Converting Odds Ratio to Relative Risk with Partial Data Information

Zhu Wang
Connecticut Children's Medical Center
University of Connecticut School of Medicine

Abstract

In medical and epidemiological studies, odds ratio is a commonly applied measure to approximate the relative risk. It is well known such an approximation is poor and can generate misleading conclusions, if the incidence rate of a study outcome is not rare. In the literature, there are times that the incidence rate is not directly available, thus using odds ratio as an approximation of relative risk can lead to potentially questionable conclusions. Motivated by real applications, this paper presents methods to convert odds ratio to relative risk when published data offers limited information. Specifically, the proposed new methods can convert odds ratio to relative risk, if an odds ratio and a confidence interval as well as the sample sizes for the treatment and control group are available. The objective is novelly mapped into a constrained nonlinear optimization problem, which is solved with both a grid search and nonlinear optimization algorithm. The methods are implemented in R package **orsk** which contains R functions and a Fortran subroutine for efficiency. The proposed methods and software are illustrated with real data applications.

Keywords: odds ratio, relative risk, nonlinear optimization, grid search, multiple roots, R.

1. Introduction

Investigators of medical and epidemiological studies are often interested in comparing a risk of a binary outcome between a treatment and control group, or between exposed and unexposed. Such an outcome can be an onset of a disease or infection. A risk is commonly measured by the odds ratio which evaluates whether the probability of a study outcome is the same for two groups. An odds ratio is a positive number which can be 1 (the outcome of interest is similarly to occur in both groups), or greater than 1 (the outcome is more likely to occur in the treatment group), or less than 1 (the outcome is less likely to occur in the treatment group). A related quantity is the relative risk, which is a more direct measure than the odds ratio. The relative risk is the ratio of the probability of the outcome occurring in the treatment group versus a control group, and is best estimated using a population sample. In addition, it can be easily shown that the odds ratio is a good approximation to the relative risk when the incidence rate is low, for instance, in rare diseases, and can largely overestimate the relative risk when the outcome is common in the study population (Zhang and Yu 1998; Robbins *et al.* 2002). Although it is well-known that the two measures evaluate different quantities in general, the odds ratio has been mis-interpreted as relative risk in some studies, and thus led to incorrect conclusions (Schulman *et al.* 1999; Schwartz *et al.* 1999; Holcomb *et al.* 2001). For

this reason, many methods have been proposed to adjust the odds ratio estimates, particularly in the logistic or other multiple regression models. For instance, see a popular adjusted odds ratio in Zhang and Yu (1998). The formula in Zhang and Yu (1998) requires the proportion of control subjects who experience the outcome. Specifically, derived from the definition of odds ratio and relative risk, the adjusted odds ratio $=\frac{\text{odds ratio}}{1-\text{risk}_0+\text{risk}_0\times\text{odds ratio}}$, where risk_0 is the risk of having a positive outcome in the control or unexposed group. The formula can also be employed to adjust the lower and upper limits of the confidence interval.

This paper deals with a completely different issue which may or may not involve a logistic regression: how to estimate the relative risk when the required data such as risk_0 is not available? Methodologies have not been proposed to address this question, based on the author's best knowledge. This question is practically important and provided challenges in Holcomb *et al.* (2001). To determine how often the odds ratio differs substantially from the relative risk estimates and to investigate whether the difference in these measures implies misinterpretation of clinical research results, Holcomb *et al.* assessed 112 clinical research articles in obstetrics and gynecology during 1998-1999. Because five articles didn't contain the information about risk_0, these investigators had to skip them when computing the adjusted odds ratio using the formula in Zhang and Yu (1998). More importantly, it remains unclear whether the odds ratio exaggerates a risk association or a treatment effect in these studies. Thus, the methods proposed here not only can convert odds ratio to relative risk, but also can be further utilized to estimate risk_0. In this sense, we extend the work in Zhang and Yu (1998) to the case where risk_0 is not directly available.

To motivate the proposed methods, a concrete example is presented below. Lee *et al.* (2010) investigated the effects of preoperative, broad-spectrum antibiotics for treatment of nonperforated appendicitis in children. Some of the results are reproduced in Table 1, which was originally extracted from a Cochrane review. Cochrane reviews are systematic reviews of primary research in human health care and health policy, and the evidence of the effects of healthcare interventions are summarized. Cochrane reviews are often recognized as the highest standard in evidence-based health care.

Clearly, Table 1 suggests that preoperative antibiotics significantly reduced the risk of wound infection compared to placebo. However, no information was provided in either Table 1 or Lee *et al.* (2010) regarding the incidence rate of wound infection, thus one might wonder how close the odds ratio approximates the relative risk. In this paper, we develop methods to address this question and implement the methods in R (R Development Core Team 2011) package **orsk** (odds ratio to relative risk.)

The paper is organized as follows. Section 2 proposes a nonlinear objective function which measures the closeness between the calculated odds ratio and the reported odds ratio. We also provide two methods to solve the nonlinear objective function. Section 3 outlines the

Table 1: Summary of Cochrane Database Review regarding use of antibiotics for nonruptured appendicitis.

	Odds ratio	95% confidence interval
Wound infection		
Placebo (n=2707)	Reference	Reference
Antibiotics (n=2610)	0.37	0.30-0.46

implementations in the package **orsk**. Section 4 illustrates the capabilities of **orsk** with real data in Table 1. Finally, Section 5 concludes the paper.

2. Methods

To assess a risk, we typically have a contingency table as Table 2 displays. From Table 2, the odds of outcome in the treatment group is n_{11}/n_{10} and the odds of outcome in the control group is n_{01}/n_{00}, then the odds ratio is

$$\theta = \frac{n_{11}n_{00}}{n_{10}n_{01}}. \tag{1}$$

With asymptotic assumptions, a $(1-\alpha)$ confidence interval (CI) for the log odds ratio is $\log(\theta) \pm z_{\alpha/2}SE$, where $z_{\alpha/2}$ is the $\alpha/2$ upper critical value of the standard normal distribution and the standard error SE can be estimated by $\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{10}}+\frac{1}{n_{01}}+\frac{1}{n_{00}}}$. The lower bound of the odds ratio can thus be mapped to $\theta_L = \exp(\log(\theta) - z_{\alpha/2}SE)$. Therefore,

$$\theta_L = \theta \exp\left[-z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{10}}+\frac{1}{n_{01}}+\frac{1}{n_{00}}}\right]. \tag{2}$$

Similarly, the upper bound of the odds ratio is

$$\theta_U = \theta \exp\left[z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{10}}+\frac{1}{n_{01}}+\frac{1}{n_{00}}}\right]. \tag{3}$$

Now, the problem to be solved can be stated as follows. Suppose $x, y, \theta, \theta_L, \theta_U$ and α are fixed and known, the aim is to estimate (n_{01}, n_{11}) and subsequently estimate the relative risk by $\frac{n_{11}}{n_{11}+n_{10}}/\frac{n_{01}}{n_{01}+n_{00}}$ with the corresponding confidence interval. In the layout of Table 1, we have $x = 2707, y = 2610, \theta = 0.37, \theta_L = 0.30, \theta_U = 0.46, \alpha = 0.05$. Subject to rounding errors, the task is approximately equivalent to solving different sets of nonlinear equations for two unknowns (n_{01}, n_{11}) given that $n_{01} + n_{00} = x$ and $n_{11} + n_{10} = y$: (i) Equations (1) and (2); (ii) Equations (1) and (3); (iii) Equations (1) to (3); (iv) Equations (2) and (3). Since (iv) is contained in (iii), we don't treat it separately. The proposal is to choose (n_{01}, n_{11}) through minimizing the sum of squared logarithmic deviations between the reported estimates $\theta, \theta_L, \theta_U$ and the corresponding would-be-estimates based on assumed n_{01} and n_{11}. In mathematical form, consider a sum of squares SS in three scenarios incorporating different combinations of Equations (1) to (3):

Table 2: Compute odds ratio.

Group	Number of outcome	Number of outcome free	Total
Control	n_{01}	n_{00}	x
Treatment	n_{11}	n_{10}	y

- Odds ratio and lower confidence interval:

$$SS(n_{01}, n_{11}) = \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} - \log(\theta) \right\}^2 + \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} - z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{y - n_{11}} + \frac{1}{n_{01}} + \frac{1}{x - n_{01}}} - \log(\theta_L) \right\}^2, \tag{4}$$

- Odds ratio and upper confidence interval:

$$SS(n_{01}, n_{11}) = \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} - \log(\theta) \right\}^2 + \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} + z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{y - n_{11}} + \frac{1}{n_{01}} + \frac{1}{x - n_{01}}} - \log(\theta_U) \right\}^2, \tag{5}$$

- Odds ratio and two-sided confidence interval:

$$SS(n_{01}, n_{11}) = \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} - \log(\theta) \right\}^2 + \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} - z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{y - n_{11}} + \frac{1}{n_{01}} + \frac{1}{x - n_{01}}} - \log(\theta_L) \right\}^2 + \left\{ \log \frac{n_{11}(y - n_{01})}{(x - n_{01})n_{01}} + z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{y - n_{11}} + \frac{1}{n_{01}} + \frac{1}{x - n_{01}}} - \log(\theta_U) \right\}^2, \tag{6}$$

The goal now is to solve the following problem:

$$\min_{n_{01}, n_{11}} SS(n_{01}, n_{11}) \text{ for integer } n_{01}, n_{11}, 1 \le n_{01} \le x - 1, 1 \le n_{11} \le y - 1. \tag{7}$$

Apparently SS will be very close to 0 for the true value of (n_{01}, n_{11}), and a smaller SS implies a better solution. Thus SS plays a role similar to the residual sum of squares in the linear regression. Implementing different objective functions (4) to (6) provides a means of cross-checking results. Ideally, the solutions should be consistent when minimizing any one of the objective functions. However, sometimes data are corrupted and inconsistent results may occur. Indeed, an application of different objective functions discovered a suspicious odds ratio and confidence interval in Lee *et al.* (2010) (not Table 1), which was formally reported in Wang (2011).

To solve the constrained optimization problem, we consider two approaches: the exhaustive grid search and a numerical optimization algorithm. In the first algorithm, the minimization can be performed as a two-way grid search over the choice of (n_{01}, n_{11}). In other words, we can evaluate all the values $SS(n_{01}, n_{11})$, for $n_{01} \in \{1, 2, ..., x - 1\}, n_{11} \in \{1, 2, ..., y - 1\}$. This will result in a total number of $(x - 1)(y - 1)$ of SS for comparison. To reduce the computational demand when $(x - 1)(y - 1)$ is large as in Table 1, we adopt a filtering procedure. Specifically, we filter out SS if $SS > \delta$ for a prespecified small threshold value δ, with a default value 10^{-4}. Apparently, a smaller threshold value δ can lead to sparser solutions; however, the algorithm may fail to obtain a solution if δ is too close to 0. The problem can also be solved by applying

numerical optimization techniques. Here we consider a spectral projected gradient method implemented in R package **BB** (Varadhan R 2009). This package can solve for large scale optimization with simple constraints. It takes a nonlinear objective function as an argument as well as basic constraints. In particular, the package can find multiple roots if available, with user specified multiple starting values. To this end, starting values for n_{01} are randomly generated from 1 to $x-1$. Similarly, starting values for n_{11} are randomly generated from 1 to $y-1$. We then form $\min(x-1, y-1)$ pairs of random numbers and select 10% as the starting values to find multiple roots. Once the solutions (n_{01}, n_{11}) are determined, the odds ratio and relative risk can be computed, and selected results can be arranged in the order of the magnitude of SS. It is worth emphasizing that the calculated odds ratios are for the scenarios created with different numbers of events in both treatment and control group that lead to comparable results for the reported odds ratio and confidence interval.

3. Implementation

The above methods have been implemented in R package **orsk**. To make the grid search algorithm computationally efficient, a Fortran subroutine is utilized. Several supporting R functions are available to extract or calculate useful statistics, such as the reported odds ratio, estimated odds ratio and relative risk, with confidence intervals. The function `orsk` returns an object of class `orsk`, for which `print` and `summary` method are available. A detailed description of these functions is available in the online help files. The optimization objective functions (4) to (6) can be called with argument `type="lower"`, `type="upper"` and `type="two-sided"`, respectively, and the default value is `"two-sided"`. With argument `method="grid"`, the grid search algorithm is called. Otherwise, the constrained nonlinear optimization algorithm is employed. The results can be illustrated using the `summary` function and argument `nlist` controls the maximum number of solutions displayed (the default value is 5). The source version of the **orsk** package is freely available from the Comprehensive R Archive Network (http://CRAN.R-project.org). The reader can install the package directly from the R prompt via

```
R> install.packages("orsk")
```

All analyses presented below are contained in a package vignette. The rendered output of the analyses is available by the R-command

```
R> library("orsk")
R> vignette("orsk_demo", package = "orsk")
```

To reproduce the analyses, one can invoke the R code

```
R> edit(vignette("orsk_demo", package = "orsk"))
```

4. Example

The data in Table 1 are used to illustrate the capabilities of **orsk**. These analyses were conducted using R version 2.14.0 (2011-10-31) and the operating system `i686-pc-linux-gnu (32-bit)`.

We applied both grid search and optimization algorithms for minimizing objective function (6) and the solutions are similar for (4) or (5). As seen below, the output includes two parts: setup and results. The setup describes the configurations of the optimization problem and the results include the solution n_{01} and n_{11}, named as `cont_yes` and `trt_yes`, respectively. In the ascending order of SS, the output also includes the estimated odds ratio with confidence interval derived from the estimate (n_{01}, n_{11}), along with the known x and y. The estimated odds ratios and confidence intervals in the output are very close to the reported value $0.37(0.30, 0.46)$. However, the derived relative risks and confidence intervals can be dramatically different. The results show that the estimated relative risks are clustered around 0.40 or 0.92. The confidence intervals can also be roughly clustered into two modes. These two clusters correspond to distinct assumptions: the former is a low incidence of wound infection (the 2nd, 4th and 5th solution), for which the odds ratio is expected to approximate the relative risk very well; on the contrary, the latter assumes a common occurrence of wound infection (the 1st and 3rd solution), for which the odds ratio poorly approximates the relative risk. In this example, the latter assumption is not realistic. In the situations under consideration it can be expected that there is often no unique solution. As such, the user should carefully investigate the results from running R package **orsk**. It can be possible that it is not clear at all which of the computational results can be taken for further analysis. But this is not unusual for an exploratory study. On the other hand, one may reasonably hope that a subject matter expert can provide valuable insights to the situation and may help make a decision.

```
R> library("orsk")

R> res1 <- orsk(x = 2707, y = 2610, a = 0.37, al = 0.3,
+      au = 0.46, method = "grid")
R> summary(res1)

         Converting odds ratio to relative risk

Call:
orsk(x = 2707, y = 2610, a = 0.37, al = 0.3, au = 0.46, method = "grid")

type:  two-sided               method:  grid
threshold value:  1e-04        maximum number of solution listed:  5
The reported odds ratio: 0.37, confidence interval 0.3, 0.46

estimated results. The calculated odds ratios and relative risks are for
 the scenarios created with different numbers of events in both control and
 treatment group that lead to comparable results for the reported odds ratio
 and confidence interval.
  ctr_yes ctr_no trt_yes trt_no       SS    OR OR_lower OR_upper
1    2579    128    2302    308 1.07e-05 0.371    0.300    0.459
2     323   2384     125   2485 1.25e-05 0.371    0.300    0.460
3    2578    129    2300    310 1.35e-05 0.371    0.300    0.459
4     321   2386     124   2486 1.38e-05 0.371    0.299    0.460
5     328   2379     127   2483 1.41e-05 0.371    0.300    0.459
```

```
     RR RR_lower RR_upper
1 0.926    0.911    0.941
2 0.401    0.329    0.490
3 0.925    0.910    0.941
4 0.401    0.328    0.489
5 0.402    0.330    0.489
```

When applying the optimization algorithm, the estimated results typically have larger SS than the grid search algorithm. Note the solutions may not be replicated since the starting values are randomly generated as described in Section 2. Similarly to the grid search algorithm, the estimated relative risks are clustered around 0.40 or 0.92. To show this, one may have to enlarge `nlist` in the R function `summary`.

```
R> require("setRNG")
R> old.seed <- setRNG(list(kind = "Mersenne-Twister", normal.kind = "Inversion",
+      seed = 1234))
R> res2 <- orsk(x = 2707, y = 2610, a = 0.37, al = 0.3,
+      au = 0.46, method = "optim")
R> summary(res2)

         Converting odds ratio to relative risk

Call:
orsk(x = 2707, y = 2610, a = 0.37, al = 0.3, au = 0.46, method = "optim")

type:  two-sided              method:  optim
threshold value:  NA          maximum number of solution listed:  5
The reported odds ratio: 0.37, confidence interval 0.3, 0.46

estimated results. The calculated odds ratios and relative risks are for
 the scenarios created with different numbers of events in both control and
 treatment group that lead to comparable results for the reported odds ratio
 and confidence interval.
  ctr_yes ctr_no trt_yes trt_no       SS    OR OR_lower OR_upper
1    2577    130    2298    312 2.24e-05 0.372    0.301    0.459
2     330   2377     128   2482 2.39e-05 0.371    0.301    0.459
3    2582    125    2309    301 2.41e-05 0.371    0.299    0.461
4     316   2391     122   2488 2.42e-05 0.371    0.299    0.461
5    2581    126    2306    304 2.46e-05 0.370    0.299    0.459
     RR RR_lower RR_upper
1 0.925    0.910    0.940
2 0.402    0.330    0.490
3 0.928    0.913    0.943
4 0.400    0.327    0.490
5 0.927    0.912    0.942
```

We now compare the computing speed between the two estimating methods. With the grid search and optimization algorithm in the above example, it took 3.4 and 2.3 seconds, respec-

tively, on an ordinary desktop PC (Intel Core 2 CPU, 1.86 GHz). Although the optimization method has some computational advantage, the grid search method can generate more accurate results with smaller SS. In the light of the computing time difference, there is no real benefit of using the optimization based method. From the code development perspective, the optimization based method is useful since it provides the solutions to which the grid method can be compared.

5. Conclusion

In this article we have outlined the methods and algorithms for converting the odds ratio to the relative risk when only partial data information is available. As an exploratory tool, R package **orsk** can be utilized for this purpose. In addition, the methods can be used in the formula in Zhang and Yu (1998) to adjust the odds ratio obtained from the logistic regression, when risk_0 (risk of having a positive outcome in the control or unexposed group) is not directly available but can be estimated by applying **orsk**.

References

Holcomb WL, Chaiworapongsa T, Luke DA, Burgdorf KD (2001). "An odd measure of risk: use and misuse of the odds ratio." *Obstetrics & Gynecology*, **98**, 685–688.

Lee SL, Islam S, Cassidy LD, Abdullah F, Arca MJ (2010). "Antibiotics and appendicitis in the pediatric population: An American Pediatric Surgical Association Outcomes and Clinical Trials Committee Systematic Review." *Journal of Pediatric Surgery*, **45**(11), 2181–2185.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robbins AS, Chao SY, Fonseca VP (2002). "What's the relative risk? A method to directly estimate risk ratios in cohort studies of common outcomes." *The Annals of Epidemiology*, **12**, 452–454.

Schulman KA, Berlin JA, Harless W, Kerner JF, Sistrunk S, Gersh BJ, Dube R, Taleghani CK, Burke JE, Williams S, Eisenberg JM, Escarce JJ (1999). "The effect of race and sex on physicians' recommendations for cardiac catheterization." *New England Journal of Medicine*, **340**, 618–626.

Schwartz LM, Woloshin S, Welch HG (1999). "Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization." *New England Journal of Medicine*, **341**, 279–283.

Varadhan R GP (2009). "**BB**: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function." *Journal of Statistical Software*, **32(4)**. URL http://www.jstatsoft.org/v32/i04/.

Wang Z (2011). "Letter to the editor. 'Antibiotics and appendicitis in the pediatric population: an American Pediatric Surgical Association Outcomes and Clinical Trials Committee Systematic Review'." *Journal of Pediatric Surgery*, **46**(4), 787–788.

Zhang J, Yu KF (1998). "What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes." *Journal of the American Medical Association*, **280**, 1690–1691.

Affiliation:

Zhu Wang
Department of Research
Connecticut Children's Medical Center
Department of Pediatrics
University of Connecticut School of Medicine
Connecticut 06106, USA
E-mail: zwang@ccmckids.org