EXPERIENTIA

Vol. 31 – Fasc. 1 Pag. 1 – 136 15. 1. 1975

Extensions of the Relative Risk Concept*

by H. ROSALIE BERTELL
Roswell Park Memorial Institute, 666 Elm Street, Buffalo (New York 14203, USA).

1. Interrelations Between Basic Population Parameters

1. *Introduction*

In the present climate of ecological awareness, the statistical practice of comparing persons with a given disease with persons not having the disease, with respect to exposure to certain environmental factors thought to be harmful, has become a major methodological technique. It is important for biostatisticians and biomedical personnel to reexamine the fundamental assumptions and models basic to this method, and the conclusions which they validly suggest for the data. Present refinements of thought can not be haphazardly superimposed on a methodology developed prior to such sophistication. It was encouraging to note that an editorial by SWARM and PFITZER[1] in December 1973, has echoed this call for more precise updating of concepts.

The usual statistical analysis of environmental hazards yields a number, $\hat{r}$, called the estimated relative risk of the disease given the exposure[2]. This statistic is also frequently referred to as an odds ratio. It is an empirical estimate of the true relative risk of disease, that is, the ratio of the incidence rate in the exposed population to the incidence rate in the population not exposed.

The primary purpose of this paper is to present a graphic basis for the mathematical understanding of the relative risk statistic. This statistic may be misleading when used in isolation. The graphic model helps to clarify its relationship to other statistical quantities, such as the proportion of the population exposed to the hazard, the observed incidence rate of the disease, and the rarity of the disease being considered. The presentation is informal in the hope of reaching a biomedical audience, which must use and make realistic decisions based on reported statistical results.

Secondarily, the question of prediction of incidence of disease relative to increase or decrease in exposure to the suspected hazard in the general population, is considered. It should become evident that effective public health response to reported risk situations is more efficient and more effective when a broader presentation of interrelated parameters is made. The proportion of cases attributable to the exposure is also considered relative to the other parameters[3].

A dynamic population model allowing for different relative risk values within subgroups of the population is given. Such variations within age, sex or geographical grouping are more realistic in view of our present knowledge of host defense mechanisms, than is the assumption of an underlying common risk. Some thoughts on clinical detection of other less obvious high risk subgroups of the population are presented.

2. *Method of arriving at a relative risk estimate*

Presentation of a case-control approach only will be given here, as a cohort approach does not essentially differ from this and might serve to confuse the basic ideas. Data from a random sample of cases and controls (non-cases) may be arranged in a 2×2 table as follows:

	With exposure	Without exposure	Sum
Cases	η_{11}	η_{12}	$\eta_{1\cdot}$
Controls	η_{21}	η_{22}	$\eta_{2\cdot}$
Sum	$\eta_{\cdot 1}$	$\eta_{\cdot 2}$	$\eta_{\cdot\cdot}$

We are assuming that cases have at some point had a medically reliable diagnosis of the disease under consideration, and that the reported exposure to the environmental hazard was prior to this diagnosis. In this model we are considering exposure vs. non-exposure as a simple yes-no alternative. The pertinent sample statistics are:

$$\hat{p}_1 = \frac{\eta_{11}}{\eta_{1\cdot}} \quad \text{and} \quad p_2 = \frac{\eta_{21}}{\eta_{2\cdot}}$$

[1] R. L. SWARM and E. A. PFITZER, *J. Human Pathology*, Vol. 4, No. 4, p. 601 (1973).
[2] J. CORNFIELD, J. natn. Cancer Inst. *11*, 1269 (1950–51).
[3] M. LEVIN, Unio Internationalis Contra Cancrum: Acta Volume *9*, 110 (1953).

* This investigation was supported by Public Health Service Research Grant No. CA-11531 from the National Cancer Institute.

which estimate the relative frequency of exposure in cases and controls, and:

$$\hat{r} = \frac{\eta_{11}\,\eta_{22}}{\eta_{12}\,\eta_{21}} = \frac{\hat{p}_1\,(1-\hat{p}_2)}{(1-\hat{p}_1)\,\hat{p}_2}$$

which estimates the risk of disease in the exposed group relative to the unexposed group. (If the reader is interested in a mathematical discussion of this statistic and its many modifications, he is referred to[2,4–7].)

An expression of the interrelation of these three variables which has proven valuable for clarification of the theory is:

$$\hat{p}_1 = \frac{\hat{r}\,\hat{p}_2}{1+(\hat{r}-1)\,\hat{p}_2}$$

It shows that the proportion of cases with the given exposure, $\hat{p}_1$, is a function of the proportion of controls exposed, $\hat{p}_2$, but is not a simple product of $\hat{r}$ and $\hat{p}_2$. The tilda is used over the r, p_1, and p_2 to indicate that they are the observed estimates of the unknown parameters as measured from a random sample of the population.

3. *What this methodology assumes about the population*

When considering the population from which the sample has been drawn, there are two distinct approaches based on the two possible choices for the first dichotomizing of the data. These two approaches give rise to two conceptually different population models.

In Model I we consider the total population, N, as divided into two groups: those who have been exposed to the suspected environmental hazard, ωN, and those who have not been exposed $(1-\omega)N$, where ω is a number between zero and one. In each of these groups we look at the incidence rates of the particular disease being studied α_1 and α_2, to see if the exposed group shows an elevated incidence rate of disease. The numbers α_1 and α_2 are usually expressed as cases per 100.000 in the population. The quantity $\alpha_1\,(\omega N)$ gives the actual number of cases of the disease in the total population with exposure, and $\alpha_2(1-\omega)N$ gives the number of cases of the disease in the total population without exposure.

In a second approach, the population is divided into two groups: those having the disease, χN, called cases, and those not having the disease, $(1-\chi)N$, called controls. The number χ represents the incidence rate of the disease in the total population in cases per 100,000. It lies between the two incidence rates α_1 and α_2, which represent the two extreme rates: everyone exposed and no one exposed. Each of these groups, cases and controls, are examined for the proportions of persons, p_1 and p_2, reporting exposure to the suspected hazard. If the exposure is hazardous, p_1 is greater than ω and p_2 is less than ω, i.e. the cases would report exposure more frequently than would the toal population.

Numerically $p_1\chi N$, the number of persons having the disease who were exposed to the hazard, is exactly the same as $\alpha_1\omega N$, the number of persons among those exposed to the hazard who actually developed the disease.

The usual cross ratio technique for estimating relative risk employed on each model yields a number:

$$r' = \frac{\alpha_1\,(1-\alpha_2)}{(1-\alpha_1)\,\alpha_2} = \frac{p_1\,(1-p_2)}{(1-p_1)\,p_2}$$

The reader might also note that this statistic, r', is independent of N, the total population size, or the sizes of the two sub-populations, cases and controls.

It should be noted that while the two cross ratios produce identifical numbers, the order of magnitude of α_1 and α_2 (incidence of disease) is normally quite different from that of p_1 and p_2 (proportion of the population exposed to the hazard). For example, in a paper on childhood leukemia[8], STARK and OLEINICK report that in a study of white males between birth and 20 years of age, the death rate for leukemia among the urban population was 4.64×10^{-5} $(\hat{\alpha}_1)$ and among the rural population was 3.23×10^{-5} $(\hat{\alpha}_2)$. If we consider exposure to urban environment as the hazard, then of the total number of white male cases studied, $N =$ 10,781, 0.71 N or 7,671 white males were exposed to this environment. This 0.71 would be $\hat{p}_1$. The proportion of non-cases exposed to the urban environment was 0.63, the $\hat{p}_2$ parameter.

[4] J. B. S. HALDANE, Ann. hum. Genet. *20*, 309 (1955).
[5] B. WOOLF, Ann. Hum. Genet. *19*, 251 (1954–55).
[6] N. MANTEL and W. HAENSZEL, J. natn. Cancer Inst. *22*, 719 (1959).
[7] H. R. BERTELL, J. Med. *5*, in press (Karger, Basel 1974).
[8] C. R. STARK and A. OLEINICK, J. natn. Cancer Inst. *37*, 369 (1966).

Model I

	Have disease	Do not have disease	Sum
Exposed population	$\alpha_1\omega N$	$(1-\alpha_1)\omega N$	ωN
Population not exposed	$\alpha_2(1-\omega)N$	$(1-\alpha_2)\,(1-\omega)N$	$(1-\omega)\,N$
Sum	$\alpha_1\omega + \alpha_2(1-\omega)\,N$	$(1-\alpha_1)\omega + (1-\alpha_2)\,(1-\omega)\,N$	N

The quantities of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are by definition estimates of the incidence rates in the exposed and non-exposed groups, hence the estimate of the relative risk of disease is:

$$\hat{r} = \frac{\hat{\alpha}_1}{\hat{\alpha}_2} = \frac{(1-\hat{\alpha}_1)r'}{(1-\hat{\alpha}_2)}$$

where r' is the cross ratio calculation as described above. The reader can verify that the quantity

$$r' = \frac{\hat{p}_1(1-\hat{p}_2)}{(1-\hat{p}_1)\hat{p}_2} = 1.4379$$

in this case, is a satisfactory estimate for $\hat{\alpha}_1/\hat{\alpha}_2 = 1.4365$, when accuracy to two decimal places is sufficient.

The rarity of the disease here assures that the error factor will be negligible. In general, assuming the exposure to be hazardous, α_1 is greater than α_2, and r' is greater than r. Where the disease is rare, this bias is negligible. Clarification of the term 'rare' will be made in section 5. Knowledge of either parameter set, disease incidence in the exposed or not exposed group, or proportion exposed in the cases and controls, will give rise to comparable estimates of relative risk. Ordinarily, because of their magnitude, the latter measurements are more easily obtained.

4. *Graphic interpretation*

The relationship introduced in section 2 for the quantities $\hat{r}$, $\hat{p}_1$ and $\hat{p}_2$, holds true for the unknown population parameters r', α_1 and α_2, as well as for r', p_1 and p_2, i.e.:

$$\alpha_1 = \frac{r'\alpha_2}{1+(r'-1)\alpha_2} \quad \text{and} \quad p_1 = \frac{r'p_2}{1+(r'-1)p_2}$$

Using the ordinary Euclidean coordinates, those pairs of points, (p_2, p_1), which yield a constant r' value, lie on the line determined by the equation:

$$p_1 = \frac{r'p_2}{1+(r'-1)p_2}.$$

We are considering here the abstract relation between the variables, apart from their interpretation in the real life situation. Hence we have considered both (α_2, α_1) and (p_2, p_1) as points in the unit square lying on the same line determined by the value of r', even though naming of the horizontal and vertical axes change as we focus on one or the other. Figure 1 shows these lines for $r' = 1$ and $r' = 3$.

In the example cited in section 3, the two points, $(\hat{\alpha}_2, \hat{\alpha}_1) = (3.23\times10^{-5}, 4.64\times10^{-5})$ and $(\hat{p}_2, \hat{p}_1) = (0.63, 0.71)$, both lie on the line determined by $r' = 1.44$.

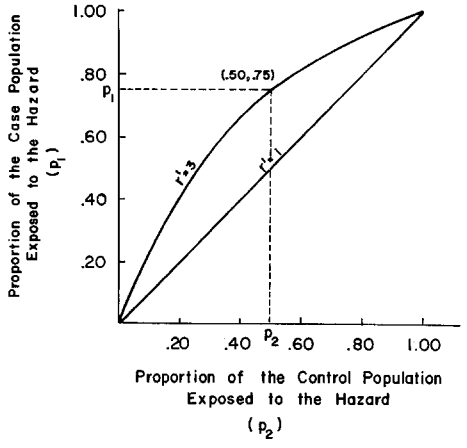

Fig. 1. Proportion of Population exposed to the hazard (p_1 and p_2).

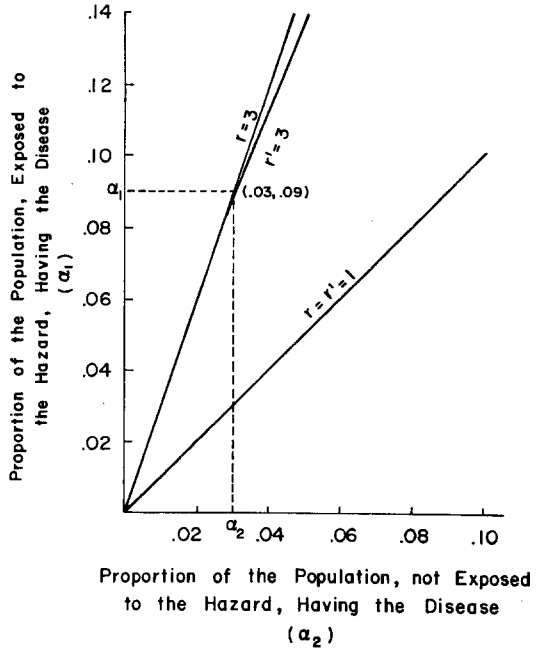

Fig. 2. Proportion of population having the disease (α_1 and α_2).

Model II

	With exposure	Without exposure	Sum
Case	$p_1\chi N$	$(1-p_1)\chi N$	χN
Control	$p_2(1-\chi)N$	$(1-p_2)(1-\chi)N$	$(1-\chi)N$
Sum	$p_1\chi + p_2(1-\chi)N$	$(1-p_1)\chi + (1-p_2)(1-\chi)N$	N

This line would lie totally between the lines determined when $r' = 1$ and $r' = 3$.

It is to be noted that $r' = 1$ is equivalent to the statement that $p_1 = p_2$. This indicates a simple means for identifying environmental hazards in pilot studies. A diseased population is matched for age, sex and geographical location with a random sample of the general population. If the diseased population shows a higher proportion exposed to some environmental factor than is true for the control group, this factor deserves closer study and refinement of hypothesis.

It can easily be shown that p_1 is a strictly increasing function of p_2 for constant r' value. Each of the lines determined by specific values of r', pass through the two points (0,0) and (1,1).

Figure 2 shows a magnification of the Euclidean grid in the neighborhood of the origin, with the horizontal axis used to plot α_2, the disease incidence rate in the population not exposed to the hazard, and the vertical axis used to plot α_1, the incidence rate in the exposed population. It will be recalled that r' approximates the actual relative risk value $r = \alpha_1/\alpha_2$ when α_1 and α_2 are sufficiently small (i.e. the disease 'rare'). Figure 2 contains the lines determined when $r' = 1$ and $r = 1$, wich coincide for all values of p_2 or α_2. It also contains the line determined when $r = 3$, which may be written: $\alpha_1 = 3\alpha_2$. This line is tangent to the curved line determined when $r' = 3$, at the origin, and very nearly approximates the curve for values of α_2 less than 0.03 (or $3{,}000 \times 10^{-5}$ in the usual cases per 100,000 expression).

II. Standard uses of the concept

5. *How rare must a disease be?*

The statistical technique which uses $\hat{p}_1$ and $\hat{p}_2$ as estimates of p_1 and p_2, determines a point in the unit square which lies on the line determined by $\hat{r}$.

In turn, $\hat{r}$ estimates r', the line on which both (p_2, p_1) and (α_2, α_1) lie (in the abstract sense). Since we have noticed that r' is greater than r which is assumed to be greater than one, the true relative risk of disease, when α_1 is greater than α_2, we may ask how rare a disease must be (or how small α_2 must be) to assure the fact that $(r' - r)/r$, the relative error, is sufficiently small.

Since:

$$\frac{r' - r}{r} = (r' - 1)\alpha_2$$

we have a condition dependent upon r' and α_2 (unless, of course, $r = r' = 1$). This leads to: $(r' - r)/r < \varepsilon$, where ε is the maximum relative error permitted, if and only if:

$$\alpha_2 < \frac{\varepsilon}{(r' - 1)}, \quad \text{where} \quad r' > 1$$

Table I gives the maximum value for the incidence rate α_2, given in cases per 100,000, which assures a relative error in estimate of r less than 0.10, 0.05 or 0.01, for some common values for r'. A relative risk estimate made in the usual way for a disease not considered rare in this precise sense can give misleading information. Of note here would be for example, reports of the relative risk of smokers contracting a common cold, or the relative risk of coffee drinkers over 65 years having heart disorders. Critical interpretation is required in order to understand the precise meaning of such statements when the disease in question is obviously not rare.

Some interesting results occur when r' lies between 1 and 2, seeming to demand little in the way of rarity of disease.

Table I. Quantifying rare disease

Relative risk estimate or r' value	Maximum relative error		
	0.10	0.05	0.01
2	10,000	5,000	1,000
3	5,000	2,500	500
4	3,333	1,667	333
5	2,500	1,250	250
6	2,000	1,000	200
7	1,667	833	167
8	1,429	714	143
9	1,250	625	125
10	1,111	556	111

Table values are in number of cases per 100,000 persons in the population. These values are multiplied by 10^{-5} to obtain α_2, the incidence rate in the general population not exposed to the hazard.

While these figures give a generous margin to the rare disease category, it will, I am sure, be conceded that an estimate of r' accepted without knowledge of the approximate incidence of the disease being studied, may be very misleading. In dealing with a disease which is not rare, if the incidence rate, α_2, in the non-exposed population can be estimated, the relative risk estimate, r' may be modified to give a more reasonable estimate of the actual risk r:

$$r \sim \frac{r'}{1 + (r' - 1)\alpha_2}$$

6. *Design of experiments*

Experimental designs based on Model I, where the exposed population and the unexposed population are studied in the hopes of discovering a higher incidence rate of disease in the former group, are really designed to estimate the parameters α_1 and α_2. In rare diseases these tests often understandably fail to detect a difference since the size of the population which must be studied becomes prohibitive. This is probably why, for example, studies of persons exposed to sick pets usually fail to show an increased incidence rate of leukemia. Unfortunately the saying: 'with statistics you can disprove anything you want,' is often reinforced by this type of faulty but convincing argument. As soon as an environmental factor is isolated as detrimental to health, persons, often those with vested interests, undertake studies based on this model and conclude that there is no significant difference in the incidence rates of disease in the exposed and non-exposed groups. The truth is, rather, that the rates are too small to allow for a detectable difference using this technique. It is like using the naked eye to see the fine inner structure of a plant cell.

The studies rooted in Model II, which attempt to obtain good statistical estimates for p_1 and p_2 are more easily carried out. These values serve to magnify the relation between α_1 and α_2, and can be used to determine much useful information about the relative risk of the disease. Studies of this type begin either with case and control groups randomly selected or with almost total case population at a given time and place together with a random control population. Determination of the proportions of each group having the given exposure can then be made.

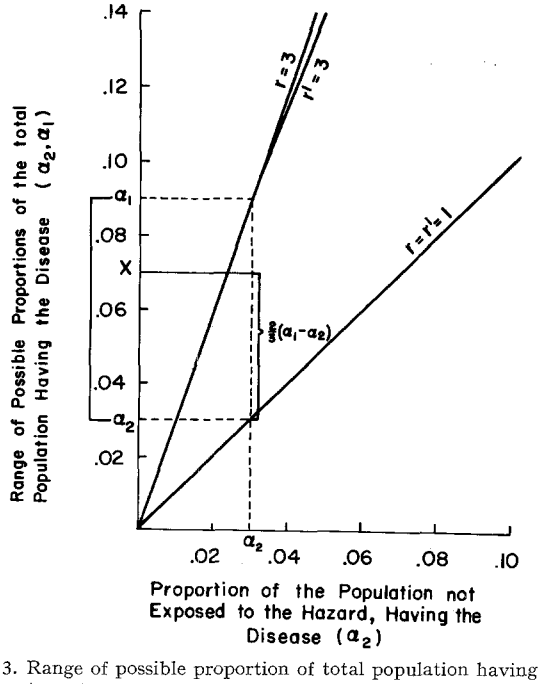

Fig. 3. Range of possible proportion of total population having the disease (α_2, α_1).

As a variation of Model II, some experiments are designed to focus on a subgroup of the population, which because of geographical location, occupation or cultural environment seems to exhibit 'more' exposure to the suspected hazard. We can distinguish two basic designs for this type: 1. that in which the crucial factor is either exposure or non-exposure, and 'more' exposure means an increase in the proportion of the general population exposed; and 2. an exposure which admits of degrees of intensity – each degree of intensity giving rise to a different incidence rate for the disease being studied in the group so exposed.

In the first case, contrary to our intuition, we would see no change in r' since the risk itself depends on the two incidence rates and is completely independent of the proportion of the population exposed. Graphically, increasing p_2 (proportion of the general population with the given exposure) simply moves the parameter point (p_2, p_1) to a new point on the line determined by r'. However we would expect an increase in the observed incidence of disease, χ, as ω, the proportion of the general population exposed, increases, in keeping with the equation:

$$\chi = \omega\alpha_1 + (1-\omega)\alpha_2$$

This relationship is expanded upon in section 7. It is readily seen that α_1 and α_2 are the two extreme incidence rates between which the observed rate fluctuates.

The second case will be dealt with in a separate paper, as more mathematical concepts are needed to understand this type of dynamic interaction[9].

7. *A basic population model*

Assume that in a given population the incidence rate of a given disease remains constant within various subgroups of the population exposed to a certain environmental hazard, and the incidence rate in similar subgroups among those not so exposed is also constant. This would, for example, be the case if incidence rate was the same for both sexes, and for all age groupings within the respective populations. Then:

$$r' = \frac{\alpha_1(1-\alpha_2)}{(1-\alpha_1)\alpha_2}$$

would be constant for this population.

As can be seen by comparing Model I with Model II, the incidence rate in the general population is a weighted average of these two incidence rates:

$$\chi = \omega\alpha_1 + (1-\omega)\alpha_2 = \alpha_2 + \omega(\alpha_1 - \alpha_2)$$

where ω measures the proportion of the population exposed to the hazard. This relationship is shown in Figure 3, which is the same as Figure 2 except for this

[9] H. R. Bertell, On estimating leukemia risk per x-ray plate (unpublished).

additional concept. In this example, ω is assumed to be two thirds.

This simple model proposes that the variation in incidence rate for disease, which might be noticed because of geographical location, occupation, etc. is a simple function of the proportion of the population exposed to the hazard. It is important to note that although observed incidence rate varies, relative risk remains constant. If no one is exposed to the hazard, the incidence rate would be α_2. If everyone were exposed, it would be α_1.

In designing a test based on this type of hypothesis, subgrouping of the data is into categories with similar proportions of persons exposed to the hazard (i.e., by geographical area, occupation, socio-economic status, etc.). These sub-groups yield pairs of values: $(\hat{p}_{21}, \hat{p}_{11})$, $(\hat{p}_{22}, \hat{p}_{12})$, ..., $(\hat{p}_{2j}, \hat{p}_{1j})$ each lying on a curve: $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_j$. These values are estimates of the unknown constant r' (and of r if α_1 and α_2 are small). Some method of weighting the average of the $\hat{r}_i$'s (such as in the WOOLF-HALDANE method)[4,5] is appropriate for estimating r'.

If information on χ, the incidence rate of disease in the general population, is available, the experimenter may wish to establish a series of weights $m_i, i = 1, 2, \ldots, k$, for each of the k subcategories, reflecting the proportion of the total population which each represents, where $\sum_{i=1}^{k} m_i = 1$.

Then: $\hat{\omega} = \sum_{i=1}^{k} m_i \hat{p}_{2i}$ is a reasonable estimate for ω, the proportion of the general population exposed to the suspected hazard.

Using the two relationships:

$$\chi = \omega\alpha_1 + (1-\omega)\alpha_2 \quad \text{and} \quad \alpha_1 = \frac{r'\alpha_2}{1+(r'-1)\alpha_2}$$

together with estimates of ω and r', a second degree equation in α_2 can be obtained:

$$(1-\omega)(r'-1)\alpha_2^2 + (r'-1)(\omega-\chi) + 1\,\alpha_2 - \chi \sim 0.$$

This can be solved for reasonable estimates of α_2 and α_1, giving the two extreme incidence rates. The excess of cases due to this exposure can then be estimated, and predictions of future trends can be made.

If the conditions for rarity of the disease are met, the substitution: $\alpha_1 \sim r'\alpha_2$ simplifies the calculations considerably.

LEVIN[3] has introduced a statistic for estimating the proportion of cases attributable to the given environmental hazard, S, which uses only the sample estimates of r and p_2:

$$S = \frac{(\hat{r}-1)\hat{p}_2}{1+(\hat{r}-1)\hat{p}_2}$$

Using the relationship between $\hat{p}_2$ and $\hat{p}_1$, given in section 2, we also have:

$$S = \frac{\hat{p}_1 - \hat{p}_2}{1-\hat{p}_2}$$

or equivalently,

$$S = \frac{(\hat{r}-1)\hat{p}_1}{\hat{r}}$$

Hence without determining α_1 and α_2, we can obtain an estimate of the proportion of cases attributable to the given exposure. It is important to understand that the relative risk factor in isolation gives us no clue as to the impact of this factor on actual incidence rates. Any combination of two estimates, of the three pertinent variables $(r', p_1)\, p_2)$ will suffice to estimate S, attributable proportion of cases. Again, the public health impact of S will be determined by the acutal incidence rate of disease, χ. 10% of 500 cases is of greater medical concern than 10% of 10 cases.

Again, this type of information is of more use in actually assessing the health hazard involved in the exposure than is the relative risk estimate alone.

8. *Constant relative risk model*

A more usual assumption is that the general population may be divided into sub-categories where the incidence rates vary independently of the proportion of persons experiencing the exposure, as for example variation of incidence rate of a disease by sex or age. It is often assumed that the relative risk of disease, r, is constant for the exposed group in each sub-grouping.

Mathematically, this yields a series of points: $(\alpha_{21}, \alpha_{11})$, $(\alpha_{22}, \alpha_{12})$ ... $(\alpha_{2k}, \alpha_{1k})$ for the k sub-groups such that: $\alpha_{1i} = r\alpha_{2i}$ for $i = 1, 2, \ldots, k$.

Again, using the size criteria given in section 5, it is sometimes reasonable to assume that r', as well as r, will remain constant. Graphically, we are assuming that the curve determined by r', differs insignificantly from the straight line relationship assumed above when the range of points is 'near' the (0, 0) point. If this size criterion for α_{1i} and α_{2i} is not maintained, the data statistics, $\hat{r}_i$, are correspondingly less reliable as estimates for r.

In the study of rare diseases, this assumption of constant relative risk has led to the WOOLF-HALDANE method[4,5] of using a weighted average of the r_i values, $i = 1, 2 \ldots, k$, as a reasonable estimate for the unknown constant r[3,4]. Here the weighting system is related to the sub-sample sizes and the variability of the estimates they give for p_{1i} and p_{2i}.

Under this assumption, the data is grouped into subcategories determined by differences in incidence rate. For finer analysis, a second division of each subcategory into groups having roughly the same propor-

tion of persons exposed to the given environmental hazard can be made (see section 7). Analysis might combine the techniques in section 7 with section 8.

The Model describing this situation in real life would involve an incidence range for each χ_i, $i = 1, 2, \ldots, k$. The actual value assumed by χ_i within this interval would depend on the value of ω_i, the proportion of persons in the ith sub-category having the given exposure. Different incidence predicting models can easily be developed with varying levels of ω_i.

This model is well developed in the existent literature[2,4–6].

III. Development of a dynamic model

9. *Problems with the assumption of constant risk*

Aside from the more obvious problem of limiting the above techniques to those studies where the rarity of disease meets the requirements given in section 5, there are other problems inherent in this standard method of analysis.

The fine division of data described in sections 7 and 8 frequently result in reduction of data within individual categories to zero, making the mathematical procedures impossible.

It may also be very unrealistic to assume that the relative risk of diseases given a certain exposure is constant over all subgroups. The relative risk may well be a function of age or sex, or other pertinent variable. For example, X-ray exposure seems to involve a greater relative risk for leukemia when given to children with a history of allergies and asthma[10]. Given our present knowledge of the interaction between viral exposure and host defense, it seems likely that the risk of disease may vary widely between groups of persons within the population. Much research effort is presently being directed toward determining high risk sub-groups, since from a public health point of view, these groups must be shielded against hazards to a greater extent than the general population.

A method of data analysis which makes no assumptions about the constancy of the underlying risk value, and which also avoids the zero cell problem and allows for rather fine sub-grouping of the data, was first presented in an article by MANTEL and HAENSZEL[6], and has been elaborated on in a more recent paper[7].

At this point, since the concepts become more complicated, the data presented in the above article[7], will be considered. What is needed for the conceptualization of the method will be presented here. For further information the reader is referred to the article itself.

The data referred to consist of leukemia cases and controls, 15 years of age or more, together with a report on whether or not these persons have been exposed to sick pet birds (canaries or parakeets) within the year prior to diagnosis of leukemia for cases or date of interview for controls. Data are compiled from the Tri-State Leukemia Survey, 1959–1962[11].

The data were first subdivided into 6 categories dependent on 2 variables; sex and age. Each subgroup, S_{ij}, where $i = 1, 2$ indicates male or female respectively, and $j = 1, 2, 3$ indicates 15–44 years, 45–64 years and 65 + years of age, was presumed to be reasonably homogeneous for incidence rate of leukemia.

It was then noticed that there is geographical variation in amount of pet exposure – probably due to sampling from urban or rural populations. Using a third subscript $k = 1, 2, 3$, to indicate that the subject was from New York, Baltimore or Minnesota, the data was further subdivided into 18 separate subgroups, S_{ijk}.

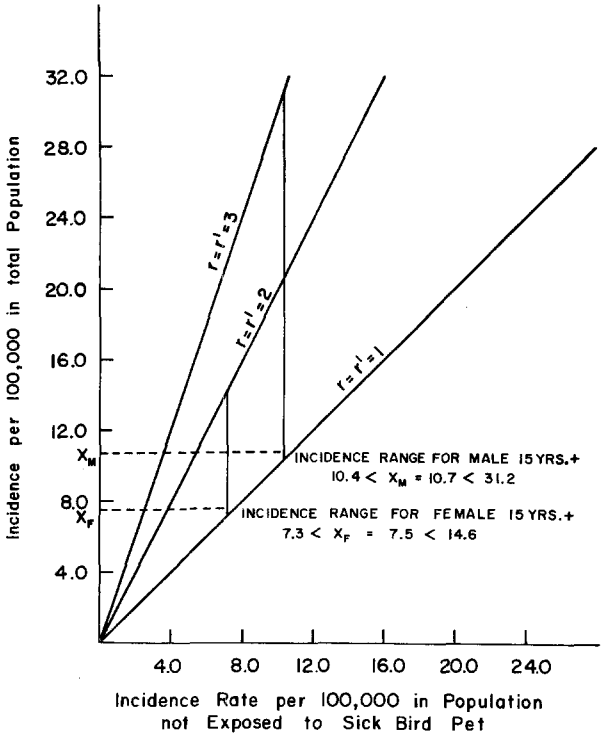

Fig. 4. Incidence rate (per 100,000) of leukemia. Based on reports of cases in Connecticut 1963–65. Cancer Incidence in Five Continents (Springer Verlag 1970, vol. 2) and Tri-State Leukemia Data on Exposure to Sick Bird Pets.

10. *A model for the analysis*

In the analysis, we are assuming variations in both the incidence rates of disease, $\alpha_{1ij.}$ and $\alpha_{2ij.}$, in the six basically homogeneous subgroups, and also variations

[10] I. BROSS and N. NATARAJAN, New Engl. J. Med. *287*, 107 (1972).
[11] S. GRAHAM, M. LEVIN, A. LILIENFELD, J. DOWD, L. SCHUMAN, R. GIBSON, L. H. HEMPELMANN and P. GERHARDT, Ann. N. Y. Acad. Sci. *107*, 557 (1963).

in the proportions of the case and control populations reporting exposure to the environmental hazard, p_{1ijk} and p_{2ijk}.

Figures 4 and 5 are a hypothetical population model which might be proposed to explain the variations observed in the data. The first diagram shows the r' lines characteristic of the male and female samples when analyzed separately. The curved line r' and its tangent r are indistinguishable this close to the origin and at this degree of decimal precision. The relative risks, weighted over age and geographical location, were found to be 2 for females and 3 for males (see Table III).

Cancer Incidence in Five Continents (Springer-Verlag, 1970, Vol. 2) was used to approximate actual incidence rates at the time of the survey. Data from Connecticut, 1963–1965 was closest to the survey time and location (half of the participants were from New York State). This report indicated that incidence rates adjusted for age were: 7.5 for the female and 10.7 for the male. Knowing this, it was desired to reconstruct a diagram similar to diagram III, giving the range of values the incidence rates could assume for the females and males. We also knew that 3% of the control females and 1.4% of the control males had reported exposure to sick bird pets, hence since:

$$\text{Rate} = (1-\omega)\alpha_2 + \omega r\alpha_2 = \alpha_2 + \omega(r-1)\alpha_2,$$
$$7.5 = \alpha_2 + 0.03\,(\alpha_2)$$
or $\alpha_2 = 7.3$ for the female,
and $10.7 = \alpha_2 + 0.014\,(2\alpha_2)$
or $\alpha_2 = 10.4$ for the male.

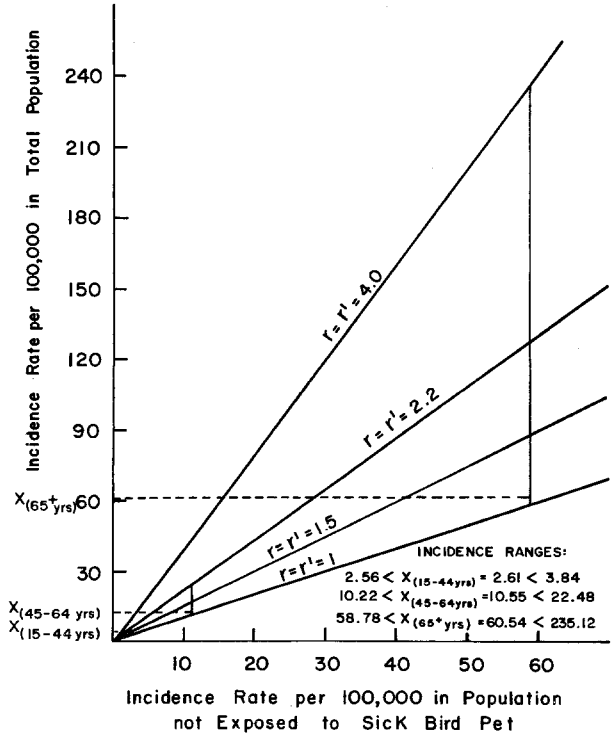

Fig. 5. See legend to Figure 4.

Interestingly, although the females showed more exposure to ill pet birds, the exposure held less risk for them and raised the indicence rate only 0.2 per 100,000 population. The males, with higher no exposure incidence rate and greater relative risk, showed an increase of 0.3 per 100,000 with roughly half the exposure by females. Figure 5 shows a similar construction of incidence ranges for the three age groups, weighted over sex and geographical location. The 15–44-year-old group is lost on this scale, because of its relative insignificance when compared with the over 65 group.

While the incidence ranges for males and females, as shown in Figure 4, overlapped, and hence could allow for equality or inequality in either direction for observed incidence rates, the age groupings have a distinctly different pattern. Increased incidence rate with increased age group will occur regardless of the exposure or lack of exposure to sick pet birds. The dramatic effect of exposure when there is a combination of high relative risk factor and increased basic incidence rate, as in the age group over 65 years, should be obvious.

This model may be used to predict incidence rate of disease in different age groups or different sexes, with changes in ω, the proportion of the population exposed to the hazard.

For example, if in a given area 20% of the population is exposed to sick pet birds and we forsee that this proportion will increase to 50%, we may predict the incidence rates in each of the three age categories to increase as follows:

Table II.

Age (years)	I [a] ($\omega = 0$)	I ($\omega = 0.20$)	I ($\omega = 0.50$)
15–44	2.56	2.82	3.20
45–64	10.22	12.67	16.35
65 +	58.78	94.05	146.95

[a] I = Incidence rate per 100,000.

The author does not wish to preclude an alternate hypothesis that incidence rates of disease also vary geographically. The reader will see the flexibility of the model and its assumptions, allowing realistic hypotheses to be drawn up to fit individual problems. Section 11 shows the results of the analysis based on this dynamic population model.

11. *Applying the dynamic model*

Ideally, the case and control populations should be matched for size in each of the relevant subdivisions. Practically, this seldom happens, but such a matched

population can be constructed using the estimates of $\hat{p}_{2ijk}$ to generate the expected number of persons having a given exposure in a control population matched for size with the case population. The mechanics of this method are explained in [6] and [7].

By pooling the data on one or two variables, the zero cell problem introduced by refinement of the classifications is avoided. For example, we can consider all male cases and controls, where the controls are size-adjusted to match cases for age and geographical area. The same can be done for each of the other variables or pairs of variables, once the basic 2×2 tables are constructed. This method of analysis differs from the WOOLF-HALDANE method in that values of r' are not calculated for each individual table and then a weighted average of estimates taken. Rather, the pooling of data takes place prior to the calculation of r'. The results of this analysis are reproduced from [7] below.

The composite relative risk value obtained by this method gives a realistic composite estimate of the relative risk in the general population, resulting from exposure to this environmental hazard. It does not pretend to estimate some constant but unknown risk.

This method of analysis as used on the Tri-State Survey data exposes the higher risk for males over females, and the progressively increasing risk with age.

In general, this composite relative risk estimate seems more in keeping with the real life biomedical situation than is the traditional assumption of a constant underlying risk.

The public health implications of this particular study are discussed in a recent report from Roswell Park Memorial Institute[12].

12. *Identifying susceptible subgroups*

If we accept the hypothesis that the relative risk of disease given some environmental hazard varies among subgroups of the population, identification of those subgroups becomes important. Some progress along this line has been made, for example, in public sensitivity to the increased risk involved in radiation exposure for pregnant women. The excessive risk to the aged of exposure to sick pet birds may also be noteworthy.

Identification of susceptible sub-groups can be very difficult, but a simple model may suggest an approach reasonable for the clinician and helpful for gaining clues for further research.

Let us assume that persons may be divided into 4 categories: 1. highly susceptible to the disease and exposed to the hazard, 2. highly susceptible to the disease and not exposed to the hazard, 3. not highly susceptible to the disease, but exposed to the hazard, and 4. not highly susceptible to the disease and not exposed to the hazard.

We may assume that the persons in the first category develop the disease, and those in the fourth do not. It is groups two and three which should give us clues about the medical history or characteristics which would make a person highly susceptible to a given disease. We all know persons with lung cancer who have never smoked, as well as inveterate smokers who have no sign of the disease. It is important to identify the personal characteristics which operate to maintain health in the latter case and to bring about breakdown in defense in the former case.

Medical personnel most often come into contact with the person who has contracted a disease, and who has had no exposure recognized as related to increased risk of the disease. We may assume that this person's susceptibility level is quite high and look for relevant indicators of this state. Much research in this area is called for if any sensible public health measures are to be taken to protect the population from potential health hazards. Any other path will lead to trying to protect everyone from everything!

13. *Conclusion*

It is hoped that through sharing of insights between mathematical and biomedical disciplines mutually constructive suggestions for facing the medical prob-

[12] I. BROSS, H. R. BERTELL, R. GIBSON, Am. J. publ. Hlth. *62*, 1520 (1972).

Table III. Complete analysis of relative risk of sick pet bird exposure vs. no pet bird exposure, using the constructed population method

Age (years)	Sex	Geographical area	Lkm. Type	Relative risk	Probablity less than
15–44	Wtd. over	Wtd. over	All	1.45	0.16
45–64	Wtd. over	Wtd. over	All	2.13	0.00
65+	Wtd. over	Wtd. over	All	3.85	0.00
Wtd. over	Male	Wtd. over	All	3.05	0.00
Wtd. over	Female	Wtd. over	All	1.88	0.00
Wtd. over	Wtd. over	New York	All	2.92	0.00
Wtd. over	Wtd. over	Baltimore	All	1.90	0.02
Wtd. over	Wtd. over	Minnesota	All	1.80	0.04
Summary: Wtd. over all variables				2.43	0.00

lems of our age will emerge. There is much more than an arithmetic increase when two or more disciplines pool their efforts.

If the reader desires to delve more deeply into the mathematical theory of point versus interval estimates for relative risk, he is encouraged to explore the articles by MIETTINEN[13,14] of the Departments of Epidemiology and Biostatistics, Harvard School of Public Health.

Dr. IRWIN BROSS and members of the biostatistics department at Roswell Park Memorial Institute have been most helpful in providing the necessary scientific atmosphere which has spurred on this new look at the relative risk statistic.

Zusammenfassung.

Seit 25 Jahren hat sich der Begriff «relatives Risiko» bewährt, um die auf Umweltbelastungen zurückgehenden gesundheitlichen Schäden quantitativ zu erfassen. In diesem Beitrag werden die mathematischen Grundlagen dargelegt und die Beziehungen zwischen dem berechneten und dem tatsächlichen Risiko untersucht. Bei kleinem Risiko wird der Fehler geschätzt. Beziehungen zwischen dem relativen Risiko und dem Verhältnis der Häufigkeit aller Erkrankungen zu der Häufigkeit der durch Umweltbelastung hervorgerufenen Erkrankungen werden algebraisch und graphisch dargestellt. Ein dynamisches Populationsmodell erlaubt, die Unterschiede des relativen Risikos innerhalb der Population zu studieren, im Gegensatz zum bisher benützten uniformen Risiko. Dieses Modell ist flexibler und hat sich bei immunologischen Studien als realistischer erwiesen.

[13] O. S. MIETTINEN, Biometrics *26*, 75 (1970).
[14] O. S. MIETTINEN, Am. J. Epidemiol. *99*, 325 (1974).

SPECIALIA

Les auteurs sont seuls responsables des opinions exprimées dans ces brèves communications. – Für die Kurzmitteilungen ist ausschliesslich der Autor verantwortlich. – Per le brevi comunicazioni è responsabile solo l'autore. – The editors do not hold themselves responsible for the opinions expressed in the authors' brief reports. – Ответственность за короткие сообщения несёт исключительно автор. – El responsable de los informes reducidos, está el autor.

Detection of Zinc in CHAMPY-MAILLET's Histological Stain by Electron Probe Analysis

MAILLET[1] has proposed the use of an osmium tetroxide zinciodide (ZIO) mixture, to reveal, with some degree of specificity, peripheral nervous structures at histological and cytological levels.

The reducing process of this solution allows simultaneously both fixation and staining of histological preparations: neurites and nerve endings appear stained in black on a yellowish background. The reducing chemical process is still imperfectly known. One of us (J.G.) has suspected that Zinc could also be constitutive of the stain, on account of its amphoteric properties; preliminary chemical tests sustain this view. It seemed therefore interesting to investigate the distribution of Zn, Os and I in stained zones of histological preparations, by means of the X-ray electron probe.

Procedure. 10–15 μm thick histological sections are prepared conventionally and mounted on glass slides with gelatin. The sections are deparafinized and covered with a thin Vestopal W film. After polyester hardening for 12 h at 60 °C, the sections are covered with carbon film by means of a Edwards vacuum coating unit model E 12 E.

Operations of the electron probe. An AMX electron probe (ARL – Glendale, Calif.) is used at an accelerating potential of 20 kV with a specimen current on brass, of 25 nA. A LiF crystal spectrometer is focused on Osmium Lα, Iodine Lα and Zinc Kα rays by means of an aqueous solution of OsO_4 (1%) and ZnI_2 (4%) evaporated on a slide and covered by Vestopal W. Discrimination of Zn Kα and Os Lα spectra is distinct with this technique (Figure 1).

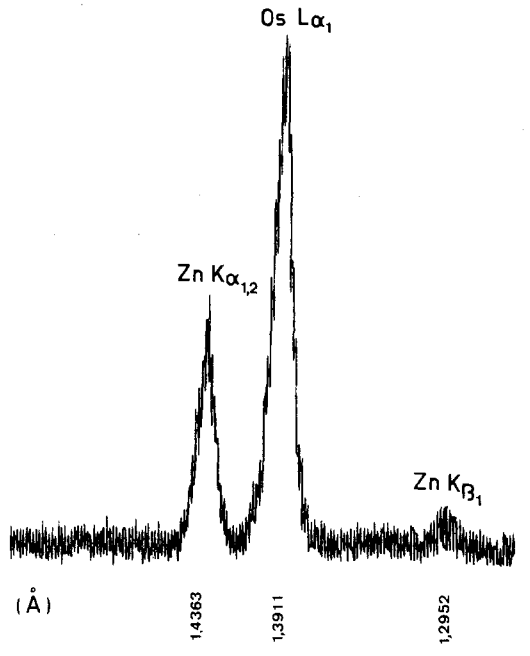

Fig. 1. Discrimination of Zn and Os in X-ray spectrometric analysis. Angström wavelenghts shown for each peak.

[1] M. MAILLET, C.r. Soc. biol., Paris *153*, 939 (1959).