OPEN ACCESS Freely available online

PLOS ONE

# A Regression-Based Method for Estimating Risks and Relative Risks in Case-Base Studies

**Tina Tsz-Ting Chui**[1], **Wen-Chung Lee**[1,2]*

**1** Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, **2** Research Center for Genes, Environment and Human Health, College of Public Health, National Taiwan University, Taipei, Taiwan

## Abstract

Both the absolute risk and the relative risk (RR) have a crucial role to play in epidemiology. RR is often approximated by odds ratio (OR) under the rare-disease assumption in conventional case-control study; however, such a study design does not provide an estimate for absolute risk. The case-base study is an alternative approach which readily produces RR estimation without resorting to the rare-disease assumption. However, previous researchers only considered one single dichotomous exposure and did not elaborate how absolute risks can be estimated in a case-base study. In this paper, the authors propose a logistic model for the case-base study. The model is flexible enough to admit multiple exposures in any measurement scale—binary, categorical or continuous. It can be easily fitted using common statistical packages. With one additional step of simple calculations of the model parameters, one readily obtains relative and absolute risk estimates as well as their confidence intervals. Monte-Carlo simulations show that the proposed method can produce unbiased estimates and adequate-coverage confidence intervals, for ORs, RRs and absolute risks. The case-base study with all its desirable properties and its methods of analysis fully developed in this paper may become a mainstay in epidemiology.

**Citation:** Chui TT-T, Lee W-C (2013) A Regression-Based Method for Estimating Risks and Relative Risks in Case-Base Studies. PLoS ONE 8(12): e83275. doi:10.1371/journal.pone.0083275

**Editor:** Momiao Xiong, University of Texas School of Public Health, United States of America

**Received** August 6, 2013; **Accepted** November 11, 2013; **Published** December 12, 2013



**Funding:** This paper is partly supported by grants from National Science Council, Taiwan (NSC 102-2628-B-002-036-MY3) and National Taiwan University, Taiwan (NTU-CESRP-102R7622-8). No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: wenchung@ntu.edu.tw

## Introduction

Both the absolute and the relative disease risks have a crucial role to play in epidemiology. The relative risk (RR) is the ratio of the disease risk for individuals at one specific exposure level to the disease risk for those at a reference level. Under the rare-disease assumption, RR is approximated by the odds ratio (OR), which in turn can be conveniently estimated in a case-control study. While an index such as RR or OR may be adequate for etiologic inferences, it is actually only part of a story. Once a factor has been demonstrated to be a risk factor for the disease, we will often be asked to predict the disease risk of an individual having a specific level of an exposure—the absolute risk. But unfortunately, the conventional case-control study does not provide an estimate for it.

Kupper et al [1] introduced a hybrid (part case-control, part cohort) design in a defined population (the 'study base')—the 'case-base' study later coined by Miettinen [2]. In contrast to the case-control study which samples the non-diseased subjects in the study base as the control group, the case-base study samples the entire study base with no regard to disease status. With such sampling scheme, the case-base study readily produces an RR estimate without resorting to the rare-disease assumption. Note that the case-base study should not be confused with the 'case-cohort' study introduced by Prentice [3]. The former, like the case-control study, is a *retrospective* design which ascertains the exposure statuses of subjects in a population retrospectively, while the latter is a *prospective* cohort study with all the time-to-event information available.

While the case-cohort study has been gaining popularity over the years [3–9], the case-base study remained little noticed since its introduction forty years ago. Miettinen [2] derived a variance formula for RR in a case-base study. Sato [10,11] later proposed a more efficient estimator for RR, which is based on maximum likelihood estimation theory. However, these researchers only considered one dichotomous exposure and did not elaborate on how to estimate absolute risks in a case-base study. Without a general-purpose regression method for analyzing data, it is no wonder that most practicing epidemiologists would not consider the case-base design when planning a study.

In this paper, we develop a logistic model for the case-base study. The model is flexible enough to admit multiple exposures in any measurement scale—binary, categorical or continuous. It can be easily fitted using common statistical packages. With one additional step of simple calculations of the model parameters, one readily obtains relative and absolute risk estimates as well as their confidence intervals. We will use Monte-Carlo simulations to study the statistical properties of the proposed method.

## Methods

Let the exposure profile of a subject be denoted by a $1 \times m$ row vector $\mathbf{z}$. Each element of $\mathbf{z}$ can be in either binary, categorical or continuous scale. Let $D$ represents the disease status of a subject, with $D=1$ for diseased and $D=0$ for non-diseased. We assume that the disease risk in the study population follows a logistic model:

$$\log\left[\frac{\Pr(D=1|\mathbf{z})}{\Pr(D=0|\mathbf{z})}\right]=\mu+\mathbf{z}\boldsymbol{\beta}, \tag{1}$$

where $\exp(\mu)$ is the baseline disease odds (the disease odds for those with an exposure profile of $\mathbf{z}=0$ in the population) and $\boldsymbol{\beta}$ is a $m\times 1$ column vector of parameters of interest [$\exp(\boldsymbol{\beta})$ is a column vector of odds ratios].

In a case-base study, the 'cases' are randomly selected from all the incident diseased subjects in the population. Let $S_1=1$ indicate that a diseased subject is recruited in the case sample, $S_1=0$, otherwise. Such a case sampling scheme implies that

$$\begin{cases}\Pr(S_1=1|D=1,\mathbf{z})=\phi_1\\ \Pr(S_1=1|D=0,\mathbf{z})=0,\end{cases} \tag{2}$$

or more concisely,

$$\Pr(S_1=1|D,\mathbf{z})=\phi_1 D, \tag{3}$$

where $\phi_1$ is a constant between 0 and 1. The 'controls' of a case-base study are randomly selected from all subjects in the population without regard to their disease status. Let $S_0=1$ indicate that a subject is recruited in the control sample, $S_0=0$, otherwise. Such a control sampling scheme implies that

$$\Pr(S_0=1|D,\mathbf{z})=\phi_0, \tag{4}$$

where $\phi_0$ is a constant between 0 and 1. The two sampling schemes are independent to each other, that is,

$$\begin{aligned}\Pr(S_0=1,S_1=1|D,\mathbf{z})=&\Pr(S_0=1|D,\mathbf{z})\times\\ &\Pr(S_1=1|D,\mathbf{z})=\phi_0\phi_1 D,\end{aligned} \tag{5}$$

The event of $S_0+S_1\geq 1$ indicates that a subject is recruited in a case-base study through case sampling, control sampling or both. The recruitment probability of a subject with a disease status of $D$ and an exposure profile of $\mathbf{z}$ is

$$\begin{aligned}\Pr(S_0+S_1\geq 1|D,\mathbf{z})=&\Pr(S_0=1|D,\mathbf{z})+\Pr(S_1=1|D,\mathbf{z})\\ &-\Pr(S_0=1,S_1=1|D,\mathbf{z})=\phi_0\\ &+(1-\phi_0)\phi_1 D.\end{aligned} \tag{6}$$

Let $\pi$ be the probability that a diseased subject in a case-base study is recruited in the control sample, that is,

$$\begin{aligned}\pi&=\Pr(S_0=1|D=1,\mathbf{z},S_0+S_1\geq 1)\\ &=\frac{\Pr(S_0=1|D=1,\mathbf{z})}{\Pr(S_0+S_1\geq 1|D=1,\mathbf{z})}=\frac{\phi_0}{\phi_0+(1-\phi_0)\phi_1}.\end{aligned} \tag{7}$$

$\pi$ is an important parameter to be used later.

From equations 1–7, we show below that the disease risk in a case-base sample also follows a logistic model as the one in the population (model 1), albeit with a different intercept:

$$\begin{aligned}&\log\left[\frac{\Pr(D=1|\mathbf{z},S_0+S_1\geq 1)}{\Pr(D=0|\mathbf{z},S_0+S_1\geq 1)}\right]=\\ &\log\left[\frac{\Pr(S_0+S_1\geq 1|D=1,\mathbf{z})\times\Pr(D=1|\mathbf{z})\times\Pr(\mathbf{z})}{\Pr(S_0+S_1\geq 1|D=0,\mathbf{z})\times\Pr(D=0|\mathbf{z})\times\Pr(\mathbf{z})}\right]=\\ &\log\left[\frac{\Pr(S_0+S_1\geq 1|D=1,\mathbf{z})}{\Pr(S_0+S_1\geq 1|D=0,\mathbf{z})}\right]+\log\left[\frac{\Pr(D=1|\mathbf{z})}{\Pr(D=0|\mathbf{z})}\right]=\\ &\log\left[\frac{\phi_0+(1-\phi_0)\phi_1}{\phi_0}\right]+\mu+\mathbf{z}\boldsymbol{\beta}=-\log\pi+\mu+\mathbf{z}\boldsymbol{\beta}=\mu^*+\mathbf{z}\boldsymbol{\beta}\end{aligned} \tag{8}$$

Suppose that there are a total of $n$ subjects recruited in a case-base study, who are indexed by $i$ ($i=1,\ldots,n$). For the $i^{\text{th}}$ subject, the exposure profile, the disease status, and the control and the case sampling statuses are $\mathbf{z}_i$, $D_i$, $S_{0,i}$, and $S_{1,i}$, respectively. Given the exposure status of the subjects recruited in the case-base study, each of the subjects provides the information of disease and sampling statuses. The likelihood function is therefore

$$\begin{aligned}L(\pi,\phi_1,\mu^*,\beta^{\mathrm{t}})&=\prod_{i=1}^{n}\Pr(D_i,\,S_{0,i},\,S_{1,i}|\mathbf{z}_i,S_{0,i}+S_{1,i}\geq 1)\\ &=\prod_{i=1}^{n}\left[\begin{array}{l}\Pr(S_{1,i}|D_i,\mathbf{z}_i,S_{0,i},S_{0,i}+S_{1,i}\geq 1)\\ \times\Pr(S_{0,i}|D_i,\mathbf{z}_i,S_{0,i}+S_{1,i}\geq 1)\\ \times\Pr(D_i|\mathbf{z}_i,S_{0,i}+S_{1,i}\geq 1)\end{array}\right]\\ &=\prod_{D_i\times S_{0,i}=1}\Pr(S_{1,i}|D_i=1,\mathbf{z}_i)\\ &\quad\times\prod_{D_i=1}\Pr(S_{0,i}|D_i=1,\mathbf{z}_i,S_{0,i}+S_{1,i}\geq 1)\\ &\quad\times\prod_{i=1}^{n}\Pr(D_i|\mathbf{z}_i,S_{0,i}+S_{1,i}\geq 1)\\ &=\prod_{D_i\times S_{0,i}=1}\vartheta_1^{S_{1,i}}\times(1-\vartheta_1)^{1-S_{1,i}}\\ &\quad\times\prod_{D_i=1}\pi^{S_{0,i}}\times(1-\pi)^{1-S_{0,i}}\\ &\quad\times\prod_{i=1}^{n}\frac{[\exp(\mu^*+\mathbf{z}_i\boldsymbol{\beta})]^{D_i}}{1+\exp(\mu^*+\mathbf{z}_i\boldsymbol{\beta})}\\ &=L_1(\vartheta_1)\times L_2(\pi)\times L_3(\mu^*,\boldsymbol{\beta}^{\mathrm{t}}).\end{aligned} \tag{9}$$

Because equation 9 is composed of three terms, the three sets of parameters ( $\phi_1$ in $L_1$, $\pi$ in $L_2$, and$\mu^*$ and $\boldsymbol{\beta}^{\mathrm{t}}$ in $L_3$) are mutually independent (the second derivatives of the log-likelihood with respect to parameters in different sets are zero).

Both $L_1$ and $L_2$ in equation 9 are binomial likelihoods. Therefore the maximum likelihood estimates of $_1$ and $\pi$, and their variances are:

$$\phi_1 = \frac{n_{\mathrm{D}}^{\mathrm{Both}}}{n_{\mathrm{D}}^{\mathrm{CN}}}, \tag{10}$$

$$\widehat{\pi} = \frac{n_{\mathrm{D}}^{\mathrm{CN}}}{n_{\mathrm{D}}}, \tag{11}$$

$$\mathrm{Var}\left(\widehat{\phi_1}\right) = \frac{\widehat{\phi_1}(1-\widehat{\phi_1})}{n_{\mathrm{D}}^{\mathrm{CN}}}, \tag{12}$$

and

$$\mathrm{Var}(\widehat{\pi}) = \frac{\widehat{\pi}(1-\widehat{\pi})}{n_{\mathrm{D}}}, \tag{13}$$

where $n_{\mathrm{D}}^{\mathrm{CN}}$ is the number of diseased subjects recruited in control sample, $n_{\mathrm{D}}^{\mathrm{Both}}$, the number of diseased subjects recruited in both the case and the control sample, and $n_{\mathrm{D}}$, the total number of diseased subjects recruited in the case-base study.

The $L_3$ in equation 9 is a likelihood for a logistic regression model. To obtain the maximum likelihood estimates of $\mu^*$and $\boldsymbol{\beta}^{\mathrm{t}}$, we can fit a logistic regression (model 8) to the case-base data. Note that the dependent variable of this logistic regression is the binary disease status with the diseased subjects coded as '1' and the non-diseased subjects as '0', regardless of their being recruited through case sampling, control sampling or both. Any statistical package that performs logistic regression analysis can obtain the estimates $\widehat{\mu^*}$and $\widehat{\boldsymbol{\beta}}^{\mathrm{t}}$, together with the variance-covariance matrix of $(\mu^*,\boldsymbol{\beta}^{\mathrm{t}})$. This variance-covariance matrix is denoted by $\boldsymbol{\Sigma}$, which is an $(m+1)\times(m+1)$ matrix.

The $\widehat{\boldsymbol{\beta}}^{\mathrm{t}}$ above readily provides the maximum likelihood estimates for the logarithms of ORs. As detailed below, the $\widehat{\pi}$ and $\widehat{\mu^*}$ above are to be further combined to provide estimates for risks and RRs. First from model 8, an estimate for $\mu$ in model 1 is

$$\widehat{\mu} = \log\widehat{\pi} + \widehat{\mu^*}. \tag{14}$$

An estimate of the disease risk for subjects in the population with an exposure profile vector $\mathbf{u}$ (a $1\times m$ row vector ) is therefore

$$\widehat{\mathrm{risk}}_{\mathbf{u}} = \frac{\exp(\widehat{\mu}+\mathbf{u}\widehat{\boldsymbol{\beta}})}{1+\exp(\widehat{\mu}+\mathbf{u}\widehat{\boldsymbol{\beta}})} = \frac{\exp\left[\log\widehat{\pi}+\widehat{\mu^*}+\mathbf{u}\widehat{\boldsymbol{\beta}}\right]}{1+\exp\left[\log\widehat{\pi}+\widehat{\mu^*}+\mathbf{u}\widehat{\boldsymbol{\beta}}\right]}. \tag{15}$$

The variance of the estimate (in logit scale) is

$$\begin{aligned}\mathrm{Var}\left[\mathrm{logit}\left(\widehat{\mathrm{risk}}_{\mathbf{u}}\right)\right] &= \mathrm{Var}\left(\widehat{\mu}+\mathbf{u}\widehat{\boldsymbol{\beta}}\right) = \\ &\mathrm{Var}\left(\log\widehat{\pi}+\widehat{\mu^*}+\mathbf{u}\widehat{\boldsymbol{\beta}}\right) = \\ &\mathrm{Var}(\log\widehat{\pi}) + \mathrm{Var}\left(\widehat{\mu^*}+\mathbf{u}\widehat{\boldsymbol{\beta}}\right) + \\ &2\times\mathrm{Cov}\left(\log\widehat{\pi},\widehat{\mu^*}+\mathbf{u}\widehat{\boldsymbol{\beta}}\right) = \\ &\frac{1}{\widehat{\pi}^2}\times\mathrm{Var}(\widehat{\pi}) + [1 \quad \mathbf{u}]\Sigma[1 \quad \mathbf{u}]^{\mathrm{t}} \\ &+0 = \frac{1-\widehat{\pi}}{n_{\mathrm{D}}^{\mathrm{CN}}} + \mathbf{v}\Sigma\mathbf{v}^{\mathrm{t}},\end{aligned} \tag{16}$$

where $\mathbf{v} = [1 \quad \mathbf{u}]$ is a $1\times(m+1)$ row vector. An estimate of the RR comparing those with an exposure profile vector $\mathbf{u}_1$ with those with $\mathbf{u}_0$ is

**Table 1.** Simulation results for a binary exposure.

| | Methods | | |
|---|---|---|---|
| | **The present method** | **Sato** | **Miettinen** |
| Estimate [true value] | | | |
| logOR [0.9163] | 0.9191 | - | - |
| logRR [0.8128] | 0.8148 | 0.8149 | 0.8149 |
| logit($\mathrm{risk}_0$) [−2.5465] | −2.5559 | - | - |
| logit($\mathrm{risk}_1$) [−1.6303] | −1.6369 | - | - |
| Variance (×100) | | | |
| logOR | 1.8297 | - | - |
| logRR | 1.3984 | 1.3984 | 1.5017 |
| logit($\mathrm{risk}_0$) | 2.5622 | - | - |
| logit($\mathrm{risk}_1$) | 3.0710 | - | - |
| Coverage probability of 95% CI | | | |
| logOR | 0.9521 | - | - |
| logRR | 0.9518 | 0.9518 | 0.9518 |
| logit($\mathrm{risk}_0$) | 0.9512 | - | - |
| logit($\mathrm{risk}_1$) | 0.9497 | - | - |
| Average length of 95% CI | | | |
| logOR | 0.5324 | - | - |
| logRR | 0.4657 | 0.4657 | 0.4825 |
| logit($\mathrm{risk}_0$) | 0.6220 | - | - |
| logit($\mathrm{risk}_1$) | 0.6818 | - | - |

doi:10.1371/journal.pone.0083275.t001

**Table 2.** Simulation results for an exposure with four levels.

| | Methods | | |
|---|---|---|---|
| | **The present method** | **Sato** | **Miettinen** |
| Estimate [true value] | | | |
| logOR comparing adjacent levels [0.9163] | 0.9189 | - | - |
| $\log RR_1$ [0.8629] | 0.8655 | 0.8654 | 0.8654 |
| $\log RR_2$ [1.6569] | 1.6615 | 1.6648 | 1.6668 |
| $\log RR_3$ [2.3203] | 2.3253 | 2.3278 | 2.3297 |
| $\text{logit}(\text{risk}_0)$ [−3.2708] | −3.2845 | - | - |
| $\text{logit}(\text{risk}_1)$ [−2.3545] | −2.3656 | - | - |
| $\text{logit}(\text{risk}_2)$ [−1.4383] | −1.4468 | - | - |
| $\text{logit}(\text{risk}_3)$ [−0.5220] | −0.5279 | - | - |
| Variance (×100) | | | |
| logOR comparing adjacent levels | 0.4854 | - | - |
| $\log RR_1$ | 0.4586 | 2.4588 | 2.5149 |
| $\log RR_2$ | 1.5899 | 3.6685 | 4.0080 |
| $\log RR_3$ | 2.6760 | 2.9777 | 3.4950 |
| $\text{logit}(\text{risk}_0)$ | 2.9127 | - | - |
| $\text{logit}(\text{risk}_1)$ | 2.3802 | - | - |
| $\text{logit}(\text{risk}_2)$ | 2.8184 | - | - |
| $\text{logit}(\text{risk}_3)$ | 4.2274 | - | - |
| Coverage probability of 95% CI | | | |
| logOR comparing adjacent levels | 0.9536 | - | - |
| $\log RR_1$ | 0.9533 | 0.9563 | 0.9556 |
| $\log RR_2$ | 0.9530 | 0.9487 | 0.9493 |
| $\log RR_3$ | 0.9518 | 0.9526 | 0.9523 |
| $\text{logit}(\text{risk}_0)$ | 0.9518 | - | - |
| $\text{logit}(\text{risk}_1)$ | 0.9504 | - | - |
| $\text{logit}(\text{risk}_2)$ | 0.9505 | - | - |
| $\text{logit}(\text{risk}_3)$ | 0.9505 | - | - |
| Average length of 95% CI | | | |
| logOR comparing adjacent levels | 0.2731 | - | - |
| $\log RR_1$ | 0.2657 | 0.6243 | 0.6319 |
| $\log RR_2$ | 0.4952 | 0.7478 | 0.7814 |
| $\log RR_3$ | 0.6437 | 0.6783 | 0.7330 |
| $\text{logit}(\text{risk}_0)$ | 0.6677 | - | - |
| $\text{logit}(\text{risk}_1)$ | 0.6011 | - | - |
| $\text{logit}(\text{risk}_2)$ | 0.6531 | - | - |
| $\text{logit}(\text{risk}_3)$ | 0.8007 | - | - |

doi:10.1371/journal.pone.0083275.t002

$$\widehat{\mathrm{RR}}_{\mathbf{u}_1/\mathbf{u}_0} = \widehat{\mathrm{risk}}_{\mathbf{u}_1} \Big/ \widehat{\mathrm{risk}}_{\mathbf{u}_0} = \frac{\exp\left[\log\hat{\pi} + \hat{\mu}^* + \mathbf{u}_1\hat{\boldsymbol{\beta}}\right]}{1+\exp\left[\log\hat{\pi} + \hat{\mu}^* + \mathbf{u}_1\hat{\boldsymbol{\beta}}\right]} \Big/ \frac{\exp\left[\log\hat{\pi} + \hat{\mu}^* + \mathbf{u}_0\hat{\boldsymbol{\beta}}\right]}{1+\exp\left[\log\hat{\pi} + \hat{\mu}^* + \mathbf{u}_0\hat{\boldsymbol{\beta}}\right]}. \quad (17)$$

Using the delta method, the variance of the estimate (in log scale) is

$$\mathrm{Var}\left[\log\left(\widehat{\mathrm{RR}}_{\mathbf{u}_1/\mathbf{u}_0}\right)\right] = \left(\widehat{\mathrm{risk}}_{\mathbf{u}_1} - \widehat{\mathrm{risk}}_{\mathbf{u}_0}\right)^2 \times \left(\frac{1-\hat{\pi}}{n_{\mathrm{D}}^{\mathrm{CN}}} + \mathbf{w}\boldsymbol{\Sigma}\mathbf{w}^{\mathrm{t}}\right), \quad (18)$$

**Table 3.** Simulation results for two binary exposures.

| | Methods | | |
|---|---|---|---|
| | **The present method** | **Sato** | **Miettinen** |
| Estimate [true value] | | | |
| $\log OR_1$ [0.9163] | 0.9206 | - | - |
| $\log OR_2$ [1.0986] | 1.1017 | - | - |
| $\log RR_{10}$ [0.8536] | 0.8571 | 0.8580 | 0.8585 |
| $\log RR_{01}$ [1.0159] | 1.0184 | 1.0193 | 1.0197 |
| $\log RR_{11}$ [1.7678] | 1.7724 | 1.7741 | 1.7754 |
| $\text{logit}(risk_{00})$ [−3.0995] | −3.1087 | - | - |
| $\text{logit}(risk_{10})$ [−2.1832] | −2.1880 | - | - |
| $\text{logit}(risk_{01})$ [−2.0008] | −2.0070 | - | - |
| $\text{logit}(risk_{11})$ [−1.0846] | −1.0863 | - | - |
| Variance (×100) | | | |
| $\log OR_1$ | 2.0187 | - | - |
| $\log OR_2$ | 1.8573 | - | - |
| $\log RR_{10}$ | 1.7228 | 3.2565 | 3.3754 |
| $\log RR_{01}$ | 1.5893 | 2.4743 | 2.5707 |
| $\log RR_{11}$ | 3.0231 | 3.0867 | 3.3906 |
| $\text{logit}(risk_{00})$ | 3.1880 | - | - |
| $\text{logit}(risk_{10})$ | 3.5971 | - | - |
| $\text{logit}(risk_{01})$ | 3.0930 | - | - |
| $\text{logit}(risk_{11})$ | 3.8039 | - | - |
| Coverage probability of 95% CI | | | |
| $\log OR_1$ | 0.9490 | - | - |
| $\log OR_2$ | 0.9503 | - | - |
| $\log RR_{10}$ | 0.9492 | 0.9508 | 0.9509 |
| $\log RR_{01}$ | 0.9508 | 0.9510 | 0.9486 |
| $\log RR_{11}$ | 0.9484 | 0.9487 | 0.9532 |
| $\text{logit}(risk_{00})$ | 0.9481 | - | - |
| $\text{logit}(risk_{10})$ | 0.9470 | - | - |
| $\text{logit}(risk_{01})$ | 0.9465 | - | - |
| $\text{logit}(risk_{11})$ | 0.9487 | - | - |
| Average length of 95% CI | | | |
| $\log OR_1$ | 0.5534 | - | - |
| $\log OR_2$ | 0.5323 | - | - |
| $\log RR_{10}$ | 0.5114 | 0.7034 | 0.7161 |
| $\log RR_{01}$ | 0.4923 | 0.6149 | 0.6257 |
| $\log RR_{11}$ | 0.6788 | 0.6862 | 0.7224 |
| $\text{logit}(risk_{00})$ | 0.6875 | - | - |
| $\text{logit}(risk_{10})$ | 0.7300 | - | - |
| $\text{logit}(risk_{01})$ | 0.6767 | - | - |
| $\text{logit}(risk_{11})$ | 0.7525 | - | - |

doi:10.1371/journal.pone.0083275.t003

where $\mathbf{w}=[1 \quad \mathbf{y}]$ is a $1\times(m+1)$ row vector with $\mathbf{y}=\dfrac{\left(\widehat{\text{risk}}_{\mathbf{u}_1}-1\right)\mathbf{u}_1-\left(\widehat{\text{risk}}_{\mathbf{u}_0}-1\right)\mathbf{u}_0}{\widehat{\text{risk}}_{\mathbf{u}_1}-\widehat{\text{risk}}_{\mathbf{u}_0}}$.

Exhibit S1 shows that Sato's formulas [10,11] of RR estimate and its variance in log scale are a special case of our formulas of equation 17 and 18 when there is only one single binary exposure.

Note that if $n_D^{CN}=0$ (no diseased subject is recruited in the control sample), $\widehat{\pi}$ (in equation 11) is not estimable. Therefore, $\widehat{\mu}$ (in equation 14), $\widehat{\text{risk}}_{\mathbf{u}}$ (in equation 15) and $\widehat{\text{RR}}_{\mathbf{u}_1/\mathbf{u}_0}$(in equation 17) are not estimable either. Under such setting, only the odds ratios, $\exp(\widehat{\boldsymbol{\beta}}^t)$, can be estimated in a case-base study. At the other extreme when $n_D^{CN}=n_D$ (all the diseased subjects are recruited in the control sample), we have $\widehat{\pi}=1$ and $\widehat{\mu}=\widehat{\mu}^*$, and therefore the case-base data can be analyzed as a cohort data. As for $n_D^{Both}$ (number of diseased subject recruited in both the case and the control sample), if it is zero the $\widehat{\phi_1}$ (in equation 10) is not estimable. This has no bearing whatsoever on the current context of estimating risks and relative risks however, since it is a nuisance parameter anyway.

We perform Monte-Carlo simulations to examine the statistical properties of the proposed method. We consider three scenarios for the exposure. In the first scenario, we assume a binary exposure ($E=0, 1$). The exposure prevalence (for $E=1$) is set at 0.3. We assume that the OR comparing $E=1$ subjects with $E=0$ subjects is 2.5 ($\boldsymbol{\beta}=\log\text{OR}=0.9163$). The disease prevalence in the study population is set at 0.1. Thus, the disease risk for $E=0$ subjects ($\text{risk}_0$) is 0.0727, the disease risk for $E=1$ subjects ($\text{risk}_1$) is 0.1638, and RR is 2.2543 ($\log\text{RR}=0.8128$).

In the second scenario, we assume an exposure with four levels ($E=0, 1,2,3$). The exposure prevalence is set at 0.3 (for $E=1$), 0.1 (for $E=2$), and 0.1 (for $E=3$), respectively. The OR comparing adjacent levels is set at 2.5 ($\boldsymbol{\beta}=\log\text{OR}=0.9163$). Again, we assume a disease prevalence of 0.1. Therefore, the four disease risks are $\text{risk}_0=0.0366$, $\text{risk}_1=0.0867$, $\text{risk}_2=0.1918$, and $\text{risk}_3=0.3724$, respectively, and the RRs are (with $E=0$ as the reference level) $\text{RR}_1=2.3699$ ($\log\text{RR}_1=0.8629$), $\text{RR}_2=5.2430$ ($\log\text{RR}_2=1.6569$), and $\text{RR}_3=10.1787$ ($\log\text{RR}_3=2.3203$), respectively.

In the third scenario, we assume two binary exposures ($E_1$ and $E_2$). The exposure prevalence is set at 0.3 for $E_1$, and 0.4 for $E_2$. The OR comparing $E_1=1$ subjects with $E_1=0$ subjects is 2.5 ($\boldsymbol{\beta}_1=\log\text{OR}_1=0.9163$), and the OR comparing $E_2=1$ subjects with $E_2=0$ subjects is 3 ($\boldsymbol{\beta}_2=\log\text{OR}_2=1.0986$). For simplicity, we assume that $E_1$ and $E_2$ are independent of each other in the population and that there is no multiplicative interaction between $E_1$ and $E_2$ in causing the disease. The disease prevalence in the study population is set at 0.1. Thus, the four disease risks are $\text{risk}_{00}=0.0431$ (for $E_1=0,E_2=0$), $\text{risk}_{10}=0.1013$ (for $E_1=1$, $E_2=0$), $\text{risk}_{01}=0.1191$ (for $E_1=0,E_2=1$), and $\text{risk}_{11}=0.2527$ (for $E_1=1,E_2=1$), respectively. The RRs are (with $E_1=0$, $E_2=0$ as the reference level) $\text{RR}_{10}=2.3481$ ($\log\text{RR}_{10}=0.8536$),$\text{RR}_{01}=2.7618$ ($\log\text{RR}_{01}=1.0159$), and $\text{RR}_{11}=5.8578$ ($\log\text{RR}_{11}=1.7678$), respectively.

The disease probabilities of subjects in the study population are assumed to follow the logistic model in model 1 with the parameter settings given in the preceding paragraphs. A case-base study is conducted in a study population of size 100000 with a case sampling probability ($\phi_1$) of 0.05 and a control sampling probability ($\phi_0$) of 0.005. Under such sampling scheme, the case-base study is expected to recruit a total of 500 distinct diseased and 500 distinct non-diseased subjects. We use the proposed method to calculate the point estimates and 95 confidence intervals (CIs) for ORs, RRs and risks. For a comparison, Sato's [10,11] and Miettinen's [2] methods are also performed.

The simulation was done for 10,000 times for each setting. The mean of the estimates for ORs (in log scale), RRs (in log scale) and risks (in logit scale) are calculated. The variance of an estimate is calculated as the sample variance of the estimates. We also

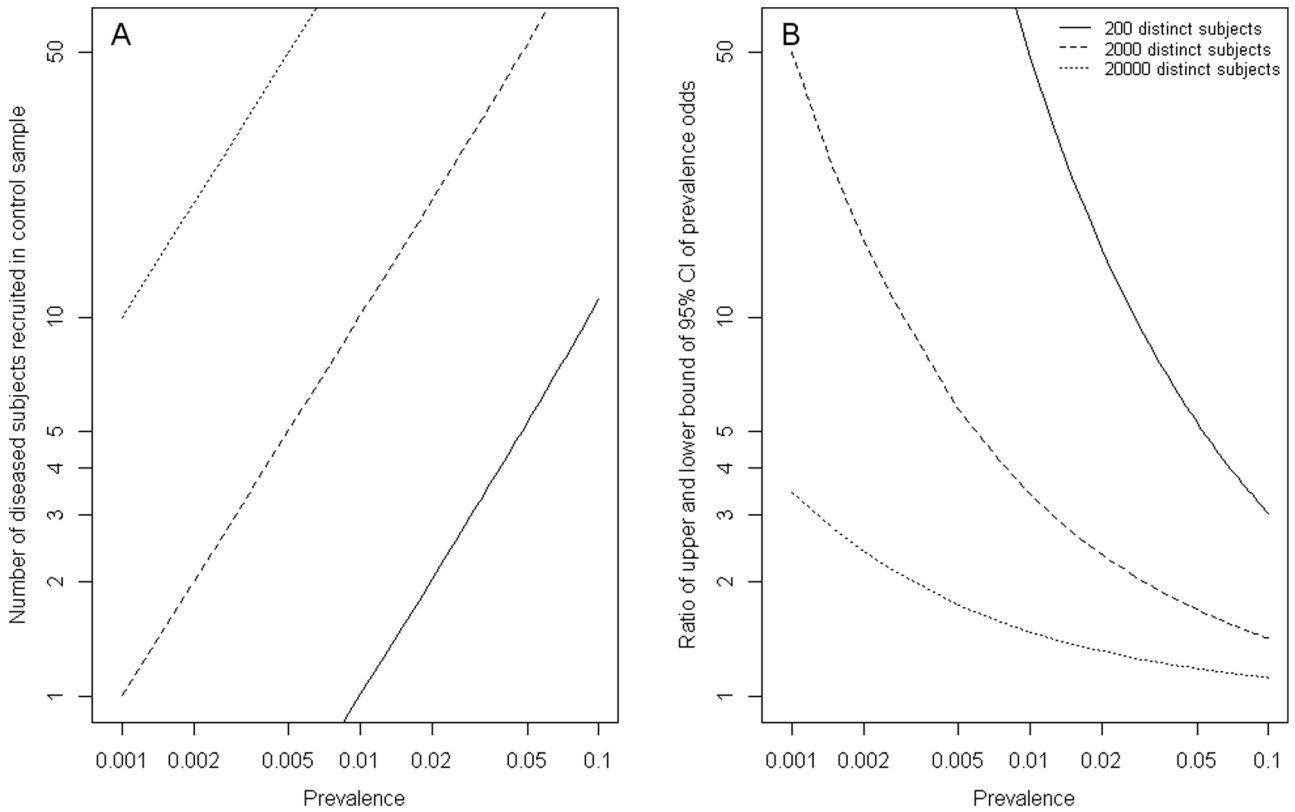


**Figure 1. Number of diseased subjects recruited in control sample (A); Ratio of upper and lower bound of 95% confidence intervals of prevalence odds (B), in a case-base study of 200 distinct subjects (solid lines), 2000 distinct subjects (dashed lines) and 20000 distinct subjects (dotted lines).**
doi:10.1371/journal.pone.0083275.g001

calculate the coverage probability and the average length of the 95% CIs for the estimates.

## Results

Table 1 shows the simulation results for a binary exposure. For all methods, the RR estimates are approximately unbiased and the 95% CIs achieve adequate coverage probabilities. However, the variance and the length of 95% CIs for our method are much smaller than those for Miettinen's methods. (Sato's method for the case of one binary exposure is exactly the same as our method.) Only our method can produce estimates for OR and risks additionally. From Table 1, we see that these estimates are approximately unbiased and their 95% CIs achieve adequate coverage probabilities.

Table 2 presents the simulation results for an exposure with four levels. It can be seen that our method can produce unbiased estimates and adequate-coverage 95% CIs for ORs, RRs, and risks. Sato's and Miettinen's methods can only produce estimates and 95% CIs for RRs. These two methods do not exploit the constancy in OR per unit change in the exposure variable. Therefore we see that though unbiased and with adequate coverage, they produce considerably larger variances and average length of 95% CIs as compared to our method. Exhibit S2 presents the simulation results for an exposure with four levels but without the constant OR assumption. We see that our method is still unbiased and with adequate coverage. The RR estimates are now the same as those using Sato's method, though. Exhibit S3 shows that our method can produce unbiased estimates and adequate-coverage 95% CIs for ORs, RRs, and risks, when the exposure is in a continuous scale.

Table 3 presents the simulation results for two binary exposures. Similarly, only our method can produce unbiased estimates and adequate-coverage 95% CIs for ORs, RRs, and risks. Sato's and Miettinen's methods can produce unbiased estimates and with adequate coverage 95% CIs for RRs only. These two methods do not exploit the assumption of no interaction between the two exposures. Therefore, we see that the variances and average length of 95% CIs for the two methods are much larger as compared to our method. Exhibit S4 presents the simulation results when there is an interaction effect between the two exposures. We see that our method can produce unbiased estimates and adequate-coverage 95% CIs for ORs, RRs, and risks, if an interaction term (cross-product term) is incorporated into the regression model. Exhibit S5 presents the simulation results for a confounder. We see that without adjusting for the confounder, one gets estimates that are biased and 95% CIs that are under-coverage. The problems can be easily fixed by performing a logistic regression analysis with both the study exposure and the confounder as its covariates.

Exhibit S6 examines the situations when the disease prevalence is lower: 0.05 and 0.01, respectively. The conclusions about method comparisons remain the same, except that the precisions for RRs and risks are compromised across all methods.

## Discussion

Logistic regression is a standard technique for analyzing case-control data. It is also the method of choice for analyzing cohort

data if time-to-event information is not available. However, the ORs that it estimates are approximating the RRs only under the rare-disease assumption. As such, there have been many methodologies/recommendations proposed to date regarding the estimation of RRs in cohort studies for common outcomes [12–17]. For example, Diaz-Quijano [17] described a novel regression-based method for estimating RRs in cohort studies. In his method, all the diseased subjects in the study are to be duplicated, and the duplicated subjects are to be re-labeled as the non-diseased. (For case-base studies, we can duplicate and re-label the diseased subjects recruited in the control sample.) Then, a logistic model is fitted to the expanded dataset, and the resulting regression coefficients are the estimates for logRRs. For case-base study, we found that such a data expansion approach produces an unbiased RR estimate for a binary exposure, but with a larger variance and a wider CI than our method; for a four-level exposure, the approach produces biased estimates and CIs with inadequate coverage (results not shown). For cohort study without time-to-event information, one can also apply our method to estimate ORs, RRs, and risks, except that the $\pi$ (equation 7) now is exactly one and is no longer a parameter to be estimated.

In addition to the usual ORs, a case-base study also provides estimates for risks (equation 15) and RRs (equation 17). From equations 16 and 18, we see that the precision of the estimation is inversely proportional to $(1-\widehat{\pi})/n_{\mathrm{D}}^{\mathrm{CN}} \approx 1/n_{\mathrm{D}}^{\mathrm{CN}}$, that is, the larger the $n_{\mathrm{D}}^{\mathrm{CN}}$ (number of diseased subjects recruited in control sample), the more precise the estimate of a risk or a RR. The value of $n_{\mathrm{D}}^{\mathrm{CN}}$ depends on the disease prevalence in the population and the sample size of the case-base study (Figure 1A). For a common disease (prevalence $>0.05$), a case-base study of 200 distinct subjects (with equal number of diseased and non-diseased subjects) is expected to have an $n_{\mathrm{D}}^{\mathrm{CN}}$ larger than 5, producing an estimate of disease odds with the upper 95% confidence bound being roughly 5 times its lower bound (Figure 1B). If the disease prevalence is lower (say, prevalence $=0.005$), one needs to increase the sample size of the case-base study (2000 subjects) to achieve comparable precision. If the registry system (for the diseased and the general population as well) in a population is readily available, the sample size then is no longer a limiting factor. In such setting, a case-base study can produce estimates for risks and RRs with reasonable precision, even if the disease is very rare (eg., $n_{\mathrm{D}}^{\mathrm{CN}} \approx 10$ and $\mathrm{upperbound}/\mathrm{lowerbound} \approx 3.5$ when sample size $=20000$ in a population with disease prevalence of 0.001).

In many respects, a case-base design is better than (or at least as good as) the commonly used case-control design. First, as just mentioned, a case-base study provides estimates not only for ORs but also for risks and RRs with reasonable accuracy (if $n_{\mathrm{D}}^{\mathrm{CN}} \geq 5$). Second, the control sampling scheme of a case-base study is a simple random sampling of all subjects in the study population without regard to disease status. This means that a researcher can initiate the control recruitment process much earlier in a case-base design (at the outset of the study) than in a case-control design (at the end of the study). Third, although there could be some people sampled more than once in a case-base study, the sampling itself incurs minimal cost. The real cost constraint is usually the total number of *distinct* subjects that are actually recruited. And with the same total number of distinct subjects, a case-base study and a case-control study have exactly the same statistical efficiency, when it comes to estimating an OR. Finally, as shown in this study, the analysis of a case-base study is no more complicated than a case-control study—one needs only to fit a logistic regression model to the data and then do one extra step of simple calculations of the model parameters.

## Supporting Information

**Exhibit S1** Comparison of Sato's formulas and the formulas derived in this paper when there is only one single binary exposure.
(DOC)

**Exhibit S2** Simulation results for an exposure with four levels but without the constant OR assumption.
(DOCX)

**Exhibit S3** Simulation results when the exposure is in a continuous scale.
(DOCX)

**Exhibit S4** Simulation results when there is an interaction effect between the two exposures.
(DOCX)

**Exhibit S5** Simulation results for a confounder.
(DOCX)

**Exhibit S6** Simulation results when the disease prevalence is lower.
(DOCX)

## Author Contributions

Conceived and designed the experiments: WCL. Performed the experiments: TTC. Analyzed the data: TTC. Contributed reagents/materials/analysis tools: WCL. Wrote the paper: TTC WCL.

## References


1. Kupper LL, McMichael AJ, Spirtas R (1975) A hybrid epidemiologic study design useful in estimating relative risk. J Am Stat Assoc 70: 524–528.
2. Miettinen OS (1982) Design options in epidemiologic research: an update. Scand J Work Environ Health 8 (Suppl. 1): 7–14.
3. Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11.
4. Self SG, Prentice RL (1988) Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Stat 16: 64–81.
5. Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50: 1064–1072.
6. Barlow WE, Ichikawa L, Rosner D, Izumi S (1999) Analysis of case-cohort designs. J Clin Epidemiol 52: 1165–1172.
7. Scheike TH, Martinussen T (2004) Maximum likelihood estimation for Cox's regression model under case-cohort sampling. Scand J Stat 31: 283–293.
8. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009) Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol 169: 1398–1405.
9. Marti H, Chavance M (2011) Multiple imputation analysis of case-cohort studies. Stat Med 30: 1595–1607.
10. Sato T (1992) Maximum likelihood estimation of the risk ratio in case-cohort studies. Biometrics 48: 1215–1221.
11. Sato T (1994) Risk ratio estimation in case-cohort studies. Environ Health Persp 102 (Suppl. 8): 53–56.
12. Zhang J, Yu KF (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. J Am Med Assoc 280: 1690–1691.
13. McNutt LA, Wu C, Xue X, Hafner JP (2003) Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol 157: 940–943.
14. Carter RE, Lipsitz SR, Tilley BC (2005) Quasi-likelihood estimation for relative risk regression models. Biostatistics 6: 39–44.
15. Lumley T, Kronmal R, Ma S (2006) Relative risk regression in medical research: models, contrasts, estimators and algorithms. University of Washington Biostatistics Working Paper Series, Working Paper 293. Available: http://www.bepress.com/uwbiostat/paper293. Accessed July 2006.
16. Marschner IC, Gillett AC (2012) Relative risk regression: reliable and flexible methods for log-binomial models. Biostatistics 13: 179–192.
17. Diaz-Quijano FA (2012) A simple method for estimating relative risk using logistic regression. BMC Med Res Meth 12: 14–19.