Logistic mixed-effects model analysis with pseudo-observations for estimating risk ratios in clustered binary data analysis

Hisashi Noma, PhD*
Department of Interdisciplinary Statistical Mathematics, The Institute of Statistical Mathematics, Tokyo, Japan
The Graduate Institute for Advanced Studies, The Graduate University for Advanced Studies (SOKENDAI), Tokyo, Japan
ORCID: http://orcid.org/0000-0002-2520-9949

Masahiko Gosho, PhD
Department of Biostatistics, Institute of Medicine, University of Tsukuba, Tsukuba, Japan

*Corresponding author: Hisashi Noma
Department of Interdisciplinary Statistical Mathematics
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
TEL: +81-50-5533-8440
e-mail: noma@ism.ac.jp

Abstract

Logistic mixed-effects model has been a standard multivariate analysis method for analyzing clustered binary outcome data, e.g., longitudinal studies, clustered randomized trials, and multi-center/regional studies. However, the resultant odds ratio estimator cannot be directly interpreted as an effect measure, and it is only interpreted as an approximation of the risk ratio estimator when the frequency of events is small. In this article, we propose a new statistical analysis method that enables to provide risk ratio estimator in the multilevel statistical model framework. The valid risk ratio estimation is realized via augmenting pseudo-observations to the original dataset and then analyzing the modified dataset by the logistic mixed-effects model. The resultant estimators of fixed effect coefficients are theoretically shown to be consistent estimators of the risk ratios. Also, the standard errors and confidence intervals of the risk ratios can be calculated by bootstrap method. All of the computations are simply implementable by using R package "glmmrr". We illustrate the effectiveness of the proposed method via applications to a cluster-randomized trial of maternal and child health handbook and a longitudinal study of respiratory disease. Also, we provide simulation-based evidence for the accuracy and precision of estimation of risk ratios by the proposed method.

Key words: clustered data; generalized linear mixed-effects model, pseudo-observation; case-cohort design, risk ratio.

1. Introduction

Logistic mixed-effects model has been a standard multivariate analysis method for analyzing clustered binary outcome data in clinical and epidemiological studies, e.g., longitudinal studies, clustered randomized trials, and multi-center/regional studies [1-3]. However, the resultant odds ratio estimator cannot be directly interpreted as an effect measure, and it is only interpreted as an approximation of the risk ratio estimator when the frequency of events is small [4,5]. Similar problems have been widely discussed for unclustered data analyses, e.g., the modified Poisson regression [6] is one of the representative methods that enables unbiased estimation of risk ratios in the multivariate analyses. The Poisson regression approach can be extended to the generalized estimating equation (GEE) framework for analyzing clustered binary data [7], but the estimating equation-based semiparametric inference method cannot be straightforwardly extended to the fully parametric multilevel modelling framework. A remarkable advantage of the multilevel modelling is that can be implemented by the direct likelihood principle for analysis of incomplete data [1,8]; the influence of missing data is ignorable if the missing data mechanism is missing-at-random (MAR) [9]. For this reason, the mixed-effects models are also widely used for primary analyses of longitudinal clinical trials [10]. The ignorability is not fulfilled for the GEE approach, and currently, no effective statistical methods are available for estimating risk ratios directly on the multilevel modelling framework.

In this article, we propose a new effective method in estimating risk ratios for clustered binary outcome data in the multilevel modelling framework. Especially, we theoretically show the regression coefficient estimates obtained from a logistic mixed-effects model fitting to a modified dataset, that some pseudo-observations are augmented, can be interpreted as unbiased estimates of risk ratios. This method can be simply implementable using standard statistical packages of the ordinary logistic mixed-effects

model, and the confidence intervals of risk ratios can be calculated by bootstrap method. In addition, the Little-Rubin's ignorability for missing data [9] is fulfilled for the proposed method. We illustrate the effectiveness of the proposed method via applications to a cluster-randomized trial of maternal and child health handbook [11] and a longitudinal clinical study of respiratory disease [12]. We also provide simulation-based evidence for the accuracy and precision of estimation of risk ratios by the proposed method.

2. The pseudo-observations approach

2.1 Target statistical model and estimands

We consider Y_{ij} as the jth binary outcome ($= 1$: event occurred, $= 0$: event did not occur) on cluster i, where $i = 1, \dots, N$, and $j = 1, \dots, n_i$. For instance, Y_{ij} is a repeated binary outcome on jth visit of ith individual for a longitudinal study, or outcome variable of jth participant of ith cluster for a cluster-randomized trial. For analyzing the clustered binary data, the logistic mixed-effects model is one of the most widely used model in the generalized linear mixed-effects model (GLMM) that uses the canonical link of binomial outcome data [8,13],

$$\text{logit}\{E[\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i]\} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\eta}_i$$

$$\boldsymbol{\eta}_i \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{\Omega})$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$ design matrices, $\boldsymbol{\beta}$ is the $p \times 1$ vector containing the fixed effects, and $\boldsymbol{\eta}_i$ is the $q \times 1$ vector containing the random effects; $\boldsymbol{\eta}_i$ is assumed to be normally distributed with mean zero and covariance matrix $\boldsymbol{\Omega}$ ($q \times q$ symmetric matrix). Conditional on the random effects $\boldsymbol{\eta}_i$, the responses $Y_{i1}, \dots, Y_{in_i}$ are independent and follow a multivariate Bernoulli distribution with a mean vector $E[\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i]$. The regression coefficients of $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Omega}$ are estimated by the (restricted) maximum likelihood (ML) method, and the resultant

estimators are consistent and the most precise ones under large sample settings [8,13]. The computations are implementable using standard statistical packages of GLMM, e.g., lme4 package [14] of R (R Foundation for Statistical Computing, Vienna, Austria), PLOC GLIMMIX of SAS (SAS Institute, Inc, Cary, NC). Also, potential influences from missing data are ignorable owing to the direct likelihood principle [8,9]. However, the regression coefficients of $\boldsymbol{\beta}$ are interpreted as log-transformed odds-ratios as like the ordinary logistic regression, and only be interpreted as approximations of log risk ratios when the event frequency is small [4,5].

To overcome the interpretability issue, an alternative mixed-effects model that is based on the log-linear-type binomial mixed-effects model can be considered,

$$\log\{E[\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i]\} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\eta}_i \qquad (*)$$

$$\boldsymbol{\eta}_i \sim \mathrm{MVN}(\boldsymbol{0},\boldsymbol{\Omega})$$

that is an extension of the log-linear binomial regression model [13] to the GLMM framework. The marginal regression model of jth component of $\boldsymbol{Y}_i$ is expressed as

$$\log\{\Pr(Y_{ij}=1|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij})\} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{\eta}_i$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are the jth row vectors of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. The log-linear binomial regression model has a substantial computational problem because the values of the regression function can become > 1, and the ML estimates of regression coefficients cannot be obtained in many cases [15,16]. The GLMM model also has the same computational problem, and the ML estimates cannot be often calculated. To resolve the computational issue of the log-linear binomial regression model, the modified Poisson regression [6] is developed based on the semiparametric estimating equation theory [17,18]. However, the Poisson regression approach cannot be straightforwardly generalized to the fully parametric mixed-effects model, because the variance function is misspecified. In the following sections, we will show an alternative pseudo-likelihood method can resolve

this issue, and we can provide a consistent estimator of $\boldsymbol{\beta}$ in the log-linear mixed-effects model (*) via simple computations by the ordinary GLMM software packages.

2.2 A pseudo-observations approach to provide risk-ratio estimator

To provide the new method, we first review the logistic regression analysis in case-cohort studies that provide a consistent risk ratio estimator [19,20]. Case-cohort studies a subtype of case-control studies, which uses random samples from a specific cohort (subcohort) as the control [21,22]. Schouten et al. [19] proposed fitting a logistic regression model to the selected cases and subcohort samples, whose duplicated samples are involved in both of cases and subcohort samples; the resultant regression coefficients estimators are shown to be consistent estimators for the log-linear binomial regression model of the target population.

We show the principle of risk-ratio estimation by a logit-link model can be applied to the GLMM analysis. Suppose formally adapting the case-cohort-type sampling with sampling probabilities of 100% (both for cases and subcohort samples) to the individual units of the cohort data regardless of the cluster structure, respectively. From this artificial sampling, individual units with $Y_{ij} = 1$ are sampled to both of cases and subcohort samples, and those with $Y_{ij} = 0$ are sampled only to the subcohort samples. Then, create an indicator variable D_{ij} that indicates being a case sample on the selected case and/or subcohort samples. We propose using D_{ij} for the outcome variable of the clustered binary outcome data analysis instead of Y_{ij}. Through this processing, the original dataset is modified that the individual units with $Y_{ij} = 1$ are duplicated to two individual units with $D_{ij} = 0$ and 1, i.e., the individual units with $D_{ij} = 0$, which do not actually exist in the real world, are added to the original dataset. Then, the probability being $D_{ij} = 1$ for (i, j) unit in the modified dataset with the *pseudo-observations* is written as

$$\Pr\left(D_{ij}=1\middle|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij}\right)=\frac{\Pr\left(Y_{ij}=1\middle|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij}\right)}{1+\Pr\left(Y_{ij}=1\middle|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij}\right)}=\frac{\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{\eta}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{\eta}_i\right)}$$

This relationship shows the probability of $D_{ij}=1$ is expressed as the logistic transformation of $\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{\eta}_i$. Also, the probability of $D_{ij}=0$ is expressed as

$$\Pr\left(D_{ij}=0\middle|\boldsymbol{X}_i,\boldsymbol{Z}_i\right)=\frac{1}{1+\Pr\left(Y_{ij}=1\middle|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij}\right)}=\frac{1}{1+\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{\eta}_i\right)}$$

Then, through fitting the logistic mixed-effects model to the modified dataset whose outcome variable is D_{ij}, we can obtain the consistent estimators of regression coefficients and covariance parameters in $\{\boldsymbol{\beta},\boldsymbol{\Omega}\}$ of the log-linear mixed-effects model (*) by those estimates from the logistic mixed-effects model analysis, i.e., the regression coefficients estimator $\widehat{\boldsymbol{\beta}}$ can be interpreted as consistent log-transformed risk ratio estimators. Note the modified dataset is no longer a case-cohort dataset, and the relationship between Y_{ij} and D_{ij} plays the essential role in the statistical inference; we can consider the estimating function is calibrated to be unbiased. Also, in general case-cohort settings, the subcohort sampling probability is set to relatively small, and then, the intercept parameter in the log-linear model is not consistently estimated by the Schouten's estimating method [19]. However, in this case, since the subcohort sampling probability is 100%, the intercept is consistently estimated. Also, this estimating method does not correspond to the ML estimation, and the asymptotic efficiency is not fulfilled; this property is the same with the pseudo-likelihood estimating method by the modified Poission regression approach [6]. However, the ML estimation in the log-linear mixed-effects model (*) is often infeasible, and this pseudo-likelihood approach provides a realistic and effective statistical inference framework for estimating the interpretable effect measures.

The formal algorithm is summarized as follows:

Algorithm 1 (Logistic mixed-effects model with pseudo-observations)

1. For the clustered binary outcome dataset, create pseudo-observations for all of the individual units with $Y_{ij} = 1$ that have the same cluster indicators, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$; the outcomes of the pseudo-observations should be 0. Then, add the pseudo-observations to the original dataset.

2. Fit the logistic mixed-effects model to the augmented dataset that the pseudo-observations are added. The resultant estimators of regression coefficients and covariance parameters in $\{\boldsymbol{\beta}, \boldsymbol{\Omega}\}$ of the logistic mixed-effects model are consistent estimators of those of the log-linear mixed effects models (*), and the exponential-transformed regression coefficients estimators $\exp(\widehat{\boldsymbol{\beta}})$ can be interpreted as consistent risk ratio estimators.

2.3 Statistical inference for the model parameters

In the original method of Schouten et al. [19], they noted the standard errors can be estimated by the sandwich variance estimator because of the estimating equation theory [23]. However, the case-cohort data involves artificial duplicated data, that do not actually exist, thus the ordinary estimating equation theory does not work for this case [20]. Noma [20] showed the sandwich variance estimator has serious bias even under large sample settings. For the proposed pseudo-observations approach, the standard errors cannot also be estimated by the sandwich variance estimator. However, we can quantify the statistical errors simply adapting the bootstrap method to the clustered data. Various parametric and nonparametric bootstrap approaches have been proposed to the GLMM framework [24], and as shown in the simulation studies in Section 4, the most common cluster-based nonparametric bootstrap could provide stable and valid inference results generally.

Algorithm 2 (Nonparametric bootstrap for the pseudo-observations method)

1. Resample N clusters from the original dataset with replacement, and re-label the cluster indicators in the resampled dataset as $1'^{(b)}, 2'^{(b)}, \ldots, N'^{(b)}$ $(b = 1, \ldots, B)$.

2. For the resampled N cluster data, create the pseudo-observations as described in Algorithm 1. Then, fit the logistic mixed-effects model to the augmented dataset. The resultant estimates $\widehat{\boldsymbol{\beta}}^{(1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}$ and $\widehat{\boldsymbol{\Omega}}^{(1)}, \ldots, \widehat{\boldsymbol{\Omega}}^{(B)}$ are the bootstrap samples of the pseudo-likelihood estimators of $\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}}\}$.

Using the bootstrap samples, confidence intervals of the elements of $\{\boldsymbol{\beta}, \boldsymbol{\Omega}\}$ can also be calculated, e.g., the 2.5th and 97.5th percentiles in the bootstrap samples can be used for calculating 95% confidence limits. The bootstrap P-values for statistical testing can also be calculable from the resampled data. The nonparametric bootstrap can effectively circumvent the parametric distribution assumptions and can provide stable inference results generally [25]. For more complicated multilevel models involving non-nested hierarchical sampling models, parametric bootstrap techniques are also applicable [24]. Note the number of resampling B should be ≥ 1,000 for providing stable confidence limits and P-values, generally. The logistic mixed-effects model requires numerical integrations for calculating ML estimates, and the bootstrap process might be required large computational efforts, but under the recent computational environment, it is not realistically problematic; the parallel computation techniques can be effectively applied (as provided in the example R code in the Supporting Information materials).

2.4 Ignorability on missing data analysis

For the pseudo-observations approach, the Little-Rubin's ignorability on missing data analysis [9] is straightforwardly fulfilled. We use a different notation in this section for

convenience. Let D denote the set of complete data in the modified dataset that the pseudo-observations are added, and R a set of indicator variables indicating whether individual components of D are observed (=1) or missing (=0). Also, let $D = (D_{\text{obs}}, D_{\text{mis}})$, where D_{obs} is the observed part of D and D_{mis} is the missing part of D. Then, the joint probability functions of D and R for the logistic mixed-effects model is expressed as

$$p(D, R|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(D|\boldsymbol{\theta})p(R|D, \boldsymbol{\phi})$$

where $\boldsymbol{\theta}$ is the parameter set of the data model and $\boldsymbol{\phi}$ is that of the missing data mechanism model. Then the fully pseudo-likelihood function based on the observed data is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}|D_{\text{obs}}, R) = \int p(D_{\text{obs}}, D_{\text{mis}}|\boldsymbol{\theta})p(R|D_{\text{obs}}, D_{\text{mis}}, \boldsymbol{\phi})dD_{\text{mis}}$$

The MAR mechanism assumes the independency between D_{mis} and the missing data mechanism, i.e., $p(R|D_{\text{obs}}, D_{\text{mis}}, \boldsymbol{\phi}) = p(R|D_{\text{obs}}, \boldsymbol{\phi})$, and the observed data pseudo-likelihood is expressed as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}|D_{\text{obs}}, R) = p(R|D_{\text{obs}}, \boldsymbol{\phi}) \int p(D_{\text{obs}}, D_{\text{mis}}|\boldsymbol{\theta})dD_{\text{mis}} \propto p(D_{\text{obs}}|\boldsymbol{\theta})$$

Thus, we can conduct valid statistical inference for $\boldsymbol{\theta}$ using the observed data pseudo-likelihood function and can ignore the missing data mechanism under the MAR assumption. This theoretical property indicates that the proposed modified logistic mixed-effects model analysis can be used for incomplete data analysis as like the ordinary likelihood-based analyses using the mixed-effects models [8,10].

2.5 Software

We developed an R package, `glmmrr` (https://github.com/nomahi/glmmrr/), to perform

the proposed methods via simple commands. Example R code for implementing these methods is provided in the Supporting Information materials.

3. Applications

3.1 Cluster-randomized trial of maternal and child health handbook

Mori et al. [11] conducted a cluster-randomized trial to assess the effectiveness of maternal and child health handbook in Mongolia. They randomly allocated the uses of the maternal and child health handbook to pregnant women in 18 districts, and assessed the effectiveness in promoting antenatal care attendance, especially whether the number of antenatal care visits were ≥ 6 times or not. 253 women in 9 districts were assigned to the intervention group, and 248 women in 9 districts were assigned to the control group. To assess the intervention effect, the intra-cluster correlations should be adequately adjusted, and the multilevel models have been one of the primary approaches [2]. In assessing the intervention effect, the logistic mixed-effects models are the standard approach for analyzing the clustered binary outcome data, but the resultant odds ratio estimates are not interpretable effect measures if the frequency of events were not small [4,5]. The crude event frequencies were 81.7% for the intervention group and 70.6% for the control group, and the logistic mixed-effects models would not be an adequate method for this case. In Table 1, we provide the odds ratio estimate by the logistic mixed-effects model involving random intercepts corresponding to 18 districts (the heterogeneity variance was denoted as τ^2); the odds ratio estimate was 2.42 (95% confidence interval [CI]: 0.45, 13.10). The point estimate of the odds ratio indicated a large intervention effect, but due to the large event frequency, it could overestimate the actual relative risk, at least it was markedly large compared with the ratio of crude event frequencies. Mori et al. [11] also performed subgroup analyses by a summary index that quantified social-economic status of

individual participants by the Demographic and Health Survey guidelines [26]. They divided the population into 5 subgroups by quartiles of the social-economic index. We also provide the subgroup analysis results in Table 1. The effect sizes were especially large for the 1st and 2nd quartile subgroups (the richest 2 subgroups), and the odds ratio estimates were 7.48 (95%CI: 1.30, 42.97) and 35.25 (95%CI: 1.18, 1048.65), respectively. Besides, for the remaining subgroups, the intervention effects were not clearly shown, but the odds-ratio estimates themselves were certainly large (1.76, 1.55, and 0.77). These subgroup analyses indicated the effect modifications by the social-economic status; the impact of the maternal and child health handbook was larger for relatively wealthy and well-educated participants. However, the odds-ratio estimates could not quantify the intervention effects adequately.

As an alternative strategy, we performed the same analyses using the modified logistic mixed-effects model analyses with pseudo-observations. We augmented the original datasets with the pseudo-observations and simply fitted the logistic mixed-effects models with random intercepts corresponding to 18 districts to the modified datasets. The confidence intervals were calculated by the bootstrap method with 1,000 resamplings. The resultant risk ratio estimates and 95%CIs are presented in Table 1. At first, the risk ratio estimate for the overall population was 1.16 (95%CI: 0.85, 1.56). The magnitude of relative risk estimate was quite small compared with the odds ratio estimate, but would be considered as plausible comparing the crude event frequencies of the two groups. For the subgroup analyses by the social-economic index, the risk ratio estimates were also moderate compared to the odds ratio estimates. For the 1st and 2nd quartile subgroups, the risk ratio estimates were 1.52 (95%CI: 0.99, 1.79) and 1.52 (95%CI: 1.09, 4.71) respectively, and there also were great discrepancies with the odds ratio estimates. Also, for the 3rd, 4th and 5th quartile subgroups, the risk ratio estimates were 1.05 (95%CI:

0.63, 1.63), 1.02 (95%CI: 0.73, 1.53) and 0.90 (95%CI: 0.53, 1.23); these estimates indicated quite small between-groups differences. All of these results indicated the odds-ratio estimates could provide misleading evidence that overestimated the actual relative risks for the high frequency events. The pseudo-observations approach would provide more precise information for the intervention effects.

3.2 Longitudinal clinical study for respiratory disease

Statistical analyses of longitudinal studies also require appropriate adjustments of correlations for repeated measurements, and the multilevel models are one of the primary statistical tools for them. Also, the direct likelihood or Bayesian approaches can address the influences of missing data by the MAR mechanism owing to the Little-Rubin's ignorablity principle [9]. Koch et al. [12] provided a longitudinal clinical trial data for respiratory disease, where 111 participants from two centers were randomly assigned to active treatment or placebo. The primary outcome was dichotomized respiratory status (good or poor), and the respiratory statuses were planned to be measured at baseline and four visits after randomization. The crude event frequencies were 68.1% for the active treatment group and 44.3% for the placebo group. We firstly fitted the logistic mixed-effects model with random intercepts corresponding to the individual participants (the heterogeneity variance was denoted as τ^2) and involving treatment (active vs. placebo), baseline respiratory status, center, gender (male vs. female), and age as the fixed effects to the clinical study data. The results are presented in Table 2. The odds ratio estimate for the treatment was 7.55 (95%CI: 2.61, 21.84) and indicated a remarkable treatment effect. Also, that for the baseline symptom was 18.36 (95%CI: 5.92, 56.96) and indicated a strong prognostic effect. For the remaining variables, the 95%CIs of odds ratios covered the null value (=1), but the point estimates indicated certainly large prognostic effects.

To assess the risk ratios directly, we fitted the logistic mixed-effects model augmented the pseudo-observations to the large event frequency data. The resultant risk ratio estimates and 95%CIs are presented in Table 2. The risk ratio estimate of the treatment was 1.61 (95%CI: 1.22, 2.17); significant treatment effect was indicated, but the magnitude of the effect measure estimate was much smaller than that of the odds ratio. Also, the risk ratio estimate of the baseline symptom was 2.04 (95%CI: 1.56, 2.85), and the magnitudes of size of the estimate were quite different. For the remaining variables, all of the risk ratio estimates became moderate. For the longitudinal study with high frequency events, the overall results and interpretations would be entirely changed through using the risk ratio estimation method.

4. Simulation studies

To illustrate the operating characteristics of the proposed method, we performed simulation studies. For data generation, to replicate the correlation structures of real-world data, we resampled the participant level data from the longitudinal clinical study of respiratory disease in the previous section. Then, we supposed the binomial log-linear mixed effects model (*) for the outcome data generation,

$$\begin{aligned}\log\{\Pr(Y_{ij} = 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})\} \\ &= \eta_{i0} + \beta_0 + \beta_1 \times Treatment + \beta_2 \times Baseline\ symptom \\ &+ \beta_3 \times Center + \beta_4 \times Gender + \beta_5 \times Age\end{aligned}$$

where $\eta_{i0} \sim N(0, \tau^2)$ was the random intercept and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_5)^T$ was the fixed effect regression coefficients $(i = 1, \dots, N; j = 1, \dots, 4)$. We considered several scenarios that chanted $N =$ 100, 200 and 500, and $\tau =$ 0.02 and 0.05. Also, the log risk ratio β_1 that expressed the treatment effect was set to $0.47\ (= \log(1.60))$ and $0.69\ (= \log(2.00))$; the other coefficients were set to the estimates obtained from the analysis of

the original data, $\beta_2 = 0.71, \beta_3 = 0.28, \beta_4 = -0.05,$ and $\beta_5 = -0.01$. β_0 was set to an appropriate value so that the overall event frequency ($\mathrm{P_{event}}$) became 0.20 and 0.40. We performed 1,000 simulations for each scenario, and the simulated datasets were analyzed by the ordinary logistic mixed-effects model and the proposed modified logistic mixed-effects model. For calculating bootstrap CIs, we performed 1,000 resampling consistently. For the evaluation measures, we calculated the means and standard deviations of the odds ratio and risk ratio estimates corresponding to β_1 obtained from 1,000 simulations. Also, for the proposed method, we calculated mean of the standard error estimates by bootstrap and root mean squared error (RMSE) of the estimates of β_1, and the coverage rate and expected width of 95% CIs of β_1.

The simulation results are presented in Table 3. For the ordinary logistic mixed-effects model analyses, the mean of the estimates of odds ratio for the treatment effect was seriously overestimated and biased from the true risk ratio, because of the large event frequencies. Besides, using the pseudo-observations approach, the risk ratios were accurately estimated under all scenarios. These results indicate that the ordinary logistic mixed-effects model could not provide interpretable effect measure estimates, and the pseudo-observations approach clearly improved this relevant issue. Also, the bootstrap standard error estimates for the proposed method accurately estimated the actual standard errors; the accuracy depended on the sample size, and it was relatively improved if the sample size became larger. The RMSEs were nearly equivalent to the standard errors of the log risk ratio estimates, because the biases were generally to be 0. For the bootstrap confidence intervals, the coverage rates were consistently around the nominal level (=0.95) and could provide valid interval estimates generally. The expected widths depended on the parameter settings, especially N and $\mathrm{P_{event}}$.

5. Discussion

Multilevel statistical modelling has been one of the standard statistical tools in clinical and epidemiological studies [1-3]. However, the resultant effect measure estimates directly influence to the interpretation of the evidence and subsequent decision making for public health. The logistic mixed-effects model is mathematically convenient model that uses the canonical link function of binomial distribution [13], but the resultant odds ratio estimates have difficulties in the interpretations as shown in the numerical studies. Recently, for non-clustered data analyses, to overcome the practical issue of odds ratio, the modified Poisson regression [6] have been widely adopted in practice that can provide risk ratio estimates directly. However, similar effective solutions have not been proposed in the statistical analyses using multilevel statistical models. The pseudo-observations approach would provide effective and realistic solutions for this issue preserving the nice property of the Little-Rubin's ignorability for missing data [9].

The practical advantage of the proposed methods is the computational conveniences. The pseudo-observation method can be implemented only adding pseudo-observations to the original dataset and then fitting the ordinary logistic mixed-effects model to the modified dataset. This process does not require special software packages and can be implemented using the standard GLMM packages. For computing confidence intervals, the bootstrap calculations are needed, but they can be performed by simple commands. The computational time is not generally so problematic, because parallel computation tools are available under current computational environments.

In addition, the risk ratio estimation methods are useful for more broad settings. For example, clinical trials for drug development often require correlated data analyses and the logistic mixed-effects model has been widely adopted, e.g., longitudinal clinical trials involving repeated outcome measurements [10], multicenter/regional clinical trials [27], and

crossover trials [28]. However, for these cases, the event frequencies are often not small, and the resultant odds ratio estimates have difficulty in measuring the magnitude of the treatment effects. The proposed risk ratio estimation method would also be effectively applied to these practical situations.

Similar ideas that modify the logistic regression analysis in ordinary binary outcome data analyses have been discussed in Dias-Quijano [29] and Dwivedi et al. [30]. Our proposed method can be seen as the generalization of their approaches, but their confidence intervals depend on modified model variance estimates or sandwich variance estimates. The variance estimators can have serious biases [20] and bootstrapping approaches would be needed. Also, the generalization to the GEE framework of their methods would have difficulties, because the working correlation structures among the pseudo-observations and the other units in the same clusters are difficult to consider; additional theoretical assumptions are needed due to the artificial augmented data. In the GEE analysis setting, we consider the modified Poisson regression approach [7] would be more effective. However, the Little-Rubin's ignorablity for missing data [9] does not fulfill for the semiparametric method, and the pseudo-observations approach would have practical advantages in some aspects. Furthermore, another representative effect measure for binary outcome is risk difference [31], and there are no effective methods to treat this measure by multilevel statistical modelling framework. The development of new methods for inferences of risk difference in analyzing clustered binary data by multilevel models is one of relevant issues for future studies.

In conclusion, the pseudo-observations approach using logistic mixed-effects model would provide new insights in clinical and epidemiological studies that require multilevel statistical models and would be an effective statistical tool for presenting more precise evidence for public health research and clinical practice.

Acknowledgements

This study was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (grant numbers: JP23K11931, JP22H03554, JP24K21306, and JP23H03063).

Data availability statements

R package `glmmrr` for implementing the proposed methods is available at GitHub (https://github.com/nomahi/glmmrr/). Also, the original datasets are involved in the R package. Example R codes for replicating the results of the analyses are provided in the Supporting Information materials.

References

1. Fitzmaurice GM, Laird NM, Ware J. *Applied Longitudinal Analysis*. Hoboken: John Wiley & Sons; 2011.
2. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. 2nd ed. Boca Raton: CRC Press; 2022.
3. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th ed. Philadelphia: Wolters Kluwer; 2020.
4. Greenland S. Interpretation and choice of effect measures in epidemiologic analysis. *Am J Epidemiol*. 1987;125(5):761-768.
5. Nurminen M. To use or not to use the odds ratio in epidemiologic analyses. *Eur J Epidemiol*. 1995;11(4):365-371.
6. Zou GY. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702-706.
7. Zou GY, Donner A. Extension of the modified Poisson regression model to

prospective studies with correlated binary data. *Stat Methods Med Res*. 2011;22(6):661-670.

8. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. New York: Springer; 2005.
9. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. 3rd ed. New York: Wiley; 2019.
10. Mallinckrodt CH, Lipkovich I. *Analyzing Longitudinal Clinical Trial Data: A Practical Guide*. Boca Raton: CRC Press; 2016.
11. Mori R, Yonemoto N, Noma H, et al. The Maternal and Child Health (MCH) handbook in Mongolia: a cluster-randomized, controlled trial. *PloS One*. 2015;10(4):e0119772.
12. Kock GG, Carr GJ, Amara IA, Stokes ME, Uryniak TJ. Categorical data analysis. In: Berry DA, ed. *Statistical Methodology in Pharmaceutical Sciences*. Boca Raton: CRC Press; 1990.
13. Agresti A. *Categorical Data Analysis*. 3rd ed. Hoboken: Wiley; 2013.
14. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Software*. 2015;67:1.
15. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157(10):940-943.
16. Wallenstein S, Bodian C. Epidemiologic programs for computers and calculators. Inferences on odds ratios, relative risks, and risk differences based on standard regression programs. *Am J Epidemiol*. 1987;126(2):346-355.
17. Uno S, Noma H, Gosho M. Firth-Type Penalized Methods of the Modified Poisson and Least-Squares Regression Analyses for Binary Outcomes. *Biom J*. 2024;66(7):e202400004.

18. Noma H, Kitano T. Modelling nonlinear effects in risk ratio and risk difference using the Poisson and Gaussian additive regression models. *Stats*. 2024;7(4):1473-1482.
19. Schouten EG, Dekker JM, Kok FJ, et al. Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. *Stat Med*. 1993;12(18):1733-1745.
20. Noma H. Variance estimation for logistic regression in case-cohort studies. *J Epidemiol*. 2024;34(1):38-40.
21. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73(1):1-11.
22. Noma H, Tanaka S. Analysis of case-cohort designs with binary outcomes: Improving the efficiency using whole cohort auxiliary information. *Stat Methods Med Res*. 2017;26(2):691-706.
23. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13-22.
24. Thai HT, Mentre F, Holford NH, Veyrat-Follet C, Comets E. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharm Stat*. 2013;12(3):129-140.
25. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: CRC Press; 1994.
26. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan*. 2006;21(6):459-468.
27. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Stat Med*. 1999;18(15):1905-1942.
28. Senn S. *Cross-Over Trials in Clinical Research*. 2nd ed. New York: Wiley; 2002.
29. Diaz-Quijano FA. A simple method for estimating relative risk using logistic

regression. *BMC Med Res Methodol*. 2012;12:14.

30. Dwivedi AK, Mallawaarachchi I, Lee S, Tarwater P. Methods for estimating relative risk in studies of common binary outcomes. *J Appl Stat*. 2014;41(3):484-500.
31. Cheung YB. A modified least-squares regression approach to the estimation of risk difference. *Am J Epidemiol*. 2007;166(11):1337-1344.

Table 1. Results of the ordinary logistic mixed-effects model analysis and the pseudo-observations approach for the cluster randomized trial of maternal and child health handbook ($N = 500$).

	Logistic mixed-effects model		Pseudo-observations approach	
	OR estimates (95%CI)	τ estimate	RR estimates (95%CI)	τ estimate
All population	2.42 (0.45, 13.10)	1.66	1.16 (0.85, 1.56)	0.08
Subgroups by the wealth index				
1st quartile	7.48 (1.30, 42.97)	0.55	1.52 (0.99, 1.79)	0.00
2nd quartile	35.25 (1.18, 1048.65)	2.06	1.52 (1.09, 4.71)	0.00
3rd quartile	1.76 (0.04, 71.62)	3.26	1.05 (0.63, 1.63)	0.00
4th quartile	1.55 (0.22, 11.04)	1.38	1.02 (0.73, 1.53)	0.00
5th quartile	0.77 (0.06, 9.83)	1.85	0.90 (0.53, 1.23)	0.00

Table 2. Results of the ordinary logistic mixed-effects model analysis and the pseudo-observations approach for the longitudinal clinical study of respiratory disease ($N = 111$).

	Logistic mixed-effects model OR estimates (95%CI)	Pseudo-observations approach RR estimates (95%CI)
Treatment (ref: placebo)	7.55 (2.61, 21.84)	1.61 (1.22, 2.17)
Baseline symptom	18.36 (5.92, 56.96)	2.04 (1.56, 2.85)
Center	2.66 (0.92, 7.71)	1.32 (1.01, 1.75)
Gender (ref: female)	0.78 (0.21, 2.92)	0.95 (0.65, 1.44)
Age	0.97 (0.94, 1.01)	0.99 (0.98, 1.00)
τ estimate	1.98	0.00

Table 3. Results of the simulation studies: Comparison of the ordinary logistic mixed-effects model and the pseudo-observations approach †.

				Linear mixed-effects model (log odds ratio)		Pseudo-observations approach (log risk ratio)					
N	P_{event}	τ	β_1	Mean	SE	Mean	SE	$\widehat{SE}$	RMSE	CR	EW
100	0.20	0.02	0.47	0.635	0.284	0.479	0.216	0.223	0.216	0.950	0.873
100	0.20	0.02	0.69	0.931	0.295	0.696	0.225	0.227	0.225	0.951	0.891
100	0.20	0.05	0.47	0.637	0.277	0.480	0.211	0.225	0.211	0.959	0.883
100	0.20	0.05	0.69	0.942	0.302	0.704	0.231	0.228	0.231	0.948	0.892
100	0.40	0.02	0.47	0.932	0.251	0.477	0.132	0.134	0.132	0.951	0.526
100	0.40	0.02	0.69	1.437	0.281	0.709	0.143	0.140	0.144	0.944	0.548
100	0.40	0.05	0.47	0.940	0.256	0.481	0.136	0.135	0.136	0.952	0.528
100	0.40	0.05	0.69	1.423	0.266	0.701	0.132	0.138	0.132	0.960	0.543
200	0.20	0.02	0.47	0.621	0.198	0.471	0.151	0.152	0.151	0.953	0.595
200	0.20	0.02	0.69	0.933	0.204	0.700	0.153	0.155	0.153	0.953	0.606
200	0.20	0.05	0.47	0.621	0.202	0.470	0.154	0.153	0.154	0.949	0.599
200	0.20	0.05	0.69	0.944	0.203	0.708	0.155	0.155	0.155	0.941	0.608
200	0.40	0.02	0.47	0.922	0.174	0.476	0.092	0.092	0.093	0.947	0.361
200	0.40	0.02	0.69	1.401	0.184	0.694	0.093	0.095	0.093	0.959	0.371
200	0.40	0.05	0.47	0.931	0.173	0.476	0.091	0.092	0.091	0.945	0.360
200	0.40	0.05	0.69	1.396	0.182	0.692	0.094	0.095	0.094	0.956	0.371
500	0.20	0.02	0.47	0.618	0.119	0.470	0.092	0.093	0.092	0.957	0.366
500	0.20	0.02	0.69	0.916	0.125	0.690	0.096	0.096	0.096	0.952	0.375
500	0.20	0.05	0.47	0.627	0.117	0.476	0.090	0.094	0.090	0.960	0.367
500	0.20	0.05	0.69	0.933	0.127	0.703	0.098	0.096	0.098	0.953	0.376
500	0.40	0.02	0.47	0.911	0.110	0.471	0.057	0.057	0.057	0.954	0.223
500	0.40	0.02	0.69	1.400	0.117	0.696	0.060	0.059	0.060	0.954	0.231
500	0.40	0.05	0.47	0.912	0.112	0.469	0.059	0.057	0.059	0.945	0.224
500	0.40	0.05	0.69	1.390	0.120	0.692	0.061	0.059	0.061	0.941	0.231

† Mean, SE: Mean and standard deviation of the estimates of β_1, $\widehat{SE}$: Mean of the standard error estimates of β_1, RMSE: Root mean squared error of the estimates of β_1, CR, EW: Coverage rate and expected width of the 95% confidence intervals of β_1.