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 epidemiology studies, odds ratio is a commonly applied measure to approximate 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 some, but limited information. The implemented R package **orsk** can convert odds ratio to relative risk, if an odds ratio estimate and a confidence interval as well as the total sample sizes of treatment and control group are available. The objective is novelly mapped into a constrained nonlinear optimization problem, which is solved with both nonlinear optimization and grid search algorithm. The package contains R functions which interface underlying Fortran routines for efficiency. A couple of real data applications illustrate the proposed methods and software.

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

1. Introduction

Investigators of medical and epidemiology 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. The odds ratio is a commonly reported measure in these applications. The odds ratio is a way of comparing whether the probability of a study outcome is the same for two groups. An odds ratio of 1 indicates that the interested outcome is similarly to occur in both groups. An odds ratio greater than 1 indicates that the outcome is more likely to occur in the treatment group. And an odds ratio less than 1 indicates that the outcome is less likely to occur in the treatment group. In medical and epidemiology studies, the relative risk is a more direct measure comparing the risks 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. The relative risk is best estimated using a population sample, but 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. However, when the outcome is common in the study population, the odds ratio can largely overestimate the relative risk (Zhang and Yu 1998; Robbins, Chao, and Fonseca 2002). Although it is well-known that they measure different quantities which are only close to each other in case of rare events, the odds ratio has been mis-interpreted as relative risk in some studies, thus led to misleading conclusions when the outcome is not rare (Schulman, Berlin, Harless, Kerner, Sistrunk, Gersh, Dube, Taleghani, Burke, Williams, Eisenberg, and Escarce 1999; Schwartz, Woloshin, and Welch 1999; Holcomb, Chaiworapongsa, Luke, and Burgdorf 2001). There are methods to adjust the odds ratio when the outcome is not rare. Zhang and Yu (1998) proposed a method to estimate the relative risk from the odds ratio (also see McNutt, Wu, Xue, and Hafner (2003) for controversial observation). The formula in Zhang and Yu (1998) requires the proportion of control subjects who experience the outcome. A natural question to ask is: can we estimate the relative risk when such information is not available? The answer is practically important. For instance, due to lack of this information, Holcomb *et al.* (2001) had to ignore some studies: "Articles in which this information was missing could not be used for risk ratio estimates". For clarity, a concrete example is presented below. A study evaluates if children with nonperforated appendicitis should receive preoperative, broad-spectrum antibiotics (Lee, Islam, Cassidy, Abdullah, and Arca 2010), and some of the results are reproduced in Table 1.

Table 1: Summary of Cochran 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

Clearly, Table 1 suggests that preoperative antibiotics significantly reduced the risk of wound infection compared to placebo. Since the paper provided no information regarding the incidence rate of wound infection, one would 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**.

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

2. Methods

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

In 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

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

A 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 $SE = \sqrt{1/n_{11} + 1/n_{10} + 1/n_{01} + 1/n_{00}}$. The lower bound of odds ratio can be thus mapped to $\theta_L = \exp(\log(\theta) - z_{\alpha/2}SE)$. Therefore,

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

Similarly, the upper bound of odds ratio is

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

Suppose x, y, θ, θ_L and α are fixed and known, as in Table 1, the objective is to estimate (n_{01}, n_{11}) and subsequently estimate the relative risk. If it weren't for rounding errors, the task would be equivalent to solving two equations (1) and (2) for two unknowns given that $n_{01} + n_{00} = x$ and $n_{11} + n_{10} = y$. Alternatively, different sets of equations can be solved: equation (1) and (3), or equation (2) and (3). Although the paper is attempting to recover unpublished information, the problem can be interpreted from a sample size perspective as well. Confidence interval-based sample size determination methods have been proposed in which the upper and lower confidence limits are treated as random variables (Cornfield 1956; Satten and Kupper 1990). Instead, we are seeking sample size requirements for requested odds ratio with fixed confidence limits by solving equation (1) and (2). Because of rounding errors, the equations do not solve exactly. The paper thus proposals a different approach by choosing n_{01} and n_{11} through minimizing the sum of squared deviations between the estimates θ and θ_L and the corresponding would-be-estimates based on assumed n_{01} and n_{11}. Specifically, consider a sum of squares SS:

$$\begin{aligned} SS(n_{01}, n_{11}) =& \{n_{11}(y - n_{01})/(x - n_{01})n_{01} - \theta\}^2 \\ &+ \left\{\theta \exp\left[-z_{\alpha/2}\sqrt{1/n_{11} + 1/(y - n_{11}) + 1/n_{01} + 1/(x - n_{01})}\right] - \theta_L\right\}^2, \end{aligned} \tag{4}$$

and we aim to solve the following problem:

$$\min_{n_{01}, n_{11}} SS(n_{01}, n_{11}) \text{ for integer } n_{01}, n_{11}, 1 \leq n_{01} \leq x - 1, 1 \leq n_{11} \leq y - 1. \tag{5}$$

Alternatively, SS can be defined based on the upper bound θ_U. It is clear that SS will be very close to 0 for the true value of (n_{01}, n_{11}) from which θ, θ_L are computed. In general, a smaller SS implies a better solution of (n_{01}, n_{11}). Thus SS serves similar to the residual sum of squares in the linear regression. To solve the constrained optimization problem, we consider two approaches: the exhaustive grid search and a numerical optimization algorithm. For the grid search method, the minimization can be conducted 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 total number of $(x-1)(y-1)$ of SS. Next, we filter out SS if $SS > \delta$ for a prespecified small threshold value δ. Apparently, a smaller threshold value δ can lead to sparser solutions; if δ is too close to zero, however, the algorithm may fail to obtain a solution. The relationship between the choice of δ and the number of solutions will be investigated below, and it is found in an empirical study that with a few choice of δ, the range of number of solutions can be explored. For each of the selected (n_{01}, n_{11}), the relative risk and its confidence interval are computed, along with the odds ratio and its confidence interval. The results are rearranged by the order 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.

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. The solutions are round to integers and the odds ratios $\hat{\theta}$ are computed afterwards. Next, we adopt a filtering procedure: the solutions are remained only if $|\hat{\theta} - \theta|/\theta \leq \delta$.

3. Implementation

The above methods have been implemented in R package **orsk**. To make the grid search algorithm computational efficient, R package `orsk` calls Fortran subroutines. 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 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. With argument `method` equal to `"grid"`, the grid search algorithm will be called. Otherwise, the constrained nonlinear optimization algorithm will be employed with `method="optim"`. The results can be illustrated using `summary` function. The source version of **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. Examples

The data in Table 1 and in Berg-Lekas, Hogberg, and Winkvist (1998) are used to illustrate the capabilities of **orsk**. These analyses were conducted using R version 2.10.1 (2009-12-14) and the operating system `i686-pc-cygwin`.

We applied **orsk** with both optim and grid search methods to Table 1. As seen below, the output includes (n_{01}, n_{11}), named as `cont_yes` and `trt_yes`, respectively. The results also include the corresponding estimated odds ratio with confidence interval. The estimated numbers are very close to the reported value 0.37 and its confidence interval $(0.30, 0.46)$. 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 $0.91 - 0.92$. The confidence intervals can also be roughly clustered into two modes. These two clusters correspond to distinct assumptions: the former is low incidence of wound infection (n_{01}/x and n_{11}/y are small), in which the odds ratio is expected to approximate the relative risk very well; the latter is for common occurrence of wound infection (n_{01}/x and n_{11}/y are large), for which the odds ratio poorly approximates the relative risk. In this context, the latter assumption is not realistic. In general, the decision for an appropriate scenario can be easily made with subject knowledge.

However, the analysis here does provide an example that the odd ratio itself can not provide the complete risk assessment, unlike the relative risk.

```
R> library("orsk")

R> res1 <- orsk(x = 2707, y = 2610, a = 0.37, al = 0.3,
+      au = 0.46, method = "optim", d = 1.1e-05)
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 = "optim",
    d = 1.1e-05)

Method:  optim
Threshold value:  1.1e-05
The reported odds ratio: 0.37, confidence interval 0.3, 0.46
The 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    2556    151    2249    361 7.30e-06 0.368    0.302    0.449
2     310   2397     119   2491 1.04e-05 0.369    0.297    0.460
3     389   2318     152   2458 1.09e-05 0.368    0.303    0.448
     RR RR_lower RR_upper
1 0.913    0.896    0.929
2 0.398    0.325    0.488
3 0.405    0.339    0.485

R> res2 <- orsk(x = 2707, y = 2610, a = 0.37, al = 0.3,
+      au = 0.46, method = "grid", d = 1e-07)
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 = "grid",
    d = 1e-07)

Method:  grid
Threshold value:  1e-07
The reported odds ratio: 0.37, confidence interval 0.3, 0.46
The 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    RR
1    2573    134    2288    322 4.15e-08 0.37      0.3    0.456 0.922
2     336   2371     130   2480 5.09e-08 0.37      0.3    0.456 0.401
  RR_lower RR_upper
1    0.907    0.938
2    0.330    0.488
```

We now compare the computing speed between the two estimating methods. With optim and grid search method in the above specifications, it took 1.8 seconds, and 3.2 seconds, respectively, on an ordinary desktop PC (Intel Core 2 CPU, 1.86 GHz). Although the optimization method has some computation advantage, the grid search method can generate more accurate results since the ratio of SS is less than 0.6% from the corresponding best estimation results for the two methods.

To study the relationship between the threshold value δ and the number of solution, we use the same data with varying δ. Figure 1 demonstrates that the number of solution converges to zero very quickly for the optim and grid search method. This implies that one can explore the whole scope of solutions with only a few varying δ. Notice that the horizontal axis has different ranges between the optim and grid search method since the former method can fail with the comparable precision of the latter.

Next, we consider an example in which the numbers of events (n_{01}, n_{11}) have been published, which allows a direct comparison between the estimated results with the reported values. In a study of the familial occurrence of dystocia (see Table 3), the authors conclude that "the risk is increased more than 20-fold (odds ratio 24.0, 95% interval 1.5 to 794.5) if one twin sister had dystocia..." (Berg-Lekas *et al.* 1998; Holcomb *et al.* 2001).

Table 3: Dysfunctional labor in primiparous women among 40 female twin couples.

	Odds ratio	95% confidence interval
Women whose twin sisters had a normal delivery (n=36, $n_{01} = 4$)	Reference	Reference
Women whose twin sisters had dystocia (n=4, $n_{11} = 3$)	24	1.99-289.6

Here we used the odds ratio confidence interval (1.99, 289.6) since (1.5, 794.5) is possibly

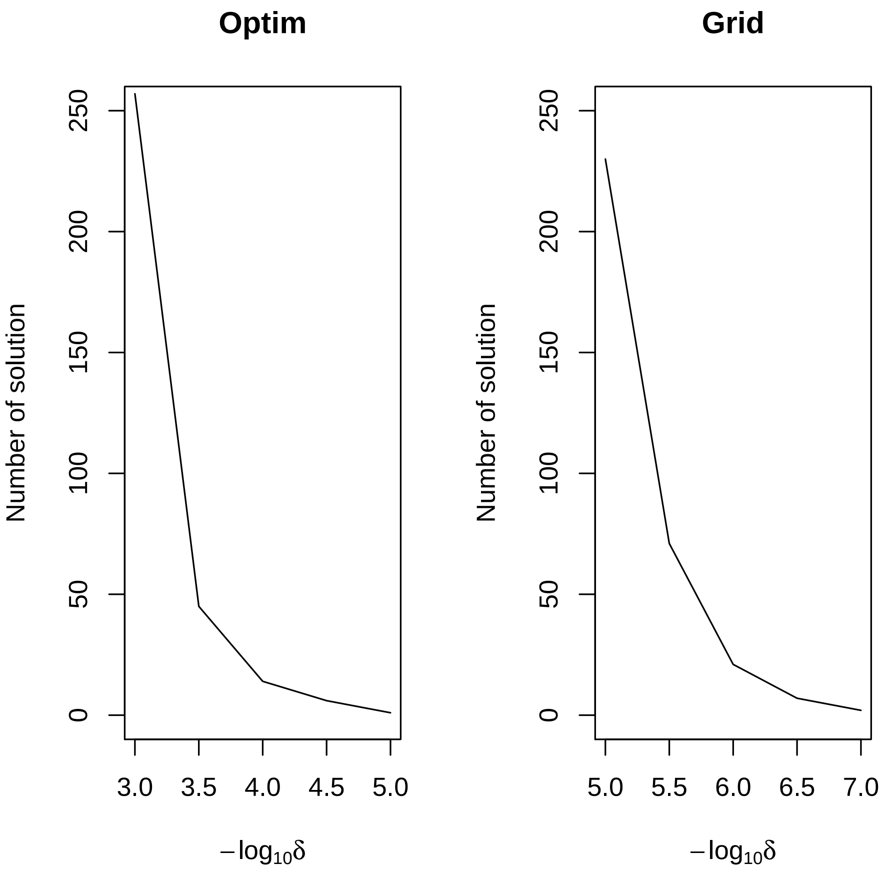

Figure 1: Threshold value and number of solution in optim and grid search method.

due to an error. The results below suggest that the proposed two methods correctly identify $(n_{01}, n_{11}) = (3, 4)$, and the odds ratio overestimate the relative risk 6.75 since the rate of dystocia was high (75%) in women whose twin sisters had dystocia, as correctly pointed out by Holcomb *et al.* (2001).

```
R> res5 <- orsk(x = 36, y = 4, a = 24, al = 1.99, au = 289.6,
+      method = "optim")
R> summary(res5)

         Converting odds ratio to relative risk

Call:
orsk(x = 36, y = 4, a = 24, al = 1.99, au = 289.6, method = "optim")

Method:  optim
Threshold value:  1e-04
The reported odds ratio: 24, confidence interval 1.99, 289.6
The 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   RR
1       4     32       3      1 1.13e-06 24     1.99      290 6.75
  RR_lower RR_upper
1     2.28     19.9

R> res6 <- orsk(x = 36, y = 4, a = 24, al = 1.99, au = 289.6,
+      method = "grid")
R> summary(res6)

         Converting odds ratio to relative risk

Call:
orsk(x = 36, y = 4, a = 24, al = 1.99, au = 289.6, method = "grid")

Method:  grid
Threshold value:  1e-04
The reported odds ratio: 24, confidence interval 1.99, 289.6
The 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   RR
1       4     32       3      1 1.13e-06 24     1.99      290 6.75
  RR_lower RR_upper
1     2.28     19.9
```

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.

References

Berg-Lekas ML, Hogberg U, Winkvist A (1998). “Familial occurrence of dystocia.” *American Journal of Obstetrics and Gynecology*, **179**(1), 117–121.

Cornfield J (1956). “A statistical problem arising from retrospective studies.” In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Statistics*, p. 135. University of California Press.

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.

McNutt LA, Wu C, Xue X, Hafner JP (2003). "Estimating the relative risk in cohort studies and clinical trials of common outcomes." *American Journal of Epidemiology*, **157**, 940–943.

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.

Satten G, Kupper L (1990). "Sample size requirements for interval estimation of the odds ratio." *American Journal of Epidemiology*, **131**(1), 177.

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/.

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, United States of America
E-mail: zwang@ccmckids.org