Determination of the effect size of an observed factor based on a multi-variate model for evaluating practical significance of differences between two groups of case-control design

Liu Hui

College of Medical Laboratory, Dalian Medical University, Dalian 116044, China

Address:

Professor Liu Hui

College of Medical Laboratory, Dalian Medical University, Dalian 116044, P.R. China

E-mail: liuhui60@sina.com

Secondary E-mail: liuhui60@dlmedu.edu.cn

Tel: 86-411-86110390

Abstract

Background: To determine the effect size of an observed factor for a disease by using consistency in a cohort study (CRC) for evaluating practical significance of differences between two groups of case-control design.

Methods: A model of multiple pathogenic factors was established by analyzing the number and distribution of observed factors in a study population. The difference in the incidence between two groups (exposed and unexposed) was calculated according to the model as CRC. The relationship of Youden's index and true and false-positive ratio (TFR) in case-control design were observed with CRC.

Results: The CRC was able to correctly reflect the number of factors combined in the models, and therefore, indicates that CRC is a reasonable indicator of effect size. Difference scores <0.25 indicate that one of four or more factors plays a role in a disease; scores >0.50 indicate one of two factors plays a role in disease and implies a high intensity level of the factor. TFR could correctly reflect CRC. Accordingly, a factor with an effect size (i.e., TFR) less than 6.0 should not be considered a clinically significant factor, even if the observed difference is statistically significant.

Conclusions: A CRC over 0.25 OR TFR over 6.0 is suggested as an indicator of a substantial effect size.

Keywords: case-control design; analysis model; cohort study; absolute risk transformation; outcomes

Background

Complex events, those in which many factors exert synergetic effects, are frequently observed not only in medical practice but also in our everyday lives. Currently, the pathogenesis of most diseases is related to interactions among extrinsic and intrinsic suspected factors [1-3]. The case-control study is a common method for examining etiological factors associated with rare diseases. In case-control studies, the potential relationship between a suspected factor is examined by comparing difference of frequencies of this factor between the diseased and non-diseased subjects with using statistical methods. However, a statistical difference does not indicate the strength of the effect of an observed factor on a disease.

Quantitative variations in a particular event are normally distributed in terms of changes in ratios [4-6] and absolute values [7-9], such as odds ratio (OR) [10] and Youden's index (Y) [11]. When the values of cardinal numbers are relatively small, an increase in a ratio may be very high although the absolute increase may not be. In contrast, when the values of cardinal numbers are relatively large, an increase in a ratio may not be high but the absolute increase may be highly significant. Thus, ratios and absolute values are not comparable [12]. Therefore, it is important to evaluate which effect size of an observed factor indicator (OR or Y) is a better measure of the association of the suspected factor with the disease.

Cohort studies, which observe the association between a specific factor and a disease, are considered to be the most reliable form of scientific evidence in the hierarchy of epidemiological evidence [13-15]. In such a study, a putative suspected factor is used as an exposure variable, the exposed and unexposed study participants are observed until they

develop the outcome of interest. This can be done by comparing the difference in the frequency of disease occurrence between the exposed and unexposed groups in cohort studies, which indicates the strength of the effect of the observed factor (its effect size) on the disease. Here, we propose a multiple risk-factor model for cohort studies to evaluate more reliable measures of the strength of the association, or functional intensity, between suspected factors and outcomes. We believe such a model has the potential to solve the aforementioned problem.

Methods

Multiple risk-factor model

The basic assumptions of the analytical model are: (1) the prevalence of the different observed factors is independent of each other and play a role in a superimposed manner, regardless of interaction or weight function; and (2) a chronic disease is a continuous process of the superimposed manner of suspected factors.

A four-factor model simulating pathogenic data was established. Four sets of random numbers with binomial distributions (P = 0.5, N = 100,000) were generated using SPSS statistical software. The four sets of data, which were independent of each other, were named A, B, C, and D. By adding the four sets of data to create group results for the ABCD group, group A can be regarded as a factor of the ABCD group, as shown in Figure 1. The highest value of ABCD was used as the denominator to convert value of ABCD from 0 to 1. A higher number of suspected factors indicates a higher probability of disease. Hence, A can be regarded as a cause of ABCD; this model was named the four-factor model, which has a probability of 0.5.

Figure 1

In the same way, four-factor models simulating pathogenic data in which the probability of the suspected factor was 0.01 and 0.001 in the study population (four-factor models with 0.01 and 0.001) were established to observe the influence of suspected-factor distribution on differences in incidence between the two groups.

Evaluation of effect size

In a similar way, a three-factor model with 0.5 (A vs ACB) and a two-factor model with 0.5 (A vs AB) were established to evaluate the differences in the magnitude of the associations between the observed factors and outcomes. Using the A group as the cause group and ABCD, ABC, and AB as results, a cohort study was established to generate simulated results. The difference in the observed occurrence of disease between the two groups was then calculated to evaluate the effect sizes.

Evaluation of odds ratio and Youden's index

We assumed that the frequencies of a genetic marker (gene) were distributed in disease and control groups as shown in Table 1.

Table 1

The following equations were used to determine the Youden's index (Y) and odds ratios (OR)[10, 11]

Y = a+d-1=a+(1-b)-1=a-b

OR = (a*d)/(b*c)=[a*(1-b)]/[b*(1-a)]

Meanwhile, we also suggested the true and false-positive ratio (TFR) in a case-control study as follows:

TFR = a / b

The basic principle of the analysis model is to comprehensively consider which of Y, OR and TFR could correctly reflect consistency in a cohort study (CRC).

The CRC is the sum of the incidence in the exposure group and the healthy rate in the non-exposure group minus 1 as follows:

CRC = Pe-(Pn-1)-1 = Pe-Pn

where Pe and Pn represent the incidence in the exposed group and non-exposed group, respectively, from the cohort study.

Evaluation of Y, OR and TFR in case-control study based on CRC from cohort study was performed using special numbers. A definite relationship between cohort outcomes and that from case-control study is as follows [16]:

$$Pe = \frac{Pd * m}{Pc * (1 - m) + Pd * m}$$

$$Pn = \frac{m * (1 - Pd)}{1 - Pc * (1 - m) - Pd * m}$$

where Pe and Pn represent the incidence of the exposure group and that of non-exposure group, respectively, in the cohort study; Pd and Pc represent the frequencies of the observation factor in disease group and in the control group, respectively, in the case-control study; and “m” represents the incidence in the total population and is assigned to a value of

5% in the present study because e a chronic disease usually is a low probability event.

Results

The differences in incidence between the two groups in the four-factors model with probabilities of 0.5, 0.01, and, 0.001 are listed in Table 2. A value of approximately 0.25 was obtained from all three of the models, indicating that the distribution of observed factors in the population did not influence the differences in disease incidence between the two groups (CRC).

Table 2

The CRC derive from the four-, three-, and two-factors models are shown in Table 3. That CRCs derive from the four-, three-, and two-factor models were approximately 0.25(1/4), 0.33(1/3), and 0.50(1/2), which can be considered that CRC correspond to the number of factors combined in the models and is reasonable measures of effect sizes.

Table 3

The data generation in test set was based on Y (Y=0.4), OR (OR=4) and TFR (TFR=4) as shown in Table 4. CRCs deriving from the same TFR with different cardinal number were similar; however, that were not closed for Y and OR, indicating TFR could correspond to CRC.

Table 4

The data for the test set were generated based on deferent TFR for low-probability event as shown in Table 5. Result showed that CRC increase with increasing TFR, suggesting that TFR could reflect CRC. TFR with 20 could provided CRC value of 0.513 (when the incidence was assumed as 0.05 for a disease); When TFR=200, CRC could reach 0.910 (incidence=0.05).

Table 5

Discussion

The present study employed models of multiple pathogenic factors that examined the effects of the number factors and the distribution of factors in the population. The results of the models indicate this methodology can be used as a pragmatic, common-sense approach to intuitively understanding the roles of observed factors in complex biological events. We found the distribution of the observed factors in the population had no influence on the differences in the incidence of disease between two groups. We also found the difference in incidence between the two groups correctly reflected the number of factors combined in the models, and therefore, the difference in incidence between the two groups (CRC) can be considered a reasonable indicator of effect size, which can be used to evaluate the intensity of an observed factor.

The effect size of parents should play one of four roles in a child according to Mendelian pattern [17], therefore, we propose that an effect size (CRC): less than 0.25 indicates a weak intensity factor (which can be understood as one of more than four factors playing roles in a disease under the standard model); ranging from 0.25 to 0.50 indicates a moderate intensity factor (which implies that one or two of three factors plays a role in a disease); and over 0.50 indicates a high intensity factor (which implies that one factor mainly plays a role in a disease); values in excess of 0.75 indicate only one observed factor plays a role in a disease. Thus, the intensity of a particular observed factor can reasonably be quantified by the obtained effect size.

Youden's index is the common index used to evaluate the effectiveness of a biomarker or diagnosis made using a biomarker [11]. However, the result showed that Y does not truly reflect the intensity of a suspected factor on a disease outcome, based on the CRC deriving from multiple pathogenic-factor model and TFR could truly reflect the intensity of a suspected factor on a disease outcome. We propose that TFR provides a better method of evaluating the suspected represented by different conditions that arise in populations in case-control study, even though change in absolute values in two group of case-control design is widely used. TFR with 20 could provided CRC value of 0.513 (when the incidence was assumed as 0.05 for a disease); When TFR=200, CRC could reach 0.910 (incidence=0.05). Accordingly, we do not think a factor with an effect size (i.e., CRC) less than 6.0 should be considered a clinically significant factor, even if the observed difference is statistically significant. We suggest a TFR over 6.0 is a substantial effect size because such factors could be further investigated and that over 20 could be used for prediction.

Conclusions

A CRC over 0.25 OR TFR over 6.0 is suggested as an indicator of a substantial effect size. As disease occurrence is a small probability event, incidence usually is less than 0.05; it is difficult to find a biomarker with TFR >20. Apparently, it is necessary to use two or more markers combined. We also think that results deriving from case-control study may overestimate the effect of an observed factor on a disease. Therefore, evaluating effect sizes using TFR deriving from CRC based on a model of multiple pathogenic factors could increase our understanding of quantitative variations in measures of association using new concepts.

Abbreviations

CRC: consistency in a cohort study; OR: odds ratio; TFR: true and false-positive ratio; Y: Youden's index

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publish

Not applicable

Availability of data and materials

The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests

None declared

Funding

None

Authors' Contributions

L.H. conceived the analysis and wrote the final version of the manuscript. I have read and approved the manuscript.

Acknowledgments

We would like to thank the native English speaking scientists of Elixigen Company (Huntington Beach, California) for editing my manuscript.

References

1. Hui L. Quantifying the effects of aging and urbanization on major gastrointestinal diseases to guide preventative strategies. BMC Gastroenterol. 2018;18(1):145.
2. Hui L. Assessment of the role of ageing and non-ageing factors in death from non-communicable diseases based on a cumulative frequency model. Sci Rep. 2017;7(1):8159.
3. Wenbo L, Congxia B, Hui L. Genetic and environmental-genetic interaction rules for the myopia based on a family exposed to risk from a myopic environment. Gene. 2017; 626:305-8.
4. Lemans JVC, Wijdicks SPJ, Boot W, Govaert GAM, Houwert RM, Öner FC, Kruyt MC.

Intrawound Treatment for Prevention of Surgical Site Infections in Instrumented Spinal Surgery: A Systematic Comparative Effectiveness Review and Meta-Analysis. Global Spine J. 2019;9(2):219-230.

5. Zhu X, Wu S. Risk of hypertension in Cancer patients treated with Abiraterone: a meta-analysis. Clin Hypertens. 2019;25:5.
6. Hui L, Jun T, Jing Y, Yu W. Screening of cerebral infarction-related genetic markers using a Cox regression analysis between onset age and heterozygosity at randomly selected short tandem repeat loci. J Thromb Thrombolysis. 2012;33(4):318-21.
7. Scribani M, Norberg M, Lindvall K, Weinehall L, Sorensen J, Jenkins P. Sex-specific associations between body mass index and death before life expectancy: a comparative study from the USA and Sweden. Glob Health Action. 2019;12(1):1580973.
8. Hydes TJ, Burton R, Inskip H, Bellis MA, Sheron N. A comparison of gender-linked population cancer risks between alcohol and tobacco: how many cigarettes are there in a bottle of wine? BMC Public Health. 2019;19(1):316.
9. Nelson M. Management of "Hypertension" Based on Blood Pressure Level Versus an Absolute Cardiovascular Risk Approach. Curr Hypertens Rep. 2019;21(1):6.
10. Qi X, Yu Y, Ji N, Ren S, Xu Y, Liu H. Genetic risk analysis for an individual according to the theory of programmed onset, illustrated by lung and liver cancers. Gene. 2018;673:107-111.
11. Hui L, Liping G. Statistical estimation of diagnosis with genetic markers based on decision tree analysis of complex disease. Comput Biol Med. 2009;39(11):989-992.

12. Xiaojun J, Hui L. An Angle Compared Index with Hybrid of Changes in the Ratio and Amplitude for Quantitative Evaluation of Disease Risk, Biological Function,and Biomarker Efficacy. Biomed Res Int. 2019;2019:8693719.
13. Durr-E-Sadaf. How to apply evidence-based principles in clinical dentistry. J Multidiscip Healthc. 2019;12:131-136.
14. Wallace DK. Evidence-based medicine and levels of evidence. Am Orthopt J. 2010;60:2-5.
15. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg. 2011;128(1):305-10.
16. Liu Hui. Analysing the relationship between cohort and case-control study results based on model for multiple pathogenic factors. Comput Math Methods Med. 2019;2019:7507043.
17. Yang X, Xiaojun J, Yixuan Z, Hui L. Genetic rules for the dermatoglyphics of human fingertips and their role in spouse selection: a preliminary study. Springerplus. 2016;5(1):1396.

Tables

Table 1: Frequency distribution of genetic markers in disease and control groups (%).

Genetic marker	Groups		Total
	Disease	Control	
Carrying gene	a	b	a+b
Not carrying gene	c	d	c+d
Total	100	100	200

Table 2 Influence of the distribution of the observed factor in the population on the consistency in a cohort study (CRC)

Observed factor			Incidence		CRC	
Distribution	Number combined	Groups	Mean	Median	Mean	Median
0.500	4	Exposed	0.624	0.500	0.250	0.250
		Unexposed	0.374	0.250		
0.010	4	Exposed	0.258	0.250	0.251	0.250
		Unexposed	0.007	0.000		
0.001	4	Exposed	0.250	0.250	0.249	0.250
		Unexposed	0.001	0.000		

Table 3 Evaluation of consistency in a cohort study (CRC) with using the four-, three-, and two-factors models

		Results (ABCD)		CRC (Exposed-Unexposed)	
Model	Cause (A)	Mean	Median	Mean	Median
Foure factor	Exposed	0.624	0.500	0.250	0.250
A vs ABCD	Unexposed	0.374	0.250		
Three factor	Exposed	0.665	0.333	0.333	0.334
A vs ABC	Unexposed	0.332	0.667		
Two factor	Exposed	0.749	0	0.500	0.500
A vs AB	Unexposed	0.249	0.500		

Table 4 The simulated results in case-control study to observe relationship among Youden's index (Y), odds ratios (OR), true and false-positive ratio (TFR) and consistency in a cohort study (CRC)

Observed factor in two groups					
Disease	Control	Y	OR	TFR	CRC
0.800	0.500	0.300	**4.000**	1.600	0.057
0.400	0.143	0.257	**4.000**	2.800	0.093
0.200	0.059	0.141	**4.000**	3.400	0.109
0.050	0.013	0.037	**4.000**	3.850	0.120
0.800	0.400	**0.400**	6.000	2.000	0.078
0.700	0.300	**0.400**	5.444	2.333	0.087
0.600	0.200	**0.400**	6.000	3.000	0.111
0.500	0.100	**0.400**	9.000	5.000	0.180
0.800	0.200	0.600	16.000	**4.000**	**0.161**
0.400	0.100	0.300	6.000	**4.000**	**0.140**
0.200	0.050	0.150	4.750	**4.000**	**0.131**
0.040	0.010	0.030	4.125	**4.000**	**0.125**

The boldface numbers indicate similar values.

Table 5 Evaluation of relationship between false-positive ratio (TFR) in case-control study and consistency in a cohort study (CRC)

Observed factor in two groups			CRC	
Disease	Control	TFR	Incidence=0.05	Incidence=0.01
1.000	0.150	6.667	0.260	0.063
1.000	0.100	10.000	0.345	0.092
1.000	0.050	20.000	0.513	0.168
1.000	0.010	100.000	0.840	0.503
1.000	0.025	40.000	0.678	0.288
1.000	0.005	200.000	0.913	0.669
1.000	0.001	1000.000	0.981	0.910

Figure legends

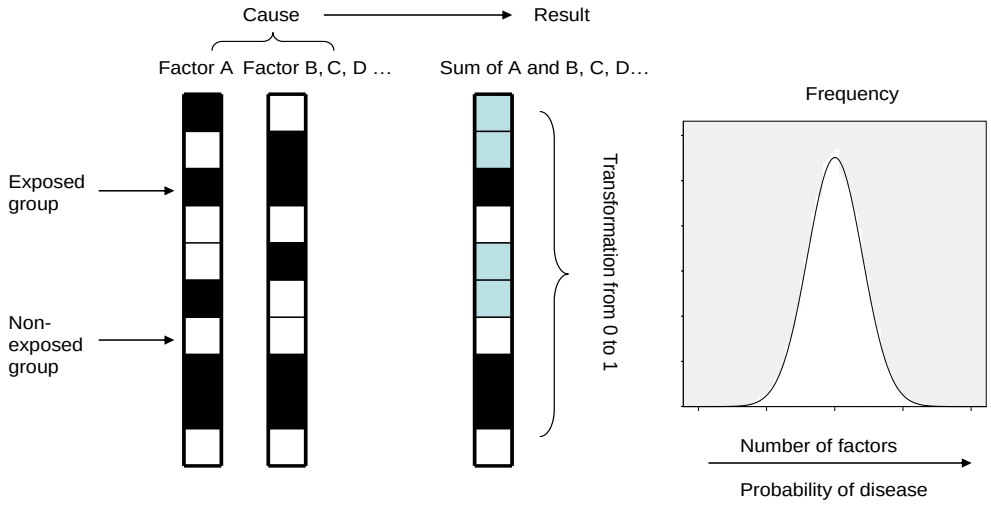

Figure 1. Model of multiple pathogenic factors to assess the association between a causal factor (observed factor) and an ABCD composite factor (disease).