International Journal of Epidemiology, 2014, 511–515
doi: 10.1093/ije/dyu022
Advance Access Publication Date: 28 February 2014

Reprints and Reflections

Limitations of the Application of Fourfold Table Analysis to Hospital Data.*,†

Joseph Berkson

Division of Biometry and Medical Statistics, Mayo Clinic, Rochester, USA

In the biologic laboratory we have a method of procedure for determining the effect of an agent or process that may be considered typical. It consists in dividing a group of animals into two cohorts, one considered the "experimental group," the other the "control." On the experimental group some variable is brought to play; the control is left alone. The results are set up as in table 1-a. If the results show that the ratio $a{:}a+b$ is different from the ratio $c{:}c+d$, it is considered demonstrated that the process brought to bear on the experimental group has had a significant effect.

A similar method is prevalent in statistical practice, which I venture to think has come into authority because of its apparent equivalence to the experimental procedure. In Biometrika it is referred to as the fourfold table and it is used as a paradigm of statistical analysis. The usual arrangement is that given in table 1-b. The entries, *a, b, c* and *d* are manipulated arithmetically to determine whether there is any correlation between *A* and *B*. A considerable number of indices have been elaborated to measure this correlation. Pearson has given the formula for calculating the product-moment correlation coefficient from a fourfold table on the assumption that the distribution of both variates is normal; Yule has an index of association for the fourfold table; there are the chi-square test and others. In essence, however, all these indices measure in different ways whether and how much, in comparison with the variation of random sampling, the ratio $a{:}a+b$ differs from the ratio $c{:}c+d$. If the difference departs significantly from zero, there is said to be correlation, and the correlation is *the* greater the greater the difference.

Now there is a distinction between the method as used in the laboratory and as applied in practical statistics. In the experimental situation, the groups, *B* and not *B,* are selected *before* the subgroupings, *A* and not *A,* are effected; that is, we start with a total group of unaffected animals. In the statistical application, the groupings, *B* and not *B,* are made *after* the subgroupings, *A* and not *A,* are already determined; that is, all the effects are already produced *before* the investigation starts. In the end, the tables of the results which are drawn up *look* alike for the two cases, but they have been arrived at differently. Correlative to this difference, a different interpretation may apply to the results, and this paper deals with a specific case of a kind that arises frequently in a medical clinic or a hospital. I take an example.

There was prevalent an impression that cholecystic disease is a provocative agent in the causation or aggravation of diabetes. In certain medical circles, the gall bladder was being removed as a treatment for diabetes. The authorities of a hospital wish to know whether their accumulated records of incidence, examined statistically, support this practice. On the face of it, it would appear that we have here the typical and elementary problem of the comparison of rates in a fourfold table. The total population of patients for a period is to be divided into two groups, "diabetes" and "no diabetes" and the rate of incidence of cholecystitis in the one compared with the rate in the other. Accordingly, table 2 was set up.

Table 2 shows a significant difference indicating positive correlation between cholecystitis and diabetes. An objection which might be brought against this particular tabulation is that the "not diabetes" group consisting as it does of all patients without diabetes, will contain a variety

*This paper was presented in somewhat different form at a meeting of the American Statistical Association in 1938. Recent inquiries have prompted its publication at this time.

†Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. Biometrics Bulletin 1946;2:47-53 Reprinted with permission.

Table 1. Fourfold Tables

a				*b*			
Typical of experimental situation				Statistical form			
Group	Effect	No Effect	Total	Group	A	Not A	Total
Experimental	a	b	a + b	B	a	b	a + b
Control	c	d	c + d	Not B	c	d	c + d
Total	a + c	b + d	a + b + c + d	Total	a + c	b + d	a + b + c + d

Table 2. Relation of cholecystitis to diabetes-hospital population

	A Cholecystitis	Not A Not Cholecystitis	Total
B: Diabetes	28	548	576
Not B: Not Diabetes	1,326	39,036	40,362
Total	1,354	39,584	40,938
Cholecystitis in diabetic group			4.86%
Cholecystitis in control group (not diabetic)			3.28%
Difference			+1.58% ± 0.5%

Table 3. Relation of cholecystitis to diabetes-hospital population, refractive errors used as control

	A Cholecystitis	Not A Not Cholecystitis	Total
Diabetes	28	548	576
Refractive errors	68	2,606	2,674
Total	96	3,154	3,250
Cholecystitis in diabetic group			4.86%
Cholecystitis in control group (refractive errors)			2.54%
Difference			+2.32%±0.5%

of diagnoses, some of which may themselves be correlated with cholecystitis, even as diabetes may be; hence the control may be considered not good. To meet this objection we do not select for the control group the entire nondiabetic population, but take a diagnosis which cannot reasonably be thought to be correlated with cholecystitis and use this as a criterion for the control group. I took, in fact, several refractive errors of the sort for which patients come to the clinic for glasses as such a diagnostic group, and table 3 was the result.

Again we see that the difference is positive and significant in comparison with the probable error, and the usual judgment would be that cholecystitis and diabetes are positively correlated. Of course, in any detailed analysis we should wish to keep age and sex constant, inquire into the reliability of the diagnoses, and so forth. But the point referred to in this paper has no relation to such questions, and for the sake of the argument we shall consider that all such factors have been adequately controlled. Even so, do the results permit any conclusion as to whether cholecystitis is biologically correlated with diabetes?

Table 4. Constitution of general population, various diseases

$n_d = p_d q_c q_r \times N = 87{,}300$	
$n_c = p_c q_a q_r \times N = 267{,}300$	
$n_r = p_r q_a q_c \times N = 960{,}300$	
$n_{dc} = p_d p_c q_r \times N = 2{,}700$	$N = 10{,}000{,}000$
$n_{dr} = p_d p_r q_c \times N = 9{,}700$	$p_d = 0.01,\ p_c = 0.03,\ p_r = 0.10$
$n_{cr} = p_c p_r q_d \times N = 29{,}700$	$q_d = 0.99,\ q_c = 0.97.\ q_r = 0.90$
$n_{dcr} = p_d p_c p_r \times N = 300$	
$n_o = q_d q_c q_r \times N = 8{,}642{,}700$	

Since the hospital population comes from the general population, let us begin there. For the sake of simplification, we shall consider only the three diseases referred to, cholecystitis, diabetes and refractive errors. If the incidence of these conditions in the general population is represented by $p_{c,}\ p_d$ *and* p_r and there is no correlation between the diseases, we have for the constitution of the population the expressions shown in table 4, in which n_d is the number having diabetes but not having cholecystitis nor refractive errors, n_{dc} those having diabetes and cholecystitis but not having refractive errors, n_{dcr} those having diabetes, cholecystitis and refractive errors, n_c those having none of these diseases, and so forth. N is the total population. If we assume for illustrative purposes, a population of 10,000,000 persons, and $p_d = 0.01$, $p_c = 0.03$, and $p_r = 0.10$, the numbers of the various constituents are given in table 4. From these figures we may set up two fourfold tables as before (table 5).

In both parts of table 5 it is seen that the difference of the pertinent ratios is zero, which is as it should be, since there is no correlation. This result, of course, could have been foreseen without this computation but I desired to establish the numbers for use later. Now suppose we follow that portion of the population which gets to the hospital. For this purpose we must develop some elementary relationships.

Table 5. Cholecystitis and diabetes, general population

	Cholecystitis	Not cholecystitis	Total		Cholecystitis	Not cholecystitis	Total
Diabetes	3,000	97,000	100,000	Diabetes	3,000	97,000	100,000
Not diabetes	297,000	9,603,000	9,900,000	Refractive errors	29,700	960,300	990,000
Total	300,000	9,700,000	10,000,000	Total	32,700	1,057,300	1,090,000
Cholecystitis in diabetic group			3%	Cholecystitis in diabetic group			3%
Cholecystitis in control group (nondiabetic)			3%	Cholecystitis in control group (refractive errors)			3%
Difference			0%	Difference			0%

We shall suppose that associated with each particular disease is a definite probability that its victims will be selected for the hospital. That is, we shall suppose that a person who has cholecystitis has a certain definite probability of being drawn to the hospital because of the presence of that disease alone, and so for other diseases. Furthermore, for simplicity we shall say that these selective probabilities operate independently, as though a person who had two diseases were like Siamese twins, each one of whom had one disease, so that the probability of the twins' coming to the hospital is the probability of either one getting there, but the presence of one disease does not affect the other in any way. Let the selective rates be represented by s_1, s_2, s_3, and so forth and their complements $(1-s)$ be represented by t_1, t_2, t_3, and so forth, the number in the general population by n and the number in the hospital by n'. Then, we have the following equations:

$$n'_1 = n_1(1-t_1) = n_1(s_1)$$
$$n'_{12} = n_{12}(1-t_1t_2) = n_{12}(s_1+s_2-s_1s_2)$$
$$\begin{aligned} n'_{123} &= n_{123}(1-t_1t_2t_3) \\ &= n_{123}(s_1+s_2+s_3-s_1s_2-s_1s_3 \\ &\quad -s_2s_3+s_1s_2s_3) \end{aligned}$$

From these relationships an interesting conclusion can at once be drawn. Suppose all the s's are equal, but small; then the following ratios will result:

$$\frac{n'_{12}}{n'_1} = \frac{n_{12}}{n_1}(2-s) = \text{approximately}, \frac{n_{12}}{n_1}x2$$
$$\frac{n'_{123}}{n'_1} = \frac{n_{123}}{n_1}(3-3s+s^2) = \text{approximately}, \frac{n_{123}}{n_1}x3$$

From these equations it is seen that the ratio of multiple diagnoses to single diagnoses in the hospital will always be greater than in the general population; for two diagnoses the ratio will be about twice that of the general population, for three diagnoses about three times, and so forth.

Let us now apply the appropriate factors of selection to the various constituents of the hypothetical general population which have been enumerated. Assuming as a simple instance that all the selective probabilities are equal and have the value 0.05, the frequencies given in tables 6 and 7 will result.

Table 6. Enumeration of hospital population for $s_d = s_c = s_r = 0.05$

General population numbers	f*	Hospital population, expected numbers
$n_d = 87{,}300$	0.05	$n'_d = 4.365$
$n_c = 267{,}300$	0.05	$n'_c = 13{,}365$
$n_r = 960{,}300$	0.05	$n'_r = 48{,}015$
$n_{dc} = 2{,}700$	0.0975	$n'_{dc} = 263$
$n_{dr} = 9{,}700$	0.0975	$n'_{dr} = 946$
$n_{cr} = 29{,}700$	0.0975	$n'_{cr} = 2{,}896$
$n_{dcr} = 300$	0.142625	$n'_{dcr} = 43$
$n_o = 8{,}642{,}700$	0	$n'_o = 0$

*The fraction of the specified individuals which is selected for the hospital under the operation of the selective forces s. It is equal to 1 minus the products of the appropriate t's; for example $f_{dcr} = 1 - t_d t_c t_r$.

We see here that though in the general population, the incidence of cholecystitis was identical among the persons who had diabetes and the persons who had refractive errors, in the hospital population the incidence was less in the diabetic group than in the control group, giving an appearance of a small negative correlation, and this in the face of the fact that we have assumed equality of selective rates for the various diseases.

In general the selective rates can be assumed to be anything but equal for different diseases. Various circumstances, such as the severity of the symptoms, the amenability of the disease to treatment by a local physician or the reputation of a particular hospital for treatment of particular diseases, will determine the probability that a specific disease will bring its victim to a particular hospital. To see the effect of a variation in selective rates, let us hypothesize some values which will differ among themselves as follows:

$s_c = 0.15$, $s_d = 0.05$, $s_r = 0.20$. The resulting numbers of the various constituents of the population that will come into the hospital are shown in table 8 and the fourfold table drawn up from these figures is given as table 9.

Table 7. Cholecystitis and diabetes, hospital population: expected numbers for $s_c = s_d = s_r = 0.05$

	Cholecystitis	Not cholecystitis	Total
Diabetes	306	5,311	5,617
Refractive errors	2,896	48,015	50,911
Total	3,202	53,326	56,528
Cholecystitis in diabetic group			5.45%
Cholecystitis in control group (refractive errors)			5.69%
Difference			−0.24%

We now find that the incidence of cholecystitis in the diabetic group is about twice that of the control. This would show, so far as the hospital population is concerned, a positive correlation between cholecystitis and diabetes, but it would be quite unrepresentative of the situation in the general population and of no biologic significance.

The relationships dealt with arithmetically in the previous tables are given algebraically as follows:

$$p'_{1.2} = \left\{ \frac{p_1q_3(1 - t_1t_2) + p_1p_3(1 - t_1t_2t_3)}{p_1q_3(1 - t_1t_2) + q_1q_3(1 - t_2) + p_1p_3(1 - t_1t_2t_3) + p_3q_1(1 - t_2t_3)} \right\}$$

$$p'_{1.3} = \frac{p_1(1 - t_1t_3)}{p_1(1 - t_1t_3) + q_1(1 - t_3)}$$

Where

$p'_{1.2}$ is the incidence in the hospital population of condition 1 among persons who have condition 2.

$p'_{1.3}$ is the incidence in the hospital population of condition 1 in the control group who have condition 3.

p_1, p_2, and p_3 are the independent probabilities in the general population of conditions 1, 2 and 3, $q = 1-p$.

t_1, t_2, and t_3 are the complements $(1-s)$ of the independent selective probabilities s_1, s_2 and s_3 applying to condition 1, 2 and 3.

Comment

The assumption made in the text that a probability can be assigned to every disease, which gives the chance that a patient suffering from that disease alone, will come to the hospital is, I think, in general accord with the actual mechanism by which such a patient is selected for the hospital population. The assumption that these probabilities operated independently in an individual who is suffering from more than one disease is doubtless oversimple. In general we may guess that if a patient is suffering from two diseases, each disease is itself aggravated in its symptoms and more likely to be noted by the patient. So far as this difference of fact from assumption goes, its effect would be to increase relatively the representation of multiple diagnoses in the hospital, and in general to increase the discrepancy between hospital and parent population, even more than if the probabilities were independent.

Table 8. Enumeration of a hospital population for $s_c = 0.15$, $s_d = 0.05$, $s_r = 0.20$

General population numbers	f*	Hospital population, expected numbers
$n_d = 87{,}300$	0.05	$n'_d = 4.365$
$n_c = 267{,}300$	0.15	$n'_c = 40{,}095$
$n_r = 960{,}300$	0.20	$n'_r = 192{,}060$
$n_{dc} = 2{,}700$	0.1925	$n'_{dc} = 520$
$n_{dr} = 9{,}700$	0.24	$n'_{dr} = 2{,}328$
$n_{cr} = 29{,}700$	0.32	$n'_{cr} = 9{,}504$
$n_{dcr} = 300$	0.354	$n'_{dcr} = 106$
$n_o = 8{,}642{,}700$	0	$n'_o = 0$

Table 9. Cholecystitis and diabetes, hospital population expected numbers for $s_c = 0.15$, $s_d = 0.05$, $s_r = 0.20$

	Cholecystitis	Not cholecystitis	Total
Diabetes	626	6,693	7,319
Refractive errors	9,504	192,060	201,564
Total	10,130	198,753	208,883
Cholecystitis in diabetic group			8.55%
Cholecystitis in control group (refractive errors)			4.72%
Difference			+3.83%

It appears from the development that it is hazardous to apply in a hospital population the method of the fourfold table analysis for an inquiry into the correlation of diseases. This applies also to other similar problems, as for instance whether the incidence of say, heart disease, is different for laborers and farmers, if it is known that laborers and farmers are not represented in the hospital in the proportion that they occur in the community. However, the formulas given indicate some special cases in which comparison is not basically invalid. If the selective rate for any particular condition is zero, the relative incidence of that condition in several disease groups may be validly examined, regardless of the selective rates affecting the other groups. This refers to inquiries in which for instance eye color or anthropologic type is examined in various disease groups to ascertain whether there is correlation between these characters and disease. If each of the disease groups examined consists of only one disease, for example, diabetes or refractive errors but not both, and if the selective rates for these two groups do not differ appreciably then also it is valid to compare the incidence in them of cholecystitis, even though the latter disease is not fairly represented in the hospital.

Except for such cases there does not appear to be any ready way of correcting the spurious correlation existing in the hospital population by any device that does not involve the acquisition of data which would themselves answer the primary question. For instance the device sometimes used of setting up in the hospital sample a one-to-one control so that both groups examined have the same number of cases and are identical as regards say, age and sex does not touch the difficulties referred to here. It is to be emphasized that the spurious correlations referred to are not a consequence of any assumptions regarding biologic forces, or the direct selection of correlated probabilities, but are the result merely of the ordinary compounding of independent probabilities. The same results as shown here would appear if the sampling were applied to randomly distributed cards instead of patients.

International Journal of Epidemiology, 2014, 515–521
doi: 10.1093/ije/dyu026
Advance Access Publication Date: 28 February 2014

Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it

Jaapjan D Snoep,[1] Alfredo Morabia,[2,3] Sonia Hernández-Díaz,[4] Miguel A Hernán[5,6] and Jan P Vandenbroucke[1]*

[1]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands, [2]Center for the Biology of Natural Systems, Queens College, City University of New York, New York, USA, [3]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA, [4]Department of Epidemiology, [5]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA, USA and [6]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

*Corresponding author. Leiden University Medical Center, Department of Clinical Epidemiology, PO Box 9600, 2300 RC Leiden, The Netherlands. E-mail: J.P.Vandenbroucke@lumc.nl

In 1946, the physician and statistician Joseph Berkson (1899–1982) pointed out that two diseases that are independent in the general population may become 'spuriously associated' in hospital-based case-control studies.[1] This spurious association was later referred to, often in lively debates,[2–14] as Berkson's fallacy, Berkson's paradox or Berkson's bias. Some authors restricted the interpretation of Berkson's fallacy to disease-disease associations,[2,5,7,8] whereas others thought that the fallacy would also apply to exposure-disease associations in hospital-based case-control studies.[10–15]

In this article we use directed acyclic graphs (DAGs) to describe the structure of Berkson's fallacy, first for disease-disease associations and then for exposure-disease associations. This permits us to understand the contentious debates and strongly differing opinions about Berkson's fallacy, and has practical implications for study design and interpretation (see Box 1).

Disease-disease associations: Berkson's fallacy

In 1946 Berkson considered the following problem.[1] Suppose a hospital wants to estimate the association between the prevalences of cholecystitis (disease 1 or D1) and diabetes mellitus (disease 2 or D2). To do so, a case-control study is conducted in which hospitalized individuals are included as cases if they have diabetes and as controls if they have ophthalmological refractive errors (disease 3 or D3). The association between cholecystitis and diabetes is then estimated by comparing the prevalence of cholecystitis D1 between cases with diabetes D2 and controls with refractive errors D3.

Berkson constructed his example so that, in the source population, the D1-D2 and D1-D3 associations were null and the probabilities of hospitalization for each of the three diseases were independent. Yet, the D1-D2 association was not null in hospitalized individuals. In the Appendix (available as Supplementary data at *IJE* online) we numerically work out the example Berkson used in his paper, and we discuss the strength and direction of the association in hospitalized individuals. Intuitively, this association arises because persons with two or more diseases have a higher probability of being hospitalized than persons with only one disease—even if these reasons are independent.