Annals of Epidemiology 26 (2016) 136–140

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Incorrect inference in prevalence trend analysis due to misuse of the odds ratio

Scott Mcclintock PhD [a,*], Zhen-qiang Ma MD, MPH, MS [b], Randall H. Rieger PhD [a]

[a] *Department of Mathematics, West Chester University, West Chester, PA*
[b] *Pennsylvania Department of Health, Harrisburg*

A R T I C L E I N F O

Article history:
Received 23 June 2015
Accepted 14 December 2015
Available online 12 January 2016

Keywords:
Epidemiologic methods
Education
Obesity
Modeling
Statistics

A B S T R A C T

Because public health agencies usually monitor health outcomes over time for surveillance, program evaluation, and policy decisions, a correct health outcome trend analysis is vital. If the analysis is done incorrectly and/or results are misinterpreted, a faulty trend analysis could jeopardize key aspects of public health initiatives such as program planning, implementations, policy development, and clinical decision making. It is essential then that accurate health outcome trend analysis be implemented in any data-driven decision-making process. Unfortunately, there continues to be common statistical mistakes in prevalence trend analysis. In this article, using recently published results from the Pediatric Nutrition Surveillance System, we will show the effect that an incorrect trend analysis and subsequent interpretation of results can have. We will also propose more appropriate statistical processes, such as the log-binomial model, for these situations.

Introduction

In public health studies, monitoring prevalence of a behavior or disease from an initial time to later time points, one can consider two primary goals of interest

1. Estimation of the true prevalence ratio for a binary heath outcome (i.e., obese or not) at the desired time points.
2. Measurement of the rate of change in prevalence ratio from the initial time point to subsequent time points and determination if that trend is statistically significant.

There has been much discussion about the appropriate statistics to use for the first goal. Both odds ratios and risk ratios are widely used as estimates of the true prevalence ratio in the population when outcomes are binary. Although the odds ratio is recommended in certain situations such as case-control studies, it has potential problems in other contexts. The odds ratio can only be considered as an approximation to the prevalence ratio when health outcomes are relatively rare (less than 10%) [1]. For health outcomes that are relatively common (>10% of the population), such as obesity and smoking, medical and public health researchers have repeatedly been advised to use a risk ratio rather than an odds ratio [2–4]. Odds ratios are systematically more extreme than their corresponding prevalence ratios, leading to a tendency to overestimate the true effect in the population [5]. Additionally, the interpretation of the odds ratio is a difficult concept for many, whereas the risk ratio can be easily interpreted as a ratio of two prevalence estimates. In spite of this, the odds ratio continues to be a popular estimator of prevalence ratio, even when the prevalence is high. This misuse of the odds ratio can be seen in many published results, some as recent as 2015 [2,5–8]. It is for these reasons that the risk ratio is generally recommended for cross-sectional, longitudinal, and trend analyses, particularly for common outcomes. These are the types of studies that our article intended to address.

Although much has been published about estimation of the prevalence ratio, there has been less attention paid to the modeling techniques used to estimate and test the trend in prevalence. Just as the choice of risk ratio versus odds ratio can make an appreciable difference in the estimation of the true prevalence ratios in the population, the choice of an incorrect model can make a substantial difference in the estimation of the trend and its significance. Logistic regression models estimate changes in odds ratios not changes in prevalence ratios. If the prevalence in the population is large enough such that the odds ratio is not a good estimate of changes in prevalence, for the same reasons, the logistic regression model should not be used. Instead, the log-binomial model [9] is a more appropriate model for estimating prevalence ratios. It uses

* Corresponding author. 25 University Avenue, West Chester, PA 19383. Tel.: +1-610-430-4963; fax: +1-610-738-0578.
E-mail address: smcclintoc@wcupa.edu (S. Mcclintock).

http://dx.doi.org/10.1016/j.annepidem.2015.12.009
1047-2797/

risk ratio estimates for testing trends in prevalence rate. Some researchers continue to use logistic regression models, even when their goal is to estimate (and test) trends in prevalence in a common outcome. There are many published studies that use logistic regression to test for trends of high prevalence conditions such as obesity and diabetes [10–13]. We will demonstrate how the use of the more appropriate log-binomial model can change the conclusions for prevalence trend studies using published obesity data from May et al. [14].

This article is structured as follows: In section 2, we discuss the methodological differences between the logistic regression model, which measures changes in odds ratios and the log-binomial model, which measures changes in risk ratios. In section 3, we further compare and contrast odds ratios and risk ratios. In section 4, we reanalyze data from the Pediatric Nutrition Surveillance System using log-binomial models. In section 5, we make recommendations for future prevalence analysis.

Section 2

Traditionally, the significance of a potential relationship between a single dichotomous dependent variable and a collection of k independent variables has been tested and measured using logistic regression. The logistic regression model is defined as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$$

This model has certain advantages. The foremost is that when solving for p, for a given set of covariates, the solution is guaranteed to be sensibly bounded between 0 and 1. In this model, beta coefficients are in terms of odds ratios rather than prevalence ratios.

Another model that has been gaining support [4,15–17] for a wide variety of analyses, including trend analysis, is the log-binomial model. It is defined as:

$$\log(p) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$$

The log-binomial model more directly estimates p, the parameter of interest. It was discussed and described as early as 1986, by Wacholder, as an alternative to logistic regression [9], but this model has been underused because it can be computationally difficult to fit. However, nowadays, with increased computational power and new algorithms and software, this is less of an issue. This model can now be fit using simple syntax in either SAS [15] or R [18], and a variety of alternatives has been offered to handle the situations where parameter estimates for the model fail to converge. These include computational improvements to the log-binomial estimation process [19] as well alternative modeling techniques such as the Cox proportional hazard model [20,21] and the Robust Poisson model [22]. Because of this, these alternatives are increasingly recommended over the logistic regression model [4,15,17–22].

Another major advantage of the log-binomial model over the logistic regression model is that the standard errors of coefficients tend to be smaller. This has been shown in a number of studies [4,16,17]. We can gain some insight into this by comparing the unadjusted odds ratios to the unadjusted risk ratios for the special case of a single binary independent variable. In this case, the model estimates become equivalent to using the estimation formulas from a 2 by 2 table. The formulas for the standard error are:

$$\text{SE of Ln (OR)} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$\text{SE of Ln (RR)} = \sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}$$

Basic algebra for this simple case, assuming all entries are nonzero, shows that the standard error of the risk ratio will always be smaller than that of the odds ratio. These have important implications. Although it is true that, for low prevalence, the odds ratio and risk ratio will be quite similar, we can still get a more precise estimate of this shared value by using the risk ratio or, equivalently, the log-binomial model without covariates.

Section 3

We can get a better sense of the differences in standard errors, as well as the discrepancies between risk ratio and odds ratio by considering a range of possible prevalence rates. Table 1 shows what happens to odds ratios and risk ratios as the prevalence increases over a time span of 11 time points. Table 1 is intended to simulate generic trend analyses, similar to that conducted by May et al., where a risk ratio is recommended over an odds ratio. We selected evenly spaced prevalence rates ranging from 2.5% to 35% to illustrate differences between odds ratios and risk ratios for both small and large prevalence rates. In this case, we are generating odds ratios and risk ratios by comparing prevalence rates at the time point of interest to those at time point one. The sample size of 56,800 is taken from the Pediatric Nutrition Surveillance System as the sample size of Maryland, a medium-sized state. At the bottom of the table, we also include standard error estimates for the risk ratio and odds ratio at both the original sample size and one fourth of the original sample size. This reduced sample size would be representative of a smaller state such as Hawaii.

The table clearly shows that as the prevalence increases the difference between the odds ratio and risk ratio also increases. This ever-increasing divergence is further shown by comparing the two lines in Figure 1.

The divergence between odds ratios and risk ratios, noted above, is a widely cited phenomenon. Unfortunately, much less attention has been paid to the difference in standard errors, although standard errors are a critical component of statistical inference. An incorrect estimate of standard error can lead to incorrect conclusions—both type I and type II errors in hypothesis tests and confidence intervals that are either too wide or too narrow. In Table 1, for small prevalence rates, the standard errors are quite different, with the risk ratio having a smaller standard error than the odds ratio. However, this difference diminishes as the prevalence rate increases. These differences are further illustrated in Figure 2. It is important to note that it is within the generally recommended range of prevalence values ($<$10%) for estimating the prevalence ratio with the odds ratio that we are seeing the largest difference in standard errors.

Another important observation comes from the lower part of Table 1. The differences between odds ratios and risk ratios do not depend on the sample size but only on the prevalence rates. As we can see from the table; however, the differences between standard error estimates do depend on sample sizes. Comparing time point 1 to time point 2 for the original sample size, we see that the standard error of the odds ratio is 0.0461, whereas the standard error of the risk ratio is 0.0305. When we look at the same prevalence rates with a reduced sample size, the standard errors increase so that the standard error of the odds ratio is 0.0923 and that of the risk ratio is