General Psychiatry

On modelling relative risks for longitudinal binomial responses: implications from two dueling paradigms

Tuo Lin ,[1] Rongzhe Zhao,[1] Shengjia Tu,[2] Hao Wu,[3] Hui Zhang,[4] Xin M Tu[1]

To cite: Lin T, Zhao R, Tu S, *et al*. On modelling relative risks for longitudinal binomial responses: implications from two dueling paradigms. *General Psychiatry* 2023;**36**:e100977. doi:10.1136/gpsych-2022-100977

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gpsych-2022-100977).

Received 01 December 2022
Accepted 10 February 2023

[1]Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, UC San Diego, La Jolla, California, USA
[2]College of Environmental Science and Engineering, Tongji University, Shanghai, China
[3]Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA
[4]Division of Biostatistics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

Correspondence to
Mr Tuo Lin;
tulin@health.ucsd.edu

ABSTRACT

Although logistic regression is the most popular for modelling regression relationships with binary responses, many find relative risk (RR), or risk ratio, easier to interpret and prefer to use this measure of risk in regression analysis. Indeed, since Zou published his modified Poisson regression approach for modelling RR for cross-sectional data, his paper has been cited over 7 000 times, demonstrating the popularity of this alternative measure of risk in regression analysis involving binary responses. As longitudinal studies have become increasingly popular in clinical trials and observational studies, it is imperative to extend Zou's approach for longitudinal data.

The two most popular approaches for longitudinal data analysis are the generalised linear mixed-effects model (GLMM) and generalised estimating equations (GEE). However, the parametric GLMM cannot be used for the extension within the current context, because Zou's approach treats the binary response as a Poisson variable, which is at odds with the Bernoulli distribution for the binary response. On the other hand, as it imposes no mathematical model on data distributions, the semiparametric GEE is coherent with Zou's modified Poisson regression. In this paper, we develop a GEE-based longitudinal model for binary responses to provide inference about RR.

INTRODUCTION

Logistic regression is widely used to model binary responses. However, many find relative risk (RR), or risk ratio easier to interpret and prefer to model regression relationships with inference about RR, rather than odds ratio (OR) as in logistic regression. Indeed, since Zou[1] published his modified Poisson regression approach for inference about RR, his paper has been cited 7 128 times, demonstrating the popularity of using RR in modelling binary responses. However, his approach isn't applied to longitudinal data. Moreover, there is no one-to-one relationship between RR and OR for regression analysis.[2] As longitudinal studies have become increasingly the standard in clinical trials and observational studies, it is imperative to develop statistical models for longitudinal binary responses with inference based on RR to fill the critical gap.

The two most popular paradigms to extend models for cross-sectional data to longitudinal data are the generalised linear mixed-effects model (GLMM) and generalised estimating equations (GEE). The parametric GLMM explicitly models the within-subject correlation using random effects, while the semiparametric, or distribution-free GEE implicitly accounts for such correlations using sandwich variance estimates.[3] Since Zou's approach treats binary responses as count variables and derives estimators of RR under the Poisson distribution, GLMM cannot be used to extend his approach to longitudinal data within the current context. As his approach is essentially a semiparametric log-linear model, a simplified version of GEE for cross-sectional data, GEE provides a coherent paradigm to develop to extend his approach to longitudinal data.

In the Models for Relative Risks for Longitudinal Binary Responses section, we first review semiparametric regression models for cross-sectional and longitudinal binary responses under the logit and log link for inference about the respective log of OR and log of RR. We then discuss a GEE-based approach for longitudinal binary responses for inference about RR by leveraging semiparametric log-linear models. In the Application section, we use real and simulated data to illustrate the proposed approach. In the Discussion section, we give our concluding remarks.

MODELS FOR RELATIVE RISKS FOR LONGITUDINAL BINARY RESPONSES

We start with a brief review of Zou's approach for inference about RR when modelling binary responses in cross-sectional data.

Cross-sectional data

Consider a study with n subjects indexed by $(1 \le i \le n)$. Let y_i denote a binary response of interest and let $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ with $x_{i0} \equiv 1$ denote a $(p+1) \times 1$ vector of explanatory, or independent, variables from the ith subject $(1 \le i \le n)$. The popular logistic regression model is defined by a generalised linear model (GLM) with the logit link as Tang *et al*[3]:

$$\begin{aligned} &y_i|\mathbf{x}_i \overset{i.d.}{\sim} \text{Bernoulli}(\mu_i), \ \mu_i = \mu(\mathbf{x}_i) = E(y_i|\mathbf{x}_i), \\ &\text{logit}(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\gamma} = \gamma_0 + \gamma_1 x_{i1} + \ldots + \gamma_p x_{ip}, \ 1 \le i \le n, \end{aligned} \tag{1}$$

where *i.d.* denotes independently distributed, Bernoulli (μ_i) denotes the *Bernoulli* distribution with mean μ_i, logit denotes the logit link function and $\boldsymbol{\gamma}$ is the vector of model parameters or coefficients. Under logistic regression, each regression coefficient γ_k has the log OR interpretation per unit change in x_{ik} for $k = 1, \ldots, p$.[3] Inference about $\boldsymbol{\gamma}$ is generalised based on maximum likelihood.[3]

For γ_k to have the RR interpretation, we need to change the logit link to the log link function to express (1) as:

$$\begin{aligned} &y_i|\mathbf{x}_i \overset{i.d.}{\sim} \text{Bernoulli}(\mu_i), \ \mu_i = E(y_i|\mathbf{x}_i), \\ &\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \ 1 \le i \le n. \end{aligned} \tag{2}$$

For differentiating log OR from log RR, we use a different symbol $\boldsymbol{\beta}$ in (2) to denote the model coefficients. Under (2), each coefficient β_k has the log RR interpretation. For example, consider one unit increase in x_{ik} from x_{ik} to $x_{ik}+1$. Denote the change in the mean of y_i in response to the change in x_{ik} by:

$$\begin{aligned} \mu_{1k} &= \mu(x_{i0}, x_{i1}, \ldots, (x_{ik}+1), \ldots, x_{ip}), \\ \mu_{0k} &= \mu(x_{i0}, x_{i1}, \ldots, x_{ik}, \ldots, x_{ip}) \end{aligned}.$$

Then, it follows from (2) that the log of RR, RR_k, for the unit change in x_{ik} from x_{ik} to $x_{ik}+1$ is:

$$\begin{aligned} \log(RR_k) &= \log\left(\frac{\mu_{1k}}{\mu_{0k}}\right) \\ &= \log(\mu_{1k}) - \log(\mu_{0k}) \\ &= \beta_k(x_{ik}+1) - \beta_k x_{ik} \\ &= \beta_k. \end{aligned}$$

The two GLMs in (1) and (2) are quite similar except for the different link functions. Under logit link in (1), the conditional mean μ_i is constrained between 0 and 1, while under the log link in (2), μ_i is confined only to positive values. Since μ_i may exceed 1, the upper bound for a probability quantity, estimates based on maximising the Bernoulli likelihood may not converge under the log link.[4 5] To alleviate this problem, we may switch the Bernoulli distribution in (2) to the Poisson, that is,

$$\begin{aligned} &y_i|\mathbf{x}_i \overset{i.d.}{\sim} \text{Poisson}(\mu_i), \ \mu_i = E(y_i|\mathbf{x}_i), \\ &\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \ 1 \le i \le n, \end{aligned} \tag{3}$$

Since the logic restriction of positive values on μ_i is consistent with the mean of the Poisson, fitting the model (3) to observed data will not be an issue. For rare diseases, μ_i will be close to 0 and y_i may be viewed as a count, frequency, or response with mean μ_i, in which case the Poisson-based (3) is a reasonable approximation. In general, with increased μ_i, (3) may not provide valid inference, since the binary y_i will not have a Poisson distribution in this case. Zou discussed the use of the sandwich variance estimator as an alternative to estimate the variance of the estimator of $\boldsymbol{\beta}$. This approach is essentially a semiparametric regression, or restricted moment model, in which only the model for the conditional mean of y_i given $\mathbf{x}_i$ in (3) is assumed:

$$\mu_i = E(y_i|\mathbf{x}_i), \ \log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \ 1 \le i \le n. \tag{4}$$

Thus, unlike (3), the semiparametric log-linear model above does not assume Poisson or any other parametric distribution for y_i. Different from a parametric model, a semiparametric model leverages estimating equations to play the role of the likelihood to provide inference.[3] Unlike maximum likelihood estimation, inference based on estimating equations is consistent regardless of the distribution of y_i, so long as the assumed conditional mean in (4) is correct.[3] Thus, even if y_i does not have a Poisson distribution, inference about $\boldsymbol{\beta}$ in (4) is still correct when based on the estimating equations.

Within the current context, the estimating equations for inference about $\boldsymbol{\beta}$ have the form:

$$\begin{aligned} &\mathbf{w}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{w}_{ni}(\boldsymbol{\beta}) = 0, \ \mathbf{w}_{ni}(\boldsymbol{\beta}) = D_i V_i^{-1} S_i, \\ &S_i = y_i - \mu_i, \ D_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mu_i \mathbf{x}_i \end{aligned} \tag{5}$$

where $V_i = Var(y_i|\mathbf{x}_i)$ is the conditional variance of y_i given $\mathbf{x}_i$. Under (4), S_i and D_i are readily evaluated. However, V_i is not determined by the semiparametric log-linear model in (4), since it only specifies the conditional mean μ_i. Within the current context, y_i follows the Bernoulli (μ_i), in which case we have $V_i = Var(y_i|\mathbf{x}_i) = \mu_i(1-\mu_i)$. We obtain the estimate $\hat{\beta}$ of $\boldsymbol{\beta}$ by solving (5) for $\boldsymbol{\beta}$. Unlike linear regression, $\hat{\beta}$ cannot be evaluated in closed form but is readily computed numerically.[3]

The estimator $\hat{\beta}$ has an asymptotically normal distribution with mean $\boldsymbol{\beta}$ and variance Σ_β:

$$\Sigma_\beta = B^{-1}\Sigma_U B^{-1}, \ \Sigma_U = E\left(D_i V_i^{-2} S_i^2 D_i^\top\right), \ B = E\left(D_i V_i^{-1} D_i^\top\right) \tag{6}$$

where B^{-1} denotes the inverse of B. We can estimate Σ_β by the following sandwich variance estimator $\hat{\Sigma}_\beta$:

$$\hat{\Sigma}_\beta = \left(\frac{1}{n}\sum_{i=1}^n \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 \mathbf{x}_i \mathbf{x}_i^\top\right) \left(\frac{1}{n}\sum_{i=1}^n \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \tag{7}$$

Note that unlike likelihood-based inference for parametric models, inference based on the estimating equations in (5) for semiparametric models is always valid, regardless of the distribution of y_i. In particular, instead of $V_i = \mu_i(1-\mu_i)$, we may also set V_i to any function of $\mathbf{x}_i$ such as $V_i = \mu_i$ (by treating y_i as a Poisson with mean μ_i) for valid inference about $\boldsymbol{\beta}$. This is why we can model a binary y_i using a semiparametric log-linear model for a count response.

Longitudinal data

We now consider extending the semiparametric log-linear model above to longitudinal data.

Suppose that the subjects are assessed repeatedly over T time points $t(1 \le t \le T)$. Let y_{it} and $\mathbf{x}_{it}$ denote the same response and explanatory variables as in the cross-sectional data setting, but with t indicating their dependence on the time of assessment $(1 \le i \le n,\ 1 \le t \le T)$. By applying the semiparametric log-linear model in (4) to each assessment t, we obtain an extension of the semiparametric log-linear model for the association of longitudinal y_{it} and $\mathbf{x}_{it}$:

$$\mu_{it} = E(y_{it}|\mathbf{x}_{it}),\ \log(\mu_{it}) = \mathbf{x}_{it}^{\top}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},\ 1 \le i \le n,\ 1 \le t \le T. \tag{8}$$

Thus, we do not explicitly model correlations among the repeated y_{it} 's. Inference about $\boldsymbol{\beta}$ is based on extending the estimating equations in (5) to the correlated y_{it} 's.

Let

$$\mu_i = (\mu_{i1}, \ldots, \mu_{iT})^{\top},\ \mathbf{y}_i = (y_{i1}, \ldots, y_{iT})^{\top},$$
$$D_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}},\ S_i = \mathbf{y}_i - \mu_i,\ 1 \le i \le n.$$

The estimating equations, which are often called the generalised estimating equations (GEE) in the literature, for inference about $\boldsymbol{\beta}$ have the form:

$$\mathbf{w}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{w}_{ni}(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i V_i^{-1}(\mathbf{y}_i - \mu_i) = 0 \tag{9}$$

where $V_i = Var(\mathbf{y}_i|\mathbf{x}_i)$ is the conditional variance of $\mathbf{y}_i$ given $\mathbf{x}_i$. As in the cross-sectional case, we can readily evaluate D_i and V_i under (8) and set $Var(y_{it}|\mathbf{x}_{it}) = \mu_{it}(1-\mu_{it})$ for each $t(1 \le t \le T)$. However, the conditional covariance between y_{is}, y_{it} given $\mathbf{x}_{is}, \mathbf{x}_{it}$ is quite complex. In almost all applications of GEE, we use a working correlation $R(\boldsymbol{\alpha})$ to approximate the true correlation $Corr(y_{is}, y_{it}|\mathbf{x}_{is}, \mathbf{x}_{it})$, where $R(\boldsymbol{\alpha})$ is a $T \times T$ correlation matrix with its entries defined by a parameter vector $\boldsymbol{\alpha}$.[3] Popular choices of $R(\boldsymbol{\alpha})$ are the working independence, with $R = \mathbf{I}_T$, and working exchangeable, with $R(\rho) = C_T(\rho)$, model, where $\mathbf{I}_T$ denotes the $T \times T$ identity matrix and ρ is a parameter.

Under a specific $R(\boldsymbol{\alpha})$, we have $V_i = A_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}$, where $A_i = diag_t(Var(y_{it}|\mathbf{x}_{it}))$ denotes a diagonal matrix with $Var(y_{it}|\mathbf{x}_{it})$ on its tth diagonal. As in the case of cross-sectional data, inference is always valid even if $R(\boldsymbol{\alpha})(V_i)$ is not the true correlation (variance) of $\mathbf{y}_i$ given $\mathbf{x}_i$. In (9), $\mathbf{w}_n(\boldsymbol{\beta})$ also depends on $\boldsymbol{\alpha}$, though we have suppressed this dependence to highlight the fact that (9) is the equation for estimating $\boldsymbol{\beta}$. Thus, $\boldsymbol{\alpha}$ must be estimated (except for the working independence model) to solve (9) for $\boldsymbol{\beta}$. We can either assign a value to or estimate $\boldsymbol{\alpha}$ together with $\boldsymbol{\beta}$. For example, under $R(\rho) = C_T(\rho)$, we may set ρ to any value between 0 and 1 or estimate ρ using correlated residuals $y_{it} - \hat{\mu}_{it}$, with $\mu_{it} = \exp\left(\mathbf{x}_{it}^{\top}\hat{\boldsymbol{\beta}}\right)$. Inference about $\boldsymbol{\beta}$ is based on the asymptotic normal distribution of the GEE estimator $\hat{\boldsymbol{\beta}}$, which has mean $\boldsymbol{\beta}$ and variance Σ_{β}:

$$\boldsymbol{\Sigma}_{\beta} = B^{-1} E\left(D_i V_i^{-1} Var(\mathbf{y}_i|\mathbf{x}_i)\, V_i^{-1} D_i^{\top}\right) B^{-\top},\ B = E\left(D_i V_i^{-1} D_i^{\top}\right) \tag{10}$$

where $B^{\top}$ denotes the transpose of B. We can estimate Σ_{β} by the sandwich variance estimator $\hat{\Sigma}_{\beta}$, which is obtained by:

$$\hat{\Sigma}_{\beta} = \hat{B}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{S}_i \hat{S}_i^{\top} \hat{V}_i^{-1} \hat{D}_i^{\top}\right) \hat{B}^{-\top},$$
$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{D}_i^{\top} \tag{11}$$

where $\hat{D}_i$, $\hat{V}_i$ and $\hat{S}_i$ denote substituting $\hat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}$ for the respective quantity D_i, V_i and S_i.

Popular software packages all support semiparametric regression models for both cross-sectional and longitudinal data. For example, PROC GEE in SAS and geeglm() in the geepack package in R[6] can be used to fit the semiparametric log-linear models in (4) for cross-sectional and (8) for longitudinal data.

APPLICATION

We illustrate our considerations with both real and simulated data. In all the examples, we set the statistical significance at $\alpha = 0.05$. All analyses are carried out using the geeglm() function in the geepack package in R.[6]

Simulation study

We consider modelling regression associations of a single time-invariant binary explanatory variable x_i with a binary response y_{it} in a longitudinal study with three assessments. To simulate the correlated y_{it}, we use a Gaussian copula with the marginal y_{it} given x_i following a Bernoulli[7]:

$$y_{it}|x_i \overset{i.d.}{\sim} \text{Bernoulli}(\mu_i),\ \log(\mu_i) = \beta_0 + x_i\beta_1,\ 1 \le t \le 3,$$
$$x_i \overset{i.i.d.}{\sim} \text{Bernoulli}\left(\tfrac{1}{2}\right). \tag{12}$$

For our simulation, we set $\beta_0 = -2$ and $\beta_1 = 1$ and an exchangeable correlation $C_3(\rho)$ in the trivariate normal with $\rho = 0.5$.

We fit the semiparametric (8) to the data simulated, that is,

$$E(y_{it}|x_i) = \mu_{it},\ \log(\mu_{it}) = \beta_0 + x_i\beta_1, \tag{13}$$

using the GEE in (9) under the working independent correlation model. Shown in table 1 are the estimates of $\boldsymbol{\beta}$ along with their standard errors (SEs) (both asymptotic

Table 1 Parameter estimates, SEs (asymptotic and empirical) and type I errors from GEE model with 1 000 MC replications

Estimates of GEE					
	Estimate	SE		Hypothesis testing	
True value	$\hat{\beta}$	Empirical	Asymptotic	H_0	Type I error
β_0=–2	–2.01	0.109	0.110	β_0=–2	0.048
β_1=1	1.01	0.123	0.125	β_1=1	0.049

GEE, generalised estimating equations; MC, Monte Carlo; SE, standard error.

and empirical), over 1 000 Monte Carlo (MC) replications under a sample size $n = 500$. The estimates $\hat{\boldsymbol{\beta}}$ were quite close to their true values, and the asymptotic SEs were quite close to their empirical counterparts. Also, shown in table 1 are type I error rates from testing the null hypothesis $H_0: \beta_0 = -2$ and $H_0: \beta_1 = 1$. We estimate the type I errors using MC iterations. Let $T^{(m)}$ denotes the Wald statistic at the mth MC replication, the type I error rate for testing H_0 is estimated by: $\hat{\alpha} = \frac{1}{1000}\sum_{m=1}^{1000} I_{\left\{T_s^{(m)} \geq q_{1,0.95}\right\}}$, where $q_{1,0.95}$ is the 95th percentile of a χ_1^2 distribution, a χ^2 distribution with 1 df. As seen, the type I error rates were close the normal values $\alpha = 0.05$.

Real study

Smoking is the chief avoidable cause of morbidity and mortality in the USA, exacting a substantive financial burden as well.[8] Smoking rates among persons with serious mental illness are exceptionally high, contributing to significant medical morbidity and mortality in this population, with many unlikely to live beyond their 50th birthday. Persons with mental illness spend nearly one-third of their monthly public assistance income on cigarettes instead of buying needed food, clothing and shelter.[9] A study was conducted to evaluate the effect of a multicomponent smoking cessation programme adapted to patients with serious psychiatric disorders within an outpatient psychiatric clinic at the University of Rochester Medical Center. This study, sponsored by the New York State Department of Health Tobacco Control Program, capitalises on packaging multiple evidence-based components to achieve a better outcome than when each practice is individually implemented in a number of clinical venues, for example, central line–associated bloodstream infections and ventilator-associated pneumonia.[10] Among the 276 participating subjects, 99 also participated in a formal evaluation, in which interviews were conducted at the point of enrolment (baseline), prior to intervention and again at 3, 6 and 12 months.

For illustrative purposes, we model the binary abstinence outcome, defined as the 7-day point prevalence (ie, abstinent from smoking for 7 days in a row), from preintervention at baseline, $t = 0$, to each of the three postintervention assessments, $t = 1, 2, 3$, at 3, 6 and 12 months, using data from 99 subjects. We create three time-varying dummy variables x_{1it}, x_{2it} and x_{3it} to indicate intervention effects at $t = 1, 2, 3$:

$$x_{1it} = \begin{cases} 1 & \text{if } t = 1 \\ 0 & \text{if } t \neq 1 \end{cases}, \quad x_{2it} = \begin{cases} 1 & \text{if } t = 2 \\ 0 & \text{if } t \neq 2 \end{cases}, \quad x_{3it} = \begin{cases} 1 & \text{if } t = 3 \\ 0 & \text{if } t \neq 3 \end{cases}.$$

Let $y_{it} = 1$ if the ith subject is abstinent for 7 days consecutively and $y_{it} = 0$ otherwise. The semiparametric GEE for change of abstinence rates over time is given by:

$$E\left(y_{it}|x_{it}\right) = \mu_{it}, \ \log\left(\mu_{it}\right) = \beta_0 + x_{1it}\beta_1 + x_{2it}\beta_2 + x_{3it}\beta_3, \quad t = 0, 1, 2, 3, \ 1 \leq i \leq 99. \tag{14}$$

We fit (8) to the 7-day point prevalence data using the GEE in (9) under the working independent correlation model.

Shown in table 2 are the estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \beta_2, \beta_3\right)^\top$ and associated SEs, p values for testing the null $H_0: \beta_t = 0$ and RRs (exponentiated $\hat{\beta}_t$) at each assessment $\left(1 \leq t \leq 3\right)$. The results show a RR greater than 1 for all three postintervention assessments, though only statistically significant at months 3 and 6. The intervention did have a significant effect on reducing

Table 2 Estimates of parameters, SEs, p values and relative risks over time from GEE model to the Smoking Cessation Study data

Estimates, SEs, p values and estimates' relative risk				
Parameter	Estimate	SE	P value	Relative risk
Baseline (β_0)	–2.156	0.339	<0.001	0.081
Month 3 (β_1)	0.754	0.354	0.033	2.125
Month 6 (β_2)	0.865	0.354	0.014	2.375
Month 12 (β_3)	0.486	0.380	0.201	1.625

GEE, generalised estimating equations; SE, standard error.

smoking in this study sample, though the effect diminished 12 months after the intervention.

DISCUSSION

We extended the popular approach for modelling RRs for binary responses to longitudinal data by leveraging the semiparametric GEE. Like the original approach in Zou,[1] the parameters of the proposed log-linear model have the log of RR interpretation and, thus, with appropriately defined explanatory variables, can be used for inference about RRs when modelling longitudinal regression relationships with binary responses. We also illustrated the proposed approach using both real and simulated longitudinal data.

The proposed GEE-based approach provides valid inference under the missing completely at random (MCAR) mechanism.[3 11] In many real studies, missing data follow the missing at random (MAR) mechanism,[3 11] in which case the lowest patterns done by the proposed approach generally yield biased estimates of RR. We can readily extend the approach to provide valid inference under MAR by employing the weighted generalised estimating equations (WGEEs).[11] Under WGEE, we also model the missingness of the binary response over time using GLMs for binary responses such as logistic regression and estimate its parameters and the parameters of the log-linear model in (8) together using a set of estimating equations that extend (9) to include the additional parameters.[3]

Contributors All authors participated in the discussion of the statistical issues and worked together to develop this paper. HZ and XMT suggested the topic, and TL, RZ, ST and HW reviewed the literature. All authors discussed the conceptual and analytical issues with modelling RRs for longitudinal data using the parametric and semiparametric models. RZ, ST and HW developed the simulation settings, algorithms and associated R codes and performed the simulation study under the direction of TL. TL, HZ and XMT drafted the manuscript, while TL, RZ, ST and HW provided all the technical details and derivations, along with completing the application section. All authors worked together to finalise the manuscript.

Funding The project described was partially supported by the National Institutes of Health (grant UL1TR001442) of Georgia Clinical and Translational Science Alliance funding.

Competing interests None declared.

Provenance and peer review Commissioned; externally peer-reviewed.

ORCID iD
Tuo Lin http://orcid.org/0000-0002-4495-8865

REFERENCES

1 Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 2004;159:702–6.
2 Feng C, Wang H, Wang B, *et al*. Relationships among three popular measures of differential risks: relative risk, risk difference, and odds ratio. *Shanghai Arch Psychiatry* 2016;28:56–60.
3 Tang W, He H, Tu XM. *Applied categorical and count data analysis*. Hall/CRC, FL: Chapman, 2012.
4 McNutt L-A, Wu C, Xue X, *et al*. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol* 2003;157:940–3.
5 Wallenstein S, Bodian C. Inferences on odds ratios, relative risks, and risk differences based on standard regression programs. *American Journal of Epidemiology* 1987;126:346–55.
6 R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. Available: www.R-project.org/ [Accessed 15 Dec 2022].
7 Yan J. Enjoy the joy of copulas: with a package copula. *J Stat Softw* 2007;21:1–21.
8 Fiore MC, Jaen CR, Baker TB, *et al*. Treating tobacco use and dependence: 2008 update. clinical practice guideline. the U.S. surgeon general'S world wide web web site. 2008. Available: www.surgeongeneral.gov/tobacco/treating_tobacco_use08.pdf [Accessed 15 Dec 2022].
9 Steinberg ML, Williams JM, Ziedonis DM. Financial implications of cigarette smoking among individuals with schizophrenia. *Tob Control* 2004;13:206.
10 Institute for Health Care Improvement. Introduction to evidence based practices and bundling. 2009. Available: www.premierinc.com/quality-safety/tools-services/safety/topics/bundling [Accessed 15 Nov 2022].
11 Little RJA, Rubin DB. *Statistical analysis with missing data*. New York: Wiley, 1987.

Tuo Lin is a fifth-year PhD student in Biostatistics at the University of California, San Diego (UCSD) in the USA. He obtained his master's degree in Statistics at UCSD in 2018. He is currently working as a graduate student researcher in the division of Biostatistics and Bioinformatics of Herbert Wertheim School of Public Health and Human Longevity Science at UCSD. He has also been working at Altman Clinical and Translational Research Institute (ACTRI) in the USA for many years, helping with study designs and data analyses. His main research interests include survey sampling and methods, causal inference and longitudinal data analysis in psychiatry studies.