Comparison of regression models for binary outcome variables in clinical trials

Adhin Bhaskar and Chinnaiyan Ponnuraja*

Department of Statistics, ICMR-National Institute for Research in Tuberculosis, No. 1, Mayor Sathiyamoorthy Road, Chetpet, Chennai 600 031, India

The widely used logistic regression may not be suitable to model binary outcomes in clinical trials. The present study compares and assesses various binary regression models such as logistic, log-binomial, Poisson and Cox proportional models for clinical trials. A dataset obtained from a clinical trial conducted on tuberculosis patients is used to illustrate the models. The estimated odds ratios from logistic regression severely overestimated the relative risks, thereby overestimating the overall relationship. Log-binomial and Poisson with sandwich covariance estimator were found to be suitable for estimating adjusted relative risks in clinical trials.

Keywords: Binary outcomes, clinical trials, regression models, relative risk.

THE variables of interest in many studies are likely to have a binary nature on the outcome. The variables may include death status, disease status, cure from disease or relapse, etc. When these arise as dependent variables in regression analysis, many studies have used logistic regression irrespective of the study design[1,2]. Many clinical trials have also used logistic regression to estimate the relationship between the binary response variable and the independent variables[3,4]. Logistic regression gives the odds ratio as an estimate of the relationship. Whereas, relative risk is considered as the ideal measure of relationship in prospective studies such as clinical trials and cohort studies. The logistic regression is suitable for clinical trials or cohort studies only if the occurrence of the event is rare. It is well known that when the outcome is common, odds ratio overestimates the relative risk. In general, logistic regression is suitable only for case-control studies[5]. However, due to popularity and easier convergence, logistic regression has been widely used in clinical trials. Usually, in treatment trials for tuberculosis (TB), the cure rate is relatively high. Hence, the use of logistic regression for this scenario may result in overestimating the effect of the treatment or other variables.

Though they are less used, methods have been developed to estimate the adjusted relative risk. Regression-based approaches are also available to estimate the adjusted relative risk when data contain confounders. The major regression-based approaches for estimating relative risk are log-binomial regression, Poisson regression, etc. Cox regression also estimates the relative risk if the time variable is constant[6].

The present study is aimed at comparing these regression models and logistic regression for a dataset obtained from a clinical trial, with a special focus on TB trials. Data obtained from the clinical trial have been used to illustrate the models. The models compared in this study are logistic regression, log-binomial regression, Poisson regression, Poisson regression with sandwich covariance estimator and Cox proportional regression.

Logistic regression falls under a broader generalized linear modelling framework and it relates log of odds with the linear function of independent variables. The model of logistic regression with k predictors can be written as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \quad (1)$$

where π_i is the probability of the event and $\frac{\pi_i}{1-\pi_i}$ is the odds, i.e. the ratio of the probability of occurrence to non-occurrence. The odds ratio of logistic regression can be obtained by taking the exponential of the regression coefficients, i.e. $\exp(\beta_k)$. The parameters of the logistic regression are estimated using the maximum likelihood method.

Similar to logistic regression, log-binomial regression also assumes a binomial distribution for the binary outcome variable. The major difference between logistic and log-binomial regression is the way the outcome is related to the independent variables. Logistic regression uses logit function whereas log-binomial regression uses a log link function to establish the relationship. The model formula of log-binomial regression can be written as

$$\log(\pi_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \quad (2)$$

The use of log link function equips the log-binomial regression to give adjusted relative risk by taking the exponential of regression coefficients. However, the use of log link function induces some constraints on log-binomial regression. For the probability of the event to be less than one, values of the linear combination of independent variables should be zero or negative. Hence, in order to avoid estimating probability greater than one, the parameter space has to be constrained. However, if the estimates are on the boundary of the parameter space, the often-used optimization methods may not converge[7]. The failure in converge occurs usually when the events are common or the model contains continuous or multinomial covariates. An adaptive logarithmic barrier algorithm has been proposed to estimate the parameters in log-binomial regression that has high convergence rate compared to other optimization techniques[8].

*For correspondence. (e-mail: cponnuraja@nirt.res.in)

Poisson regression is a widely used approach under the generalized linear modelling framework. The exponential of regression coefficients in Poisson regression directly gives the adjusted relative risk. In Poisson regression, the response variable is related to the explanatory variables using a log link function. Poisson regression is basically a count regression approach; however, it is suitable for binary variables if the outcome is rare. The model is written as

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \quad (3)$$

When the outcome is common, Poisson regression is likely to overestimate the binomial standard error. In order to avoid overestimation of standard errors, the variance–covariance matrix can be computed using robust covariance estimator, which is also known as sandwich covariance estimator. The sandwich estimator is derived from gradient and hessian. The estimator is named so, since the outer product of the gradient is enclosed within two hessians as HgH[9]. Zou[10] has demonstrated Poisson regression with robust covariance estimator for prospective studies.

Cox proportional regression is often used for modelling time to event data with censoring. It models the hazard rate as a function of the predictors. The model can be written as

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_k x_k), \quad (4)$$

where $h_0(t)$ is the baseline hazard function. Cox regression for binary data was first suggested by Lee and Chia[11]. They used it to estimate the prevalence ratio in cross-sectional studies. Cox regression directly gives the adjusted relative risk and confidence interval for prospective studies, if constant risk period is assigned to each subject[6].

The secondary data used for the study consist of 412 TB patients admitted in a randomized controlled clinical trial, and it was ethically approved. The trail consisted of an experimental treatment and a control regimen for patients diagnosed with pulmonary TB. The event of interest is time to sputum conversion during the treatment period. For the illustration of binary regression models, the sputum conversion status (yes/no) has been used as the response variable. Other covariates are also included like age, gender, weight, treatment and drug susceptibility test.

The data used for the study contain 412 patients diagnosed with TB consisting of 66.50% ($n = 274$) males and 33.5% ($n = 138$) females. Among the participants, 50.49% ($n = 208$) received experiential treatment and 49.51% ($n = 204$) received control treatment (Table 1). The primary outcome for the study was the time to sputum conversion, i.e. recovery from the disease. Among the participants, 92.96% ($n = 383$) experienced the outcome. Since the rate of event is high in the study, the use of logistic regression is likely to overestimate the relative risk. Log-binomial, Poisson regression or Cox proportional regression models are ideal for this situation. In this study we compare the performance of various binary regression models using the illustrative examples.

The unadjusted relative risks for gender and treatment were computed using contingency table and binary regression models such as logistic regression, log-binomial regression, Poisson regression, Poisson regression with sandwich covariance estimator and Cox proportional regression. Table 2 presents the results of unadjusted relative risks and confidence intervals. The estimated odds ratio of logistic regression severely overestimated the relative risk for gender and treatment. The odds ratio for gender was 1.85 times larger than the relative risk obtained from the contingency table that resulted in severe overestimation of the relationship between gender and treatment outcome. Moreover, the confidence interval was also wider. The unadjusted odds ratio obtained for gender was 1.949 (0.912, 4.165), whereas the relative risk was 1.05 (0.988, 1.120). The odds ratio (OR = 2.034, CI = (0.940, 4.660)) corresponding to treatment also overestimated the relative risk (RR = 1.050, CI = (0.995, 1.107)). The odds ratio corresponding to treatment was 1.94 times larger than the relative risk computed from the contingency table.

In contrast, log-binomial estimated relative risk similar to that obtained from the contingency table for treatment (RR = 1.050, CI = (0.995, 1.107)) as well as gender (RR = 1.052, CI = (0.988, 1.120)). Poisson regression also estimated relative risk accurately. However, as the variance of Poisson distribution is larger than the binomial variance, the estimated standard errors were larger. This resulted in wider confidence interval of the relative risk for treatment (RR = 1.050, CI = (0.859, 1.283)) as well as gender (RR = 1.052, CI = (0.851, 1.307)) in Poisson fit. The overestimation of the standard error was rectified by computing the covariance matrix of Poisson fit using the sandwich estimator, which resulted in estimating a confidence interval similar to that obtained with the standard formulae.

Table 1. Socio-demographic and clinical profile of patients

Variable		*n* (%)
Age (yrs)[a]		32.75 (12.09)
Gender	Male	274 (66.50)
	Female	138 (33.50)
Treatment	Experimental	208 (50.49)
	Control	204 (49.51)
Drug sensitivity	Yes	367 (89.08)
	No	45 (10.92)
Sputum conversion	Yes	383 (92.96)
	No	29 (7.04)
Baseline weight (kg)[a]		38.59 (6.26)

[a]Mean (SD).

Table 2. Unadjusted relative risk and 95% confidence interval for gender and treatment on outcome

Model	Gender (male)	Treatment (experimental)
Cross tab	1.052 (0.988, 1.120)	1.050 (0.995, 1.107)
Logistic[a]	1.949 (0.904, 4.184)	2.034 (0.940, 4.660)
Log-binomial	1.052 (0.988, 1.120)	1.050 (0.995, 1.107)
Poisson	1.052 (0.851, 1.307)	1.050 (0.859, 1.283)
Poisson-sandwich	1.052 (0.988, 1.120)	1.050 (0.995, 1.107)
Cox proportional	1.153 (0.930, 1.428)	1.148 (0.939, 1.403)

[a]Odds ratio.

Though Cox regression estimates relative risk at a constant duration of the risk, the estimated relative risk from Cox regression overestimated the relationship. The estimated relative risk from Cox regression was 1.093 and 1.096 times larger than that obtained from the contingency table for treatment and gender respectively, with slightly wider confidence intervals. The unadjusted relative risk for gender and treatment obtained from Cox proportional regression was 1.153 (0.930, 1.428) and 1.148 (0.939, 1.403) respectively.

Regression models have been fitted to obtain the adjusted effect of the independent variables on the outcome. The adjusted relative risks have been calculated from various multivariable binary regression models using the predictors – gender, age, treatment, sensitivity and weight at baseline (Table 3). The log-binomial regression and Poisson with sandwich estimator gave similar estimates and confidence intervals for adjusted relative risk. Whereas the confidence interval obtained from the standard Poisson regression was wider than the log-binomial and Poisson with the sandwich estimator. The adjusted odds ratio obtained from multivariable logistic regression was much larger than the log-binomial and Poisson models. Especially, the odds ratio corresponding to gender (OR = 1.953, CI = (0.775, 4.931)), treatment (OR = 1.951, CI = (0.887, 4.528)) and sensitivity (OR = 1.734, CI = (0.483, 11.134)) was much larger compared to log-binomial and Poisson estimates. As seen in the estimates obtained from the unadjusted model, the adjusted relative risk obtained from Cox regression was larger compared to the other models. The estimated standard errors were found to be larger than the log-binomial regression and Poisson regression models.

Odds ratio is ideal for case-controls studies, as the prevalence or incidence of the outcome is not estimated. The adjusted odds ratio for case-control studies can be obtained by means of logistic regression. In prospective studies, incidence of the disease can be estimated directly. Hence the relationship between binary outcome and exposure is often assessed using relative risk. The unadjusted relative risk can be computed using a contingency table. However, many a time, there is a need for estimating relative risk after adjusting the effect of confounding variables. Mantel Hansen methods can provide adjusted relative risks after adjusting the effect of one or two extraneous variables. However, computation is difficult if there are many variables to be adjusted. Regression-based approaches are more suitable and easier to estimate the relative risk in such a situation.

This study has compared the performance of various regression models in estimating the relative risk in clinical trials using an empirical dataset. The results have shown the unsuitability of logistic regression for clinical trials, though it is widely used. On comparison with the unadjusted relative risk obtained from the crossstab, Poisson and log-binomial regression estimated relative risks accurately. Poisson regression showed the tendency to overestimate the standard errors of relative risk, which resulted in a wider confidence interval. The overestimation of standard error can be corrected using Poisson regression with robust covariance estimation. The results of Poisson with sandwich covariance estimator and log-binomial regression were found to be similar. These models estimated unadjusted as well as adjusted relative risk with more precision. The estimates of Cox regression were found to be less efficient compared to log-binomial and Poisson regression with robust covariance estimator. Odds ratios obtained from logistic regression were substantially larger than the relative risks obtained from Poisson and log-binomial regression for unadjusted as well as adjusted effects. A review study showed that 19 out of 24 clinical trials that used logistic regression for the binary outcome variable had at least one OR 100% larger than the RR on log scale[12].

The log-binomial regression reported failure to converge due to restricted range of the parameter space. The regression fails especially when there are too many categories and continuous independent variables, or the incidence rate is high[13]. The IBM SPSS software uses Newton–Raphson, Fisher scoring, or a hybrid method of iteration to estimate the parameters. The default iteration method used is the hybrid method in which Fisher scoring iterations are performed before switching to the Newton–Raphson method. The most commonly used 'glm' function of 'stats' package of the R-software employs iteratively reweighted least squares method to estimate parameters. These iteration methods used in the IBM SPSS and R-software failed to estimate the parameters of the log-binomial regression fitted to the current data due to failure in convergence. The introduction of more effective

Table 3. Multivariable regression models fitted to predict factors affecting sputum conversion

Variable	Logistic OR (95% CI)	Log-binomial RR (95% CI)	Poisson RR (95% CI)	Poisson-sandwich RR (95% CI)	Cox proportional RR (95% CI)
Age (yrs)	0.983 (0.952, 1.017)	0.999 (0.997, 1.001)	0.999 (0.990, 1.008)	0.999 (0.997, 1.001)	0.996 (0.988, 1.005)
Gender (male)	1.953 (0.775, 4.931)	1.044 (0.973, 1.119)	1.051 (0.820, 1.353)	1.051 (0.986, 1.120)	1.146 (0.888, 1.478)
Baseline weight (kg)	1.033 (0.964, 1.115)	1.001 (0.999, 1.003)	1.002 (0.984, 1.021)	1.002 (0.998, 1.006)	1.006 (0.987, 1.025)
Treatment (experimental)	1.951 (0.887, 4.528)	1.049 (0.999, 1.101)	1.047 (0.853, 1.285)	1.047 (0.991, 1.105)	1.143 (0.931, 1.403)
Drug sensitivity (yes)	1.734 (0.483, 11.134)	1.008 (0.970, 1.049)	1.03 (0.739, 1.399)	1.03 (0.960, 1.106)	1.085 (0.788, 1.490)

OR, Odds ratio; RR, Relative risk; CI, Confidence interval.

iteration procedures through R-packages has ensured a higher convergence rate of the model compared to other optimization techniques that makes the log-binomial regression more efficient[14]. However, Poisson regression with sandwich estimator can be used in case of nonconvergence of log-binomial regression. In addition, Poisson regression tends to be robust to various model misspecifications[15]. A simulation study comparing various binary regression models showed that Poisson regression with robust variance estimator gives unbiased estimates of the relationship[16]. However, Poisson regression may not be suitable if the interest of the researcher is to construct a predictive model rather than estimating accurate relative risk, as the estimated probability might become larger than one.

The results of this study illustrate the unsuitability of logistic regression for prospective studies. In clinical trials, use of logistic regression might result in overestimating the effect of treatment, especially when the outcome is common. Thus, the present study recommends Poisson with robust variance estimator and log-binomial regression as the most suitable models for clinical trials in modelling binary variables.

Conflict of interest: The authors declare no conflict of interest.

1. Vondran, M. *et al*., Evaluation of risk factors for a fulminant *Clostridium difficile* infection after cardiac surgery: a single center, retrospective cohort study. *BMC Anesthesiol.*, 2018, **18**(1), 133.
2. Adem, K. S., Janakiraman, K., Gebremeskel, B. F., Chala, M. B., Gelaw, A. Y. and Alemu. K., Epidemiology and factors associated with peripheral neuropathy among HIV infected patients in Gondar, Ethiopia: a cross-sectional study. *PLoS ONE*, 2019, **4**, e0211354.
3. Hanioka, T., Ojima, M., Tanaka, H., Naito, M., Hamajima, N. and Matsuse, R., Intensive smoking-cessation intervention in the dental setting. *J. Dent. Res.*, 2010, **89**, 66–70.
4. Menezes-Silva, R. *et al.*, Randomized clinical trial of class II restoration in permanent teeth comparing ART with composite resin after 12 months. *Clin. Oral Invest.*, 2019, **23**(9), 3623–3635.
5. Lee, J., Tan, C. S. and Chia, K. S., A practical guide for multivariate analysis of dichotomous outcomes. *Ann. Acad. Med. Singapore*, 2009, **38**, 714–719.
6. Diaz-Quijano, F. A., A simple method for estimating relative risk using logistic regression. *BMC Med. Res. Methodol.*, 2012, **12**.
7. Barros, A. J. D. and Hirakata, V. N., Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med. Res. Methodol.*, 2003, **3**, 1–13.
8. de Andrade, B. B. and Andrade, J. M. L., Some results for maximum likelihood estimation of adjusted relative risks. *Commun. Stat. Theory Methods*, 2018, **47**, 5750–5769.
9. Hilbe, J. M., *Negative Binomial Regression*, Cambridge University Press, New York, 2011, 2nd edn.
10. Zou, G. A., Modified Poisson regression approach to prospective studies with binary data. *Am. J. Epidemiol.*, 2004, **159**, 702–706.
11. Lee, J. and Chia, K. S., Estimation of prevalence rate ratios for cross-sectional data: an example in occupational epidemiology. *Br. J. Ind. Med.*, 1993, **50**, 861–862.
12. Knol, M. J., Duijnhoven, R. G., Grobbee, D. E., Moons, K. G. M. and Groenwold, R. H. H., Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials. *PLoS ONE.*, 2011, **6**(6), e21248.
13. Knol, M. J., Le Cessie, S., Algra, A., Vandenbroucke, J. P. and Groenwold, R. H. H., Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *CMAJ*, 2012, **184**(8), 895–899.
14. Donoghoe, M. W. and Marschner, I. C., Logbin: an R package for relative risk regression using the log-binomial model. *J. Stat. Softw.*, 2018, **86**(9).
15. Chen, W., Qian, L., Shi, J. and Franklin, M., Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification. *BMC Med. Res. Methodol.*, 2018, **18**(1), 1–12.
16. Huang, F. L., Alternatives to logistic regression models in experimental studies. *J. Exp. Educ.*, 2019, 1–16; doi:10.1080/00220973.2019.1699769.

ACKNOWLEDGEMENT. We thank all the doctors and supporting staff of ICMR-National Institute for Research in Tuberculosis, Chennai, associated with this clinical trial.

Received 3 January 2020; revised accepted 22 October 2020

doi: 10.18520/cs/v119/i12/2010-2013