the ratio of  $P$  to  $Q$  can be estimated with negligible bias from a sample of a fixed size by adding  $\frac{1}{2}$  to the observed frequencies and obtaining the natural log of the ratio of the two numbers. We may apply Haldane's results to the particular situation by noting that the category size is fixed, and that the number of obverse or inverse pairs is binomially distributed as a consequence of the model we have described. Thus, letting  $n_k$  equal the observed number of obverse pairs in the  $k$ th category and  $m_k$  equal the observed number of inverse pairs,

$$y_k = \ln [(n_k + \frac{1}{2}) / (m_k + \frac{1}{2})]$$

is an approximately unbiased estimate of the natural log relative risk in the  $k$ th category. The variance of this statistic, worked out to a close approximation by Haldane in the same article, is given by

$$s_{y_k}^2 = \frac{1}{n_k + 1} + \frac{1}{m_k + 1}.$$

The above expressions will be used in our illustrative analysis of breast cancer cases and controls. (See columns 5, 6 and 7, Table 1.)

But by hypothesis, we also have the relative risk,  $R_{j12}$ , which equals the odds for an obverse pair, for every  $j$ . The hypothesized natural log relative risk is, therefore,

$$\ln R_{j12} = (x_{j1} - x_{j2}) \ln b = d_j \ln b.$$

(Note that  $d_j$  must be positive since  $x_{j1} > x_{j2}$  by definition.) Within a given category,  $k$ , the average of hypothesized log relative risks is thus

$$(\overline{\ln R})_k = \bar{d}_k \ln b.$$

This average is slightly lower than the logarithm of the expected relative risk. This is because variation exists among the hypothesized log relative risks within the  $k$ th category. However, because the category contains neighboring hypothetical values, this variation is generally slight, and in any event the variation of hypothetical log relative risks within categories is minute in comparison to the random variation of observations. Thus, by hypothesis, we have that  $\bar{d}_k \ln b$  is approximately equal to the natural log relative risk in the conditional domain defined by the observed values of  $x$  in  $k$ .

The model for the general population has been reduced to a linear model applicable to retrospective data. Consequently, we are in a position to analyze the variance of the observed log relative risks, obtain a least squares estimate of  $\ln b$ , test  $\ln b$  for significance, and use the residual variation of observed log relative risk in a "goodness of fit" test of the model. This is illustrated in the next section.

## 6. ILLUSTRATIVE ANALYSIS

The data for this study come from hospital admissions to Roswell Park Memorial Institute from April 17, 1955 to April 17, 1957. The cases consist of all female breast cancer patients admitted during that interval. From the remaining females admitted without breast cancer or cancer of the genitals, a matching control for each case was selected at random from the class of admissions corresponding to the case with respect to age (5 year age groups), race (white, non-white), nativity (native, foreign-born), religion (Protestant, Catholic, Jewish), marital status (married, single) and parity (number of live births). The reader will note that there is only an approximate conformity between practical matching and theoretical concepts, particularly in connection with age and parity matching. Also the reader will appreciate the fact that a new opportunity for the risk hypothesis to go wrong has been introduced by the choice of hospital patients rather than a complete or random sample of the population. (Ideally, we mean the whole human population.) Let it suffice here to say that these are the practical problems which make the results of any single study inconclusive and which warrant attention in further studies. (See discussion.)

Table 2 shows the data for 331 case-control pairs divided into 22 categories, each (except the last) of size 15. A few pairs for which the difference between case and control patients was zero have been dropped. This was done because it is logically true that the relative risk in identical sub-classes must be 1, and therefore no data are required to establish this truth.

Column (2) shows the categorical limits of the *absolute* difference

TABLE 2  
 PAIRED BREAST CANCER DATA ACCORDING TO RANKED CATEGORIES OF ABSOLUTE CASE-CONTROL DIFFERENCES IN MENSTRUAL YEARS

(1) Category (k)	(2) Categorical Limits of Absolute Case-Control Differences (d <sub>j</sub> ) (menstrual years)	(3) Average Absolute Case-Control Differences (d <sub>k</sub> )	(4) Number of Pairs		(5) Observed Relative Risk (n <sub>k</sub> + 1/2)/(m <sub>k</sub> + 1/2)	(6) Observed Natural Log Relative Risk (y <sub>k</sub> )	(7) Variance of $s_k^2 = \frac{1}{n_k + 1} + \frac{1}{m_k + 1}$
			Obverse (Case > Control) (n <sub>k</sub> )	Inverse (Case < Control) (m <sub>k</sub> )			
1	0.1 to 0.7	0.36	9	6	1.46	.378	.243
2	0.7 to 1.0	0.91	8	7	1.13	.122	.236
3	1.0 to 1.1	1.01	7	8	0.88	-.128	.236
4	1.1 to 1.7	1.35	7	8	0.88	-.128	.236
5	1.7 to 2.0	1.83	6	9	0.68	-.386	.243
6	2.0 to 2.3	2.09	8	7	1.13	.122	.236
7	2.3 to 2.9	2.59	7	8	0.88	-.128	.236
8	2.9 to 3.0	2.99	4	11	0.39	-.942	.283
9	3.0 to 3.7	3.33	7	8	0.88	-.128	.236
10	3.7 to 4.1	3.97	7	8	0.88	-.128	.236
11	4.1 to 4.7	4.42	8	7	1.13	.122	.236
12	4.7 to 5.2	4.96	9	6	1.46	.378	.243
13	5.2 to 5.8	5.55	13	2	5.40	1.686	.404
14	5.8 to 6.3	6.04	9	6	1.46	.378	.243
15	6.3 to 7.0	6.85	11	4	2.56	.940	.283
16	7.0 to 7.9	7.59	12	3	3.57	1.273	.326
17	7.9 to 9.0	8.55	9	6	1.46	.378	.243
18	9.0 to 10.7	9.82	9	6	1.46	.378	.243
19	10.7 to 12.0	11.40	9	6	1.46	.378	.243
20	12.0 to 14.0	13.08	12	3	3.57	1.273	.326
21	14.0 to 16.3	15.05	11	4	2.56	.940	.283
22	16.3 to 27.7	21.99	12	4	2.78	1.023	.276
TOTAL			194	137			

between case and control. Note that the range of variation is usually less than 1 year within a given class, the only exceptions being at the high end of the scale.

The average *absolute* difference,  $\bar{d}_k$ , between cases and controls has been entered in the third column. It is important to note that the average absolute difference is used, because it has been shown that the hypothesized natural logarithm of relative risk in these categories is proportional to  $\bar{d}_k$ , not the average of signed differences. The factor of proportionality is  $\ln b$  which is to be estimated from the data.

In column (4) we have the observed number of obverse and inverse pairs, and from this is calculated the entries in columns (5), (6) and (7). Note that there is a general tendency for the observed relative risk and natural log relative risk in columns (5) and (6), respectively, to increase as  $\bar{d}_k$  increases. The analysis of computed chi-squares is shown in Table 3.

The component of variation due to the least-squares estimated slope is highly significant and the residual is far from significantly large. This means that the exponential hypothesis has passed the test. Variations from hypothesis in the "goodness of fit" test of the model are of the very same magnitude as would be expected of random variations, as shown by the residual chi-square of 17.061 with 21 degrees of freedom. Also, more detailed inspection of deviations in the various categories reveals no meaningful pattern. At the same time the value of  $\ln b$  has been found to be significantly different from 0, with a probability less than .0001.

The least squares estimate of  $\ln b$  is given by

$$\sum w_k y_k \bar{d}_k / \sum w_k \bar{d}_k^2 = +.058.$$

It is positive, as hypothesized. The standard error of  $\ln b$  is given by

$$1 / \sqrt{\sum w_k \bar{d}_k^2} = .014.$$

TABLE 3

Analysis of Variance of Log Relative Risk*			
Component of Variation	S. S.	d. f.	Probability
Total ( $\sum w_k y_k^2$ )	34.829	22	
Due to Slope [ $(\sum w_k y_k \bar{d}_k)^2 / \sum w_k \bar{d}_k^2$ ]	17.768	1	< .0001
Residual	17.061	21	> .50

\* $w_k = 1/s_k^2$

Thus, approximate 95% confidence limits put the differential rate of change in risk of breast cancer during menstruating phases, as compared to "other" phases, somewhere between +3.1% and +8.5% per year. This is analogous to a person paying somewhere between an extra 3.1% to 8.5% in continuously compounded interest on a loan, except that here the accumulation is in terms of the risk of breast cancer.

## 7. DISCUSSION

Several questions arise in connection with the foregoing analysis. Among these are questions dealing with risk functions of other than exponential form, extensions to more than one variable, and hospital selection. These will be discussed in this section.

It is evident that the exponential risk hypothesis was particularly convenient in the analysis of retrospective matched pair data on breast cancer. The reader will recall that, once the risk function,  $r_{ix}$ , was specified, the door was opened to analysis by virtue of the fact that hypothesized relative risk in sub-class  $x_1$  versus  $x_2$  was given by

$$R_{i12} = r_{ix_1}/r_{ix_2} .$$

When  $r_{ix}$  is in the exponential form,

$$r_{ix} = C_i b^x ,$$

the relative risk becomes

$$R_{i12} = b^{x_1 - x_2} ,$$

and

$$\ln R_{i12} = (x_1 - x_2) \ln b .$$

Since the natural logarithm of hypothesized relative risk is proportional to the difference between case and control measures of  $x$ , analysis of the natural logarithm of observed relative risk is quite simple.

While it is true that the exponential hypothesis for breast cancer seems to be appropriate in view of prior results such as Lillienfeld's, we can imagine other studies where it might make more sense to deal with, say, a linear risk function. The simplest non-null linear risk hypothesis might be one of proportionality:

$$r_{ix} = C_i x .$$

Then,

$$R_{i12} = x_1/x_2 .$$

Consequently, the conditional probability of an obverse pair would be given by

$$P_{i12} = \frac{x_1/x_2}{(x_1/x_2) + 1} = \frac{x_1}{x_1 + x_2}$$

and

$$Q_{i12} = 1 - P_{i12} = \frac{x_2}{x_1 + x_2},$$

is the probability of an inverse pair. In other words,  $x$  appears in the hypothetical probability expressions just as if it were an observed frequency rather than a measured variable. In this case, one could order all the observed pairs according to  $R_{i12}$  as the ordering principle, divide the array into a number of categories, count the number of obverse and inverse pairs in each category and test these observed frequencies against the hypothesized (expected) frequencies of obverse and inverse pairs in an approximate chi-square goodness of fit test. (When this is done with the breast cancer data, using  $x$  equal to menstrual years, the fit is rather poor, with  $P$  somewhere in the neighborhood of .10 to .03, depending somewhat on what rule is used to form the categories.) Chi-square here would have as many degrees of freedom as there are categories.

But under a more general linear hypothesis,

$$r_{ix} = C_i(a + bx),$$

the hypothesized relative risk is

$$R_{i12} = (a + bx_1)/(a + bx_2),$$

which evidently depends on the relative magnitudes of  $a$  and  $b$ . So, too, the conditional probabilities of obverse and inverse pairs would depend on the relative magnitude of  $a$  and  $b$ . In principle, it would be possible to specify the range of  $a/b$  such that the data fit the hypothesis within specified probability limits. But as a practical matter, such an analysis, which might require repeated chi-square tests, would be quite tedious unless an electronic computer were available. On the other hand, if the investigator were willing to develop a more specific hypothesis, in which the relative levels of  $a$  and  $b$  would be specified, then the analysis would reduce to the same type as under the proportionality hypothesis. This restriction in the applicability of the analysis seems to hold generally: in order to test risk hypotheses conveniently it appears desirable to reduce the hypotheses, possibly through specifying certain parameters, to either a proportionality or exponential function.

Note that we have dealt with testing non-null hypotheses. We can, of course, consider the null hypothesis as a particular case. Under the null hypothesis, however, the ordering principle on which categories

are based is lost. No categories can be specified on this principle. But note that if we merely count the total number of obverse and inverse pairs and test the null hypothesis with one degree of freedom, we have a conventional sign test.

The exponential hypothesis seems to be most suited to the analysis because one not only obtains a goodness of fit test of the model, but at the same time obtains a test of the significance of the growth or decay constant. It is natural to wonder, therefore, whether the exponential hypothesis can be extended to more than one variable. That is, can hypotheses such as

$$r_{ixy} = C_i a_1^x a_2^y$$

be tested?

Under this hypothesis, no difficulty is encountered in forming the relative risk:

$$R_{i12} = a_1^{x_1 - x_2} a_2^{y_1 - y_2}.$$

The natural logarithm of  $R_{i12}$  is a linear form

$$\ln R_{i12} = (x_1 - x_2) \ln a_1 + (y_1 - y_2) \ln a_2.$$

And so, if a suitable ordering principle can be specified so as to produce an array of pairs which can be classified into ordered categories, the analysis can proceed without difficulty. But this seems to be the big problem here, that no clear-cut ordering principle seems to be available. Unless the relative magnitude of  $a_1$  and  $a_2$  are somehow implied, one does not know in which order to place the pairs. Of course, it is sometimes possible to circumvent this difficulty in practical situations where one of the variables, say  $x$ , has an established relevance and the other does not. In this case, a null hypothesis with respect to  $a_2$  can be entertained and the ordering principle becomes based on  $x$  alone. Then  $\ln a_1$  can be estimated by the method illustrated in this paper. Then the pairs can be re-ordered according to  $(y_1 - y_2)$  and classified into categories. Using the estimated value of  $\ln a_1$ ,  $\ln R_{i12}$  can be calculated for each pair and summed over all pairs in each category. Then a goodness of fit test of this null hypothesis can be made to see whether deviations of the number of obverse and inverse pairs are significantly great. Perhaps the most relevant component of chi-square in this test would be the linear component with respect to  $y$ . Also, perhaps, in some practical situations, this component of variation might be used to estimate  $\ln a_2$ . But such a method of estimation would clearly be subject to possible bias in the estimation of both  $a_1$  and  $a_2$ .

The final question to be discussed is hospital selection. Customarily,

if an hypothesis holds in the more general population of a city or state, it is taken to carry more weight than if it holds only in a hospital population. Of course, there is nothing sacred about a city or state-wide population, for there may well be selection in this population when reference is made to an even more general super-population extending further through time and space. But, as implied before, the burden of proof tends to pass to the counter-hypothesis (selection) when the original hypothesis holds repeatedly in a variety of circumstances. By the same token, results which have passed only a single test in a restricted situation such as a hospital should be viewed with some reserve.

The practical problem which arises is whether a hospital study provides a *prima facie* case for further studies on perhaps a larger scale. It seems that the case in favor of further study should generally be stronger when an hypothesis has passed a test even in a restricted setting. If, furthermore, the restricted study has shown agreement with a predicted *pattern*, as a result of a constructed risk function, then the case in favor of the hypothesis should be further strengthened. At least, since successfully predicted patterns of effects must tend to elicit comparably complex patterns from counter-hypotheses, an anonymous cry of "selection" will not suffice to explain away the results.

#### ACKNOWLEDGEMENTS

I wish to thank Dr. Morton L. Levin and Dr. Saxon Graham for their stimulating discussions and criticisms on both theoretical and practical levels, and to thank Mr. Oliver Glidewell for his assistance in the preparation of the data.

#### REFERENCES

- Cornfield, J. [1951]. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* 2, 1269-75.
- Haldane, J. B. S. [1955]. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics* 20, 309-11.
- Kraus, A. S. [1960]. Comparison of a group with a disease and a control group from the same families, in the search for possible etiologic factors. *American Journal of Public Health* 50, 303-11.
- Lilienfeld, A. M. and Johnson, E. A. [1955]. The age distribution in female breast and genital cancers. *Cancer* 8, 875-82.
- Lilienfeld, A. M. [1956]. The relationship of cancer of the female breast to artificial menopause and marital status. *Cancer* 9, 927-34.
- Woolf, B. [1954]. On estimating the relation between blood group and disease. *Annals of Human Genetics* 19, 251-3.
- Wynder, E. L., Bross, I. J., and Hirayama, T. [1960]. A study of the epidemiology of cancer of the breast. *Cancer* 13, 559-601.