Application of different statistical methods to estimate relative risk for self-reported health complaints among shoe factory workers exposed to organic solvents and plastic compounds

Khaldoun Nijem[1], Petter Kristensen[2,3], Awni Al-Khatib[4] and Espen Bjertness[3]

1) Department of Biology, Faculty of Science and Technology, Hebron University, Hebron, West Bank
2) Department of Occupational Medicine, National Institute of Occupational Health, Oslo, Norway
3) Section for Preventive Medicine and Epidemiology, Institute of General Practice and Community Medicine, University of Oslo, Oslo, Norway
4) Department of Chemistry, Faculty of Science and Technology, Hebron University, Hebron, West Bank

ABSTRACT

Objectives: Prevalence odds ratio (POR) is commonly used as a surrogate for relative risk (RR) in cross-sectional studies. When prevalences are high, POR may be a poor approximation for RR. Prevalence ratios (PRs) are more easily interpretable when evaluating exposure effects. Our objectives were to compare estimates of PRs and corresponding 95% confidence intervals (CIs) using three different statistical methods on a real data set, furthermore, to report possible practical problems in applying the methods.

Methods: Two statistical methods were compared: log-binomial regression and Cox regression. We examined selected high prevalence symptoms: headache, tingling of limbs, and breathing difficulty, and their association with solvent-exposed work tasks in 164 Hebron shoe factory workers.

Results: The two methods estimated identical crude point PR estimates and quite similar adjusted estimates. CIs were wider in Cox regression than in log-binominal regression, as exemplified by adjusted estimates for the association between participation in cleaning tasks and tingling of limbs in log-binomial regression (PR=1.78; CI=1.25–2.54), Cox regression (PR=1.76; CI=1.01–3.06). When we used Cox regression with robust variance we obtained narrower CIs (PR=1.76; CI=1.19–2.60). In the log-binomial regression analysis we had to exclude a few subjects with a predicted risk exceeding one.

Conclusions: Log-binomial regression is appropriate from a theoretical viewpoint. However, some individuals had a predicted risk larger than one, which caused the computation to abort. Cox regression could produce heavy ties when adjusted for confounders and yielded rather wide CIs, however, by using robust variance we will obtain narrow CIs. In conclusion, the two suggested methods have certain limitations and difficulties. However, Cox regression encountered less serious problems than in the other methods, and is also widely available.

INTRODUCTION

The prevalence odds ratio (POR) is commonly used in cross-sectional studies to assess associations between exposures and outcome. PORs can be estimated by logistic regression whenever the health outcome is dichotomous and the data needs covariate adjustment.

POR can be used as an approximation of prevalence ratio (PR) and interpreted as a relative risk (RR) in the case of rare diseases assumption (e.g. prevalence of outcomes below 0.1) (1-3). However, since many health outcomes are common, the interpretation of an odds ratio as a relative risk is often questionable (4). Lee and Chia (5) proposed the use of prevalence ratio (PR) instead of POR in cross-sectional studies of common diseases. According to Lee (6) PR is easier to communicate than POR and its meaning is more transparent. Others point out that the POR is commonly interpreted incorrectly as a relative risk in cross sectional studies dealing with common diseases such as for example musculoskeletal complaints (4) and other high prevalence outcomes (7).

Several methods have been proposed to estimate PRs for high prevalence outcomes (7). The methods are Cox proportional hazards (5), log-binomial regression (8) and a General Estimating Equations (GEE-logistic regression model) (9-12).

Skov et al. (7) applied these methods to simulated data sets and concluded that the point estimates of the models were close to the true parameters, but Cox regression produced too wide confidence intervals. However, Cox regression with robust variance can produce more appropriate confidence intervals (8). For the other methods, confidence intervals were generally considered to be correct (7). Zochetti et al. (13) concluded that the log-binomial model is preferable.

Our objective was to compare estimates of PRs and corresponding 95% confidence intervals, using the different methods (Cox regression, and log-binomial regression) on a real data set with high prevalence outcomes. The data set contained information about health complaints among shoe factory workers exposed to organic solvents and plastic compounds (14). PORs will also be presented to illustrate differences compared with PR estimates using the two methods. Finally, we will report possible practical problems in applying the methods.

SUBJECTS AND METHODS

Study Population

A sample of 164 male shoe factory workers in Hebron City who had worked more than one year were interviewed in 1996-97. The study population and methods are described in more detail elsewhere (14).

Questionnaire

Health complaints among shoe factory workers were collected in a structural interview (14). Questions relating to neuropsychiatric symptoms were obtained from a Swedish neuropsychiatric symptom questionnaire (Q16) (15).

Other questions included symptoms representing potential peripheral nervous system effects (tingling of limbs), mucous membrane irritation (sore eyes and breathing difficulty), in addition to work tasks, cumulative exposure, age, socio-demographic characteristics (smoking, marital status, and education) (14).

Exposure

The workers were exposed to organic solvents and plastic compounds, depending on work tasks and the type of production. Cumulative exposure was estimated for workers by calculating total months of work in four tasks (gluing, cleaning, varnishing and plastic molding). Adhesive work was categorized into four exposure subgroups (0, 1-12, 13-72, >72 months), cleaning into three subgroups (0, 1-24, >24 months), whereas varnishing and plastic molding was dichotomized (0, ≥1 month).

Statistical analysis

We applied two methods using the S-Plus 2000 software: Cox regression, and log-binomial regression. We estimated PRs for associations between exposed work tasks and selected high prevalence outcomes: headache, tingling of limbs and breathing difficulty. We also compared the PRs with the corresponding PORs in a standard logistic regression analysis (SPSS package for Windows, version 8).

The PRs and PORs were adjusted for categories of age (16-24, 25-29, 30-36, >36 years), education (<9, 9, >9 years), marital status (single or married), and smoking (yes, no). We calculated 95% confidence intervals (CI) for the estimated PRs and PORs.

In Cox regression the time variable was set to the same value (unity) for all individuals. An individual who reported a symptom was coded as a "death", the others as "censored". According to the recommendation of Skov et al. (7) the Breslow method for ties was used. In this model, when b_1 is the estimated coefficient corresponding to exposure, exp (b_1) is an approximation to the relative risk (RR) associated with that exposure. To obtain more correct confidence intervals for PRs estimated by Cox regression, we applied the robust variance option in STATA software (STATA/SE 8.0).

The log-binomial model is similar to logistic regression in assuming a binomial distribution of outcome. However, instead of using a logit link function, as is customary in standard logistic regression, a log link is applied. Hence, for a particular individual the relation between the risk p of an adverse outcome and the covariate values $x_1, x_2, \ldots$ is $\log(p) = b_0 + b_1x_1 + b_2x_2 + \ldots$, where $b_1, b_2, \ldots$ are the parameters to be estimated. If, say, x_1 is a dichotomous exposure then RR = exp (b_1) for that exposure. The log link binomial model is available in several statistical software packages, for example S-Plus.

RESULTS

In the log-binomial regression analysis we had to exclude a few subjects (between 1 and 8 for different work tasks and symptoms) because of predicted risks exceeding unity, causing the software to abort computation.

The prevalence of headache was 0.65. In the log-binomial regression, headache was moderately associated with exposure for >24 months in the cleaning task (adjusted PR=1.57; CI=1.17–2.10) (Table 1). Cox regression yielded a similar point estimate but a wider CI (PR=1.58; CI=0.98–2.54). It is possible to improve the situation using the robust variance estimates for the Cox regression (PR=1.58; CI=1.24–2.00). As expected, the crude PRs were identical in the two methods, but the CIs showed the same pattern as for the adjusted estimates.

The prevalence of tingling of limbs among shoe factory workers was 0.46. Association with cleaning activities showed the same pattern with similar adjusted PRs, and a wider CI in Cox regression than in log-binominal regression (Table 2).

The prevalence of breathing difficulty was 0.28. Breathing difficulty was found to be associated with exposure to adhesives and varnishing compounds in the two statistical methods (Table 3). Again, the same pattern was observed concerning crude and adjusted

Table 1. PORs, PRs and corresponding CIs between different work tasks and headache (prevalence = 65%) among Hebron shoe factory workers (n = 164), estimated by ordinary logistic regression, log-binomial regression, Cox regression, and Cox regression with robust variance.

		Statistical model			
Exposure	N	Logistic regression POR (CI)	Log-binomial regression PR (CI)	Cox regression PR (CI)	Cox regression/ robust variance PR (CI)
Adhesive					
0	51	1 (reference)	1 (reference)	1 (reference)	1 (reference)
1-12	21				
Unadjusted		0.55 (0.20–1.55)	0.79 (0.50–1.24)	0.79 (0.40–1.55)	0.79 (0.50–1.24)
Adjusted*		0.47 (0.16–1.42)	0.76 (0.48–1.20)	0.76 (0.38–1.51)	0.76 (0.48–1.19)
13-72	47				
Unadjusted		0.97 (0.42–2.23)	0.99 (0.74–1.32)	0.99 (0.61–1.62)	0.99 (0.75–1.31)
Adjusted*		0.98 (0.42–2.33)	0.97 (0.74–1.28)	0.99 (0.61–1.62)	0.99 (0.75–1.33)
>72	45				
Unadjusted		0.91 (0.39–2.10)	0.97 (0.72–1.30)	0.97 (0.59–1.59)	0.97 (0.72–1.29)
Adjusted*		1.11 (0.44–2.84)	1.02 (0.75–1.40)	1.05 (0.61–1.80)	1.05 (0.76–1.44)
Cleaning					
0	91	1 (reference)	1 (reference)	1 (reference)	1 (reference)
1-24	41				
Unadjusted		1.51 (0.7–3.26)	1.18 (0.88–1.57)	1.18 (0.74–1.87)	1.18 (0.88–1.57)
Adjusted*		1.43 (0.64–3.20)	1.08 (0.8–1.47)	1.15 (0.71–1.86)	1.15 (0.85–1.55)
>24					
Unadjusted	32	2.98 (1.03–8.66)	1.41 (1.07–1.86)	1.51 (0.94–2.40)	1.51 (1.19–1.91)
Adjusted*		3.76(1.23–11.48)	1.57 (1.17–2.10)	1.58 (0.98–2.54)	1.58 (1.24–2.00)
Plastic					
0	105	1 (reference)	1 (reference)	1 (reference)	1 (reference)
≥1	59				
Unadjusted		1.30 (0.66–2.53)	1.10 (0.87–1.38)	1.10 (0.74–1.62)	1.10 (0.87–1.38)
Adjusted*		1.39 (0.69–2.80)	1.12 (0.89–1.41)	1.12 (0.75–1.68)	1.12 (0.89–1.42)
Varnish					
0	103	1 (reference)	1 (reference)	1 (reference)	1 (reference)
≥1	61				
Unadjusted		1.25 (0.64–2.43)	1.08 (0.86–1.37)	1.08 (0.73–1.60)	1.08 (0.86–1.36)
Adjusted*		1.35 (0.68–2.68)	1.13 (0.90–1.42)	1.11 (0.74–1.65)	1.11 (0.88–1.39)

* PR and POR adjusted for categories of age, smoking, marital status, and education.

PRs and their corresponding CIs. POR estimates were invariably stronger (more distant from unity) than PRs, as expected (Tables 1-3).

DISCUSSION

In cross-sectional studies PORs are often presented and interpreted as relative risks, on the rare disease assumption. The reason could be that PORs are easily computed in logistic regression. However, the high overall prevalences of certain outcomes make POR a poor replacement for the RR. To overcome this problem, many authors have suggested directly estimating the PR, which is more easily interpreted than POR (16). We used two suggested methods for this, namely Cox regression and log-binomial regression. The log-binomial model yields the "correct" likelihood structure under the assumptions of multiplicative effects, and is thus the most appropriate method to estimate PR and corresponding CIs directly (1). However, the log-binomial model might produce prevalences greater than one (17). Although Cox regression produced approximately the same PR point estimates, it suffered from other shortcomings. Cox regression introduced heavy ties, that sometimes were difficult to correct for in the model, and the Breslow method is not particularly well suited for this, whereas other methods may produce bias (7).

Table 2. PORs, PRs and corresponding CIs between work tasks and tingling of limbs (prevalence = 46%) among Hebron shoe factory workers (n = 164), estimated by ordinary logistic regression, log-binomial regression, Cox regression, and Cox regression with robust variance.

		Statistical model			
Exposure	N	Logistic regression POR (CI)	Log-binomial regression PR (CI)	Cox regression PR (CI)	Cox regression/ robust variance PR (CI)
Adhesive					
0	51	1 (reference)	1 (reference)	1 (reference)	1 (reference)
1-12	21				
Unadjusted		0.52 (0.18–1.50)	0.68 (0.35–1.34)	0.68 (0.29–1.57)	0.68 (0.35–1.33)
Adjusted*		0.48 (0.15–1.49)	0.69 (0.38–1.94)	0.65 (0.27–1.56)	0.65 (0.33–1.27)
13-72	47				
Unadjusted		0.65 (0.29–1.44)	0.78 (0.49–1.24)	0.78 (0.43–1.43)	0.78 (0.49–1.24)
Adjusted*		0.64 (0.28–1.47)	0.76 (0.48–1.20)	0.78 (0.43–1.44)	0.78 (0.49–1.24)
>72	45				
Unadjusted		1.30 (0.58–2.91)	1.13 (0.77–1.67)	1.13 (0.65–1.97)	1.13 (0.77–1.66)
Adjusted*		0.97 (0.39–2.39)	1.01 (0.68–1.50)	0.98 (0.53–1.80)	0.98 (0.66–1.46)
Cleaning					
0	91	1 (reference)	1 (reference)	1 (reference)	1 (reference)
1-24	41				
Unadjusted		2.73 (1.28–5.82)	1.72 (1.17–2.53)	1.72 (1.01–2.93)	1.72 (1.17–2.53)
Adjusted*		3.01 (1.34–6.77)	1.78 (1.25–2.54)	1.76 (1.01–3.06)	1.76 (1.19–2.60)
>24	32				
Unadjusted		3.23 (1.40–7.44)	1.83 (1.24–2.73)	1.83 (1.05–3.22)	1.83 (1.24–2.72)
Adjusted*		2.93 (1.22–7.04)	1.61 (1.11–2.35)	1.68 (0.95–2.98)	1.68 (1.12–2.52)
Plastic					
0	105	1 (reference)	1 (reference)	1 (reference)	1 (reference)
≥1	59				
Unadjusted		3.33 (1.71–6.47)	1.83 (1.32–2.53)	1.83 (1.16–2.87)	1.83 (1.32–2.52)
Adjusted*		3.19 (1.59–6.42)	1.69 (1.22–2.33)	1.73 (1.09–2.75)	1.73 (1.25–2.39)
Varnish					
0	103	1.0 (reference)	1.0 (reference)	1.0 (reference)	1.0 (reference)
≥1	61				
Unadjusted		0.74 (0.39–1.40)	0.84 (0.59–1.21)	0.84 (0.52–1.36)	0.84 (0.59–1.21)
Adjusted*		0.65 (0.33–1.28)	0.84 (0.59–1.19)	0.80 (0.49–1.31)	0.80 (0.57–1.14)

* PR and POR adjusted for categories of age, smoking, marital status, and education.

Even though the log-binomial regression model is preferable from a theoretical point of view, it encountered numerical problems. Although the model itself may generally be appropriate, one may occasionally encounter a few individuals for whom the predicted risk is larger than one, due to a rare combination of covariates. Apart from being illogical, a predicted risk above one will often cause the software to abort computations, giving only slight clues as to the nature of the problem. The higher the prevalence, the more frequent this problem will be. To avoid this problem we rewrote the software to locate and remove those few individuals that caused the computation to crash. Clearly, this strategy is not tenable in situations with more frequent predictions above one. Ultimately, of course, if many of the predicted risks exceed one, this is a sign of a mis-specified model rather than of just a few deviating individuals.

It is worth noting that ordinary logistic regression does not suffer from any of the shortcomings of the other models. But when comparing the PORs with the PRs for high prevalence outcomes, it is clear that they differ substantially, as the POR typically overestimates the PR.

Also, there are different assumptions underlying the logistic model as compared to the log-binominal model. Whereas the logistic model assumes a constant exposure over all covariate levels, the log-binominal

Table 3. PORs, PRs and corresponding CIs between work tasks and breathing difficulty (prevalence = 28%) among Hebron shoe factory workers (n = 163), estimated by ordinary logistic regression, log-binomial regression, Cox regression, and Cox regression with robust variance.

		Statistical model			
Exposure	N	Logistic regression POR (CI)	Log-binomial regression PR (CI)	Cox regression PR (CI)	Cox regression/ robust variance PR (CI)
Adhesive					
0	51	1 (reference)	1 (reference)	1 (reference)	1 (reference)
1–12	21				
Unadjusted		1.64 (0.51–5.29)	1.46 (0.60–3.54)	1.46 (0.53–4.01)	1.46 (0.61–3.51)
Adjusted*		1.72 (0.49–6.04)	1.88 0.81–4.37)	1.47 (0.51–4.20)	1.47 (0.62–3.48)
13–72	46				
Unadjusted		2.19 (0.87–5.48)	1.77 (0.89–3.54)	1.84 (0.84–4.03)	1.84 (0.94–3.62)
Adjusted*		2.53 (0.97–6.59)	2.25 (1.08–4.69)	1.97 (0.89–4.33)	1.97 (0.99–3.92)
>72	45				
Unadjusted		1.37 (0.52–3.60)	1.27 (0.59–2.74)	1.36 (0.59–3.15)	1.36 (0.65–2.85)
Adjusted*		2.05 (0.70–6.01)	1.87 (0.81–4.28)	1.72 (0.71–4.14)	1.72 (0.78–3.79)
Cleaning					
0	91	1 (reference)	1 (reference)	1 (reference)	1 (reference)
1–24	40				
Unadjusted		2.43 (1.07–5.53)	1.90 (1.06–3.39)	1.97 (1.01–3.87)	1.97 (1.12–3.47)
Adjusted*		2.81 (1.15–6.91)	1.93 (1.07–3.46)	2.03 (1.01–4.09)	2.03 (1.15–3.58)
>24	32				
Unadjusted		2.12 (0.87–5.19)	1.74 (0.92–3.29)	1.74 (0.82–3.68)	1.74 (0.92–3.28)
Adjusted*		2.37 (0.92–6.14)	1.54 (0.80–2.96)	1.81 (0.84–3.89)	1.81 (0.95–3.44)
Plastic					
0	104	1 (reference)	1 (reference)	1 (reference)	1 (reference)
≥1	59				
Unadjusted		1.87 (0.93–3.76)	1.56 (0.95–2.55)	1.56 (0.87–2.80)	1.56 (0.95–2.55)
Adjusted*		1.86 (0.89–3.91)	1.56 (0.96–2.48)	1.53 (0.84–2.79)	1.53 (0.94–2.48)
Varnishing					
0	103	1.0 (reference)	1.0 (reference)	1.0 (reference)	1.0 (reference)
≥1	60				
Unadjusted		2.43 (1.20–4.92)	1.88 (1.14–3.11)	1.93 (1.07–3.47)	1.93 (1.18–3.16)
Adjusted*		2.75 (1.30–5.80)	2.03 (1.25–3.30)	1.99 (1.10–3.60)	1.99 (1.22–3.25)

* PR and POR adjusted for categories of age, smoking, marital status, and education.

model assumes a constant PR over all levels of adjustment. If the log-binominal model was correct, the logistic regression model should include interaction terms, and vice versa. Again, at low prevalences the difference may not be substantial, but it becomes considerable at high prevalences.

To illustrate differences within a real data set, we selected outcomes with different, high prevalences. Preferably, the estimated PRs should be similar for the two methods (7). This was true for unadjusted PRs but when we adjusted PRs for potential confounding factors, slight differences were obtained.

Confidence intervals of unadjusted and adjusted PRs obtained by Cox regression were too wide compared with those obtained by log-binomial analysis. However, when we used robust variance estimates for Cox regression we obtained appropriate confidence intervals.

As a conclusion, the two suggested methods have certain limitations and difficulties. The log-binomial model is appropriate from a theoretical viewpoint. However, Cox regression with robust variance may be a suitable method since we obtained point PR estimates with less serious problems than we experienced with the other method.

ACKNOWLEDGEMENT

The authors are grateful to Professor Håkon K. Gjessing, Section of Medical Statistics, University of Oslo, for his help with statistical analysis.

REFERENCES

1. Traissac P, Martin-Prével Y, Delpeuch F, et al. [Logistic regression vs other generalized linear models to estimate prevalence rate ratios] [in French, English summary]. *Rev Epidemiol Sante Publique* 1999; **47**: 593-604.
2. Rothman KJ. Modern epidemiology. Boston/Toronto: Little, Brown and Company, 1986.
3. Strömberg U. Prevalence odds v prevalence ratio. *Occup Environ Med* 1994; **51**: 143-144.
4. Axelson O, Fredriksson M, Ekberg K. Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. *Occup Environ Med* 1994; **51**: 574.
5. Lee J, Chia KS. Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology. *Br J Ind Med* 1993; **50**: 861-864.
6. Lee J. Prevalence odds ratio v prevalence ratio – a response. *Occup Environ Med* 1995; **52**: 781-784.
7. Skov T, Deddens J, Petersen M, et al. Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol* 1998; **27**: 91-95.
8. Barros A, Hirakata V. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence value. *BMC Med Res Meth* 2003; **3**: 21.
9. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 1986; **123**: 174-184.
10. Liang K, Zeger SL. Longitudinal data analysis using generalized linear model. *Biometrica* 1986; **73**: 13-22.
11. Schouten EG, Dekker JM, Kok FJ. Risk ratio and rate ratio estimation in case cohort design: hypertension and cardiovascular mortality. *Stat Med* 1993; **12**: 1733-1745.
12. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol* 2004; **160**: 301-305.
13. Zocchetti C, Consonni D, Bertazzi PA. Relationship between prevalence rate ratios and odds ratios in cross-sectional studies. *Int J Epidemiol* 1997; **26**: 220-223.
14. Nijem K, Kristensen P, Al-Khatib A, et al. Prevalence of neuropsychiatric and mucous membrane irritation complaints among Palestinian shoe-factory workers exposed to organic solvents and plastic compounds. *Am J Ind Med* 2001; **40**: 192-198.
15. Hane M, Hogstedt C. [Neuropsychiatric symptoms among solvent exposed workers – a questionnaire for screening]. [In Swedish, English summary]. *Läkartidningen* 1980; **77**: 437-439.
16. Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done. *Occup Environ Med* 1998; **55**: 272-277.
17. Strömberg U. Prevalence odds ratio v prevalence ratio – some further comments. *Occup Environ Med* 1995; **52**: 143.