Research article

Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio

Aluíso JD Barros* and Vânia N Hirakata

Address: Programa de Pós-graduação em Epidemiologia, Universidade Federal de Pelotas, Brazil
Email: Aluíso JD Barros* - abarro@epidemio-ufpel.org.br; Vânia N Hirakata - vnh.ez@terra.com.br
* Corresponding author

Abstract

Background: Cross-sectional studies with binary outcomes analyzed by logistic regression are frequent in the epidemiological literature. However, the odds ratio can importantly overestimate the prevalence ratio, the measure of choice in these studies. Also, controlling for confounding is not equivalent for the two measures. In this paper we explore alternatives for modeling data of such studies with techniques that directly estimate the prevalence ratio.

Methods: We compared Cox regression with constant time at risk, Poisson regression and log-binomial regression against the standard Mantel-Haenszel estimators. Models with robust variance estimators in Cox and Poisson regressions and variance corrected by the scale parameter in Poisson regression were also evaluated.

Results: Three outcomes, from a cross-sectional study carried out in Pelotas, Brazil, with different levels of prevalence were explored: weight-for-age deficit (4%), asthma (31%) and mother in a paid job (52%). Unadjusted Cox/Poisson regression and Poisson regression with scale parameter adjusted by deviance performed worst in terms of interval estimates. Poisson regression with scale parameter adjusted by $\chi^2$ showed variable performance depending on the outcome prevalence. Cox/Poisson regression with robust variance, and log-binomial regression performed equally well when the model was correctly specified.

Conclusions: Cox or Poisson regression with robust variance and log-binomial regression provide correct estimates and are a better alternative for the analysis of cross-sectional studies with binary outcomes than logistic regression, since the prevalence ratio is more interpretable and easier to communicate to non-specialists than the odds ratio. However, precautions are needed to avoid estimation problems in specific situations.

Background

Epidemiologic studies found in the literature are frequently cross-sectional, as this is a simple, fast and inexpensive design alternative. Often the outcomes are binary, and logistic regression is used for the analysis. This results in the odds ratio being frequently reported in situations where incidence or prevalence ratios are estimable, despite the fact that it is "biologically interpretable only
interpretations such as “the exposed group has a “forget” what their measure of association is and make associated with odds ratios in our experience: the authors prevalence ratio. Here resides the most common mistake with frequent outcomes, what is common in cross-sectional studies. Logistic regression was used for the analysis of 37 (34%) and 10 (22%) of these studies, respectively. We have, therefore, that an important proportion of such studies end up reporting odds ratios, the effect measure yielded by logistic regression, rather than prevalence or incidence ratios.

The use of odds ratios is absolutely correct. There is nothing intrinsically wrong with them. But, when working with frequent outcomes, what is common in cross-sectional studies, the odds ratio can strongly overestimate the prevalence ratio. Here resides the most common mistake associated with odds ratios in our experience: the authors “forget” what their measure of association is and make interpretations such as “the exposed group has a risk of illness four times greater than the non-exposed group”. The relative risk interpretation given to the odds ratio can be misleading, in theoretical and practical terms, especially if used for definition of policy priorities in conjunction with other true relative risks [1–4].

Additionally, logistic regression is often used for the sake of control of confounding and adjustment of interactions. But confounding and interaction are dependent on the measure of effect, so that controlling for confounding for the odds ratio is not the same thing as doing so for the prevalence ratio [5,6]. Therefore, interpreting the odds ratio as if it were a prevalence ratio is inadequate not only in terms of the possible overestimation, but also because confounding may not be appropriately controlled.

Several alternatives have been discussed in the literature for the analysis of binary outcomes in cross-sectional (or longitudinal) studies using the prevalence ratio rather than the odds ratio. The simplest way is to transform the odds ratios obtained by logistic regression into prevalence ratios [7–9]. Another possibility is to use a statistical model that estimates directly the prevalence ratio and its confidence interval. Alternatives explored in the epidemiological literature are Cox regression with equal times of follow-up assigned to all individuals [10], log-binomial regression (a generalized linear model with a logarithmic link function and binomial distribution for the residual) [7,11–13], Poisson regression [13] and complementary log-log model, where the link function is log(- log(1 - π)) and the distribution is binomial [13,14].

The authors that contributed to the discussion have not reached a conclusion on which would be the best approach, and only three publications make some comparison among available alternatives [4,13,15]. Possible fixes for the problems related to confidence intervals in some of the techniques proposed were dealt with to some extent by one of the papers [13], where the conclusion is that “there are no valid reasons for the systematic choice of odds ratio and of the logistic regression model to estimate prevalence rate ratios unless the type of study imperatively requires their use.”

We have applied, in the context of a cross-sectional study, log-binomial regression, Cox regression, and Poisson regression. Corrections for the standard errors were included for Cox and Poisson regressions. Also, to increase the applicability of the results, a confounder was always included in the scenarios studied, which involved outcomes of varying prevalences. The point and interval estimates obtained with each approach were compared to the standard Mantel-Haenszel-like prevalence ratios and confidence interval estimators.

Methods

Using Cox regression to analyze a binary outcome in a cross-sectional study was suggested by Lee & Chia [10], and assessed by others [4,15]. Usually, Cox regression is used to analyze time-to-event data, that is, the response is the time an individual takes to present the outcome of interest. Individuals that never get ill are assigned the total length of time of the follow-up, and are treated as censored, meaning that it is not known when they will get ill, but at least until the time of the end of the follow-up they are well. Individuals lost to follow-up are treated in a similar way. Cox regression estimates the hazard rate function that expresses how the hazard rate depends upon a set of covariates. The model formulation is

\[ h(t) = h_0(t) \exp(\beta_1 z_1 + \ldots + \beta_k z_k) \]

where \( h_0(t) \) is the base hazard function of time, \( z_i \) are covariates and \( \beta_i \) the coefficients for the \( k \) covariates. The Cox model treats \( h_0(t) \) as a nuisance function and actually does not estimate it [16].

When a constant risk period is assigned to everyone in the cohort, the hazard rate ratio estimated by Cox regression equals the cumulative incidence ratio in longitudinal studies, or the prevalence ratio in cross-sectional studies [17,18]. Although this model can produce correct point estimates, the underlying distribution of the response is Poisson. As prevalence data in a cross-sectional study follow a binomial distribution, the variance of the coefficients tends to be overestimated, resulting in wider confidence intervals compared to those based on the
binomial distribution. This is easily explained by comparing the binomial variance, $p(1-p)$, with a maximum of 0.25 when $p = 0.5$ with Poisson variance, $\lambda$, that grows steadily with the intensity of the process. That is, the variance estimated by the Poisson model will be very close to the binomial variance when the outcome is rare, but will be increasingly greater as the outcome becomes more frequent. In such a situation we have underdispersion, the opposite to the more commonly observed overdispersion, where the data is more dispersed than the model predicts.

It is possible to improve the situation using the robust variance estimates proposed by Lin & Wei [19], similar to other robust sandwich estimators proposed for parametric models, such as Huber's sandwich estimator [20]. In this paper, Cox regression with equal follow-up times was assessed, with standard and robust variance estimates.

Poisson regression is commonly used in epidemiology to analyze longitudinal studies where the outcome is a count of episodes of an illness occurring over time (e.g. episodes of diarrhea). The model formulation is

$$\log \left( \frac{n}{t} \right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$  \hspace{1cm} (2)

where $n$ is the count of events for a given individual, $t$ the time it was followed-up, and $X_i$ the covariates. The model parameters ($\beta_i$) are log relative risks. In this context, Poisson regression is equivalent to Cox regression [21], and the parameters estimated are the same.

As described for Cox regression, the prevalence ratio is directly estimated by the model, and the confidence intervals are wider than those provided by a binomial model. A simple remedy is to multiply the estimated Poisson variance by some estimate of underdispersion (or overdispersion). These estimates can be based on the deviance or the chi-square of the model, dividing these quantities by the residual degrees of freedom [22,23]. In practice, this ratio is used as a scale parameter, replacing the original Poisson value of 1. A robust variance estimate is also available for the Poisson model, based on the Huber sandwich estimate [20] (which again yields results that are equal to Cox regression with robust variance). This alternative is known to underestimate the true variability with moderately sized samples, while adjusting the scale parameter tends to overestimate it. Other alternatives would be jackknife and bootstrap variance estimates [23]. We decided not to use the latter alternatives as they are not directly available in standard statistical software. Thus, Poisson regression was used in this paper with unadjusted variances, with scale parameter adjustment for both deviance and chi-square statistics, and with robust variance estimates.

The last model assessed was the log-binomial model [15] – a generalized linear model where the link function is the logarithm of the proportion under study and the distribution of the error is binomial [4,7,11,12,15]. The measure of effect in this model is also the relative risk.

For $k$ covariates the model is written as

$$\log(\pi) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$  \hspace{1cm} (3)

where $\pi$ is the probability of success (e.g., the proportion of sick persons in a group), and $X_i$ the covariates. The relative risk estimate of a given covariate is $\exp(\beta)$.

Since $\log(\pi)$ must be in the interval $-\infty$ to 0, restrictions in the estimation process have to be used to avoid predicting probabilities out of the [0,1] interval. When estimates are on the boundaries of the valid parameter space, the estimates of the Newton-Raphson method will not converge to the maximum likelihood estimates [24]. Convergence problems in the estimation process are most likely to happen when the model contains a continuous covariate or multiple politomic covariates, or the outcome prevalence is high [12,24]. When the estimates are not on the boundary of the parameter space, convergence problems may still happen, and better starting values for the estimation process than the default used by the software will help. Most log-binomial models fitted in this paper used the default Stata estimation options, without convergence problems. In one case, when the model failed to converge, the "search" option, which makes the procedure search for a better starting value, was used [25].

The results obtained from the various models were compared to the pooled Mantel-Haenszel-like prevalence ratios (MHPR) and corresponding confidence intervals, used here as the reference results. Mantel-Haenszel estimates are easy to obtain in simple situations such as the ones dealt with in this work (one exposure and one confounder). However, for more complex situations, their estimation is more complicated and the use of statistical models more efficient.

All the analyses were performed with Stata 7.0 [25], and the actual command lines used are listed below. Each outcome-exposure-confounder combination was represented by one row in the dataset and its frequency given by the variable `freq`.

*** M-H relative risk

```stata
cs ill exposed [fweight = freq], by(confounder)
```

*** Poisson regression unadjusted
poisson ill exposed confounder \([\text{fweight = freq}], \text{irr}\)

*** Poisson regression adjusted by chi-squared

glm ill exposed confounder \([\text{fweight = freq}], \text{family(poisson)}\)
\[\text{scale(x2)}\] eform

*** Poisson regression adjusted by deviance

glm ill exposed confounder \([\text{fweight = freq}], \text{family(poisson)}\)
\[\text{scale(dev)}\] eform irls

*** Poisson regression with robust variance

poisson ill exposed confounder \([\text{fweight = freq}], \text{irr r}\)

*** Log-binomial regression

glm ill exposed confounder \([\text{fweight = freq}], \text{family(binomial)}\)
\[\text{link(log)}\] eform

*** Odds ratio from logistic regression

logistic ill exposed confounder \([\text{fweight = freq}]\)

For the comparison of the above techniques, real data from a population-based survey were used. A birth cohort was initiated in 1993, including all births happening in Pelotas, Southern Brazil [26]. These children were seen at birth, and their mothers interviewed. At 1 and 3 months of age, a sub-sample of 655 were sought for follow-up information. At 6 and 12 months, a larger sub-sample (including the 655 seen at 1 and 3 months) was sought, that comprised all children born with low birthweight and 20% of the remaining children. At these points, 1363 children were sought. The data used in this work came from a visit done between November 1997 and April 1998, when the children were 4–5 years-old. The children sought were the same as those in the 12-month revisit, and 1273 (93%) were actually interviewed. From the 90 children lost to follow-up, 61 (68%) had moved to other towns, 18 (20%) could not be found, 6 (7%) had died, and 5 (6%) refused to participate. The children were submitted to a nutritional assessment (weight and height) and their mothers answered a standardized pre-coded questionnaire including information on socioeconomic, demographic, reproductive, and health characteristics.

Three outcomes with different prevalences were used in the analyses, each in conjunction with a risk factor and a confounding factor, in a way to form 3 distinct situations. The three sets of variables used respectively as outcome, risk factor and confounder were: situation 1 – underweight (weight for age Z-score < -2), previous hospitalization and birth weight; situation 2 – asthma (asthma or bronchitis reported by the person responsible for the child in study), whether mother smoked and social class; situation 3 – status of maternal employment (whether or not in a paid job), father living with the family and social class. All variables were made dichotomous in order to simplify the comparisons and understanding of the models.

In order to widen the scenarios available, each set of variables was manipulated to increase the level of confounding. This was achieved by arbitrary changes in the prevalences of the risk and confounding factors, re-weighting the relevant strata in the data in a way to keep the sample size constant. The original and manipulated data are fully presented in the results section.

Confounding was measured by the proportional change from the crude prevalence ratio to the adjusted (Mantel-Haenszel) prevalence ratio using the expression

\[
\text{confounding} = \frac{\text{PR (M-H)} - \text{PR (crude)}}{\text{PR (crude)}} \times 100
\]

so that whenever the adjusted prevalence ratio was smaller than the crude, confounding was negative.

### Results

Underweight was the least common outcome studied, with a prevalence of 4.1%. Asthma (prevalence = 31.2%), was intermediate and mother in a paid job was the commonest (prevalence = 51.4%). In the modified situation 1, underweight prevalence was increased in the low birth weight group from 7.8% to 10.4%, changing confounding from -14 to -18%. The overall prevalence of underweight changed to 4.9% (Tables 1 and 2). In situation 2, the prevalence of asthma was reduced from 22.7% to 12.7% in the upper class and increased from 38.1% to 52.1% in the lower class. Confounding changed from -8% to -17%. Overall asthma prevalence changed from 31.2% to 34.3% (Tables 3 and 4). In situation 3, confounding was increased by changing both the prevalences of the exposure and the outcome, achieving an increase from 4% to 25%. This was the only situation where the test for heterogeneity of prevalence ratios across strata was significant, indicating an interaction (Tables 5 and 6).

Comparing the results of the different models in situation 1 (Table 7), we see that the point estimates obtained with Cox, Poisson and log-binomial models are very close to the Mantel-Haenszel prevalence ratio (MHPR) for the original and modified data. In terms of the confidence intervals, the differences between Cox, Poisson and log-binomial models and the reference were less than 5%, except for Poisson scaled by deviance, where the CIs were approximately 50% narrower.
Table 1: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the original data using underweight (weight for age Z-score < -2) as the outcome, previous hospitalization as the risk factor and low birth weight as confounder (situation 1 original).

<table>
<thead>
<tr>
<th></th>
<th>Underweight</th>
<th>Normal</th>
<th>All</th>
<th>Exp. prev. = 19.2%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever in hospital</td>
<td>8</td>
<td>4.7%</td>
<td>163</td>
<td>171</td>
</tr>
<tr>
<td>Never</td>
<td>14</td>
<td>1.9%</td>
<td>704</td>
<td>718</td>
</tr>
<tr>
<td>All</td>
<td>22</td>
<td>2.5%</td>
<td>867</td>
<td>889</td>
</tr>
</tbody>
</table>

Second stratum: Low birth weight

<table>
<thead>
<tr>
<th></th>
<th>Underweight</th>
<th>Normal</th>
<th>All</th>
<th>Exp. prev. = 31.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever in hospital</td>
<td>16</td>
<td>13.4%</td>
<td>103</td>
<td>119</td>
</tr>
<tr>
<td>Never</td>
<td>14</td>
<td>5.3%</td>
<td>250</td>
<td>264</td>
</tr>
<tr>
<td>All</td>
<td>30</td>
<td>7.8%</td>
<td>353</td>
<td>383</td>
</tr>
</tbody>
</table>

Combined strata: Normal and low birth weight

<table>
<thead>
<tr>
<th></th>
<th>Underweight</th>
<th>Normal</th>
<th>All</th>
<th>PR (crude) = 2.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever in hospital</td>
<td>24</td>
<td>8.3%</td>
<td>266</td>
<td>290</td>
</tr>
<tr>
<td>Never</td>
<td>28</td>
<td>2.9%</td>
<td>954</td>
<td>982</td>
</tr>
<tr>
<td>All</td>
<td>52</td>
<td>4.1%</td>
<td>1220</td>
<td>1272</td>
</tr>
</tbody>
</table>

* P-value for testing heterogeneity of the prevalence ratios across strata.

Table 2: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the modified data using underweight (weight for age Z-score < -2) as the outcome, previous hospitalization as the risk factor and low birth weight as confounder (situation 1 modified).

<table>
<thead>
<tr>
<th></th>
<th>Underweight</th>
<th>Normal</th>
<th>All</th>
<th>Exp. prev. = 19.2%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever in hospital</td>
<td>8</td>
<td>4.7%</td>
<td>163</td>
<td>171</td>
</tr>
<tr>
<td>Never</td>
<td>14</td>
<td>1.9%</td>
<td>704</td>
<td>718</td>
</tr>
<tr>
<td>All</td>
<td>22</td>
<td>2.5%</td>
<td>867</td>
<td>889</td>
</tr>
</tbody>
</table>

Second stratum: Low birth weight

<table>
<thead>
<tr>
<th></th>
<th>Underweight</th>
<th>Normal</th>
<th>All</th>
<th>Exp. prev. = 31.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever in hospital</td>
<td>22</td>
<td>18.5%</td>
<td>97</td>
<td>119</td>
</tr>
<tr>
<td>Never</td>
<td>18</td>
<td>6.8%</td>
<td>246</td>
<td>264</td>
</tr>
<tr>
<td>All</td>
<td>40</td>
<td>10.4%</td>
<td>343</td>
<td>383</td>
</tr>
</tbody>
</table>
Table 2: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the modified data using underweight (weight for age Z-score < -2) as the outcome, previous hospitalization as the risk factor and low birth weight as confounder (situation 1 modified). (Continued)

<table>
<thead>
<tr>
<th>Underweight</th>
<th>Normal</th>
<th>All</th>
<th>Exp. prev.</th>
<th>PR (crude)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Prev.</td>
<td>N</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Ever in hospital</td>
<td>30</td>
<td>10.3%</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>Never</td>
<td>32</td>
<td>3.3%</td>
<td>950</td>
<td>982</td>
</tr>
<tr>
<td>All</td>
<td>62</td>
<td>4.9%</td>
<td>1210</td>
<td>1272</td>
</tr>
</tbody>
</table>

* P-value for testing heterogeneity of the prevalence ratios across strata.

Table 3: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the original data using asthma as the outcome, maternal smoking as the risk factor and social class as confounder (situation 2 original).

<table>
<thead>
<tr>
<th>First stratum: High social class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (crude)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (M-H)</td>
</tr>
<tr>
<td>M-H weight</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (M-H)</td>
</tr>
<tr>
<td>M-H weight</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second stratum: Low social class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (crude)</td>
</tr>
<tr>
<td>M-H weight</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (M-H)</td>
</tr>
<tr>
<td>M-H weight</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Combined strata: High and low social class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (crude)</td>
</tr>
<tr>
<td>M-H weight</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Prev.</td>
</tr>
<tr>
<td>Exp. prev.</td>
</tr>
<tr>
<td>PR (M-H)</td>
</tr>
<tr>
<td>M-H weight</td>
</tr>
</tbody>
</table>

* P-value for testing heterogeneity of the prevalence ratios across strata.
Situation 2 (Table 8) was similar to situation 1 in terms of point estimates. Confidence intervals were strongly overestimated by unadjusted Cox/Poisson models. In the modified data, the 95%CI was overestimated by 11% by Poisson regression with scale parameter adjusted by $\chi^2$.

In situation 3 (Table 9), the outcome prevalence was highest, and there was a significant interaction between risk factor and confounder. Ignoring the interaction (i.e., using a misspecified model), the log-binomial model performed slightly worse than the Cox and Poisson models in relation to the point estimates. The latter presented a maximum difference of 2% compared to the MHPR, while the log-binomial estimates were up to 8.7% greater. In terms of interval estimates, only the Cox/Poisson models with robust variance presented differences less than 5% for both the original and modified data. An interaction term was included in the robust Poisson and log-binomial regressions. In the original situation, identical results (up to the third decimal place) were obtained from both models, matching the stratum-specific relative risks and confidence intervals. However, in the modified situation, the log-binomial model did not converge, while the robust Poisson model again reproduced the stratum-specific estimates. A common reason for non-convergence is inappropriate starting values for model parameters. Stata’s option "search", which specifies that the command "glm" should search for good starting values, solved the problem and, with this option, the results obtained from the log-binomial model were again virtually identical to Poisson regression.

A graphical summary of the results concerning the interval estimates is shown in Figure 1. It is clear from the figure that robust Poisson/Cox regression and log-binomial regression are the best performers, consistently producing the confidence intervals with amplitude closest to the reference. In third place, but with some confidence intervals nearly 20% wider than the reference, ranked Poisson regression with the scale parameter adjusted by $\chi^2$.

Table 4: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the modified data using asthma as the outcome, maternal smoking as the risk factor and social class as confounder (situation 2 modified).

<table>
<thead>
<tr>
<th>First stratum: High social class</th>
<th></th>
<th></th>
<th>Exp. prev. = 26.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asthma</td>
<td>No</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Prev.</td>
<td>N</td>
</tr>
<tr>
<td>Mother smokes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>14.6%</td>
<td>123</td>
<td>144</td>
</tr>
<tr>
<td>No</td>
<td>48</td>
<td>12.1%</td>
<td>350</td>
</tr>
<tr>
<td>All</td>
<td>69</td>
<td>12.7%</td>
<td>473</td>
</tr>
<tr>
<td>Exp. prev. = 43.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother smokes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>68.1%</td>
<td>91</td>
<td>285</td>
</tr>
<tr>
<td>No</td>
<td>149</td>
<td>39.9%</td>
<td>224</td>
</tr>
<tr>
<td>All</td>
<td>343</td>
<td>52.1%</td>
<td>315</td>
</tr>
<tr>
<td>Exp. prev. = 35.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother smokes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>215</td>
<td>50.1%</td>
<td>214</td>
<td>429</td>
</tr>
<tr>
<td>No</td>
<td>197</td>
<td>25.6%</td>
<td>574</td>
</tr>
<tr>
<td>All</td>
<td>412</td>
<td>34.3%</td>
<td>788</td>
</tr>
</tbody>
</table>

* P-value for testing heterogeneity of the prevalence ratios across strata.

<table>
<thead>
<tr>
<th>Combined strata: High and low social class</th>
<th></th>
<th></th>
<th>Exp. prev. = 35.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asthma</td>
<td>No</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Prev.</td>
<td>N</td>
</tr>
<tr>
<td>Mother smokes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>215</td>
<td>50.1%</td>
<td>214</td>
<td>429</td>
</tr>
<tr>
<td>No</td>
<td>197</td>
<td>25.6%</td>
<td>574</td>
</tr>
<tr>
<td>All</td>
<td>412</td>
<td>34.3%</td>
<td>788</td>
</tr>
</tbody>
</table>

PR (crude) = 1.96
PR (M-H) = 1.62
Confounding = -17.3%
P-value(het)* = 0.2
Discussion

The literature on the different alternatives to analyze cross-sectional or longitudinal data using prevalence (or cumulative incidence) ratios instead of odds ratios has not yet proposed a strategy that produces both point and interval acceptable estimates. To our knowledge this is the first paper to focus on different strategies and compare them to a suitable reference in terms of the prevalence ratios and confidence intervals obtained.

We have shown that there are several alternatives available that will provide very good results in terms of point estimates: Cox, Poisson and log-binomial regression. The case of interval estimates is more complicated, as some models will overestimate or underestimate them, in different situations. Even so, we are still left with three viable alternatives: log-binomial regression, Cox/Poisson regression with robust variance, and Poisson regression with scale parameter adjusted by $\chi^2$.

One limitation of this work is not having dealt with continuous covariates. The main reason was the reference used. The Mantel-Haenszel techniques work for categorical variables only. Furthermore, most epidemiological analyses involve only categorical variables. The main problem with this omission is that continuous variables are a potential cause for model misbehavior, that is, the log-binomial model not converging, and the Poisson model producing estimates of individual probabilities greater than 1. This situation happens when the estimates are on the boundary of the parameter space, and is illustrated with the artificial data presented by Deddens in a paper where a simple strategy, the COPY method, was proposed to achieve convergence when fitting log-binomial models in such a case [27].

---

Figure 1
Comparison of the relative differences between the 95% confidence intervals obtained by unadjusted Poisson/Cox regression, Poisson regression with scale factor adjusted by $\chi^2$ and deviance, Poisson/Cox regression with robust variances and log-binomial regression and the Cornfield 95% confidence interval for each of the six situations studied. S1a (outcome prevalence / confounding): 4.1% / 14%; S1b: 4.9% / 18%; S2a: 31.2% / 8%; S2b: 34% / 17%; S3a: 51% / 4%; S3b: 54% / 25%.
Table 5: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the original data using mother in a paid job as the outcome, father living with the family as the risk factor and social class as confounder (situation 3 original).

<table>
<thead>
<tr>
<th></th>
<th>First stratum: High social class</th>
<th>Second stratum: Low social class</th>
<th>Combined strata: High and low social class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mother employed</td>
<td>No</td>
<td>All</td>
</tr>
<tr>
<td>Father Present</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>66</td>
<td>79.5%</td>
<td>17</td>
</tr>
<tr>
<td>No</td>
<td>250</td>
<td>54.5%</td>
<td>209</td>
</tr>
<tr>
<td>All</td>
<td>316</td>
<td>58.3%</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father Present</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>112</td>
<td>70.9%</td>
<td>46</td>
</tr>
<tr>
<td>No</td>
<td>189</td>
<td>37.8%</td>
<td>311</td>
</tr>
<tr>
<td>All</td>
<td>301</td>
<td>45.7%</td>
<td>357</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P-value for testing heterogeneity of the prevalence ratios across strata.

Table 6: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the modified data using mother in a paid job as the outcome, father living with the family as the risk factor and social class as confounder (situation 3 modified).

<table>
<thead>
<tr>
<th></th>
<th>First stratum: High social class</th>
<th>Second stratum: Low social class</th>
<th>Combined strata: High and low social class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mother employed</td>
<td>No</td>
<td>All</td>
</tr>
<tr>
<td>Father Present</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>73</td>
<td>90.1%</td>
<td>8</td>
</tr>
<tr>
<td>No</td>
<td>230</td>
<td>50.0%</td>
<td>230</td>
</tr>
<tr>
<td>All</td>
<td>303</td>
<td>56.0%</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father Present</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>295</td>
<td>56.0%</td>
<td>232</td>
</tr>
<tr>
<td>No</td>
<td>53</td>
<td>40.2%</td>
<td>79</td>
</tr>
<tr>
<td>All</td>
<td>348</td>
<td>52.8%</td>
<td>311</td>
</tr>
</tbody>
</table>
The log-binomial regression, used without any correction to the standard errors, presented results that were equivalent to those yielded by robust Poisson/Cox regression in situations 1 and 2. In situation 3, where an interaction was ignored, the model tended to present confidence intervals that were too narrow (up to 10.4%) compared to the reference, and slightly different point estimates. This situation was included in this exercise to present a scenario with a misspecified model, situation that is bound to happen in reported analyses, as failing to look for or correctly identifying interactions is not infrequent. When the correct model (including the interaction) was fitted, the results were again equivalent to those yielded by the reference and robust Poisson/Cox regression in the original data. In the modified data the model failed to converge, what was solved by using better starting values for the estimation procedure. In situations where the estimates are on the boundary of the parameter space the model will not converge, unless a strategy such as the COPY method is used [27].

Cox regression has been suggested as an alternative to logistic regression but the problems with the variance estimates were not dealt with [4,10,15]. As expected, we showed that confidence intervals can be strongly overestimated (up to 69% in our examples using real data). The use of robust variance estimates [19], as we proposed, improved variance estimation considerably, limiting the difference relative to the reference confidence intervals to less than 3% in the studied examples. Poisson regression, as mentioned before, works similarly, and has the advantage over Cox regression of using a command syntax similar to linear and logistic regressions in Stata.

The use of Poisson regression offers still other alternatives by means of changing the scale parameter to correct the standard errors when over or underdispersion is observed [22]. In the set of situations we presented, correction by the Pearson $\chi^2$ was superior to correction by the deviance, and, although not as good as robust estimates, represented a considerable improvement in relation to the uncorrected standard errors. The maximum observed difference relative to the reference confidence intervals was 17%. Poisson regression, however, can also present problems when the estimates are on the boundaries of the parameter space, as mentioned above. It is strongly advisable that the individual probabilities are calculated (Stata’s “predict” command will do that) and examined.

We have used the robust Poisson model in the analysis of several epidemiological studies, three of which have been already published [28–30]. In all cases we have used the same modeling strategy with logistic regression and robust Poisson regression. In these real situations the final sets of selected variables were the same, and the differences in model parameters within the expected between odds ratios and prevalence ratios. Until more experience is gathered, this may be a useful strategy to help identify anomalous results with robust Poisson regression, along with assessing the predicted individual probabilities.

The rapid and continuous evolution of statistical software means that most packages will perform at least one of the analyses that performed best in this exercise. Stata 7.0, used here, and widely employed in epidemiology research groups, can perform them all. It was not possible for us to assess other packages in terms of what they can do, and how to do it, as software like SPSS, SAS, S-Plus, among others, were not available to us.

**Conclusions**

We have shown that the use of Cox or Poisson regression without any adjustment for the analysis of cross-sectional data, as suggested sometimes in the literature, may lead to large errors in interval estimates. On the other hand, taken the precautions discussed in the paper, the log-binomial

---

### Table 6: Absolute frequencies, outcome prevalences, exposure prevalences, crude and pooled prevalence ratio (PR) estimates, and relative confounding for the analysis of the modified data using mother in a paid job as the outcome, father living with the family as the risk factor and social class as confounder (situation 3 modified). (Continued)

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>All</th>
<th>Prev. exp.</th>
<th>PR ( crude)</th>
<th>PR (M-H)</th>
<th>Confounding</th>
<th>P-value(het)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Father Present</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>283</td>
<td>309</td>
<td>549</td>
<td>608</td>
<td>592</td>
<td>1.58</td>
<td>0.01</td>
</tr>
<tr>
<td>All</td>
<td>651</td>
<td>1200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prev. exp. = 50.7%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P-value for testing heterogeneity of the prevalence ratios across strata.
Table 7: Comparison of prevalence ratios and respective confidence interval estimates (obtained by unadjusted Poisson/Cox regression, Poisson regression with scale factor adjusted by \( \chi^2 \) and deviance, Poisson/Cox regression with robust variances, log-binomial regression and logistic regression) and odds ratio with confidence interval estimate (obtained by logistic regression) with the Mantel-Haenszel prevalence ratio in the analysis of the original and modified data using underweight (weight for age Z-score < -2) as the outcome, previous hospitalization as the risk factor and low birth weight as confounder (situation 1).

<table>
<thead>
<tr>
<th>Original data</th>
<th>Point estimate 95% Confidence interval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>value</td>
<td>% diff.</td>
</tr>
<tr>
<td>PR Mantel-Haenszel</td>
<td>2.48</td>
<td>--</td>
</tr>
<tr>
<td>PR Poisson/Cox (unadj)</td>
<td>2.48</td>
<td>-0.2%</td>
</tr>
<tr>
<td>PR Poisson (( \chi^2 ))</td>
<td>2.48</td>
<td>-0.2%</td>
</tr>
<tr>
<td>PR Poisson (deviance)</td>
<td>2.48</td>
<td>-0.2%</td>
</tr>
<tr>
<td>PR Poisson/Cox (robust)</td>
<td>2.48</td>
<td>-0.2%</td>
</tr>
<tr>
<td>PR log-binomial</td>
<td>2.48</td>
<td>-0.1%</td>
</tr>
<tr>
<td>OR logistic regression</td>
<td>2.64</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modified data</th>
<th>Point estimate 95% Confidence interval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>value</td>
<td>% diff.</td>
</tr>
<tr>
<td>PR Mantel-Haenszel</td>
<td>2.61</td>
<td>--</td>
</tr>
<tr>
<td>PR Poisson/Cox (unadj)</td>
<td>2.60</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR Poisson (( \chi^2 ))</td>
<td>2.60</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR Poisson (deviance)</td>
<td>2.60</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR Poisson/Cox (robust)</td>
<td>2.60</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR log-binomial</td>
<td>2.61</td>
<td>-0.1%</td>
</tr>
<tr>
<td>OR logistic regression</td>
<td>2.85</td>
<td>9.3%</td>
</tr>
</tbody>
</table>

Table 8: Comparison of prevalence ratios and respective confidence interval estimates (obtained by unadjusted Poisson/Cox regression, Poisson regression with scale factor adjusted by \( \chi^2 \) and deviance, Poisson/Cox regression with robust variances, log-binomial regression and logistic regression) and odds ratio with confidence interval estimate (obtained by logistic regression) with the Mantel-Haenszel prevalence ratio in the analysis of the original and modified data using asthma as the outcome, maternal smoking as the risk factor and social class as confounder (situation 2).

<table>
<thead>
<tr>
<th>Original data</th>
<th>Point estimate 95% Confidence interval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>value</td>
<td>% diff.</td>
</tr>
<tr>
<td>PR Mantel-Haenszel</td>
<td>1.22</td>
<td>--</td>
</tr>
<tr>
<td>PR Poisson/Cox (unadj)</td>
<td>1.22</td>
<td>0.0%</td>
</tr>
<tr>
<td>PR Poisson (( \chi^2 ))</td>
<td>1.22</td>
<td>0.0%</td>
</tr>
<tr>
<td>PR Poisson (deviance)</td>
<td>1.22</td>
<td>0.0%</td>
</tr>
<tr>
<td>PR Poisson/Cox (robust)</td>
<td>1.22</td>
<td>0.0%</td>
</tr>
<tr>
<td>PR log-binomial</td>
<td>1.23</td>
<td>0.1%</td>
</tr>
<tr>
<td>OR logistic regression</td>
<td>1.36</td>
<td>11.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modified data</th>
<th>Point estimate 95% Confidence interval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>value</td>
<td>% diff.</td>
</tr>
<tr>
<td>PR Mantel-Haenszel</td>
<td>1.62</td>
<td>--</td>
</tr>
<tr>
<td>PR Poisson/Cox (unadj)</td>
<td>1.62</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR Poisson (( \chi^2 ))</td>
<td>1.62</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR Poisson (deviance)</td>
<td>1.62</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR Poisson/Cox (robust)</td>
<td>1.62</td>
<td>-0.4%</td>
</tr>
<tr>
<td>PR log-binomial</td>
<td>1.65</td>
<td>1.7%</td>
</tr>
<tr>
<td>OR logistic regression</td>
<td>2.49</td>
<td>53.5%</td>
</tr>
</tbody>
</table>
model and the Cox or Poisson models with adjusted variances provide correct point and interval estimates. It is, therefore, not only possible, but actually easy to use other models than logistic regression to analyze cross-sectional (or longitudinal) data with binary outcomes, the advantage being the prevalence (or cumulative incidence) ratio as the measure of association, more interpretable and easier to communicate, especially to non-epidemiologists. It is for the analyst to choose among these methods, based on software availability and the analyst's training.

### Competing interests
None declared.

### Author's contributions
AB proposed the idea, carried out part of the literature review and modeling, and drafted the manuscript. VH carried out most of the literature review, analyses, and prepared the tables and figures.

### Acknowledgements
We thank the Brazilian Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for supporting this work, and Drs. Cesar G. Victora, Bernardo L. Horta and J. Norberto W. Dachs for the suggestions and encouragement.

### References

### Table 9: Comparison of prevalence ratios and respective confidence interval estimates (obtained by unadjusted Poisson/Cox regression, Poisson regression with scale factor adjusted by $\chi^2$ and deviance, Poisson/Cox regression with robust variances, log-binomial regression and logistic regression) and odds ratio with confidence interval estimate (obtained by logistic regression) with the Mantel-Haenszel prevalence ratio in the analysis of the original and modified data using mother in a paid job as the outcome, father living with the family as the risk factor and social class as confounder (situation3).

<table>
<thead>
<tr>
<th></th>
<th>Original data</th>
<th></th>
<th>Modified data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point estimate</td>
<td>95% Confidence interval</td>
<td>Point estimate</td>
<td>95% Confidence interval</td>
</tr>
<tr>
<td></td>
<td>value</td>
<td>% diff.</td>
<td>lower</td>
<td>upper</td>
</tr>
<tr>
<td><strong>PR Mantel-Haenszel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR Poisson/Cox (unadj)</td>
<td>1.69</td>
<td>--</td>
<td>1.52</td>
<td>1.87</td>
</tr>
<tr>
<td>PR Poisson ($\chi^2$)</td>
<td>1.68</td>
<td>-0.3%</td>
<td>1.41</td>
<td>2.00</td>
</tr>
<tr>
<td>PR Poisson (deviance)</td>
<td>1.68</td>
<td>-0.3%</td>
<td>1.49</td>
<td>1.90</td>
</tr>
<tr>
<td>PR Poisson/Cox (robust)</td>
<td>1.68</td>
<td>-0.3%</td>
<td>1.46</td>
<td>1.94</td>
</tr>
<tr>
<td>PR log-binomial</td>
<td>1.62</td>
<td>-4.0%</td>
<td>1.47</td>
<td>1.78</td>
</tr>
<tr>
<td>OR logistic regression</td>
<td>3.75</td>
<td>122.6%</td>
<td>2.72</td>
<td>5.17</td>
</tr>
</tbody>
</table>

Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/21/prepub

Publish with BioMed Central and every scientist can read your work free of charge
"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:
- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp