The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 5

Relative Risk Estimation in Randomized Controlled Trials: A Comparison of Methods for Independent Observations

Lisa N. Yelland, *University of Adelaide*
Amy B. Salter, *University of Adelaide*
Philip Ryan, *University of Adelaide*

Recommended Citation:
Yelland, Lisa N.; Salter, Amy B.; and Ryan, Philip (2011) "Relative Risk Estimation in Randomized Controlled Trials: A Comparison of Methods for Independent Observations," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 5.
DOI: 10.2202/1557-4679.1278

Relative Risk Estimation in Randomized Controlled Trials: A Comparison of Methods for Independent Observations

Lisa N. Yelland, Amy B. Salter, and Philip Ryan

Abstract

The relative risk is a clinically important measure of the effect of treatment on binary outcomes in randomized controlled trials (RCTs). An adjusted relative risk can be estimated using log binomial regression; however, convergence problems are common with this model. While alternative methods have been proposed for estimating relative risks, comparisons between methods have been limited, particularly in the context of RCTs. We compare ten different methods for estimating relative risks under a variety of scenarios relevant to RCTs with independent observations. Results of a large simulation study show that some methods may fail to overcome the convergence problems of log binomial regression, while others may substantially overestimate the treatment effect or produce inaccurate confidence intervals. Further, conclusions about the effectiveness of treatment may differ depending on the method used. We give recommendations for choosing a method for estimating relative risks in the context of RCTs with independent observations.

KEYWORDS: binary outcome, log binomial regression, randomized controlled trial, relative risk, simulation

Author Notes: The authors would like to thank Professor Maria Makrides (Chair) on behalf of the members of the DINO Steering Committee (Maria Makrides, Robert Gibson, Andrew McPhee, Carmel Collins, Peter Davis, Lex Doyle, Karen Simmer, Paul Colditz, Scott Morris and Philip Ryan) for granting permission to use data from the DINO Trial, Ms. Kristyn Willson for providing the datasets for analysis, and Mr. Thomas Sullivan for reviewing drafts of the paper. This work was supported by a Commonwealth funded Australian Postgraduate Award.

1 Introduction

Binary outcomes are common in randomized controlled trials (RCTs) and may relate to efficacy (for example, presence or absence of disease) or safety (for example, presence or absence of side effects). Such outcomes have traditionally been analyzed using logistic regression, which provides an estimate of the effect of treatment on the outcome expressed as an odds ratio. Rather than estimating odds ratios, recent literature suggests a growing preference for estimating relative risks; the odds ratio has been described as incomprehensible (Lee, 1994) and lacking clinical meaning (Sinclair and Bracken, 1994), whereas the relative risk is considered easier to interpret (Walter, 2000; Schechtman, 2002). As a result, the relative risk is now the effect measure of choice for many RCTs; see Ahmad et al. (2009) for a recent example.

To estimate relative risks, log binomial regression has been recommended (Skov et al., 1998). Like logistic regression, log binomial regression can be used to estimate the effect of treatment on a binary outcome while controlling for both categorical and continuous baseline covariates, however convergence problems are common (Zou, 2004). This poses a problem in RCTs, since the analysis method should be pre-specified (ICH E9 Expert Working Group, 1999) before it is known whether the model will converge. Alternative methods for estimating relative risks have been proposed (Wacholder, 1986; Flanders and Rhodes, 1987; Schouten et al., 1993; Deddens et al., 2003; Zou, 2004; Lumley et al., 2006; Localio et al., 2007) but comparisons between methods have been limited, particularly in the context of RCTs, and it is currently unclear which method should be preferred. As a result, a variety of methods are being used in RCTs in practice (see e.g. Boardman et al., 2004; Green et al., 2008; Ahmad et al., 2009; van der Meer et al., 2009).

The aims of this study were: (i) to compare the relative performance of methods for estimating relative risks in the context of RCTs with independent observations; and (ii) to make recommendations about which method(s) should be used in practice. The methods for estimating relative risks are described in Section 2, compared by simulation in Section 3, and applied to an example dataset in Section 4. We conclude with a discussion in Section 5 and recommendations in Section 6.

2 Methods

2.1 Setting and notation

Consider a two-group parallel RCT comparing a new treatment to a standard or control treatment. Let there be N independent subjects recruited and allocated to

the 'treatment' or 'control' group. Let Y_i be a binary outcome for subject i $(i = 1,\ldots,N)$, where $Y_i = 1$ if the outcome is a 'success', however this is defined, and $Y_i = 0$ otherwise. We assume that the probability of success for subject i (π_i) depends on a vector of known covariates $\mathbf{X}_i = (X_0, X_{1i}, X_{2i}, \ldots, X_{Ki})$, where $X_0 = 1$, X_{1i} is a binary indicator for treatment group, coded as $X_{1i} = 1$ for treatment and $X_{1i} = 0$ for control, and $X_{2i}, \ldots, X_{Ki}$ are $K-1$ possible baseline covariates. The aim of the analysis is to estimate the relative risk of success comparing treatment to control and to obtain a confidence interval, while potentially adjusting for pre-specified categorical and/or continuous baseline covariates.

We consider ten different methods for analyzing such data: log binomial regression, constrained log binomial regression, the COPY 1000 method, expanded logistic regression, log Poisson regression, log normal regression and four methods based on logistic regression (conditional standardization or marginal standardization for estimating the relative risk, combined with either the delta method or bootstrapping for obtaining a confidence interval). Each method will now be described.

2.2 Log binomial regression

The log binomial regression model is a generalized linear model (GLM) with a log link, written as

$$\log(\pi_i) = \mathbf{X}_i\boldsymbol{\beta}, \qquad (1)$$

where $\boldsymbol{\beta}$ is a column vector of model parameters and the errors are assumed to follow a binomial distribution. Given the model parameter estimates $\hat{\boldsymbol{\beta}}$, obtained by the method of maximum likelihood, the relative risk comparing treatment to control can be estimated simply by $\exp(\hat{\beta}_1)$, where β_1 is the coefficient of the treatment indicator X_{1i}. This model will fail to provide an estimate of the relative risk if there are convergence problems during the estimation process as a result of restrictions placed on the parameter space. These restrictions relate to the requirement that the predicted probabilities $\hat{\pi}_i = \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})$ must satisfy $0 \le \hat{\pi}_i \le 1$, implying that $\mathbf{X}_i\hat{\boldsymbol{\beta}}$ must not exceed zero for any $\mathbf{X}_i$, $i = 1,\ldots,N$. If convergence problems occur, choosing different starting values for the parameter estimates may help. However, if the maximum likelihood estimate is on the boundary of the parameter space then convergence will not occur (Barros and Hirakata, 2003;

Deddens et al., 2003). This has prompted the development of alternative methods for estimating relative risks.

2.3 Alternative methods

Wacholder (1986) proposed a constrained iterative estimation procedure to estimate relative risks, referred to hereafter as constrained log binomial regression. This procedure involves checking whether any predicted probabilities exceed some maximum allowable value, such as 0.99, at each iteration. Any predicted probabilities exceeding this maximum are set to the maximum before parameter estimates are updated and the process iterates until convergence.

Deddens et al. (2003) suggested performing log binomial regression on a modified dataset containing $c-1$ copies of the original dataset and one copy of the original dataset with the outcomes reversed (i.e. successes become failures and vice versa). The authors suggested using $c=1000$ in practice and called this the 'COPY 1000 method'. The maximum likelihood parameter estimates in the new dataset approximate those in the original dataset but lie in the interior of the parameter space, rather than possibly lying on the boundary. The standard errors of the parameter estimates need to be multiplied by the square root of c to correct for the additional data. Rather than making physical copies of the data, appropriate weights can be applied to a single copy of the original dataset combined with a single copy of this dataset where the outcomes are reversed (Lumley et al., 2006; Petersen and Deddens, 2008; Yu and Wang, 2008).

Schouten et al. (1993) recognized that by manipulating the data, logistic regression can be used to estimate relative risks directly. An expanded dataset is created, where successes are duplicated and the outcome is changed to a failure for these duplicates. The probability of success in the original dataset then equals the odds of success in the expanded dataset. Fitting a logistic regression model to the expanded dataset results in consistent estimates of the parameters in the log binomial regression model, apart from possibly the intercept, however the standard errors will be incorrect and robust variance estimates are therefore recommended. A potential disadvantage of this expanded logistic regression approach is that it allows invalid predicted probabilities (i.e. $\hat{\pi}_i > 1$) to occur.

McNutt et al. (2003) suggested estimating relative risks using log Poisson regression. Like log binomial regression, this involves fitting a GLM with a log link and hence has the same form as (1), however the errors are assumed to follow a Poisson distribution. As a result, this method tends to overestimate the standard errors when applied to binary data (McNutt et al., 2003) which can be corrected using robust variance estimation (Barros and Hirakata, 2003; Zou, 2004; Carter et al., 2005). This method has the advantage of avoiding the convergence problems

of log binomial regression (Zou, 2004) but invalid predicted probabilities can occur (Blizzard and Hosmer, 2006)

Lumley et al. (2006) noted that a log normal regression model could be used to estimate relative risks, which corresponds to fitting model (1) and assuming a normal error distribution. As for expanded logistic regression and log Poisson regression, invalid predicted probabilities are possible. This method will produce consistent estimates of relative risk but standard errors may be over- or underestimated and hence robust variance estimation was suggested.

The logistic regression model is a GLM with a logit link and a binomial error distribution, commonly described as $\log\{\pi_i/(1-\pi_i)\} = \mathbf{X}_i\boldsymbol{\beta}$. The predicted probability of success for subject i, given by $\hat{\pi}_i = 1/\{1+\exp(-\mathbf{X}_i\hat{\boldsymbol{\beta}})\}$, will lie between 0 and 1 for any $\mathbf{X}_i$ and hence there are no restrictions on the parameter space for this model (Wacholder, 1986). As a result, logistic regression does not suffer from the same convergence problems as log binomial regression. The calculation of relative risks from logistic regression is more complex however, since logistic regression models assume a constant odds ratio, rather than a constant relative risk. Flanders and Rhodes (1987) presented formulae for calculating standardized relative risks using parameter estimates obtained from fitting a logistic regression model. If conditional standardization (CS) is used, the estimated relative risk comparing treatment to control is given by the ratio of the predicted probabilities for treatment and control, calculated conditional on user-specified reference values $X_2^*,\ldots,X_K^*$ for the baseline covariates, i.e.

$$\frac{1/[1+\exp\{-(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2X_2^*+\ldots+\hat{\beta}_KX_K^*)\}]}{1/[1+\exp\{-(\hat{\beta}_0+\hat{\beta}_2X_2^*+\ldots+\hat{\beta}_KX_K^*)\}]}.$$

Alternatively, if marginal standardization (MS) is used, the estimated relative risk is the ratio of the average predicted probability over a standard population assigning all N subjects to the treatment group, to the average predicted probability over the same population treating all N subjects as controls, i.e.

$$\frac{\frac{1}{N}\sum_{i=1}^{N}1/[1+\exp\{-(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2X_{2i}+\ldots+\hat{\beta}_KX_{Ki})\}]}{\frac{1}{N}\sum_{i=1}^{N}1/[1+\exp\{-(\hat{\beta}_0+\hat{\beta}_2X_{2i}+\ldots+\hat{\beta}_KX_{Ki})\}]}.$$

Although CS and MS both produce estimates of relative risk, it is important to note that they estimate different relative risks with different

interpretations. CS estimates the relative risk conditional on the chosen reference values which may not apply to other covariate patterns. MS estimates the average relative risk in the standard population which may differ for other populations. In practice, the choice between CS and MS depends on the question of interest (Localio et al., 2007).

Once the relative risk has been estimated using CS or MS, confidence intervals can be obtained using the delta method or bootstrapping (Localio et al., 2007; Kleinman and Norton, 2009). This results in four methods based on fitting a logistic regression model, referred to hereafter as CS + delta method, CS + bootstrap, MS + delta method and MS + bootstrap.

2.4 Simulations

The relative performance of the ten methods for estimating relative risks was studied by simulation. We considered a single binary (144 scenarios), a single continuous (144 scenarios) or both a binary and continuous baseline covariate (108 scenarios). For each scenario, 1000 datasets were generated containing 200 or 500 subjects. Treatment allocation was performed using blocked randomization, either overall or within strata defined by a baseline covariate. Four strata defined by three cut points ($\mu-\sigma$, μ and $\mu+\sigma$) were used in the case of stratification based on a continuous covariate with mean μ and standard deviation σ. In the two covariate simulations, stratification was based on either the binary or continuous covariate but not both simultaneously. Subjects were randomly assigned an outcome with probability $\pi_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$, where X_{1i}, X_{2i} and X_{3i} (where applicable) were the treatment group indicator and the baseline covariate(s) respectively. If π_i exceeded 1 for given values of the treatment indicator and baseline covariate(s), a new value was generated for the continuous covariate until $\pi_i < 1$, effectively truncating the distribution of the continuous covariate. New values rarely had to be generated in practice and, in the most extreme case, the covariate was truncated at approximately two standard deviations above the mean.

For the single covariate scenarios, the prevalence of the binary covariate was 0.5 or 0.75 and the continuous covariate was normally distributed with mean 0.5 and variance 0.05 or 0.25. This choice of mean combined with the larger variance allows direct comparison with results obtained using a binary covariate with prevalence 0.5, since the mean and variance of the binary covariate are also 0.5 and 0.25 respectively in this case. The baseline risk was 0.1 ($\beta_0 = -2.30$) and the following combinations of treatment and covariate effects were considered: a treatment relative risk of 1 ($\beta_1 = 0$) with a covariate relative risk of 1 or 2 ($\beta_2 = 0$

or 0.69); a treatment relative risk of 1.25 ($\beta_1 = 0.22$) with a covariate relative risk of 1, 1.25 or 2 ($\beta_2 = 0$, 0.22 or 0.69); and a treatment relative risk of 2 ($\beta_1 = 0.69$) with a covariate relative risk of 1, 1.25, 2 or 3 ($\beta_2 = 0$, 0.22, 0.69 or 1.10).

For the two covariate scenarios, the prevalence of the binary covariate was 0.5 and the continuous covariate was normally distributed with mean 0.5 and variance 0.25. The baseline risk was 0.1 ($\beta_0 = -2.30$) and the treatment and covariate relative risks were all 1 or 2 (i.e. $\beta_k = 0$ or 0.69 for $k = 1, 2, 3$), with at least one of the covariates having an effect for all scenarios.

For each simulated dataset, analyses were performed using each of the ten methods described in Sections 2.2 and 2.3. For the single covariate simulations, both unadjusted and adjusted analyses were performed. For the two covariate simulations, adjustment was made for one or both of the covariates. All analyses were performed using SAS (Cary, NC, USA) version 9.1.3 or later. Log binomial regression, the COPY 1000 method, expanded logistic regression, log Poisson regression, log normal regression and logistic regression were performed using PROC GENMOD. Robust variance estimates were obtained for expanded logistic regression, log Poisson regression and log normal regression using the REPEATED statement (Zou, 2004; Spiegelman and Hertzmark, 2005). Constrained log binomial regression was performed using a SAS macro referred to in Carter et al. (2005) and obtained from the first author. Sample means of the baseline covariate(s) were used as reference values for CS and the simulated dataset was used as the standard population for MS. The delta method was implemented using PROC IML and confidence limits were based on a logarithmic transformation (Flanders and Rhodes, 1987). Bootstrap confidence intervals were obtained using the bias corrected and accelerated method which allows asymmetric confidence intervals (Carpenter and Bithell, 2000). This approach used SAS macros contained in the file jack-boot.sas (available from http://support.sas.com/kb/24/982.html) with 2000 bootstrap samples.

The methods for estimating relative risks were compared based on the following properties, determined for each simulation scenario: convergence rate, calculated as the percentage of simulated datasets where the model converged; type I error rate, calculated as the percentage of Wald tests which resulted in a rejection of the null hypothesis of no treatment effect at the 5% level when the null hypothesis was true; power, calculated as the percentage of Wald tests which resulted in a rejection of the null hypothesis of no treatment effect at the 5% level when the null hypothesis was false; median percent relative bias in the estimated relative risk, where relative bias was calculated as the estimated relative risk minus the true relative risk, divided by the true relative risk; median width of the 95% confidence interval for the relative risk as a measure of precision; coverage

rate, calculated as the percentage of 95% confidence intervals containing the true relative risk; the percentage of simulated datasets where the conclusion about whether treatment had an effect on the outcome (based on whether the p-value from a Wald test was less than 0.05) differed across methods; and the maximum prevalence of invalid predicted probabilities (expanded logistic regression, log Poisson regression and log normal regression only). Changes in these properties resulting from adjusting for an irrelevant covariate (relative risk=1) or failing to adjust for an important covariate (relative risk>1) were compared between methods. Differences in these properties for a binary and continuous covariate were also compared between methods in the subset of simulation scenarios involving a single binary covariate with prevalence 0.5, or a single continuous covariate with mean 0.5 and variance 0.25. Where a method failed to converge for a particular dataset, results from that dataset were excluded for that method when making comparisons.

Given the large number of simulation scenarios considered (396 in total), it is not possible to present the full results of the simulation study in this article (full results are available from the first author on request). Instead, we summarize results across all simulation scenarios by calculating the percentage of scenarios where convergence problems and invalid predicted probabilities occurred, where the (equal) minimum and maximum power and confidence interval width were attained, and where the type I error and coverage rates differed significantly from the nominal level. For 1000 simulated datasets, type I error rates less than 3.6% or greater than 6.4% differ significantly ($p<0.05$) from the nominal level of 5% based on a normal approximation test for a proportion. Similarly, coverage rates less than 93.6% or greater than 96.4% differ significantly from the nominal level of 95%. To quantify the magnitude of any problems that were observed, the median and range were calculated for the type I error rate, median percent relative bias and coverage rate across all scenarios. The median and range were also calculated for the convergence rate across scenarios where convergence problems occurred, and for the maximum prevalence of invalid predicted probabilities across scenarios where invalid predicted probabilities occurred. To determine whether differences between methods were important or negligible, the difference between the best and worst method was calculated for power, median percent relative bias and median confidence interval width for each scenario; the median and range of these differences was then determined across all scenarios. Results are presented both overall and in two subsets defined by whether log binomial regression converged. In the former subset, all simulation scenarios were included but results from any simulated dataset where log binomial regression failed to converge were excluded. In the latter subset, only scenarios where log binomial regression failed to converge for a minimum of 50 simulated datasets were included to avoid conclusions being drawn based on very small samples.

2.5 Sensitivity analyses

Limited additional simulations were conducted to determine the sensitivity of the simulation results to the particular scenarios considered. Firstly, the number of covariates was increased, since consideration of more than one or two covariates is common in practice. Secondly, the simulation model was changed to determine how the methods perform when the true link function is the logit, probit or complementary log-log link, rather than the log link. Additional simulations were based on 500 subjects with treatment allocation performed using blocked randomization. Adjusted analyses were performed with adjustment for all covariates and for each covariate individually.

For simulations with an increased number of covariates, datasets were generated with four covariates: a binary covariate with prevalence 0.5, a binary covariate with prevalence 0.75, a normally distributed covariate with mean 0.5 and variance 0.25, and a normally distributed covariate with mean 0.5 and variance 0.05. The baseline risk was 0.1 ($\beta_0 = -2.30$), the treatment relative risk was 1 or 2 ($\beta_1 = 0$ or 0.69), and the relative risk was 1.25 for three covariates and 2 for the remaining covariate (i.e. $\beta_k = 0.22$ or 0.69 for $k = 2,\ldots,5$).

For simulation models with a non-log link, both CS and MS were used to determine the true relative risk for comparison with the estimated relative risks. The reference values for CS were the population means of the baseline covariates. The standard population for MS was based on the covariate values for all simulated datasets combined. This approach produced a standard population containing 500,000 subjects for each simulation scenario. The covariate distributions in such a large dataset are expected to agree closely with the covariate distributions used to generate each simulated dataset. Datasets were generated with a binary covariate with prevalence 0.5 and/or a normally distributed covariate with mean 0.5 and variance 0.25, with at least one of the covariates having an effect for the two covariate scenarios.

For simulations based on the logit link, subjects were randomly assigned an outcome with probability $\pi_i = 1/\{1+\exp(-\mathbf{X}_i\boldsymbol{\beta})\}$, where $\exp(\beta_1)$ is the true odds ratio. The baseline risk was 0.1 ($\beta_0 = -2.20$) and the treatment and covariate odds ratios were all 1 or 2 (i.e. $\beta_k = 0$ or 0.69 for $k = 1,2,3$).

For simulations based on the probit link, outcomes were generated with probability $\pi_i = \Phi(\mathbf{X}_i\boldsymbol{\beta})$, where Φ is the cumulative distribution function for the standard normal distribution. The baseline risk was 0.1 ($\beta_0 = -1.28$), the risk in the treatment group when the covariate(s) were 0 was 0.1 or 0.2 ($\beta_1 = 0$ or 0.44),

and the risk in the control group at the covariate mean was 0.1 or 0.2 (i.e. $\beta_k = 0$ or 0.88 for $k = 2,3$).

For simulations based on the complementary log-log link, outcomes were generated with probability $\pi_i = 1 - \exp\{-\exp(\mathbf{X}_i\boldsymbol{\beta})\}$. The baseline risk was 0.1 ($\beta_0 = -2.25$), the risk in the treatment group when the covariate(s) were 0 was 0.1 or 0.2 ($\beta_1 = 0$ or 0.75), and the risk in the control group at the covariate mean was 0.1 or 0.2 (i.e. $\beta_k = 0$ or 1.50 for $k = 2,3$).

3 Simulation Results

3.1 Convergence

There were no convergence problems when unadjusted analyses were performed. When adjustment was made for the baseline covariate(s), convergence problems were observed for log binomial regression, constrained log binomial regression, the COPY 1000 method and log normal regression but not for other methods. Log binomial regression performed poorly, with convergence problems occurring for 33.3% of scenarios and convergence rates as low as 48.4% for a given scenario (Table 1). Performance was also quite poor for constrained log binomial regression and the COPY 1000 method, with convergence problems occurring in 16.7% and 30.1% of scenarios respectively, although convergence rates did not fall below 88.5% and 85.9% respectively for any scenario. In contrast, log normal regression converged for 99.8% of simulated datasets in two scenarios where the sample size was 200, and for 100% of simulated datasets otherwise. When log binomial regression converged, the COPY 1000 method always converged but constrained log binomial regression and log normal regression sometimes failed to converge.

Table 1: Number (%) of simulation scenarios where convergence problems occurred and median (minimum) convergence rate for these scenarios.

Analysis method [a)]	N (%)	Median (minimum)
Log binomial regression	132 (33.3)	95.9 (48.4)
Constrained log binomial regression	66 (16.7)	98.9 (88.5)
COPY 1000 method	119 (30.1)	98.3 (85.9)
Log normal regression	2 (0.5)	99.8 (99.8)

[a)] Methods that are not listed had convergence rates of 100% for all 396 simulation scenarios.

3.2 Type I error

All methods produced some type I error rates which differed significantly from the nominal level. This mostly occurred when the sample size was 200 and/or the continuous covariate had an effect. Type I error problems were not severe, however, as the smallest overall type I error rate observed was 2.7% and the largest was 8.6% across all methods (Table 2). Larger type I error rates did occur in the subset of simulations where log binomial regression failed to converge, although these results were generally based on relatively few simulated datasets per scenario. Overall, log normal regression showed the best performance, with type I error rates only differing significantly from the nominal level in 11.9% of the 118 scenarios considered where the treatment did not have an effect. Other methods had type I error problems in 14.4% to 18.6% of scenarios. When log binomial regression converged, the type I error rate differed significantly from the nominal level in 17.8% of scenarios and only log normal regression achieved a better result (10.2%). When log binomial regression failed to converge, log Poisson regression had the fewest problems with type I error (26.1%). In contrast, constrained log binomial regression performed poorly, producing type I error problems in 65.2% of scenarios.

3.3 Power

Logistic regression methods performed best in terms of power overall, achieving the (equal) highest power in 69.4% of the 278 scenarios considered where the treatment had an effect (Table 3). The worst performing method was the COPY 1000 method which had (equal) minimum power for 65.1% of scenarios. Differences in power between methods were typically small however; the difference in power between the best and worst method ranged between 0% and 6.5%, with a median of only 0.5%. In the subset of simulations where log binomial regression converged, logistic regression methods again performed best. When log binomial regression failed to converge, the (equal) highest power was most often achieved by constrained log binomial regression.

Table 2: Number (%) of simulation scenarios where the type I error rate differed significantly from the nominal level, and median (range) for the type I error rate, both overall and by convergence status of the log binomial regression model.

	Overall (118 scenarios)		Log binomial converged (118 scenarios)		Log binomial did not converge (23 scenarios)	
Analysis method	N (%)	Median (range)	N (%)	Median (range)	N (%)	Median (range)
Log binomial regression	21 (17.8)	4.3 (2.8, 6.6)	21 (17.8)	4.3 (2.8, 6.6)	N/A	N/A
Constrained log binomial regression	20 (16.9)	4.4 (2.8, 8.6)	21 (17.8)	4.3 (2.8, 6.5)	15 (65.2)	11.7 (3.9, 23.8)
COPY 1000 method	22 (18.6)	4.2 (2.7, 6.4)	26 (22.0)	4.2 (2.7, 6.4)	7 (30.4)	8.7 (0.0, 22.2)
Expanded logistic regression	17 (14.4)	4.4 (3.0, 6.2)	25 (21.2)	4.3 (2.8, 6.0)	7 (30.4)	7.5 (3.1, 13.7)
Log Poisson regression	17 (14.4)	4.3 (3.0, 6.0)	21 (17.8)	4.3 (2.8, 6.0)	6 (26.1)	7.4 (2.0, 11.8)
Log normal regression	14 (11.9)	4.7 (2.9, 7.8)	12 (10.2)	4.6 (2.9, 8.0)	7 (30.4)	8.1 (3.1, 12.5)
Logistic regression methods [a)]	17 (14.4)	4.5 (3.2, 6.4)	22 (18.6)	4.5 (3.1, 6.4)	9 (39.1)	8.1 (2.3, 12.8)

Table 3: Number (%) of simulation scenarios where (equal) minimum and maximum values were attained for power, both overall and by convergence status of the log binomial regression model.

	Overall (278 scenarios)		Log binomial converged (278 scenarios)		Log binomial did not converge (37 scenarios)	
	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Analysis method	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Log binomial regression	114 (41.0)	59 (21.2)	101 (36.3)	68 (24.5)	N/A	N/A
Constrained log binomial regression	98 (35.3)	84 (30.2)	104 (37.4)	74 (26.6)	10 (27.0)	32 (86.5)
COPY 1000 method	181 (65.1)	42 (15.1)	188 (67.6)	44 (15.8)	16 (43.2)	14 (37.8)
Expanded logistic regression	126 (45.3)	58 (20.9)	134 (48.2)	60 (21.6)	21 (56.8)	11 (29.7)
Log Poisson regression	114 (41.0)	66 (23.7)	117 (42.1)	67 (24.1)	11 (29.7)	13 (35.1)
Log normal regression	127 (45.7)	102 (36.7)	135 (48.6)	106 (38.1)	18 (48.6)	12 (32.4)
Logistic regression methods [a)]	39 (14.0)	193 (69.4)	40 (14.4)	199 (71.6)	13 (35.1)	13 (35.1)

[a)] Type I error rate and power are the same for all logistic regression methods.

3.4 Bias

When unadjusted analyses were performed, the median percent relative bias in the estimated relative risk was the same for all methods, except for the COPY 1000 method which generally produced slightly smaller median relative risk estimates. When adjusted analyses were performed, CS methods often substantially overestimated the relative risk, with median percent relative biases as large as 15.1% (Table 4), corresponding to a median bias of 0.3 when the true relative risk was 2. In contrast, the median percent relative bias did not exceed 5.1% for any other method. Underestimation of the relative risk also occurred and the minimum value for the median percent relative bias was around -5% for all methods. Similar results were observed when log binomial regression converged. When it failed to converge, all methods tended to overestimate the true relative risk, especially CS methods. Overall, the difference in the median percent relative bias between the best and worst method was generally small (median 0.6%) but could be substantial, depending on the scenario (range 0% to 17.0%).

Table 4: Median (range) for the median percent relative bias in the estimated relative risk, both overall and by convergence status of the log binomial regression model.

Analysis method	Overall (396 scenarios)	Log binomial converged (396 scenarios)	Log binomial did not converge (60 scenarios)
Log binomial regression	0.0 (-4.7, 3.2)	0.0 (-4.7, 3.2)	N/A
Constrained log binomial regression	0.0 (-4.7, 5.1)	0.0 (-4.7, 3.2)	4.3 (-7.8, 17.3)
COPY 1000 method	-0.2 (-4.9, 2.9)	-0.3 (-4.9, 2.9)	2.1 (-13.6, 23.8)
Expanded logistic regression	0.0 (-4.7, 3.2)	0.0 (-4.7, 3.2)	3.5 (-7.8, 15.8)
Log Poisson regression	0.0 (-4.7, 3.2)	0.0 (-4.7, 3.2)	3.6 (-8.1, 15.8)
Log normal regression	0.0 (-4.7, 4.1)	0.0 (-4.7, 4.2)	2.1 (-6.0, 12.0)
CS methods [a)]	0.0 (-4.7, 15.1)	0.0 (-4.7, 14.7)	9.6 (-7.3, 24.0)
MS methods [a)]	0.0 (-4.7, 3.2)	0.0 (-5.4, 3.2)	2.9 (-6.6, 16.0)

[a)] Bias is the same for CS methods and for MS methods.

3.5 Precision

Precision was generally poor for CS + bootstrap and MS + bootstrap compared to other methods. CS + bootstrap achieved the widest median confidence interval for the vast majority of scenarios, both overall and by convergence status of log binomial regression, while constrained log binomial regression showed the best performance (Table 5). The variability between methods in median confidence

interval width could be large, with differences between the best and worst method of up to 1.05 (median difference 0.15).

3.6 Coverage

The percentage of simulation scenarios where coverage rates differed significantly from the nominal level varied greatly between methods (Table 6). Overall, CS + bootstrap and MS + bootstrap failed to achieve acceptable coverage in 34.3% and 29.0% of simulation scenarios respectively. By comparison, coverage problems occurred for between 6.6% and 13.9% of scenarios for other methods. Observed coverage rates ranged between 85.9% and 97.9%, indicating that undercoverage could be substantial. The COPY 1000 method produced the best results both overall and when log binomial regression converged, only failing to maintain acceptable coverage for 6.6% and 7.1% of scenarios respectively. When log binomial regression failed to converge, the top performing methods were log Poisson regression and MS + bootstrap, each with coverage problems in 13.3% of scenarios. Constrained log binomial regression performed very poorly in this subset, with coverage rates significantly different from the nominal level in 63.3% of scenarios.

3.7 Differing conclusions

The conclusion about the effectiveness of treatment differed depending on the method used in up to 13.6% of simulated datasets for a given scenario. Most inconsistencies occurred for adjusted analyses, especially in the two covariate scenarios when adjustment was made for both covariates. Inconsistencies were also common when treatment had an effect on the outcome. In some cases, the different conclusions resulted from p-values that were marginally smaller or larger than 0.05, however large differences in p-values were also often observed. The largest differences in p-values occurred when adjustment was made for a continuous covariate and treatment had no effect (data not shown). Log binomial regression, constrained log binomial regression and the COPY 1000 method generally led to the same conclusions, as did expanded logistic regression, log Poisson regression and logistic regression methods.

Table 5: Number (%) of simulation scenarios where (equal) minimum and maximum values were attained for median confidence interval width, both overall and by convergence status of the log binomial regression model.

	Overall (396 scenarios)		Log binomial converged (396 scenarios)		Log binomial did not converge (60 scenarios)	
Analysis method	Minimum N (%)	Maximum N (%)	Minimum N (%)	Maximum N (%)	Minimum N (%)	Maximum N (%)
Log binomial regression	3 (0.8)	0 (0.0)	9 (2.3)	0 (0.0)	N/A	N/A
Constrained log binomial regression	224 (56.6)	0 (0.0)	219 (55.3)	0 (0.0)	35 (58.3)	0 (0.0)
COPY 1000 method	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	22 (36.7)	0 (0.0)
Expanded logistic regression	57 (14.4)	0 (0.0)	59 (14.9)	0 (0.0)	0 (0.0)	0 (0.0)
Log Poisson regression	14 (3.5)	0 (0.0)	13 (3.3)	0 (0.0)	1 (1.7)	0 (0.0)
Log normal regression	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.7)	0 (0.0)
CS + delta method	66 (16.7)	0 (0.0)	66 (16.7)	0 (0.0)	0 (0.0)	1 (1.7)
CS + bootstrap	0 (0.0)	319. (80.6)	0 (0.0)	319 (80.6)	0 (0.0)	59 (98.3)
MS + delta method	98 (24.7)	0 (0.0)	96 (24.2)	0 (0.0)	1 (1.7)	0 (0.0)
MS + bootstrap	0 (0.0)	77 (19.4)	0 (0.0)	77 (19.4)	0 (0.0)	0 (0.0)

DOI: 10.2202/1557-4679.1278

Table 6: Number (%) of simulation scenarios where the coverage rate differed significantly from the nominal level, and median (range) for the coverage rate, both overall and by convergence status of the log binomial regression model.

	Overall (396 scenarios)		Log binomial converged (396 scenarios)		Log binomial did not converge (60 scenarios)	
Analysis method	N (%)	Median (range)	N (%)	Median (range)	N (%)	Median (range)
Log binomial regression	42 (10.6)	94.9 (91.1, 96.7)	42 (10.6)	94.9 (91.1, 96.7)	N/A	N/A
Constrained log binomial regression	53 (13.4)	95.5 (85.9, 97.2)	41 (10.4)	95.6 (92.6, 97.2)	38 (63.3)	89.4 (76.2, 98.0)
COPY 1000 method	26 (6.6)	95.1 (92.4, 96.8)	28 (7.1)	95.1 (92.4, 97.8)	13 (21.7)	93.4 (76.9, 100)
Expanded logistic regression	39 (9.8)	95.6 (93.2, 97.0)	48 (12.1)	95.6 (93.2, 97.2)	10 (16.7)	94.0 (86.3, 98.0)
Log Poisson regression	39 (9.8)	95.6 (93.2, 97.0)	38 (9.6)	95.6 (93.2, 97.2)	8 (13.3)	94.2 (88.2, 98.6)
Log normal regression	34 (8.6)	95.3 (92.2, 97.1)	32 (8.1)	95.4 (92.0, 97.5)	11 (18.3)	93.2 (87.5, 98.0)
CS + delta method	55 (13.9)	95.6 (87.8, 97.0)	52 (13.1)	95.6 (87.8, 97.1)	26 (43.3)	92.1 (84.1, 98.0)
CS + bootstrap	136 (34.3)	95.9 (87.5, 97.9)	135 (34.1)	96.0 (87.5, 98.2)	20 (33.3)	93.1 (84.3, 98.0)
MS + delta method	43 (10.9)	95.6 (93.1, 97.0)	43 (10.9)	95.6 (93.1, 97.1)	9 (15.0)	93.9 (88.2, 98.0)
MS + bootstrap	115 (29.0)	96.0 (94.0, 97.7)	114 (28.8)	96.0 (94.0, 97.7)	8 (13.3)	94.9 (88.2, 98.6)

3.8 Invalid predicted probabilities

Invalid predicted probabilities only occurred for expanded logistic regression, log Poisson regression and log normal regression in scenarios where adjustment was made for a continuous covariate. When invalid predicted probabilities did occur they were rare, with a median prevalence of only 0.5% for all methods (Table 7). Log normal regression had a lower maximum prevalence of invalid predicted probabilities across all scenarios compared to expanded logistic regression and log Poisson regression (4.5% vs. 6.5% and 6.0% respectively). However, log normal regression produced invalid predicted probabilities in more scenarios (25.5%), compared to expanded logistic regression (21.2%) and log Poisson regression (20.2%). The maximum prevalence of invalid predicted probabilities was the same or lower for log Poisson regression compared to expanded logistic regression in all but one scenario.

Table 7: Number (%) of simulation scenarios where invalid predicted probabilities occurred and median (maximum) prevalence of invalid predicted probabilities for these scenarios.

Analysis method [a)]	N (%)	Median (maximum)
Expanded logistic regression	84 (21.2)	0.5 (6.5)
Log Poisson regression	80 (20.2)	0.5 (6.0)
Log normal regression	101 (25.5)	0.5 (4.5)

[a)] Methods that are not listed produced no invalid predicted probabilities for all 396 simulation scenarios.

3.9 Adjusting for an irrelevant covariate

The impact of adjusting for a covariate that had no effect on the outcome varied between methods. More convergence problems occurred when adjustment was made for an irrelevant covariate, particularly for log binomial regression and the COPY 1000 method. The latter method also suffered from more problems with type I error as a result of adjustment, while other methods were largely unaffected. Adjustment resulted in small changes in power, ranging from -3.3% to 1.8%. The median percent relative bias in the estimated relative risk often increased slightly following adjustment and the largest increases were seen for log normal regression. This method also tended to produce wider confidence intervals for adjusted analyses, along with CS methods. In contrast, confidence intervals based on adjustment were often narrower for other methods, although the maximum change in the median confidence interval width in either direction was only 0.13. Coverage problems were more common for log binomial regression

and CS + bootstrap when adjusting for an irrelevant covariate but other methods were largely unaffected.

3.10 Failing to adjust for an important covariate

Failing to adjust for a covariate that had an effect on the outcome had the advantage of avoiding many of the convergence problems observed in the adjusted analyses. A disadvantage of failing to adjust was the resulting increase in type I error problems for all methods except the COPY 1000 method. Further, lack of adjustment often resulted in power reductions which could be substantial (up to 13.7%), although small increases in power also occurred. There was generally little change in the median percent relative bias when no adjustment was made, however large reductions sometimes occurred for CS methods. Narrower confidence intervals were common for log normal regression and CS methods based on unadjusted analyses, while confidence intervals generally became wider for other methods. Unadjusted analyses resulted in more coverage problems for expanded logistic regression, log Poisson regression, log normal regression and MS + delta method but fewer coverage problems for other methods.

3.11 Binary vs. continuous covariate

There were differences in results depending on whether a binary or continuous covariate was considered. For the continuous covariate, convergence problems were observed for log binomial regression, constrained log binomial regression, the COPY 1000 method and log normal regression, while no convergence problems occurred for the binary covariate. The continuous covariate was also associated with more coverage problems. Power, median percent relative bias and median confidence interval width could be smaller or larger for the continuous covariate compared to the binary covariate and this held for all methods.

3.12 Sensitivity analyses

Increasing the number of covariates in the simulation model led to more problems with convergence, especially for log binomial regression. This method suffered from convergence problems in 16 (40.0%) scenarios and convergence rates fell as low as 39.9% for these scenarios. Type I error and coverage problems were less common compared to simulations with fewer covariates. Type I error rates ranged between 3.2% and 6.2%, while coverage rates ranged between 92.0% and 97.3%. Biases also became less severe and the median percent relative bias did not fall below -1.7% or exceed 9.0%. MS + bootstrap showed improved performance in terms of precision and coverage and was no longer identified as a poor performer

for these properties. Increasing the number of covariates in the simulation model had little impact on the results otherwise (data not shown).

Simulating datasets using a logit, probit or complementary log-log link generally produced results similar to those obtained using a log link. Convergence problems still occurred for log binomial regression, constrained log binomial regression and the COPY 1000 method. Type I error rates were again close to the nominal level, except for constrained log binomial regression and the COPY 1000 method when the probit link or complementary log-log link was used. Maximum type I error rates for these methods were 70.9% and 73.6% respectively, compared to the nominal level of 5% (Table 8). Logistic regression was still the most powerful method for the majority of simulation scenarios. Confidence intervals remained widest for bootstrapping methods, particularly when combined with CS.

When MS was used to determine the true relative risk, CS methods tended to overestimate the relative risk, as seen for the log link simulations. Bias remained small for other methods when the data were simulated using a logit link, but underestimation of the relative risk could be substantial for log binomial regression and the COPY 1000 method using a probit or complementary log-log link (Table 9). Coverage rates were generally acceptable for the logit link, ranging from 93.7% to 97.3%. In contrast, coverage rates were often too low for the probit and complementary log-log link, although expanded logistic regression, log Poisson regression and MS + delta method generally performed well and coverage rates did not fall below 93.0% for these methods (Table 10).

When CS was used to determine the true relative risk, CS methods performed well in terms of bias, while other methods tended to underestimate the relative risk. Biases were fairly small based on the logit link but could be substantial based on the probit or complementary log-log link (Table 9). Coverage rates remained fairly close to the nominal level for the logit link, ranging from 92.5% to 97.3%. For the probit and complementary log-log link however, undercoverage was common for all methods and coverage rates dropped as low as 6.5% (Table 10).

4 Example

To illustrate the different methods for estimating relative risks, we consider a RCT of fish oil supplements (a source of omega-3 fatty acids) versus placebo for preterm infants (Makrides et al., 2009). Mothers of infants born less than 33 weeks gestation were recruited from five Australian hospitals between 2001 and 2005. Treatment assignment was performed using blocked randomization within strata defined by infant birth weight ($<$1.25kg or $\geq$1.25kg), gender and recruiting hospital.

Table 8: Number (%) of simulation scenarios where the type I error rate differed significantly from the nominal level, and median (range) for the type I error rate when the data were simulated using a logit, probit or complementary log-log link.

	Logit link (17 scenarios)		Probit link (17 scenarios)		Complementary log-log link (17 scenarios)	
Analysis method	N (%)	Median (range)	N (%)	Median (range)	N (%)	Median (range)
Log binomial regression	2 (11.8)	4.8 (3.2, 5.8)	0 (0.0)	5.0 (4.0, 5.9)	0 (0.0)	4.9 (3.6, 7.9)
Constrained log binomial regression	2 (11.8)	4.8 (3.2, 5.8)	5 (29.4)	5.4 (4.0, 24.6)	5 (29.4)	5.4 (4.0, 70.9)
COPY 1000 method	2 (11.8)	4.6 (3.2, 5.8)	2 (11.8)	5.2 (4.0, 19.9)	5 (29.4)	5.2 (4.0, 73.6)
Expanded logistic regression	2 (11.8)	4.8 (3.2, 5.6)	0 (0.0)	5.1 (4.0, 5.8)	0 (0.0)	4.8 (4.0, 5.9)
Log Poisson regression	2 (11.8)	4.7 (3.2, 5.8)	0 (0.0)	5.1 (4.0, 5.8)	0 (0.0)	4.8 (4.0, 5.8)
Log normal regression	1 (5.9)	4.8 (3.2, 6.3)	0 (0.0)	5.6 (4.0, 6.0)	0 (0.0)	5.0 (4.0, 6.2)
Logistic regression methods [a)]	1 (5.9)	4.9 (3.2, 6.1)	0 (0.0)	5.3 (4.1, 5.9)	0 (0.0)	5.0 (4.2, 6.0)

Table 9: Median (range) for the median percent relative bias in the estimated relative risk when the data were simulated using a logit, probit or complementary log-log link and the true relative risk was determined using MS or CS.

	Logit link (34 scenarios)		Probit link (34 scenarios)		Complementary log-log link (34 scenarios)	
Analysis method	MS	CS	MS	CS	MS	CS
Log binomial regression	-0.1 (-1.6, 1.2)	-0.6 (-3.6, 1.2)	-0.3 (-8.5, 1.2)	-0.5 (-14.7, 1.2)	-0.2 (-10.0, 1.2)	-0.9 (-23.4, 1.2)
Constrained log binomial regression	-0.1 (-1.5, 1.2)	-0.6 (-3.4, 1.2)	-0.3 (-4.8, 1.2)	-0.5 (-11.3, 1.2)	0.0 (-4.6, 6.1)	-0.6 (-18.7, 1.2)
COPY 1000 method	-0.3 (-1.9, 1.1)	-0.6 (-3.6, 1.1)	-0.4 (-9.4, 1.1)	-0.5 (-15.6, 1.1)	-0.1 (-15.1, 1.1)	-0.4 (-27.7, 1.1)
Expanded logistic regression	0.0 (-0.8, 1.0)	-0.6 (-2.5, 1.0)	0.0 (-0.8, 2.3)	-0.4 (-6.5, 1.0)	0.1 (-1.6, 4.1)	-0.2 (-13.4, 1.0)
Log Poisson regression	0.0 (-1.0, 1.0)	-0.5 (-2.8, 1.0)	0.0 (-0.8, 1.0)	-0.5 (-6.8, 1.0)	0.0 (-1.6, 1.0)	-0.2 (-14.9, 1.0)
Log normal regression	-0.3 (-1.7, 1.8)	-0.6 (-4.2, 1.8)	-0.3 (-5.3, 1.8)	-0.7 (-11.0, 1.8)	-0.2 (-6.7, 1.8)	-0.4 (-20.6, 1.8)
CS methods [a)]	0.1 (-0.8, 4.2)	-0.1 (-2.2, 1.2)	0.0 (-0.8, 9.2)	-0.1 (-6.4, 1.8)	0.1 (-1.6, 17.8)	0.0 (-9.7, 5.8)
MS methods [a)]	0.0 (-1.0, 1.0)	-0.5 (-2.8, 1.0)	-0.1 (-0.8, 1.0)	-0.5 (-6.8, 1.0)	0.0 (-1.6, 1.0)	-0.2 (-14.9, 1.0)

[a)] Type I error rate is the same for all logistic regression methods. Bias is the same for CS methods and for MS methods.

Table 10: Number (%) of simulation scenarios where the coverage rate differed significantly from the nominal level, and median (range) for the coverage rate when the data were simulated using a logit, probit or complementary log-log link and the true relative risk was determined using (a) MS or (b) CS.

	Logit link (34 scenarios)		Probit link (34 scenarios)		Complementary log-log link (34 scenarios)	
Analysis method	N (%)	Median (range)	N (%)	Median (range)	N (%)	Median (range)
(a)						
Log binomial regression	2 (5.9)	94.8 (93.8, 96.6)	8 (23.5)	94.6 (85.6, 96.2)	7 (20.6)	94.6 (81.5, 95.6)
Constrained log binomial regression	2 (5.9)	95.3 (94.0, 96.8)	13 (38.2)	94.9 (70.5, 96.5)	13 (38.2)	94.6 (25.1, 96.4)
COPY 1000 method	2 (5.9)	95.0 (93.9, 96.6)	6 (17.6)	94.8 (81.1, 96.2)	9 (26.5)	94.8 (53.1, 95.7)
Expanded logistic regression	2 (5.9)	95.3 (94.1, 96.8)	2 (5.9)	95.2 (94.1, 96.5)	2 (5.9)	95.3 (93.0, 96.4)
Log Poisson regression	2 (5.9)	95.4 (94.2, 96.8)	2 (5.9)	95.3 (94.2, 96.5)	2 (5.9)	95.3 (93.5, 96.4)
Log normal regression	2 (5.9)	95.3 (93.7, 96.8)	7 (20.6)	94.4 (90.5, 96.5)	7 (20.6)	94.9 (75.2, 96.4)
CS + delta method	2 (5.9)	95.2 (94.2, 96.8)	5 (14.7)	94.8 (90.0, 96.5)	10 (29.4)	94.9 (69.8, 96.4)
CS + bootstrap	2 (5.9)	95.5 (94.3, 97.1)	3 (8.8)	95.1 (89.8, 96.5)	10 (29.4)	95.2 (69.4, 96.5)
MS + delta method	2 (5.9)	95.3 (94.2, 96.8)	2 (5.9)	95.2 (94.2, 96.5)	1 (2.9)	95.2 (93.7, 96.4)
MS + bootstrap	4 (11.8)	95.5 (94.4, 97.3)	4 (11.8)	95.7 (94.4, 96.8)	9 (26.5)	95.8 (94.3, 97.1)
(b)						
Log binomial regression	5 (14.7)	94.9 (92.5, 96.6)	14 (41.2)	93.9 (59.6, 95.7)	16 (47.1)	93.6 (25.9, 95.6)
Constrained log binomial regression	5 (14.7)	95.3 (92.8, 96.8)	19 (55.9)	93.3 (52.8, 96.4)	19 (55.9)	92.2 (17.8, 96.4)
COPY 1000 method	5 (14.7)	95.1 (93.0, 96.6)	13 (38.2)	94.2 (58.9, 95.7)	16 (47.1)	94.2 (7.2, 95.7)
Expanded logistic regression	2 (5.9)	95.2 (94.1, 96.8)	11 (32.4)	94.7 (90.1, 96.4)	14 (41.2)	94.8 (58.7, 96.4)
Log Poisson regression	2 (5.9)	95.3 (93.7, 96.8)	14 (41.2)	94.5 (87.8, 96.4)	14 (41.2)	94.7 (44.8, 96.4)
Log normal regression	4 (11.8)	95.3 (92.6, 96.8)	14 (41.2)	94.2 (71.2, 96.4)	14 (41.2)	94.4 (6.5, 96.4)
CS + delta method	2 (5.9)	95.3 (94.2, 96.8)	7 (20.6)	95.1 (92.0, 96.7)	7 (20.6)	94.9 (79.0, 96.4)
CS + bootstrap	2 (5.9)	95.6 (94.3, 97.1)	4 (11.8)	95.3 (92.5, 96.7)	7 (20.6)	95.1 (80.8, 96.5)
MS + delta method	2 (5.9)	95.2 (93.8, 96.8)	14 (41.2)	94.5 (88.0, 96.4)	14 (41.2)	94.6 (41.0, 96.4)
MS + bootstrap	3 (8.8)	95.5 (94.2, 97.3)	11 (32.4)	94.9 (89.5, 96.8)	18 (52.9)	95.5 (51.7, 96.8)

DOI: 10.2202/1557-4679.1278

We consider two binary outcomes of interest: whether the infant was discharged home from hospital on oxygen and whether the infant had a significant mental delay at 18 months (defined by a Mental Development Index score <70 using the Bayley Scales of Infant Development, Second Edition (Bayley, 1993)). These outcomes were analyzed using each of the ten methods for estimating relative risks described in Sections 2.2 and 2.3. Adjustment was made for one or both of the stratification variables, infant birth weight (treated as continuous) and gender. Stratification by recruiting hospital was performed for administrative purposes only and was therefore not considered in the analysis.

Since the focus of the current article is on methods for estimating relative risks based on independent data, one infant per mother was randomly selected for inclusion in the analysis to remove clustering due to multiple pregnancies. Results presented here are therefore purely for illustrative purposes and should not be used to draw conclusions about the effectiveness of fish oil supplements for preterm infants. In the subset used in this analysis, 54% of the 540 infants were male and infant birth weight was approximately normally distributed with a mean of 1.3kg and a variance of 0.18kg.

4.1 Discharged home on oxygen

The percentage of infants discharged home on oxygen was 9.36% and 10.41% for the treatment and control group respectively, producing an unadjusted relative risk (95% confidence interval (CI)) of 0.90 (0.54, 1.50). There was no evidence to suggest that gender had any effect on whether the infant was discharged home on oxygen, however the risk of being discharged home on oxygen significantly decreased as birth weight increased. This example is therefore similar to the two covariate simulation scenarios where treatment and the binary covariate had no effect on the outcome, while the continuous covariate had an effect.

Adjusted relative risks varied depending on the method and the covariate(s) included in the model, ranging between 0.84 and 0.93 (Table 11). When adjustment was made for birth weight, log binomial regression and the COPY 1000 method failed to converge until a variety of starting values were provided. These methods had the most convergence problems in the simulation study. CS + bootstrap produced the widest confidence interval when adjustment was made for birth weight and the second widest when adjustment was made for gender only. This is consistent with the simulation results which showed that CS + bootstrap had poor precision. When adjustment was made for gender only, confidence intervals became narrower for log normal regression and CS methods but wider otherwise. This same pattern was observed in the simulation study for the scenarios where no adjustment was made for an important covariate (birth weight in this case). Despite differences between methods, the conclusion was the

same in each case: there was no evidence to suggest that treatment has an effect on the risk of being discharged home on oxygen.

Table 11: Relative risk (95% CI) for treatment from analysis of whether infant was discharged home on oxygen.

Analysis method	Adjusted for gender and birth weight	Adjusted for gender only	Adjusted for birth weight only
Log binomial regression	0.88 (0.56, 1.41)	0.90 (0.53, 1.50)	0.91 (0.57, 1.45)
Constrained log binomial regression	0.88 (0.54, 1.44)	0.90 (0.54, 1.49)	0.91 (0.56, 1.47)
COPY 1000 method	0.89 (0.56, 1.41)	0.90 (0.53, 1.49)	0.91 (0.57, 1.44)
Expanded logistic regression	0.87 (0.53, 1.42)	0.90 (0.54, 1.50)	0.85 (0.52, 1.40)
Log Poisson regression	0.87 (0.53, 1.41)	0.90 (0.54, 1.50)	0.87 (0.53, 1.40)
Log normal regression	0.90 (0.55, 1.50)	0.86 (0.52, 1.44)	0.93 (0.56, 1.55)
CS + delta method	0.84 (0.47, 1.51)	0.90 (0.54, 1.50)	0.85 (0.47, 1.52)
CS + bootstrap	0.84 (0.43, 1.55)	0.90 (0.51, 1.51)	0.85 (0.45, 1.55)
MS + delta method	0.87 (0.54, 1.40)	0.90 (0.54, 1.50)	0.87 (0.54, 1.40)
MS + bootstrap	0.87 (0.50, 1.48)	0.90 (0.50, 1.54)	0.87 (0.52, 1.49)

4.2 Significant mental delay

In the treatment group, 4.03% of infants showed a significant mental delay compared to 9.49% of control group infants, resulting in an unadjusted relative risk (95% CI) of 0.43 (0.21, 0.87). There was no evidence to suggest that either gender or birth weight had any effect on the outcome. In this example, the risk was twice as high in one group compared to the other, as was the case with many of the simulation scenarios considered, however here the treatment reduced the risk, rather than increasing it.

Adjusted relative risks ranged between 0.42 and 0.48, depending on the method and the covariate(s) included in the model (Table 12). The difference in estimates between CS and other methods that were often seen in the simulation study were not observed in this example. This may be due to the fact that neither of the covariates appeared to be having an effect on the outcome. Log normal regression and MS + bootstrap produced the widest confidence intervals, followed by CS + bootstrap. Wider confidence intervals for the bootstrapping methods were expected based on the results of the simulation study. The conclusion was the same for all methods, with strong evidence to suggest that treatment reduces the risk of significant mental delay. The upper limit of the 95% confidence interval was very close to one for log normal regression however, and a slightly smaller treatment effect may have resulted in different conclusions being drawn, depending on the method.

Table 12: Relative risk (95% CI) for treatment from analysis of whether infant had a significant mental delay.

Analysis method	Adjusted for gender and birth weight	Adjusted for gender only	Adjusted for birth weight only
Log binomial regression	0.43 (0.20, 0.86)	0.43 (0.20, 0.85)	0.43 (0.20, 0.84)
Constrained log binomial regression	0.43 (0.21, 0.88)	0.43 (0.21, 0.88)	0.43 (0.21, 0.87)
COPY 1000 method	0.44 (0.21, 0.86)	0.44 (0.20, 0.86)	0.43 (0.20, 0.85)
Expanded logistic regression	0.42 (0.21, 0.85)	0.42 (0.21, 0.86)	0.42 (0.21, 0.86)
Log Poisson regression	0.43 (0.21, 0.87)	0.43 (0.21, 0.87)	0.42 (0.21, 0.87)
Log normal regression	0.48 (0.23, 0.98)	0.47 (0.23, 0.96)	0.44 (0.21, 0.90)
CS + delta method	0.42 (0.21, 0.87)	0.42 (0.21, 0.87)	0.42 (0.21, 0.87)
CS + bootstrap	0.42 (0.19, 0.87)	0.42 (0.19, 0.87)	0.42 (0.19, 0.87)
MS + delta method	0.43 (0.21, 0.87)	0.43 (0.21, 0.87)	0.42 (0.21, 0.87)
MS + bootstrap	0.43 (0.18, 0.89)	0.43 (0.18, 0.90)	0.42 (0.18, 0.88)

5 Discussion

We have compared ten different methods for estimating relative risks from independent observations in a RCT setting. The simulation results indicate that there is variability between the methods and that the best method to use depends on the statistical property considered. There was often little difference between the best method and several competing methods, suggesting that a number of methods may be reasonable for use in practice.

Few comparisons have previously been made between methods for estimating relative risks. One study reported improved convergence rates for constrained log binomial regression compared to log binomial regression but found that convergence was still poor in some settings (Carter et al., 2005), consistent with the results of this study. Previous research also shows that expanded logistic regression (Skov et al., 1998) and log Poisson regression (Barros and Hirakata, 2003; Zou, 2004; Carter et al., 2005; Petersen and Deddens, 2008) often perform similarly to log binomial regression when there are no convergence problems, as confirmed by the current study. However, there is evidence to suggest that log Poisson regression has several advantages over expanded logistic regression; parameter estimates from log Poisson regression had greater asymptotic efficiency (Lumley et al., 2006), produced invalid predicted probabilities less often and generally had smaller bias and mean squared error (Blizzard and Hosmer, 2006). These findings are consistent with the results of the current study, where the median bias and maximum prevalence of invalid predicted probabilities were often slightly smaller for log Poisson regression compared to expanded logistic regression.

Although adjustment is often made for baseline covariates in RCTs, unadjusted analyses do have several advantages. Firstly, we found no convergence problems for unadjusted analyses, suggesting there may be no need to consider alternatives to log binomial regression when only unadjusted analyses are planned. Secondly, bias actually decreased for CS methods as a result of failing to adjust for an important covariate and was largely unaffected for other methods. This is consistent with asymptotic theory which indicates that unadjusted estimates of treatment effect obtained from a GLM will be unbiased for identity and log links but potentially biased for other links (Gail et al., 1984). Adjusting for an important covariate generally resulted in narrower confidence intervals, however. Adjusted analyses may therefore be preferred to unadjusted analyses in practice. A potential risk associated with performing an adjusted analysis is that one or more of the pre-specified baseline covariates may not actually have an effect on the outcome. We found little cost associated with adjusting for an irrelevant covariate for expanded logistic regression, log Poisson regression and MS methods. In contrast, other methods suffered from more problems with convergence (log binomial regression, constrained log binomial regression and the COPY 1000 method), type I error (COPY 1000 method) and coverage (log binomial regression and CS + bootstrap), as well as reduced precision (log normal regression and CS methods) when an irrelevant covariate was included in the model.

Since the data were mostly simulated from a GLM with a log link, logistic regression methods may have been at a disadvantage compared to other methods due to use of an incorrect link function. Despite this, MS still performed relatively well in these scenarios, particularly when combined with the delta method. In contrast, CS performed poorly with median biases as large as 0.3 when the true relative risk was 2. Use of the CS method in practice could result in an important overstatement of the effect of an intervention if the log link is correct. When the logit link was correct however, both CS and MS generally performed well. If the data are truly logistic, then logistic regression methods are useful as they allow the results to be presented in a way that is easy to interpret while fitting a model with the appropriate form for the data. A similar approach may also be used to estimate relative risks based on other link functions (Cummings, 2009; Penman and Johnson, 2009). Interestingly, several of the methods that assume a log link applies (expanded logistic regression, log Poisson regression and log normal regression) performed well whether the data were simulated from a GLM with a log, logit, probit or complementary log-log link, provided the true relative risk was determined using MS. This suggests that these methods may be reasonable to use even when the log link is incorrect, provided the marginal relative risk is of interest.

Invalid predicted probabilities were occasionally produced by expanded logistic regression, log Poisson regression and log normal regression. The importance of producing valid predicted probabilities is a matter of debate (Blizzard and Hosmer, 2006; Lumley et al., 2006). We would argue that the decision between models that do or do not allow invalid predicted probabilities depends on the context. If the purpose of the model is to predict risk for individuals based on their covariate values, then use of models that allow invalid predicted probabilities should be avoided. However, if the model is used simply to estimate the overall effect of treatment, as in RCTs, then invalid predicted probabilities are less of a concern, provided the model generally fits the data well.

In the simulation study, the type I error rate was determined based on a Wald test of the null hypothesis of no treatment effect. The Wald test was chosen as it could be performed for all methods considered and led to conclusions about the effectiveness of treatment that were consistent with the 95% confidence intervals. Use of other tests for some methods may be preferable in practice, however. We investigated type I error rates based on a likelihood ratio test for log binomial regression, the COPY 1000 method and logistic regression methods, and a score test for expanded logistic regression, log Poisson regression and log normal regression. Compared to the Wald test, these tests produced fewer type I error problems but did not alter the recommendations we make in Section 6.

While we compared a large number of methods for estimating relative risks across a wide range of scenarios, the simulation study did have its limitations. First, it was assumed that all data were available for analysis and hence missing data were not considered. Missing data are a common problem in RCTs, although analysis may still be based on complete data if imputation methods are used to fill in the missing values. Second, we did not attempt to overcome convergence problems when they occurred. Choosing different starting values (Deddens et al., 2003) or estimation methods (for example, using the difficult option in Stata (Cummings, 2009)) may lead to convergence for the log binomial regression model, while altering the number of copies may improve convergence for the COPY 1000 method (Deddens et al., 2003). Although these approaches may work in some instances in practice, as illustrated in the example dataset, they were not feasible to investigate in a large simulation study. Finally, only independent observations were considered. Further research is needed to understand how the different methods for estimating relative risks compare when the data are correlated and this work is currently in progress.

6 Recommendations

Recommendation of a single method for estimating relative risks in practice is difficult, since no method performed best across all scenarios or all properties

considered in the simulation study. Further, differences observed between methods were often small and may have been due to the particular datasets generated in the simulation study. In some scenarios however, large differences were observed between methods that could have important implications in practice. We therefore make recommendations about which methods should be avoided on the basis of unacceptable performance relative to other methods in the context of RCTs with independent observations.

If a single approach for estimating relative risks is of interest, CS methods should be avoided based on the large median percent relative biases observed in some settings. Bootstrap methods should also be avoided due to their relatively poor precision and problems with coverage in certain settings, although improved performance may be possible using a larger number of bootstrap samples or a different method for obtaining confidence intervals. Log binomial regression, constrained log binomial regression and the COPY 1000 method are not recommended as convergence problems may occur. These methods also tended to have more problems compared to other methods when the log link was incorrect. Although convergence problems were possible for log normal regression, they rarely occurred and only in small samples. Given that this method generally performed well otherwise, it may still be a reasonable approach to use in practice, particularly if a large study is planned. Other methods worthy of consideration in practice are expanded logistic regression, log Poisson regression and MS + delta method. These methods generally performed well in the simulation study, even when the log link was incorrect, provided the marginal relative risk was of interest.

An alternative analysis approach, sometimes used in practice (e.g. Meade et al., 2008), is to try estimating relative risks using log binomial regression and only use another method if convergence problems occur. This approach seems reasonable, since the simulation study showed that log binomial regression generally performed well when it converged. If log binomial regression fails to converge, constrained log binomial regression and the COPY 1000 method are not recommended due to the possibility of further convergence problems. Additionally, constrained log binomial regression performed relatively poorly in terms of type I error and coverage when log binomial regression did not converge, although it was often most powerful. CS methods should also not be used based on problems with bias, precision and coverage in some settings. This leaves expanded logistic regression, log Poisson regression, log normal regression and MS methods for consideration in practice. Again, convergence problems could occur for log normal regression but given the rarity of such problems in the simulation study, we chose not to exclude this method from consideration on this basis.

Since several different methods may be appropriate for estimating relative risks based on statistical properties, it is important to consider other factors when choosing between them. Log Poisson regression and log normal regression are very simple to describe and implement, making them appealing options in practice. An additional advantage of log Poisson regression is that it is now widely used, making it a more recognizable and potentially acceptable choice. Expanded logistic regression is also quite simple to implement but a preliminary data manipulation step is required and explanation is more difficult. MS methods may not be the best choice in practice as they are complex to describe and implement. Currently they require detailed user-written programs in SAS, although this may change with future software developments. Log normal regression appears to be more sensitive to deviations from the assumed log link compared to the other methods.

Whichever approach is chosen to estimate relative risks in practice, it is important to pre-specify the approach. The results of the simulation study showed that different methods can lead to different conclusions about the effectiveness of treatment, with conclusions varying between methods for up to 13.6% of simulated datasets. Thus, pre-specification is necessary to avoid the possibility of choosing a method based on favorability of the results.

In conclusion, log binomial regression can be a useful tool for providing a clinically meaningful estimate of the effect of treatment on a binary outcome while controlling for potential confounders. This model may fail to converge however, and alternative methods for estimating relative risks are therefore required. If log binomial regression is pre-specified as the method of analysis for a binary outcome in a RCT and adjusted analyses are planned, we recommend pre-specifying an alternative approach that will be used in the event that log binomial regression fails to converge and specifying different starting values for the parameter estimates does not resolve the problem. Of the many alternative methods available, log Poisson regression would be a good choice in practice.

References

Ahmad, F., Hogg-Johnson, S., Stewart, D. E., Skinner, H. A., Glazier, R. H. and Levinson, W. (2009). Computer-assisted screening for intimate partner violence and control: a randomized trial. *Annals of Internal Medicine* 151(2): 93-102.

Barros, A. J. and Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology* 3: 21.

Bayley, N. (1993). *Manual for the Bayley Scales of Infant Development, Second Edition (BSID-II)*. San Antonio, TX, Psychological Corp.

Blizzard, L. and Hosmer, D. W. (2006). Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal* 48(1): 5-22.

Boardman, L. A., Steinhoff, M. M., Shackelton, R., Weitzen, S. and Crowthers, L. (2004). A randomized trial of the Fischer cone biopsy excisor and loop electrosurgical excision procedure. *Obstetrics and Gynecology* 104(4): 745-750.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19(9): 1141-1164.

Carter, R. E., Lipsitz, S. R. and Tilley, B. C. (2005). Quasi-likelihood estimation for relative risk regression models. *Biostatistics* 6(1): 39-44.

Cummings, P. (2009). Methods for estimating adjusted risk ratios. *Stata Journal* 9(2): 175-196.

Deddens, J., Petersen, M. R. and Lei, X. (2003). Estimation of prevalence ratios when PROC GENMOD does not converge. *Proceedings of the 28th Annual SAS Users Group International Conference*, Paper 270-28. SAS Institute Inc., Cary, NC.

Flanders, W. D. and Rhodes, P. H. (1987). Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *Journal of Chronic Diseases* 40(7): 697-704.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71(3): 431-444.

Green, B. B., Cook, A. J., Ralston, J. D., Fishman, P. A., Catz, S. L., Carlson, J., Carrell, D., Tyll, L., Larson, E. B. and Thompson, R. S. (2008). Effectiveness of home blood pressure monitoring, web communication, and pharmacist care on hypertension control: a randomized controlled trial. *Journal of the American Medical Association* 299(24): 2857-2867.

ICH E9 Expert Working Group (1999). Statistical principles for clinical trials. *Statistics in Medicine* 18(15): 1905-1942.

Kleinman, L. C. and Norton, E. C. (2009). What's the risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. *Health Services Research* 44(1): 288-302.

Lee, J. (1994). Odds ratio or relative risk for cross-sectional data. *International Journal of Epidemiology* 23(1): 201-203.

Localio, A. R., Margolis, D. J. and Berlin, J. A. (2007). Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology* 60(9): 874-882.

Lumley, T., Kronmal, R. and Ma, S. (2006). Relative risk regression in medical research: models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series*: Working Paper 293.

Makrides, M., Gibson, R. A., McPhee, A. J., Collins, C. T., Davis, P. G., Doyle, L. W., Simmer, K., Colditz, P. B., Morris, S., Smithers, L. G., Willson, K. and Ryan, P. (2009). Neurodevelopmental outcomes of preterm infants fed high-dose docosahexaenoic acid: a randomized controlled trial. *Journal of the American Medical Association* 301(2): 175-82.

McNutt, L. A., Wu, C. T., Xue, X. N. and Hafner, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology* 157(10): 940-943.

Meade, M. O., Cook, D. J., Guyatt, G. H., Slutsky, A. S., Arabi, Y. M., Cooper, D. J., Davies, A. R., Hand, L. E., Zhou, Q., Thabane, L., Austin, P., Lapinsky, S., Baxter, A., Russell, J., Skrobik, Y., Ronco, J. J., Stewart, T. E. and Study, L. O. V. (2008). Ventilation strategy using low tidal volumes, recruitment maneuvers, and high positive end-expiratory pressure for acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. *Journal of the American Medical Association* 299(6): 637-645.

Penman, A. D. and Johnson, W. D. (2009). Complementary log-log regression for the estimation of covariate-adjusted prevalence ratios in the analysis of data from cross-sectional studies. *Biometrical Journal* 51(3): 433-442.

Petersen, M. R. and Deddens, J. A. (2008). A comparison of two methods for estimating prevalence ratios. *BMC Medical Research Methodology* 8: 9.

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat - Which of these should we use? *Value in Health* 5(5): 431-436.

Schouten, E. G., Dekker, J. M., Kok, F. J., Lecessie, S., Vanhouwelingen, H. C., Pool, J. and Vandenbroucke, J. P. (1993). Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. *Statistics in Medicine* 12(18): 1733-1745.

Sinclair, J. C. and Bracken, M. B. (1994). Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* 47(8): 881-889.

Skov, T., Deddens, J., Petersen, M. R. and Endahl, L. (1998). Prevalence proportion ratios: estimation and hypothesis testing. *International Journal of Epidemiology* 27(1): 91-95.

Spiegelman, D. and Hertzmark, E. (2005). Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* 162(3): 199-200.

van der Meer, V., Bakker, M. J., van den Hout, W. B., Rabe, K. F., Sterk, P. J., Kievit, J., Assendelft, W. J. J., Sont, J. K. and SMASHING Study Group (2009). Internet-based self-management plus education compared with usual care in asthma: a randomized trial. *Annals of Internal Medicine* 151(2): 110-120.

Wacholder, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences. *American Journal of Epidemiology* 123(1): 174-184.

Walter, S. D. (2000). Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology* 53(9): 931-939.

Yu, B. B. and Wang, Z. Q. (2008). Estimating relative risks for common outcome using PROC NLP. *Computer Methods and Programs in Biomedicine* 90(2): 179-186.

Zou, G. Y. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 159(7): 702-706.