www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Method to estimate relative risk using exposed proportion and case group data

Received: 19 December 2016
Accepted: 10 April 2017
Published online: 18 May 2017

Yoichi Yada

A change in risk of an event occurring, which is affected with a factor, is a common issue in many research fields, and relative risk is widely used because of intuitive interpretation. Estimating relative risk has required data from two follow-up groups and can thus be cost and time consuming. Subjects for whom an event occurred (case group) are often observed but generally analyzed in comparison to those for whom an event did not (control group); however, estimating relative risk using case group data without approximation is hindered. In this study, an obstacle to estimate relative risk using case control data is clarified as a mathematical expression and a new equation to estimate relative risk using the exposed proportion and case group data is proposed. The proposed equation is derived without using the Bayesian methods. A method to estimate the confidence interval for the proposed estimator is also provided. The usefulness of the proposed equation, which requires neither control nor follow-up groups, is demonstrated for both theoretical and real-life examples.

A change in risk of an event occurring associated with exposure to a factor is generally studied in many fields, such as medicine and social science[1,2]. Relative risk (*RR*), also known as "rate ratio", is widely used as a measure of association and can be interpreted intuitively[3,4] because of its simple definition:

$$RR \equiv \frac{\pi_1}{\pi_0}, \tag{1}$$

where π_1 and π_0 are the probabilities of an event occurring (i.e., risks) for subjects exposed and unexposed to a factor. Estimating *RR* requires the estimators of both π_1 and π_0, such as the prevalence or cumulative incidence rate.

The probability estimators can be calculated using existing data of large-scale epidemiological studies or should be obtained from a smaller study designed for the estimation. Let N be the total number of subjects to be studied, such as population, and N_1 and N_0 be the exposed and unexposed parts of N. The N_1 is written as

$$N_1 = N - N_0 = E \cdot N, \tag{2}$$

where E is the exposed proportion. The probabilities of an event occurring can be written as

$$\pi_1 = \frac{N_{11}}{N_1} \text{ and } \pi_0 = \frac{N_{01}}{N_0}, \tag{3}$$

where N_{11} and N_{01} are the numbers of subjects for whom an event occurred among N_1 and N_0. When p_1 and p_0 are the estimators of π_1 and π_0, they should be defined as

$$p_1 \equiv \frac{n_{11}}{n_1} \text{ and } p_0 \equiv \frac{n_{01}}{n_0}, \tag{4}$$

where n_1 and n_0 are the observed numbers of exposed and unexposed subjects and n_{11} and n_{01} are the numbers of subjects for whom the event occurred among n_1 and n_0. Thus, *eRR*, which is defined as

Division of Pharmacology, Department of Biomedical Sciences, Nihon University School of Medicine, 30-1 Oyaguchi-kamicho, Itabashi City, Tokyo, 173-8610, Japan. Correspondence and requests for materials should be addressed to Y.Y. (email: yada67yoichi@gmail.com)

	All subjects			**Cohort data**			**Case control data**		**Random sample data**
	Occurred		**Total**	**Occurred**		**Total**	**Case group**	**Control group**	
	Yes	**No**		**Yes**	**No**				
Exposed	N_{11}	N_1–N_{11}	N_1	n_{11}	n_1–n_{11}	n_1	m_{11}	m_{10}	l_1
Unexposed	N_{01}	N_0–N_{01}	N_0	n_{01}	n_0–n_{01}	n_0	m_1–m_{11}	m_0–m_{10}	l–l_1
Total			N				m_1	m_0	l

Table 1. Contingency tables for all subjects, cohort, case control, and random sample data. "Subjects"(N) comprise "Exposed"(N_1) and "Unexposed"(N_0), both of which include subjects for whom an event occurred (N_{11} and N_{01}). Both of exposed and unexposed cohort (n_1 and n_0) have subjects for whom the event occurred (n_{11} and n_{01}). Exposed subjects (m_{11} and m_{10}) can be found in both of case and control group (m_1 and m_0). Exposed subjects (l_1) can be found in a random sample of the whole subjects (l).

$$eRR \equiv \frac{n_{11} \cdot n_1^{-1}}{n_{01} \cdot n_0^{-1}}, \tag{5}$$

is used as the estimator of relative risk. The groups of n_{11} and n_{01} can be found in groups of exposed and unexposed subjects, who were followed to the event occurring (called "cohort"). However, appropriate cohorts may be occasionally found in epidemiological survey results or should be obtained from a fresh study designed for the purpose (i.e., cohort study).

Unfortunately, few existing results provide appropriate cohorts and long-term observations of cohorts, for example, over several years or decades, are likely to be costly and time consuming, and thus, estimating relative risk can be burdensome for researchers. Meanwhile, because case groups are commonly observed, studies comparing them to a control group (case control study) and estimating the change in risk tend to be less costly and time consuming. Although a case control study is often conducted, estimating relative risk using case control data is hindered. To demonstrate, let m_1 and m_0 be the numbers of observed subjects in a case group and control group and m_{11} and m_{01} be the numbers of exposed subjects in the case and control groups (see Table 1). When $meRR$ is defined similarly to the estimator of relative risk as

$$meRR \equiv \frac{m_{11} \cdot (m_{11} + m_{10})^{-1}}{(m_1 - m_{11}) \cdot (m_1 - m_{11} + m_0 - m_{10})^{-1}}, \tag{6}$$

$meRR$ may be misused as an estimator of relative risk but will largely vary with observing conditions that researchers can designate, such as the size of m_1. Moreover, researchers cannot perceive the effects of those observing conditions. Thus, $meRR$ is not appropriate for the estimation. Although this obstacle for estimating relative risk caused by observation is well known to epidemiologists[1], few studies have clarified the effects of observing conditions as a mathematical expression.

According to Cornfield (1951), relative risk can be approximated using an odds ratio (OR)[5], which is defined as

$$OR \equiv \frac{\pi_1 \cdot (1 - \pi_1)^{-1}}{\pi_0 \cdot (1 - \pi_0)^{-1}} = \frac{N_{11} \cdot (N_1 - N_{11})^{-1}}{N_{01} \cdot (N_0 - N_{01})^{-1}}, \tag{7}$$

when π_0 is small (so-called "rare disease assumption"). Thus, the estimator of OR (eOR), which is defined as

$$eOR \equiv \frac{m_{11} \cdot (m_0 - m_{10})}{m_{10} \cdot (m_1 - m_{11})}, \tag{8}$$

is often computed instead of estimating relative risk. However, OR always overstates the association and the divergence of overstatement depends on RR or π_0[6,7] and thus, using eOR may be misleading.

In addition, some study designs that reduce costs and estimate relative risk were proposed[8–10], although they still require cohorts or the likes. Few studies have focused on deriving the above equations. Zhang and Yu (1998) proposed an equation that can compute relative risk from the odds ratio[11] as follows:

$$RR = \frac{OR}{1 + \pi_0 \cdot (OR - 1)} = OR + \pi_1 \cdot (1 - OR). \tag{9}$$

This equation served as a new method to estimate relative risk using case control data; however, the estimator of π_0 or π_1 is still required to perform the calculation.

Other than above, the Bayesian methods also provide an equation of relative risk. When $\mathrm{P_o}$ and $\mathrm{P_e}$ are the probabilities of finding subjects for whom an event occurred and who were exposed to a factor, the Bayes' theorem[12] can be written as

$$\pi_1 = \frac{\mathrm{P_{eo}} \cdot \mathrm{P_o}}{\mathrm{P_e}}, \tag{10}$$

where $\mathrm{P_{eo}}$ is the probability of finding subjects who were exposed to a factor among subjects for whom an event occurred. Because π_0 can be written as

$$\pi_0 = \frac{(1 - \mathrm{P_{eo}}) \cdot \mathrm{P_o}}{1 - \mathrm{P_e}}, \tag{11}$$

then *RR* is

$$RR = \frac{1 - \mathrm{P_e}}{\mathrm{P_e}} \cdot \frac{\mathrm{P_{eo}}}{1 - \mathrm{P_{eo}}}. \tag{12}$$

However, because $\mathrm{P_{eo}}$ and $\mathrm{P_e}$ will vary depending on methods of observation, precise estimation with using this equation should require follow-up data of all subjects or a carefully collected random sample of that. Moreover, because of difference in probability definitions, such as using "the probability of finding exposed subjects" rather than "the exposed proportion", there is resistance toward the Bayesian methods among some researchers, such as traditional statisticians.

This study illustrates an obstacle, which prevent relative risk from being estimated using case control data, as a mathematical expression of inconsistency in the observations and proposes a new equation to estimate relative risk, which requires case group data and the exposed proportion. The proposed equation is derived without the Bayesian methods, and do not require the probability estimators; that is, neither control groups nor cohorts are needed. Theoretical and real-life examples that demonstrate validity and wide applicability of the proposed equation are also provided.

Results

To clarify an obstacle in estimating relative risk using case control data and derive an equation to estimate relative risk, let us introduce a proportion of observed subjects among all subjects of interest (hereinafter, "observed proportion"). For example, the number of observed individuals exposed to a factor divided by the exposed population constitutes the observed proportion of exposed individuals. As a expression, the observed proportion is the same as "the sampling proportion", which is the proportion of a sample among all subjects of interest. However, the observed proportion cannot be estimated while the sampling proportion can be even assigned by researchers.

In cohort studies, the observed proportions can be defined as follows:

$$OP_{\mathrm{exp}} \equiv \frac{n_1}{N_1} = \frac{n_{11}}{N_{11}} + d_{\mathrm{exp}} \tag{13}$$

and

$$OP_{\mathrm{unexp}} \equiv \frac{n_0}{N_0} = \frac{n_{01}}{N_{01}} + d_{\mathrm{unexp}}, \tag{14}$$

where OP_{exp} and OP_{unexp} are the observed proportions of exposed and unexposed subjects and d_{exp} and d_{unexp} are constants. Cohort studies must be designed as follows:

$$\frac{n_{11}}{n_1} = \frac{N_{11}}{N_1} \left(\Leftrightarrow \frac{n_1}{N_1} = \frac{n_{11}}{N_{11}} \right) \tag{15}$$

and

$$\frac{n_{01}}{n_0} = \frac{N_{01}}{N_0} \left(\Leftrightarrow \frac{n_0}{N_0} = \frac{n_{01}}{N_{01}} \right), \tag{16}$$

such that d_{exp} and d_{unexp} are sufficiently small to be ignored. Inserting equations (13) and (14) into equation (5), we obtain

$$eRR = \frac{\{(OP_{\mathrm{exp}} - d_{\mathrm{exp}}) \cdot N_{11}\} \cdot (OP_{\mathrm{exp}} \cdot N_1)^{-1}}{\{(OP_{\mathrm{unexp}} - d_{\mathrm{unexp}}) \cdot N_{01}\} \cdot (OP_{\mathrm{unexp}} \cdot N_0)^{-1}}. \tag{17}$$

When $d_{\mathrm{exp}} = 0$ and $d_{\mathrm{unexp}} = 0$,

$$eRR = \frac{N_{11} \cdot N_1^{-1}}{N_{01} \cdot N_0^{-1}} = \frac{\pi_1}{\pi_0}. \tag{18}$$

Therefore, *eRR* can be used to estimate the relative risk in cohort studies.

In case control studies, the observed proportions may be defined as follows:

$$OP_{\mathrm{case}} \equiv \frac{m_{11}}{N_{11}} = \frac{m_1 - m_{11}}{N_{01}} + d_{\mathrm{case}} \tag{19}$$

and

$$OP_{\text{cont}} \equiv \frac{m_{10}}{N_1 - N_{11}} = \frac{m_0 - m_{10}}{N_0 - N_{01}} + d_{\text{cont}}, \tag{20}$$

where OP_{case} and OP_{cont} are the observed proportions of case group and control group and d_{case} and d_{cont} are constants. Case control studies must be designed as

$$\frac{m_{11}}{m_1 - m_{11}} = \frac{N_{11}}{N_{01}} \left(\Leftrightarrow \frac{m_{11}}{N_{11}} = \frac{m_1 - m_{11}}{N_{01}} \right) \tag{21}$$

and

$$\frac{m_{10}}{m_0 - m_{10}} = \frac{N_1 - N_{11}}{N_0 - N_{01}} \left(\Leftrightarrow \frac{m_{10}}{N_1 - N_{11}} = \frac{m_0 - m_{10}}{N_0 - N_{01}} \right), \tag{22}$$

such that d_{case} and d_{cont} should be sufficiently small to be ignored. Substituting equations (19) and (20) in equation (8), we obtain

$$eOR = \frac{OP_{\text{case}} \cdot N_{11} \cdot \{(OP_{\text{cont}} - d_{\text{cont}}) \cdot (N_0 - N_{01})\}}{OP_{\text{cont}} \cdot (N_1 - N_{11}) \cdot \{(OP_{\text{case}} - d_{\text{case}}) \cdot N_{01}\}}. \tag{23}$$

When $d_{\text{case}} = 0$ and $d_{\text{cont}} = 0$,

$$eOR = \frac{N_{11} \cdot (N_1 - N_{11})^{-1}}{N_{01} \cdot (N_0 - N_{01})^{-1}} = \frac{\pi_1 \cdot (1 - \pi_1)^{-1}}{\pi_0 \cdot (1 - \pi_0)^{-1}}. \tag{24}$$

Therefore, *eOR* can be used to estimate the odds ratio.

However, inserting equations (19) and (20) into equation (6), we must obtain

$$meRR = \frac{OP_{\text{case}} \cdot N_{11} \cdot \{OP_{\text{case}} \cdot N_{11} + OP_{\text{cont}} \cdot (N_1 - N_{11})\}^{-1}}{OP_{\text{case}} \cdot N_{01} \cdot \{OP_{\text{case}} \cdot N_{01} + OP_{\text{cont}} \cdot (N_0 - N_{01})\}^{-1}} \tag{25}$$

when $d_{\text{case}} = 0$ and $d_{\text{cont}} = 0$. Thus assuming OP_{case} is equivalent to OP_{cont}, *meRR* can estimate the relative risk. Unfortunately, the equivalence of OP_{case} and OP_{cont} cannot be estimated but must be tested.

Equation (25) is a mathematical expression that illustrates an obstacle to estimate relative risk using case control data. Thus, excluding both OP_{case} and OP_{cont} would clearly remove this obstacle in estimating relative risk. Here, let us focus on the exposure odds, which is the ratio of exposed subjects to unexposed ones. Let *EOC* be the exposure odds in a case group and defined as

$$EOC \equiv \frac{m_{11}}{m_1 - m_{11}}. \tag{26}$$

Inserting equation (19) into equation (26) leads

$$EOC = \frac{OP_{\text{case}} \cdot N_{11}}{(OP_{\text{case}} - d_{\text{case}}) \cdot N_{01}}. \tag{27}$$

When $d_{\text{case}} = 0$, substituting equations (2) and (3) into equation (27) leads

$$EOC = \frac{E}{1 - E} \cdot \frac{\pi_1}{\pi_0}. \tag{28}$$

Assume that a random sample is selected from all subjects and *eE* is the proportion of exposed subjects among the sample. Thus, *eE* can be written as

$$eE \equiv \frac{l_1}{l}, \tag{29}$$

where l is the size of a random sample and l_1 is the number of exposed subjects among the sample. The observed proportion of a random sample (that is, the sampling proportion) may be defined as

$$OP_{\text{sample}} \equiv \frac{l}{N} = \frac{l_1}{N_1} + d_{\text{sample}}, \tag{30}$$

where d_{sample} is a constant. Inserting equation (30) into equation (29),

$$eE = \frac{(OP_{\text{sample}} - d_{\text{sample}}) \cdot N_1}{OP_{\text{sample}} \cdot N} \tag{31}$$

Because the random sampling should provide

	Population			Cohort data			Case control data		Random sample data
	Developed		Total	Developed		Total	Case Group	Control Group	
	Yes	No		Yes	No				
Exposed	900	29100	30000	30	970	1000	180	97	300
Unexposed	700	69300	70000	10	990	1000	140	231	700
Total			100000				320	328	1000

Table 2. Model data: population, cohort, case control, and census data. This city, which has a population of 100000, and 30000 individuals exposed to X, includes 900 exposed and 700 unexposed patients who developed Y. Accordingly, 30 and 10 patients should be found when 1000 exposed and 1000 unexposed participants have been observed as cohorts; 180 patients and 97 participants should have been exposed when a case group of 320 and a control group of 328 are observed; and 300 exposed people should be found when 1000 individuals are randomly observed.

$$\frac{N_1}{N} = \frac{l_1}{l}, \tag{32}$$

then d_{sample} is sufficiently small to be ignored. When $d_{\text{sample}} = 0$, inserting equation (2) into equation (31) leads

$$eE = E. \tag{33}$$

Thus, let PRR be defined as

$$PRR \equiv \frac{l - l_1}{l_1} \cdot \frac{m_{11}}{m_1 - m_{11}}. \tag{34}$$

Substituting equations (26) and (29) into equation (34) leads

$$PRR = \frac{1 - eE}{eE} \cdot EOC. \tag{35}$$

Both d_{case} and d_{sample} should be sufficiently small to be ignored when a random sample is selected from all subjects of whom a case group represents an event-occurring part. When $d_{\text{case}} = 0$ and $d_{\text{sample}} = 0$, combining equations (28), (33), and (35), we must obtain

$$PRR = \frac{1 - E}{E} \cdot \left(\frac{E}{1 - E} \cdot \frac{\pi_1}{\pi_0} \right) = \frac{\pi_1}{\pi_0}. \tag{36}$$

Therefore, PRR must be an estimator of relative risk when subjects among whom a case group is observed and subjects from whom a random sample is selected are the same.

This estimator is computed from the exposure odds in a case group and those in all subjects to be studied, and thus, no control group is required. In addition, the estimation is performed without a cohort.

Equation (34) is quite similar to equation (12), but note that PRR was derived without using the Bayesian methods and can be applicable to more general data: data of a case group and a random sample.

Therefore, by considering the observed proportions, an observational inconsistency preventing relative risk from being estimated in the case control studies was clarified as a mathematical expression, and a new equation to estimate relative risk using the exposed proportion and a case group was proposed; the proposed equation requires neither control groups nor cohorts.

Application to Model Data.

Suppose the probabilities of disease Y developing among people exposed and unexposed to chemical compound X are 0.03 and 0.01 (i.e., relative risk is 3).

When the proportion of exposed people in a city, which has a population of 100000, is 30%, researchers should observe the following data: 30 patients are found among 1000 exposed participants and 10 patients among 1000 unexposed participants during a follow-up period; 180 exposed patients are observed in a case group of 320 and 97 exposed participants are observed in a control group of 328; and 300 exposed people are found in a random sample of 1000 participants (see Table 2). The observed proportions of the case and control groups, which are unavailable for the researchers, are then 1/5 and 1/300.

Thus, estimating relative risk from cohort data must be

$$eRR = \frac{30/1000}{10/1000} = 3.00. \tag{37}$$

Estimating odds ratio from case-control data is

$$eOR = \frac{180 \times 231}{140 \times 97} = 3.06 \tag{38}$$

and $meRR$ should be

$$meRR = \frac{180/(180 + 97)}{140/(140 + 231)} = 1.72. \tag{39}$$

Finally, the proposed estimator *PRR* can be computed as

$$PRR = \frac{1000 - 300}{300} \times \frac{180}{140} = 3.00. \tag{40}$$

Note that the proposed equation will estimate the relative risk as precisely as the estimation in a cohort study but does not require follow-up group data, such as cohort data.

Confidence Interval. The proposed estimator *PRR* is the ratio of two odds.

On estimating the odds ratio as $eOR = m_{11} \cdot (m_0 - m_{10}) \cdot m_{10}^{-1} \cdot (m_1 - m_{m11})^{-1}$, the following $eSE(ln\ eOR)$ is known as the maximum likelihood estimator for the standard deviation of ln *eOR*[13]:

$$eSE(\ln eOR) = \sqrt{\frac{1}{m_{11}} + \frac{1}{m_{10}} + \frac{1}{m_1 - m_{11}} + \frac{1}{m_0 - m_{10}}}. \tag{41}$$

Let us apply this formula to *PRR* for estimating confidence interval (CI).

When these two odds are nonzero, the estimator of the standard deviation of the logarithm of *PRR* will be

$$eSE(\ln PRR) = \sqrt{\frac{1}{l - l_1} + \frac{1}{l_1} + \frac{1}{m_{11}} + \frac{1}{m_1 - m_{11}}}. \tag{42}$$

Thus, the following formulas would provide the $100(1-\alpha)\%$ confidence limits for *PRR*.

$$LCL = \exp\left(\ln PRR - Z_{\alpha/2} \cdot \sqrt{\frac{1}{l - l_1} + \frac{1}{l_1} + \frac{1}{m_{11}} + \frac{1}{m_1 - m_{11}}}\right) \tag{43}$$

and

$$UCL = \exp\left(\ln PRR + Z_{\alpha/2} \cdot \sqrt{\frac{1}{l - l_1} + \frac{1}{l_1} + \frac{1}{m_{11}} + \frac{1}{m_1 - m_{11}}}\right), \tag{44}$$

where *LCL* and *UCL* are the lower and upper limits of CI and $Z_{\alpha/2}$ represents the $\alpha/2$ point of the normal distribution, such as 1.96 for 95% interval.

To prove this estimators for CI, computer simulation was conducted. It is assumed that 30% of the population 100000 was exposed. The total number of exposed and unexposed people for whom an event occurred was determined by using two sets of risks, in which the relative risk is 3: $\pi_1 = 0.03$ and $\pi_0 = 0.01$ or $\pi_1 = 0.3$ and $\pi_0 = 0.1$. Samples, exposed case-groups, and unexposed case-groups were picked from the corresponding people based on each six sets of the observed proportions, and the CI was computed each time. Each set of six proportions was chosen so that each group should be close to the size used generally in research.

Table 3 demonstrates the number of times the true relative risk was included in the 95% CI in each one million trials. It is shown that the true value (relative risk: 3) is included at a rate of approximately 95%; this method will well estimate CI.

Application to Real-Life Data. The suicide rate among the youth of Japan is considerably high and suicide accounts for nearly half of the causes of death among those in their twenties[14]. Meanwhile, unemployment is suggested to increase suicide risk[2,15].

The proposed equation was applied to the latest suicide and employment data in Japan as real-life data, and confidence intervals at 95% were also estimated. The prevalence of suicide and employment among individuals in their twenties in 2015 was obtained from a statistics report published by the Ministry of Health, Labour and Welfare[16] and the Labour Force Survey[17]. The data used are presented in Table 4. Suicide victims who were unemployed are treated as "No occupation". Although the Labour Force Survey was conducted in a specific month in 2015 using random sampling, the indicators should represent the characteristics of the Japanese population in that year.

The estimation of relative risk for unemployed women is

$$PRR = \frac{(6.21 - 0.23) \times 1000000}{0.23 \times 1000000} \times \frac{19}{621 - 19} = 0.82, \tag{45}$$

and the 95% confidence interval for this relative risk can be estimated as follows:

$$\begin{aligned} LCL &= \exp\left(\ln 0.82 - 1.96 \cdot \sqrt{\frac{1}{(6.21 - 0.23) \times 1000000} + \frac{1}{0.23 \times 1000000} + \frac{1}{19} + \frac{1}{621 - 19}}\right) \\ &= 0.52 \end{aligned} \tag{46}$$

and

Observed Proportion		Theoretical Number of Exposed Subjects/Total Subjects		Number of Times Including True Value	Rate
Sample	Case Group	Sample	Case Group		
A. ($\pi_1 = 0.03$, $\pi_0 = 0.01$)					
0.01	0.20	300/1000	180/320	953646	95.4%
0.01	0.10	300/1000	90/160	953074	95.3%
0.01	0.01	300/1000	18/32	955724	95.6%
0.10	0.20	3000/10000	180/320	969840	97.0%
0.10	0.10	3000/10000	90/160	961068	96.1%
0.10	0.01	3000/10000	18/32	958187	95.8%
B. ($\pi_1 = 0.3$, $\pi_0 = 0.1$)					
0.01	0.020	300/1000	180/320	938895	93.9%
0.01	0.010	300/1000	90/160	943709	94.4%
0.01	0.002	300/1000	18/32	953717	95.4%
0.10	0.020	3000/10000	180/320	951479	95.1%
0.10	0.010	3000/10000	90/160	951707	95.2%
0.10	0.002	3000/10000	18/32	956232	95.6%

Table 3. Number of times the true value (relative risk: 3.0) was included in 95% confidence interval in each one million trials. For a population of 100000, in which 30000 people was exposed, two sets of risk (A and B) were applied. In A, risk of exposed subjects (π_1) is 0.03 and that of unexposed subjects (π_0) is 0.03; the number of exposed and unexposed subjects for whom an event occurred is 900 and 700. In B, $\pi_1 = 0.3$ and $\pi_0 = 0.1$; 9000 exposed subjects and 7000 unexposed subjects developed an event. Sample, exposed case group, and unexposed case group were picked one million times for each of six sets of observed proportions from the corresponding subjects, and confidence limits were computed each time.

A: Employment situation			**B: Incidence of suicide**		
(million)	**Women**	**Men**	**(real number)**	**Women**	**Men**
Total population	**6.21**	**6.56**	**Total**	**621**	**1731**
Labour force	4.63	5.33	Self-employed or family workers	3	35
Employed person[a]	*4.40*	*5.02*	Employees or office workers	238	892
Unemployed person[a]	*0.23*	*0.30*	Students or pupils	82	307
Not in Labour force	1.57	1.23	No occupation	290	467
Attending school[b]	*0.77*	*1.01*	(Unemployed)[c]	(19)	(62)
Housekeeping[b]	*0.68*	*0.03*	Unknown	8	30
Other[b]	*0.11*	*0.20*			

Table 4. Employment (A) and suicide rate (B) among population aged 20–29 years in Japan, 2015. Under "A: Employment situation", the population is divided into "Labour force" and "Not in labour force". "Labour force" consists of "Employed person" and "Unemployed person" and "Not in labour force" includes "Attending school", "Housekeeping", and "Other". Under "B: Incidence of suicide", suicide victims are divided into five groups: "Self-employed or family workers", "Employees or office workers", "Students or pupils", "No occupation", and "Unknown". In B, "Unemployed" is treated as a part of "No occupation". [a]Labour force. [b]Not in Labour force. [c]No occupation.

$$UCL = \exp\left(\ln 0.82 + 1.96 \cdot \sqrt{\frac{1}{(6.21 - 0.23) \times 1000000} + \frac{1}{0.23 \times 1000000} + \frac{1}{19} + \frac{1}{621 - 19}}\right)$$
$$= 1.30. \tag{47}$$

The estimation for men can be done in the same way. Thus, the estimated relative risk is 0.82 (95% CI: 0.52–1.30) for women and 0.78 (95% CI: 0.60–1.00) for men. Unemployment did not increase the risk of suicide.

Incidentally, the proportions of victims who were classified under "No occupation" are comparatively large for both women and men, and thus, the situation of no occupation might increase risk. Let us, on trial, assume that a person who is neither employed nor attending school is the same as an individual with no occupation. The number of women in no occupation is then 1.04 million (6.21 − 4.40 − 0.77 = 1.04); the estimates of the relative risk and confidence limits for women in no occupation can be computed as follows:

$$PRR = \frac{(6.21 - 1.04) \times 1\,000\,000}{1.04 \times 1\,000\,000} \times \frac{290}{621 - 290} = 4.36, \tag{48}$$

$$\begin{aligned} LCL &= \exp\left(\ln 4.36 - 1.96 \cdot \sqrt{\frac{1}{(6.21 - 1.04) \times 1\,000\,000} + \frac{1}{1.04 \times 1\,000\,000} + \frac{1}{290} + \frac{1}{621 - 290}}\right)\\ &= 3.72 \end{aligned} \tag{49}$$

and

$$\begin{aligned} UCL &= \exp\left(\ln 4.36 + 1.96 \cdot \sqrt{\frac{1}{(6.21 - 1.04) \times 1\,000\,000} + \frac{1}{1.04 \times 1\,000\,000} + \frac{1}{290} + \frac{1}{621 - 290}}\right)\\ &= 5.10. \end{aligned} \tag{50}$$

For men, the number is 0.53 million ($6.56 - 5.02 - 1.01 = 0.53$); the estimation can be done in the same way. Thus, the relative risk would be estimated to be 4.36 (95% CI: 3.72–5.10) for women and 4.20 (95% CI: 3.78–4.67) for men.

Although the calculations were not adjusted and the definition of no occupation is tentative, these results suggest that being neither employed nor educated may substantially increase the risk of suicide among the young Japanese population. It might be also suggested that the Japanese governments should consider the indicator of unemployment.

Note that relative risks were estimated without a fresh cohort study, which is generally difficult to conduct.

Discussion

Evaluating a change in risk of an event occurring caused by exposure to (or the presence of/occupation as) a factor is generally attempted in many research fields, such as epidemiology, medicine, social science, politics, and product development. Relative risk, which is the ratio of the risks, can be easily interpreted and widely used, but has been believed to require large-scale epidemiological research or a smaller cohort study designed for the estimation. A case control study, which compares the case and control group, is more convenient than the cohort study, but relative risk cannot be estimated using case control data. The estimator of the odds ratio, which can be calculated using case control data, is often used instead of relative risk, because the former can sometimes approximate the latter. A method to calculate relative risk using the odds ratio was also proposed. Unfortunately, the odds ratio may be misleading to interpret the change in risk and calculating relative risk using the ratio still requires either estimator of risks. Furthermore, control group data are still required, burdening researchers in terms of cost and effort.

In this study, introducing the observed proportion, an observational inconsistency preventing relative risk from being estimated in case control studies was clarified as a mathematical expression; by excluding this inconsistency, a new equation that estimates relative risk using case data was proposed. The proposed equation, which serves as an estimator of relative risk itself without approximation, requires only the exposure odds of a case group and that of all subjects to be studied; no control group is then needed. The calculation is done without using risk estimators, and thus, cohorts are also not needed. Therefore, evaluating a change in risk can be easily conducted without additional costs, efforts, and time generally needed in a fresh study. Moreover, the proposed equation was derived without using the Bayesian probabilities nor the Bayes' theorem and is free from researcher's resistance toward the Bayesian methods.

A method of estimating confidence limits of the proposed estimator was also presented and proved to estimate that successfully. Although there may be a more appropriate estimation method of confidence interval, pursuing the best method is beyond the scope of this paper.

Once the exposed proportions by various characteristics are investigated, changes in every risk associated with the exposure will able to be estimated by applying the proposed equation to appropriate case group data. Even the estimation of a change in risk, which has been believed to be impossible, can be done, such as the adverse effect of a social situation on the suicide rate, the effect of a policy on birthrate, or the impact of a new drug for a pandemic on survival rate. There are two caveats: the case group must comprise subjects from whom the exposed proportion was computed and the exposure to the factor must precede the occurring event. Existing statistical methods, such as adjusting confounding factors, should be also applicable for the proposed estimator.

Although the proposed equation is quite simple, its advantages will not only reduce the costs of epidemiological studies but may also make itself a powerful tool in almost all research fields that treat risks.

References

1. Andrade, C. Understanding relative risk, odds ratio, and related terms: as simple as it can get. *J Clin Psychiatry.* **76**, 857–861, doi:10.4088/JCP.15f10150 (2015).
2. Milner, A., Page, A. & LaMontagne, A. D. Long-term unemployment and suicide: A systematic review and meta-Analysis. *PLoS One.* **8**, e51333, doi:10.1371/journal.pone.0051333 (2013).
3. Rothman, K. J., Greenland, S., & Lash, T. L. Definition in *Modern epidemiology*. (3rd ed.) 53–54 (Lippincott Williams & Wilkins, 2008).
4. Agresti, A. Definition and expression in *Categorical Data Analysis*. (3rd ed.) 44–45 (John Wiley & Sons, 2011).
5. Cornfield, J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst.* **11**, 1269–1275 (1951).
6. Davies, H. T., Crombie, I. K. & Tavakoli, M. When can odds ratios mislead? *BMJ.* **316**, 989–991, doi:10.1136/bmj.316.7136.989 (1998).

7. McNutt, L. A., Wu, C., Xue, X. & Hafner, J. P. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol.* **157**, 940–943, doi:10.1093/aje/kwg074 (2003).
8. Liddell, F. D. K., McDonald, J. C., Thomas, D. C. & Cunliffe, S. V. Methods of cohort analysis: appraisal by application to asbestos mining. *J R Stat Soc Ser A.* **140**, 469–491, doi:10.2307/2345280 (1977).
9. Maclure, M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol.* **133**, 144–153, doi:10.1093/oxfordjournals.aje.a115853 (1991).
10. Prentice, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika.* **73**, 1–11, doi:10.1093/biomet/73.1.1 (1986).
11. Zhang, J. & Yu, K. F. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* **280**, 1690–1691, doi:10.1001/jama.280.19.1690 (1998).
12. Gelman, A., Carlin, J. B., Stern, H. B., & Rubin, D. B. An equation in *Bayesian data analysis* (3rd ed.) 6–7 (Chapman & Hall/CRC, 2014).
13. Sahai, H & Khurshid, A. Equations in *Statistics in Epidemiology: Methods, Techniques and Applications* 21–22 (CRC Press LCC, 1995).
14. Ministry of Health, Labour and Welfare. *Annual Health, Labour and Welfare Report 2013–2014 (Summary)* http://www.mhlw.go.jp/english/wp/wp-hw8/dl/summary.pdf (2015).
15. Maki, N. & Martikainen, P. A register-based study on excess suicide mortality among unemployed men and women during different levels of unemployment in Finland. *J Epidemiol Community Health.* **66**, 302–307, doi:10.1136/jech.2009.105908 (2012).
16. Ministry of Health, Labour and Welfare. *Heisei 27-nenn chu ni okeru jisatsu no jyoukyou [Statistics of suicide in Japan 2015]* http://www.mhlw.go.jp/file/06-Seisakujouhou-12200000-Shakaiengokyokushougaihokenfukushibu/h27kakutei-2syou_2.pdf (2016) [in Japanese].
17. Ministry of Internal Affairs and Communications. *Annual Report on the Labour Force Survey 2015* http://www.stat.go.jp/english/data/roudou/report/2015/index.htm (2016).

Acknowledgements

The author would like to thank Editage (www.editage.jp) for English language editing and also to thank anonymous reviewers for helpful comments.

Author Contributions

Y.Y. conducted the study and wrote the paper.

Additional Information

Competing Interests: The author declares that he has no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.