REVIEW ARTICLE

Interpreting Results in 2 × 2 Tables

Part 9 of a Series on Evaluation of Scientific Publications

Wilhelm Sauerbrei, Maria Blettner

SUMMARY

Background: The findings of epidemiological studies, diagnostic tests, and comparative therapeutic trials are often presented in 2 × 2 tables. These must be interpreted correctly for a proper understanding of the findings.

Methods: The authors present basic statistical concepts required for the analysis of nominal data, referring to standard works in statistics.

Results: The relative risk and odds ratio are defined to be indices for the relationship between two binary quantities (e.g., exposure—yes/no and disease—yes/no). The topics dealt with in this article include the effect of sample size on the length of the confidence interval and the p-value, and also inaccuracies caused by measuring error. Exposures are often expressed on a three-level scale (none, low, high). The authors also consider the 2×3 table as an extension of the 2 × 2 table and discuss the categorization of continuous measurements. Typically, more than one factor is involved in the development of a disease.
The effect that a further factor can have on the observed relationship between the exposure and the disease is discussed.

Conclusions: Sample size, measurement error, categorization, and confounders influence the statistical interpretation of 2 × 2 tables in many ways. Readers of scientific publications should know the inherent problems in the interpretation of simple 2×2 tables and check that the authors have taken these into account adequately in analyzing and interpreting their data.

Key words: publications, clinical research, epidemiology, statistics, clinical trial

Cite this as: Dtsch Arztebl Int 2009; 106(48): 795–800
DOI: 10.3238/arztebl.2009.0795

Institut für Medizinische Biometrie und Medizinische Informatik, Universitätsklinikum Freiburg: Prof. Dr. rer. nat. Sauerbrei

Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsklinikum Mainz: Prof. Dr. rer. nat. Blettner

The results of epidemiological studies, diagnostic test procedures and therapeutic comparisons are often presented as 2 x 2 tables. The terms four-field table, contingency table, and cross table are also often used. For example, the British Medical Journal recently published a case control study in which the association between tea consumption and esophageal carcinoma was examined (1). Of the 300 patients with esophageal carcinoma ("cases"), 249 reported that they never, or very rarely, drank green tea; 17 (6.4%) frequently drank green tea. Of the 571 study participants without esophageal carcinoma ("controls"), 356 reported that they rarely drank green tea, in comparison with 30 subjects with regular consumption. The findings were presented in a 2 x 2 table (1). It is remarkable that some of the participants provided no information (missing data). As a second example, let us consider a clinical study with patients with metastatic breast cancer, in which factors investigated included the influence of the prior therapy (2). All patients had received taxanes. Progression was detected in 10 (28.6%) of the 35 patients with prior anthracycline treatment. There was a greater rate of progression (15 out of 26; 57.7%) in the group without prior anthracycline treatment. In the case control study, an alternative classification could be selected for tea consumption, for example "never," "moderate," or "frequent". In the therapeutic study, the tumor response is often classified as "complete remission," "partial remission," or "no change or progression".

The relative risk (RR) or the odds ratio (OR) can be calculated from the simple 2 x 2 table. This is why it is important to understand the central properties of the 2 x 2 table and to know how even simple extensions can change the analysis and interpretation. If this is overlooked, wrong conclusions may be drawn, leading to mistaken assessment of the risk, diagnosis, prognosis, or therapy for the individual patient.

Typically, more than one factor is involved in the development of a disease. For this reason, the analysis should consider more than one potential factor in most situations. For example, not only the type and temperature of the tea should be considered, but also coffee and alcohol consumption. Simple contingency tables are then no longer adequate for the analyses and presentation of the results. The evaluations must be performed

BOX

Measures of association in a 2 × 2 table

Plan and notation of a basic 2 x 2 table

		Disease present		
		Yes (D +)	No (D –)	
Exposed	Yes	a	b	a + b
	No	c	d	c + d
		a + c	b + d	n

Definitions

- Risk describes the probability of falling ill
- P_0 = probability of falling ill for non-exposed persons
- P_1 = probability of falling ill for exposed persons
- $P_0 = c / (c + d)$
- $P_1 = a / (a + b)$

Risk difference: $RD = P_1 – P_0$
Relative risk, risk ratio: $RR = P_1/P_0$

O_0 = odds;
for non-exposed persons: $O_0 = P_0/(1 – P_0)$
O_1 = odds for exposed persons; $O_1 = P_1/(1 – P_1)$
OR = odds ratio = $O_1 / O_0 = (P_1 / [1 – P_1]) / P_0 / [1 – P_0]) = (a \times d) / (b \times c)$
If (a + c)/n is “small,” RR and OR have similar values.

with multivariate models, which simultaneously consider several variables.

In the following sections, we will employ the notation of the 2 x 2 table (see *Box*) and discuss the results of a hypothetical study *(Table 1)*. We will define the terms risk, relative risk, and odds ratio and discuss the influence of the sample size on the length of the confidence interval and on the p-value. We will also explain how measurement errors can lead to bias in the result. As a simple extension, we will then consider the 2 x 3 table and explain how an additional factor can affect the observed correlation between exposure and disease. We will use the term “risk factor,” taken from epidemiology. The same considerations apply analogously to diagnosis, prognosis, and therapy. Sauerbrei and Schumacher (1999) (3) have discussed additional aspects relevant to prognosis studies. For further information, we refer the reader to Fletcher et al. (2005) (4), Altman (1991) (5), Campbell et al. (2007) (6), and Schumacher and Schulgen (2008) (7).

Definitions

Let us consider a group of n persons. We are interested in the following two properties:

- Is the person exposed or not exposed?
- Is the person ill or not?

The word “exposed” is used here to represent various characteristics, such as people exposed to a specific occupational stress, persons with a specific genetic constellation, or persons with values outside the normal range for specific laboratory parameters. In the example above, this means individuals who “frequently” drink green tea. In therapeutic studies, “exposed” can be replaced by “therapy A,” “not exposed” by “therapy B,” “ill” by “no therapeutic success,” and “not ill” by “therapeutic success”.

In *Table 1*, we consider a cohort study with 450 persons, 36 of whom are ill and 414 not ill. Two thirds of the persons (300) are exposed, while one third is not exposed. The incidence rate is 8% (36 of 450) in the total group, 10% (30 of 300) in exposed persons, and 4% (6 of 150) in non-exposed persons.

The incidence rate, risk, relative risk and odds ratio are derived from the 2 x 2 table *(Box)*.

A relative risk of 1 (RR = 1) means that exposed persons (therapy A) and non-exposed persons (therapy B) have the same risk of falling ill (“being cured”). If RR is greater than 1, this means that exposed persons have a higher risk than non-exposed persons. If RR = 1.5, this means that the risk of exposed persons is 50% greater than that of non-exposed persons. If RR = 2, this means that the risk is doubled. In other words, the risk is increased by 100% or increased to 200%. If RR = 0.5, this means that persons in the exposed group only have half the risk of persons in the non-exposed group. This can also be referred to as a “protective factor.” It is important to bear in mind which groups are used as reference. If RR = 1.5 (for example, smokers versus non-smokers), this means that the risk is increased by 50% for smokers. If smokers are used as the reference group, RR = 1/1.5 = 0.67. Thus, in comparison to smokers, the risk for non-smokers is reduced by one third (1 – 0.67 = 0.33).

Aside from the relative risk, the so-called odds ratio (OR) is often used as a measure of association *(Box)*. The odds ratio is the quotient of the chances (odds) of a disease (cure) for persons with or without exposure (therapy).

The relative risk cannot be directly calculated for case control studies. The reason for this is that the ratio of cases to controls is laid down in the design, so that (a + c)/n is fixed by the investigator. It follows that neither a/(a + b) nor c/(c + d) is a useful parameter, as they do not represent the incidence rate. The relative risk cannot then be calculated. The odds ratio may be regarded as an auxiliary construct for the relative risk. The odds ratio and relative risk are of about the same numerical size when the probabilities of disease (P_1 and P_0) are both small. A value of 1% to 5% can still be regarded as small for these calculations. It should be remembered that the odds ratio and relative risk are of about equal size in only these cases. If the relative risk is greater than 1, the odds ratio is always slightly larger than the relative risk. In our case, RR = 2.50 and OR = 2.67.

TABLE 1

Presentation of the results of a hypothetical study in a 2 × 2 table

Total (% column)	(% line)	Disease present Yes (D+)	 No (D-)	
Exposed	Yes	30 (10.0%) (83.3%)	270 (90.0%) (65.2%)	300 (66.7%)
	No	6 (4.0%) (16.7%)	144 (96.0%) (34.8%)	150 (33.3%)
		36 (8.0%)	414 (92.0%)	450

Problem 1: Sample size, confidence interval, and p-value

Aside from the relative risk, many publications give the confidence interval and p-values to summarize the association between two factors. A p-value is said to be statistically significant if it lies below the "magic" limit, which is often 5%. If the RR is fixed, the confidence interval and the p-value depend on the sample size *(Table 2)*. In our example, the estimated RR = 2.5, with a 95% confidence interval (CI) of 1.06–5.87. The test for an association between the two factors—the chi-square test for independence—gives a p-value of 0.027. If the sample size is doubled (n = 900) or halved (n = 225), the estimate is unchanged. On the other hand, the confidence interval becomes narrower and the p-value smaller as n increases from 225 (p = 0.118) to 900 (p = 0.002). At n = 225, the value 1.00 is contained in the confidence interval, so that the effect of exposure is not statistically significant. When interpreting the p-value, the estimate of the relative risk, the sample size, and the confidence interval should be considered together.

Problem 2: Effect of measurement error on the relative risk

We would like to show how an error in the classification of exposed and non-exposed persons can influence the result of the 2 x 2 table *(Table 3)*. If we assume that (only) 10% of all subjects are wrongly classified, an avarage of 30 exposed cases are wrongly classified as non-exposed subjects. Moreover, about 15 non-exposed subjects are wrongly assigned to the exposed group. We will assume that this error is independent of the status of the disease (no differential misclassification). In this case, three exposed cases (10% of the 30 misclassifications) and one non-exposed case (4% of 15 persons, giving 0.6 case, rounded up to one case) are wrongly classified. Because of this measurement error, the data in *Table 3* gives RR = 2.03 (95% CI = 0.95–4.34, p-value = 0.061). The result is therefore "non-significant." Thus we have shown that a misclassification which is the same for cases and controls (a non-differential misclassification) leads to an underassessment of RR. It is however rarely justified to assume that any misclassification is non-differential. It follows that it is in any case necessary to perform a detailed investigation of the effect of potential measurement errors.

Problem 3: Exposure in more than two steps

In the above case control study, the frequency of tea consumption was classified into three groups (never moderate, frequent) (1). We will extend our contingency table to a 2 x 3 table, with the exposure in three categories *(Table 4)*. The risk of the high exposure group relative to non-exposed persons was 2.8 and relative to the low exposure group was 2.0. The influence of the exposure on the disease is more marked than in *Table 1*. If, however, a 2 x 2 chi-square test is (wrongly) applied to a 2 x 3 table, the association is no longer significant. In the 2 x 2 table, a 2 x 2 chi-square test of independence has only one degree of freedom, whereas it has two degrees of freedom in the 2 x 3 table. For a fixed level of significance, the critical value is greater in the 2 x 3 table. However, with this procedure, the fact is ignored that the three degrees of exposure are sequential (absent, low, high). A suitable test of trend should consider this sequence. What is important is that the categories should be fixed prior to the evaluation, on the basis of objective and biometric arguments. Because of the problems with multiple testing, it is totally unacceptable to perform retrospective "searching" to attain a smaller p-value (and a "significant" result) (8).

Problem 4: Categorization of continuous variables

Although exposure is frequently measured as a continuous variable (i.e. a variable with many possible values, such as blood pressure), the evaluation is often based on categorical data (high, intermediate, low). There are many disadvantages in categorizing continuous variables by classifying class limits. Firstly, some of the

TABLE 2

Influence of the sample size N on the length of the confidence interval and on the p-value

N	RR	95% CI	p-Value	Interpretation
225	2.5	0.75–8.37	0.118	Non-significant
450	2.5	1.06–5.87	0.027	Significant
900	2.5	1.37–4.57	0.002	Highly significant

CI, confidence interval

TABLE 3

Influence of a 10% classification error (non-differential misclassification) on the 2 ×2 table as in Table 1

	D+	D–	
E+	28 (9.8%)	257 (90.2%)	285 (63.3%)
E–	8 (4.8%)	157 (95.2%)	165 (36.7%)
	36 (8.0%)	414 (92.0%)	450

TABLE 4

Classification of exposure into three categories

a) Ordinal factor and resulting 2 × 3 table

	D+	D–	
E+ high	22 (11.0%)	178 (89.0%)	200 (44,4%)
E+ low	8 (8.0%)	92 (92.0%)	100 (22,2%)
E–	6 (4.0%)	144 (96.0%)	150 (33,3%)
	36 (8.0%)	414 (92.0%)	450

b) Results of different analyses

Summary of E+	RR	95% CI	p-Value
Yes (see Tables 1 and 2)	2.5	1.06–5.87	0.027
No			0.058[a] 0.020[b]
E+ low vs. E – E+ high vs. E –	2.0 2.8	0.72–5.59 1.14–6.61	
E+ high vs. E+ low	1.4	0.64–2.98	

[a] Chi-quare test of independence in 2 × 3 table;
[b] Test of trend with score values (0 – E–, 1 – E+ low, 2 – E+ high);
CI, confidence interval

originally recorded information is not used. This loss is at its greatest if there are low numbers of categories—for example, a threshold for the classification as “high” or “low.” Moreover, the number of suitable categories and their limits must be specified. If categories are chosen for which the number of cases is too low, the estimate of the effect for these categories is unstable. If target variables are considered when specifying the limits, this can lead to a marked overestimate of the effect and to false p-values. Altman et al. (1994) (9) have shown that there are a wide variety of problems associated with the popular “optimal” cut-off point approach, in which, depending on the data, different threshold values are examined for categorization as “high” or “low”. If many different cut-off points are considered, this greatly increases the error of the first type. In other words, a significant result is found, although the factor investigated does not in fact influence the target variable. Instead of the assumed probability of error of 5%, multiple application of tests leads to a probability of error of almost 50% (9).

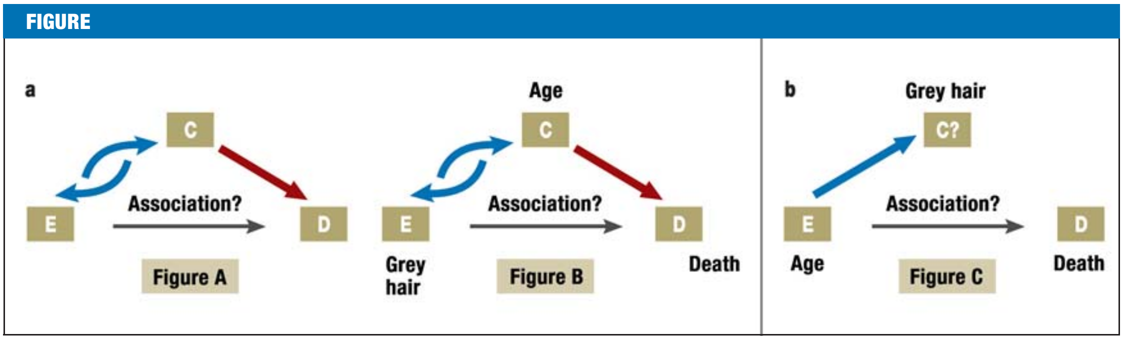
With continuous risk factors, it is better to estimate the dose-effect curve, rather than to perform categorization (10).

Problem 5: Influence of a third factor

Confounding and Simpson’s paradox—In many studies, the influence of exposure on a disease is influenced by an additional factor (*Figure*). We assume that the results given in *Table 1* were recorded in a group of non-smokers. There are however also data for a second group (smokers) (*Table 5*). RR is greater than 1 in both groups, so that the disease is more frequent after exposure.

If these groups are not considered in the evaluation, and the figures for the two groups are simply added, an estimate is obtained for the relative risk which is less than 1. This phenomenon is known as Simpson’s paradox. The reason for this is that the distribution of exposure in the two groups is different, as is the risk of disease. Thus, addition of the tables is not justified. However, when evaluating studies with several important influencing factors, it is often not evident which factors should be considered in the analysis. The association (or correlation) between the different influencing factors may be important here. Variables are known as confounders if they are correlated with both exposure and disease. The classical procedure for dealing with categorical confounders is the so-called Mantel-Haenszel statistic, based on stratification of the data according to the confounder variables. The Mantel-Haenszel statistic is a weighted mean of the odds ratios for the individual categories, with the weights depending on their sizes. It is evident that this procedure can become highly complex if there are several confounders. In particular, it may lead to some categories being occupied with only a few cases and controls. Then modelling must be performed with regression models. The logistic regression model has become the established approach in medicine for the simultaneous investigation of several factors influencing binary target variables.

FIGURE



Depiction of the influence of a potential confounder; if we investigate the association between grey hair and death, age is a logical confounder. If another scientist were to investigate the association between age and death, he would certainly be amazed by the influence of grey hair

Interaction—We have been assuming that the additional factor (here smoking) is not of primary interest, but influences the association between exposure and disease and must therefore be considered in the evaluation. It is however often the case that we are not only interested in the effects of the individual factors on the disease, but also their combined effect. It is often assumed in medical research that factors act multiplicatively. This means that if two or more factors are present, the relative risk is calculated as the product of the individual relative risks. If the result of the study is very different from the calculated product, there is an interaction between the factors. Minor deviations are always observed in real studies. A test for interaction can be used to investigate whether the deviations are random or statistically significant.

In the present example, the variable "smoking" is also of interest as second factor. If exposure E is not considered, the calculated RR for smokers is (60/150)/(36/450) = 5.0 (*Table 6a*). Conversely, *Table 5b* shows that the RR for exposure is 0.83, if the smoking status is not considered. A multiplicative effect means that subjects who not only smoke but have also been exposed to E have an increased risk of 0.83 x 5.0 = 4.15, in comparison to non-smokers without exposure E.

Table 6b shows that the risk is in fact increased by the factor of (20/40)/(6/150) = 12.5. There is thus a deviation from multiplicativity, i.e., an interaction between the two factors. The other publications contain more detailed discussions of this topic and present suitable methods for investigating interactions (11–13).

Discussion

Every publication of a clinical or epidemiological study should contain a simple descriptive presentation of the results (14). In many cases, the 2 x 2 table is a sufficiently clear method to present the principle results. On the other hand, there are some catches in interpreting this apparently simple table. The reader of a scientific publication should be aware of these and make sure that the authors have drawn proper attention to possible problems.

Conflict of interest statement
The authors declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 25 March 2009, revised version accepted on 16 June 2009.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Islami F, Pourshams A, Nasrollahzadeh D, et al.: Tea drinking habits and oesophageal cancer in a high risk area in northern Iran: population based case-control study. BMJ 2009; 338: b929. Doi: 10.1136/bmj.b929
2. Andreetta C, Puppin, C, Minisini A, et al.: Thymidine phosphorylase expression and benefit from capecitabine in patients with advanced breast cancer. Annals of Oncology 2009; 20: 265–71.
3. Sauerbrei W, Schumacher M: Aspekte der statistischen Evaluation neuer Prognosefaktoren: Illustration bei Studien in der Onkologie. Geburtshilfe und Frauenheilkunde 1999; 59: 483–7.

TABLE 5

Influence of an additional factor

a) Influence of E in 2 strata

Stratum I (non-smokers); RR = 2.5

	D +	D –	
E +	30 (10.0%)	270 (90.0%)	300 (66.7%)
E –	6 (4.0%)	144 (96.0%)	150 (33.3%)
	36 (8.0%)	414 (92.0%)	450

Stratum II (smokers); RR = 1.38

	D +	D –	
E +	20 (50.0%)	20 (50.0%)	40 (26.7%)
E –	40 (36.0%)	70 (63.6%)	110 (73.3%)
	60 (40.0%)	90 (66.0%)	150

b) Influence of E

without considering the strata; RR = 0.83

	D +	D –	
E +	50 (14.7%)	290 (85.3%)	340 (56.7%)
E –	46 (17.7%)	214 (82.3%)	260 (43.3%)
	96 (16.0%)	504 (84.0%)	600

TABLE 6

Relative risk for smokers and interactions between E and smoking (derived from Table 5a)

a) Influence of smoking; exposure E not considered; RR = 5.00

	D +	D –	
Non-smokers	36 (8.0%)	414 (92.0%)	450 (75.0%)
Smokers	60 (40.0%)	90 (60.0%)	150 (25.0%)
	96 (16.0%)	504 (84.0%)	600

b) Combined influence of smoking and exposure E; RR = 12.5; the combinations non-smoker/E+ and smoker/E– are not given

	D +	D –	
(Non-smokers, E –)	6 (4.0%)	144 (96.0%)	150
(Smokers, E +)	20 (50.0%)	20 (50.0%)	40
	26 (13.7%)	164 (86.3%)	190

4. Fletcher RH, Fletcher SW: Klinische Epidemiologie. 2. Aufl. Bern: Huber 2007.

5. Altman DG: Practical statistics for medical research. London: Chapman and Hall 1991.

6. Campbell MJ, Machin D, Walters SJ: Medical statistics—a textbook for the health sciences. 4. Aufl. Chichester: Wiley 2007.

7. Schumacher M, Schulgen G: Methodik klinischer Studien – Methodische Grundlagen der Planung, Durchführung und Auswertung. 3. Aufl. Berlin: Springer 2008.

8. Victor A, Elsäßer A, Hommel G, Blettner M: Wie bewertet man die p-Wert-Flut – Hinweise zum Umgang mit dem multiplen Testen. 2009 (Deutsches Ärzteblatt, in press)

9. Altman DG, Lausen B, Sauerbrei W, Schumacher M: Dangers of using „optimal" cutpoints in the evaluation of prognostic factors. J Net Cancer Inst 1994; 86: 829–35.

10. Royston P, Sauerbrei W: Multivariable model-building—a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley 2008.

11. Altman DG, Matthews JNS: Interaction 1: heterogeneity of effects. BMJ 1996; 313: 486.

12. Assmann SF, Pocock SJ, Enos LE, Kasten LE: Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064–9.

13. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM: Statistics in medicine—reporting of subgroup analyses in clinical trials. The New England Journal of Medicine 2007; 357: 2189–94.

14. Spriestersbach A, Gerhold-Ay A, du Prel JB, Röhrig B, Blettner M: Deskriptive Statistik. Dtsch Arztebl Int 2009; 106(36): 578–83.

Corresponding author
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI)
Klinikum der Universität Mainz, 55101 Mainz, Germany
blettner@imbei.uni-mainz.de