Prevalence proportion ratios: estimation and hypothesis testing

Torsten Skov, a,b James Deddens, b,c Martin R Petersen b and Lars Endahl a

Background Recent communications have argued that often it may not be appropriate to analyze cross-sectional studies of prevalent outcomes with logistic regression models. The purpose of this communication is to compare three methods that have been proposed for application to cross sectional studies: (1) a multiplicative generalized linear model, which we will call the log-binomial model, (2) a method based on logistic regression and robust estimation of standard errors, which we will call the GEE-logistic model, and (3) a Cox regression model.

Methods Five sets of simulations representing fourteen separate simulation conditions were used to test the performance of the methods.

Results All three models produced point estimates close to the true parameter, i.e. the estimators of the parameter associated with exposure had negligible bias. The Cox regression produced standard errors that were too large, especially when the prevalence of the disease was high, whereas the log-binomial model and the GEE-logistic model had the correct type I error probabilities. It was shown by example that the GEE-logistic model could produce prevalences greater than one, whereas it was proven that this could not happen with the log-binomial model. The log-binomial model should be preferred.

Keywords Generalized linear model, Cox regression, cross sectional study, log-binomial model, GEE-logistic model

Accepted 28 May 1997

A lively discussion about the appropriateness of estimating prevalence proportion ratios versus prevalence odds ratios in cross-sectional studies started when Lee,1 and Lee and Chia2 published letters to the Editors. Stromberg3 pointed out that under certain stationarity assumptions, and provided that the mean duration of the disease in the exposed and the unexposed group is known, the incidence rate ratio can be calculated from the prevalence odds ratio. Axelson et al.4 argued that in some cases the assumptions do not apply. For example the duration of musculoskeletal disorders may well be influenced by the exposure, and, we may add, may be difficult to define. They argued that in these cases the prevalence proportion ratio (or the prevalence ratio as they called it) is more interpretable than the prevalence odds ratio. Moreover, Axelson et al.4 showed that controlling for confounding of the prevalence odds ratio may in fact give an estimate that is further away from the prevalence proportion ratio than the unadjusted prevalence odds ratio.

When the prevalence of the outcome is low, there is little difference between the prevalence odds ratio and the prevalence proportion ratio. However, many cross-sectional studies are concerned with high-prevalence outcomes. If the prevalence proportion ratio is the parameter of interest in such a study, estimation of the prevalence odds ratio with logistic regression will be a poor approximation.

There is a need, therefore, for a method that can estimate and test prevalence proportion ratios adjusted for several confounders. Wacholder5 devised a multiplicative generalized linear model in GLIM, and Zocchetti et al.6 pointed out that computations had become easier with the advent of standard procedures like GENMOD in SAS. Schouten et al.7 proposed a method based on logistic regression and robust estimation of standard errors. Finally, Lee1 advocated Cox regression for this purpose. In the present study, we compare the three methods with simulations, and discuss some theoretical and practical aspects of their use.

Material and Methods

Terminology

Whereas there seems to be agreement on the meaning of prevalence odds and prevalence odds ratio (POR), some diversity exists with regard to the measure that is interchangeably called

a National Institute of Occupational Health, Lerse Parkalle 105, DK-2100 København Ø, Denmark.
b National Institute for Occupational Safety and Health, Division of Surveillance, Hazard Evaluations and Field Studies, 4676 Columbia Parkway, Cincinnati, OH 45226–1988, USA.
c Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA.
the prevalence rate ratio, the prevalence proportion ratio, and the relative risk. We suggest the term prevalence proportion ratio (PPR). Prevalence ratio could be used, but in the strict sense prevalence denotes the number of diseased people in a population rather than the proportion of those with disease. Furthermore, it is logical to coin the term as a ratio between proportions to distinguish it from a ratio between odds.

Prevalence rate ratio is a traditional term that should be avoided because prevalence proportions are not functions of time and, thus, are not rates. With regard to relative risk, this term is inappropriate because in the context of a cross-sectional study, the PPR may not estimate a disease risk, but rather a relative probability of randomly selecting a person having a specified symptom at the time of the study. It is assumed here that time is defined as the probability of developing a health outcome during a certain period of time. The PPR may or may not in turn estimate a relative risk.

**Suggested methods**

Lee and Chia recommended that the Cox proportional hazard model be used to estimate PPR. According to Lee, Breslow had shown that with a constant risk period (equal follow-up time for all subjects), the proportional hazard model estimates the cumulative incidence ratio. Therefore, by assuming constant risk period, the Cox model could be adapted to estimate PPR for cross-sectional data.

In the Cox model, the survival time of an individual with covariates \( X = (x_1, \ldots, x_k) \) is assumed to follow a hazard function:

\[
h(t) = h_0(t) e^{\beta_1 x_1 + \cdots + \beta_k x_k}
\]

where the \( \beta_s \) are unknown constants, and \( h_0(t) \) is the baseline hazard (when all covariates are zero). The variable \( t \) represents time, which for Lee's method is set equal to a constant.

Wacholder proposed a generalized linear model with logarithmic link function and binomial distribution function for estimating PPR. We will call this the log-binomial model. Let \( Y(0/1) \) denote the absence/presence of the symptom in an individual with covariates \( X = (x_1, \ldots, x_k) \). Then

\[
p = p(Y = 1|X) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)
\]

The model is defined only if \( \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k < 0 \) for all \( x_k \). Wacholder devised a macro using the GLIM program to fit this model. It can now be easily fit with, for example, GENMOD in SAS.

Schouten et al. suggested to fit the parameters in the log-binomial model by logistic regression on a manipulated data set. The manipulation is made by duplicating every case in the data set to a non-case observation. The new data set can be divided into three groups: cases, original non-cases, and new non-cases (resulting from the duplication of the cases). The probabilities of an observation falling into each of these groups are: \( p/(p + 1 - p + p) \), \( 1 - p/(p + 1 - p + p) \), and \( p/(p + 1 + p) \). Thus, the probability that the observation is a case in the new data set is

\[
p^* = \frac{p}{1 + p} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}
\]

which has the logistic form. Schouten et al. suggested the use of standard logistic regression to obtain the parameter estimates. However, in the original log-binomial model maximum likelihood estimates are obtained by maximizing over \( \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k < 0 \). Thus to be equivalent, the logistic must be maximized over the same parameter space, which standard logistic regression does not do. Also, because of the dependencies between the observations in the new data set, the standard errors of the estimates cannot be obtained directly from the logistic regression. Liang and Zeger proposed to find the maximum likelihood in models with correlations between the observations by solving a multivariate analogue of the quasi score function. This approach, called generalized estimation equations or GEE, requires only that a working correlation has been specified. It is not necessary to find the correct working correlation to obtain consistent and asymptotically normal estimates for the betas and robust variance estimates as long as the estimates of the correlation parameters are consistent. The variance estimate used in GEE is similar to the robust sandwich estimator described by Schouten et al. We will call this the GEE-logistic model.

**Simulations**

We compared the three estimation methods with simulations. In all simulations, the parameter for the exposure effect was called \( \beta_1 \). In some simulations another covariate was included that might or might not confound the exposure effect. The parameter for this covariate was called \( \beta_2 \). Since it did not confound the exposure effect in all simulations it is not referred to as a confounder, but as 'a potential confounder' or 'the other covariate'. The baseline prevalence among those without the exposure and without the potential confounder is given by \( \exp(\beta_0) \). Only categorical covariates were considered, and no interactions were assumed. The exposure parameter \( \beta_1 \) was the parameter of interest.

True probabilities of disease for any combination of exposure and the other covariate were computed from the chosen values of the parameter for baseline prevalence, the parameter for exposure, and the parameter for the other covariate. The true probability was used with a random number generator to produce binomially distributed data for each population. One-thousand data sets were constructed for each true value, and analysed with the three regression methods. A Wald test statistic based on the parameter estimate divided by its standard error estimate was used to calculate the proportion of rejections for a Wald test for the null hypothesis that the true parameter is equal to the chosen parameter. When the null hypothesis was that the true parameter value was zero, a likelihood ratio test for the significance of the variable was also computed for the Cox and the log-binomial models. The resulting figures are called the rejection percentages. If the method performs well, the average parameter estimate from the 1000 data sets should be equal to the chosen parameter value, and the rejection percentages should be equal to the significance level (which was chosen to be \( \alpha = 0.05 \)) when the null hypothesis is true.

The power of the test was calculated as the rejection percentage for the null hypothesis that the true parameter value was zero, when, in fact, the chosen parameter was not zero.
Five sets of simulations encompassing 14 different simulation conditions were performed:

(1) Five simulations with the prevalence of the disease ranging from 0.1 to 0.9 ($\beta_0$ ranging from −2.3 to −0.1), the exposure parameter $\beta_1 = 0$ (PPR = 1), and no other covariates. The sample size for both exposure groups was 300.

(2) Three simulations with the prevalence of the disease among non-exposed ranging from 0.05 to 0.3 ($\beta_0$ ranging from −3.0 to −1.2), the exposure parameter $\beta_1 = 0.7$ (PPR = 2), and no other covariates. The sample size for both exposure groups was 300.

(3) One simulation with the prevalence among those with exposure = 0 and covariate = 0 equal to 0.2 ($\beta_0 = −1.6$), the exposure parameter $\beta_1 = 0.7$, one covariate parameter $\beta_2 = 0.7$, and no association between the exposure and the other covariate (no confounding). The sample size for both exposure groups was 100 when the covariate = 0 and 200 when the covariate = 1.

(4) One simulation with the prevalence among those with exposure = 0 and covariate = 0 equal to 0.2 ($\beta_0 = −1.6$), the exposure parameter $\beta_1 = 0.7$, one covariate parameter $\beta_2 = 0.7$, and an association between exposure and the other covariate corresponding to a PPR of 2. The non-exposed group had 100 and 200 observations for covariate = 0 and covariate = 1 respectively, while the exposed group had these sample sizes reversed.

(5) Four simulations with the prevalence among the non-exposed equal to 0.2 ($\beta_0 = −1.6$), and the exposure parameter $\beta_1 = 0.1$, 0.3, 0.5, and 0.7. This was used to estimate the power for testing $H_0$: $\beta_1 = 0$.

In addition to the simulation data sets, one sample data set was produced that did not fulfill the assumption of additive effects.

The Cox regression models were fitted with PROC PHREG in SAS, with the Breslow method of handling ties. The other methods of handling ties all produced biased estimates (not shown) and were not considered. The log-binomial model is a generalized linear model with logarithmic link function and binomial distribution function, and was fitted with PROC GENMOD in SAS using the observed information matrix to estimate the standard errors (which is the default). The GEE-logistic model was fitted with a GEE macro (version 2.02) written for SAS. The working correlation structure was chosen to be uncorrelated.

**Results**

Table 1 contains the results of the first set of simulations. The three regression methods gave exactly the same parameter estimates. (This will always happen when there is only one 0/1 covariate). The rejection percentages of the log-binomial and the GEE-logistic methods were very nearly what was expected in all cases. However, as the prevalence increased, the Cox regression performance worsened markedly. There were only minor differences between the Wald based rejection percentages and the likelihood ratio based rejection percentages for the Cox and the log-binomial methods.

Table 2 contains the results of the second set of simulations which have a true PPR of 2.0. The findings were similar to the previous set of simulations. The average prevalences correspond closely to the prevalences in the first three rows in Table 1.

In Table 3, the model has an exposure variable and one other covariate. One simulation (set 3) is with and one (set 4) is without confounding. The point estimates are almost identical. The

<table>
<thead>
<tr>
<th>$P(Y = 1)$</th>
<th>Average estimated $\beta_1$</th>
<th>Average estimated $\beta_1$</th>
<th>Rejection %, Wald</th>
<th>Rejection %, LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y = 1) = 0.1$</td>
<td>Cox regression -0.00535</td>
<td>0.2627</td>
<td>0.037</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model -0.00535</td>
<td>0.2496</td>
<td>0.052</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model -0.00535</td>
<td>0.2496</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>$P(Y = 1) = 0.3$</td>
<td>Cox regression 0.00111</td>
<td>0.1495</td>
<td>0.021</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model 0.00111</td>
<td>0.1252</td>
<td>0.047</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model 0.00111</td>
<td>0.1252</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>$P(Y = 1) = 0.5$</td>
<td>Cox regression 0.00084</td>
<td>0.1157</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model 0.00084</td>
<td>0.0820</td>
<td>0.045</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model 0.00084</td>
<td>0.0820</td>
<td>0.045</td>
<td></td>
</tr>
<tr>
<td>$P(Y = 1) = 0.7$</td>
<td>Cox regression 0.00069</td>
<td>0.0976</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model 0.00069</td>
<td>0.0533</td>
<td>0.043</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model 0.00069</td>
<td>0.0533</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>$P(Y = 1) = 0.9$</td>
<td>Cox regression -0.00054</td>
<td>0.0861</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model -0.00054</td>
<td>0.0272</td>
<td>0.063</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model -0.00054</td>
<td>0.0272</td>
<td>0.063</td>
<td></td>
</tr>
</tbody>
</table>

SE = Standard error.
LR = Likelihood ratio.

<table>
<thead>
<tr>
<th>$P(Y = 1)$</th>
<th>Average estimated $\beta_1$</th>
<th>Average estimated $\beta_1$</th>
<th>Rejection %, Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y = 1)$</td>
<td>Cox regression 0.7250</td>
<td>0.3263</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model 0.7250</td>
<td>0.3158</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model 0.7250</td>
<td>0.3158</td>
<td>0.057</td>
</tr>
<tr>
<td>$P(Y = 1)$</td>
<td>Cox regression 0.7136</td>
<td>0.1582</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model 0.7136</td>
<td>0.1355</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model 0.7136</td>
<td>0.1355</td>
<td>0.043</td>
</tr>
<tr>
<td>$P(Y = 1)$</td>
<td>Cox regression 0.7041</td>
<td>0.1290</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>Log-binomial model 0.7041</td>
<td>0.0998</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>GEE-logistic model 0.7041</td>
<td>0.0998</td>
<td>0.058</td>
</tr>
</tbody>
</table>

SE = Standard error.
The resulting confidence intervals were too wide. The log-binomial model is also preferable. This is because the standard errors of the Cox model were larger than those for the log-binomial model, and confidence intervals which were generally of correct size. Zochetti et al. give a theoretical basis for the standard errors of the Cox model being larger than those for the log-binomial model, and also conclude that the log-binomial model is preferable.

Discussion

All three models produced point estimates close to the true parameter. Cox regression produced standard errors that were too large, especially when the prevalence of the disease was high. This resulted in the confidence intervals being too wide. The log-binomial and the GEE-logistic models give very different parameter estimates. The sum of the parameters from the GEE-logistic model is 0.097 corresponding to a prevalence among those with the exposure and the covariate), without and with confounding, 1000 samples, sample size 300 in each exposure group.

The use of a log-binomial model has been criticized on the grounds that it might produce prevalences greater than 1.11-13. Of course, the proportional hazards model avoids this problem by simply not estimating the prevalences. The above example shows that the GEE-logistic model may indeed produce prevalences above 1, at least when the model is misspecified. With regard to the log-binomial model, maximum likelihood estimates by definition are obtained by maximizing only over the parameter space. Thus, for any covariate pattern in the data, the predicted prevalence must be between 0 and 1. The only way of producing predicted values above 1 is to predict outside the realm (convex hull) of the data. (See the proof in the Appendix.) Thus, the possibility of producing predicted prevalences above 1 is not an objection to the use of the log-binomial model for estimation and hypothesis testing of PPR in cross-sectional studies.

However, it is possible that the computer program may not yield the maximum likelihood estimate. We have had this happen in simple situations (for which we could calculate the correct estimates by hand) when the maximum was on the boundary of the parameter space. We have also experienced situations in which the log-binomial model did not converge unless suitable initial parameter values were provided to the software. Inspection of the parameter estimates when the iteration stopped disclosed that the estimate of $P \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ was greater than 0. Input of suitable initial values always solved the problem. Such initial values may always be obtained from a logistic regression on a data set manipulated as suggested by Schouten et al.7 If the final parameter estimates from the log-binomial model differ substantially from the estimates from the logistic model, this may be interpreted as a sign of a misspecified model.

The simulations have been restricted to multiplicative models which are the type typically used with Cox regression and logistic regression. The use of the logarithmic link function in the log-binomial model also presupposes that the data fit a multiplicative model. The log-binomial model also assumes binomially

Table 3 Simulation results for sets 3 and 4. $\beta_0 = -1.6, \beta_1 = \beta_2 = 0.7$ (prevalence ranging from 0.2 among non-exposed to 0.82 among those with the exposure and the covariate), without and with confounding, 1000 samples, sample size 300 in each exposure group.

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Other covariate</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Other covariate</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Other covariate</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 4 Sample data set 1/0 indicates presence/absence of the disease or covariate.

Table 5 Parameter estimates and estimated standard errors (SE) from analyses of the sample data set in Table 4.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Log-binomial model</th>
<th>GEE-logistic model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>without interaction, estimate (SE)</td>
<td>with interaction, estimate (SE)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.909 (0.086)</td>
<td>-2.303 (0.300)</td>
</tr>
<tr>
<td>Exposure</td>
<td>0.388 (0.095)</td>
<td>2.079 (0.304)</td>
</tr>
<tr>
<td>Other covariate</td>
<td>0.388 (0.095)</td>
<td>2.079 (0.304)</td>
</tr>
<tr>
<td>Interaction exposure*other covariate</td>
<td>-2.057 (0.312)</td>
<td>-2.057 (0.312)</td>
</tr>
</tbody>
</table>
distributed outcome data. We have not systematically tested how the models behave when the data are not truly binomial.

Conclusion
The prevalence of the disease is usually high when estimation of PPR instead of POR is contemplated. Under these circumstances the Cox regression model produces standard errors that are too large and hence rejects false null hypotheses too seldom. The GEE-logistic model gives the same results as the log-binomial model when the data are truly binomial and additive, but when this is not the case it may produce prevalences greater than 1. The log-binomial model has the correct type I error probabilities and should be preferred.

References

Appendix
Theorem 1 The log-binomial model produces predicted values between zero and one, inclusive, for any point \( z \) in the convex hull of the observed covariates \( x_1, \ldots, x_m \).

PROOF Let \( z \) be an arbitrary point of the convex hull of the covariates (i.e. \( z = \sum_{i=1}^{m} \lambda_i x_i \), where \( \forall i : 0 \leq \lambda_i \leq 1, \sum_{i=1}^{m} \lambda_i = 1 \)).

Let \( \hat{b} \) be the maximum likelihood estimator of \( \beta \). The predicted value corresponding to the covariates \( z \) is

\[
\hat{p}_z = e^{z' \hat{b}} = e^{z' \sum_{i=1}^{m} \lambda_i \hat{b}_i} = \prod_{i=1}^{m} \left( e^{x_i \hat{b}} \right)^{\lambda_i}
\]

Because \( \hat{b} \) is the maximum likelihood estimator it must satisfy that \( \forall i : x_i' \hat{b} \leq 0 \) and therefore the predicted value is between zero and one, inclusive.