Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Rose et al., J Biom Biostat 2015, 6:4
DOI: 10.4172/2155-6180.1000253

Research Article Open Access

Bayesian Logistic Regression Modeling as a Flexible Alternative for Estimating Adjusted Risk Ratios in Studies with Common Outcomes

Charles E Rose[1]*, Yi Pan[1] and Andrew L Baughman[2]

[1]*Division of HIV/AIDS Prevention, Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA*
[2]*Division of Global HIV/AIDS, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, USA*

Abstract

Background: For cohort and cross-sectional studies, the risk ratio (RR) is the preferred measure of effect rather than an odds ratio (OR), especially when the outcome is common (>10%). The log-binomial (LB) and Poisson models are commonly used to estimate the RR; the OR estimated using logistic regression is often used to approximate the RR when the outcome is rare. However, regardless of the prevalence of the outcome, logistic regression predicted exposed and unexposed risks may be used to estimate the RR. Because maximum likelihood estimation is used to fit the logistic model, estimation of the Standard error of the RR is difficult.

Methods: To overcome difficulty in estimation of the SE of the RR and provide a flexible framework for modeling, we developed a Bayesian logistic regression (BLR) model to estimate the RR, with associated credible interval (CI_B). We applied the BLR model to a large hypothetical cross-sectional study with categorical variables and to a small hypothetical clinical trial with a continuous variable for which the LB method did not converge. Results of the BLR model were compared to those from several commonly used RR modeling methods.

Results: Our examples illustrate the Bayesian logistic regression model estimates adjusted RRs and 95% CI_{BS} comparable to results from other methods. Adjusted risks and risk differences were easily obtained from the posterior distribution.

Conclusions: The Bayesian logistic regression modeling approach compares favorably with existing RR modeling methods and provides a flexible framework for investigating confounding and effect modification on the risk scale.

Keywords: Bayesian logistic regression; Log-binomial; Poisson regression; Prevalence ratio; Risk ratio

Introduction

In epidemiology, when the study design is appropriate for estimating risk, the risk ratio (RR) comparing exposed to unexposed is the preferred measure of effect rather than an odds ratio (OR) [1]. It is usually appropriate to estimate the RR for cohort, cross-sectional (where RR is often referred to as a prevalence ratio (PR)), and randomized trials study designs [2-7]. In general, if the outcome is rare (prevalence <10%), then the RR and OR and their confidence intervals (CI) will be similar. However, if the outcome is common, then the OR can differ substantially from the RR. The OR may be said to under/overestimate the magnitude of the RR as the RR decreases/increases from 1.0 and the extent of the difference between the OR and RR increases as the prevalence increases [8]. Commonality of the outcome has become an important consideration in selecting the model to estimate the RR [2,4].

Two modeling strategies have emerged for estimating the RR of common outcomes, direct and indirect. A recent report reviewed 12 different methods for estimating RR that fall within these two strategies [9]. Direct methods model the risk or log of risk (e.g., log-binomial model) and recent studies have focused on direct methods because of the ease in obtaining the RR and CI. In contrast, indirect RR estimation methods often model the log-odds using logistic regression. The RR is then estimated using logistic regression by back transforming the predicted log-odds to the probability scale [3,10-12] and obtaining standardized predicted risks for exposed and unexposed cohorts.

Bayesian logistic regression (BLR) models have been used in epidemiology studies to estimate the OR between exposed and unexposed cohorts. The Bayesian paradigm is a flexible framework that allows us to easily generate the risks, RR, risk differences, and associated posterior distributions. A Bayesian log-binomial model has recently been proposed [13] that accounts for the inequality constraint on the parameters necessary in a log binomial model. Given the familiarity of logistic regression among epidemiologists and the popularity of Bayesian analysis, in this report we expand upon existing methods and develop a novel BLR modeling approach to indirectly estimate the median RR and associated credible interval (CI_B) from the RR posterior distribution. We compare the Bayesian RR estimates to those from several commonly used RR modeling techniques using two previously published hypothetical data sets. Our examples illustrate the ease of obtaining the RR posterior distribution and the performance of the model when adjusting for confounders, including a continuous variable.

Models and Methods

Defining risk and odds ratios

The RR is defined as the probability that a subject in the exposed group will experience the outcome relative to the probability that a

***Corresponding author:** Charles E Rose, Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road NE, Mailstop E48, Atlanta, GA 30329, Tel: +1 404-639-3028; E-mail: crose@cdc.gov

Received October 02, 2015; **Accepted** October 14, 2015; **Published** October 21, 2015

Citation: Rose CE, Pan Y, Baughman AL (2015) Bayesian Logistic Regression Modeling as a Flexible Alternative for Estimating Adjusted Risk Ratios in Studies with Common Outcomes. J Biom Biostat 6: 253. doi:10.4172/2155-6180.1000253

subject in the unexposed group will experience the same outcome, i.e.,

$$RR = \frac{P(Outcome \mid Exposed)}{P(Outcome \mid Unexposed)}.$$

An OR is defined as the odds of having the outcome given the subject is exposed relative to the odds of having the outcome given the subject is unexposed, i.e.,

$$OR = \frac{P(Outcome \mid Exposed) / (1 - P(Outcome \mid Exposed))}{P(Outcome \mid Unexposed) / (1 - P(Outcome \mid Unexposed))}.$$

Logistic regression works well if one is interested in the OR but the OR may perform poorly as an approximation of the ratio of two probabilities [14]. The assertion that logistic regression performs poorly, especially for a common outcome, presumes that the OR is used as an approximation for the RR [14,15]. The RR is commonly estimated using a direct method and less frequently using an indirect method.

Direct methods for estimating the risk ratio

The Mantel-Haenszel (MH) method can be used to directly estimate the adjusted RR across strata to assess an association between outcome and exposure [16]. In addition, the log-binomial and Poisson models are commonly used for estimating the RR, likely because of the ease of fitting these models in standard software packages and because the RR is the natural measure of effect from these models. For both log-binomial (or log-linear risk) [17] and Poisson models [18], let y_i (0 or 1) denote the outcome status for the i^{th} subject, X_1 (0 or 1) denote the exposure status of that subject and let $\boldsymbol{X}_i$ represent the entire set, i.e., including the exposure variable, of p explanatory variables. Then the probability of the i^{th} subject experiencing the outcome is given by

$$P_i\left[Y = 1 \mid X_{1i}, X_{2i}, ..., X_{pi}\right] = \mu_i,$$

where μ_i is the risk for the i^{th} subject. The log-binomial models the log of the risk on the explanatory variables, which includes the intercept, as:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2,i} + ... + \beta_p X_{p,i} = \boldsymbol{\beta} \mathbf{X}_i.$$

Hence, the RR comparing exposed to unexposed, adjusted for the explanatory variables, is given by e^{β_1}.

There may be challenges when using the log-binomial model to estimate the RR because when fitting the log-binomial model, especially given continuous variables, non-convergence may be an issue when the MLE is close to or on the boundary of the parameter space [14]. In addition, simulation studies have shown the CI obtained for the RR when using the log-binomial model may be too narrow [4,18]. If the log-binomial model fails to converge, several methods have been proposed to obtain model convergence: COPY [2], Iterative Weighted Least Squares (IWLS) truncated algorithm [19], inverse-probability-of-treatment-weighted (IPTW) [20], and a constrained Bayesian model [13].

Assuming Y is a binary response variable, the COPY method is implemented by making C copies of the data set, with C suggested to be 1,000, and appending them to the original data set [2]. In one of the copies, the outcome Y is switched to 1-Y for every observation in that copy. The log-binomial pseudo-MLE is then estimated by using C as the weight when fitting the model to the modified dataset. The IWLS truncated algorithm was developed to address the boundary issue when the log-binomial model does not converge [19]. The user chooses a threshold (T) near 1 that is used in the algorithm, and p (predicted risk) is set to min (p, T) at each iteration until convergence. The IPTW method also uses a weighted likelihood with a set of weights based on the idea of standardization [20]. The weights depend on the exposure variable probability and conditional probabilities of exposure given the other explanatory variables. The conditional probabilities can be estimated by strata defined by the other explanatory variables, and or in the case of a large number of strata that may lead to unstable estimation, through propensity scores (PS) [21]. A constrained Bayesian log-binomial model has been proposed that imposes a constraint of the parameters such that, given the data, only combinations of the parameters that result in estimates within the defined parameter space are accepted in the parameter estimation [13].

The Poisson model places no restriction on the sum of the β's given the data, which leads to an often noted limitation of the method that estimated probabilities may be greater than 1.0. In addition, because Poisson errors overestimate binomial errors when the outcome is common, the Poisson model often overestimates the standard errors (SE) of the β's leading to the RR having a CI that is too wide. To adjust the SE when using Poisson regression it has been suggested to use a robust SE, which leads to CIs that are not as conservative [18, 22, 23].

Indirectly estimating the risk ratio using logistic regression

Maximum likelihood estimation is used to fit the logistic regression model, and the predicted probabilities from the fitted model are used to estimate the RR indirectly. The logistic model is the regression of the log of the odds on the explanatory variables:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i,$$

where η_i is modeled as a linear combination of $p+1$ explanatory variables, including an indicator variable for the exposure of interest and the intercept, as

$$\eta_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_p X_{p,i}.$$

The natural measure for logistic regression for comparing the exposed to the unexposed is the OR, which is readily obtained since

$$OR = \frac{\frac{\mu_E}{1-\mu_E}}{\frac{\mu_U}{1-\mu_U}} = \frac{e^{\beta_0 + \beta_1 + \beta_2 X_2 + ... + \beta_p X_p}}{e^{\beta_0 + \beta_2 X_2 + ... + \beta_p X_p}} = e^{\beta_1}.$$

The RR is estimated by summarizing risks for the exposed and unexposed cohorts after the logistic regression model is fit. Estimating risks for the exposed and unexposed cohorts involves summarizing risks across strata by exposure cohort. This is often referred to as regression standardization in which a standardized risk is computed as a weighted average of category specific risks.

Three methods used for logistic regression standardization are conditional, stratification, and marginal [3,6,10]. The conditional method requires choosing a standard reference value for each of the covariates, and the stratification method requires standard weights for each stratum or combination of covariates. To avoid these complications and derive an internally adjusted measure, we use the marginal method, which is defined as

$$RR = \frac{\frac{1}{n}\sum_{i=1}^{n} P_i(y_i \mid \boldsymbol{X}_i, Exposed)}{\frac{1}{n}\sum_{i=1}^{n} P_i(y_i \mid \boldsymbol{X}_i, Unexposed)},$$

where n is the total sample size of the exposed and unexposed cohorts

combined, and the probabilities are estimated from the logistic regression model. Hence, this method involves estimating the risk of those not exposed as though they were exposed and vice versa. Standardizing risks across strata using regression models to obtain summarized risks is straightforward [6], but obtaining the CI about the standardized RR is challenging due to the difficulty in estimating the SE. The bootstrap and the delta method have been employed to estimate the CI for the standardized RR obtained using logistic regression [10,12,14]. However, these methods can be computationally burdensome for the model estimation, require a lot of programming, and be analytically complicated.

Bayesian logistic regression model

The Bayesian modeling framework and current software for Bayesian analysis can meet these complex challenges in a straightforward manner. Unlike SEs computed or derived for maximum likelihood estimates, a Bayesian posterior distribution of any statistic or transformation of interest can be readily obtained. Thus, we extended the logistic regression model for estimating the RR to the Bayesian framework. Using previous notation, the standard Bayesian logistic regression (BLR) model includes stochastic, systematic (linear predictor), and prior distribution components:

$$y_i \sim Bernoulli(\mu_i), \qquad \text{(stochastic)}$$

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i, \qquad \text{(systematic)}$$

$$\boldsymbol{\beta} \sim N_{p+1}\left(\boldsymbol{b}, \boldsymbol{\Sigma}^{-1}\right), \qquad \text{(prior distribution).}$$

The prior distribution for $\boldsymbol{\beta}$ is assumed to be multivariate normal with $\boldsymbol{b}$ defined as the vector of means of the coefficients of the $p + 1$ explanatory variables and $\boldsymbol{\Sigma}$ as the $p + 1$ by $p + 1$ precision matrix, i.e., the inverse of a variance-covariance matrix. Standardized marginal risks for the exposed and unexposed cohorts are calculated as before and in the Bayesian analysis, the posterior distribution of the indirectly estimated *RR* is obtained by Bayes' rule as the product of the likelihood function and the prior distribution [24]. Bayesian software can be used to simulate draws to approximate the posterior distribution of model parameters, and from the simulated values, we can estimate the posterior distribution of any quantity of interest, including the RR, which is a primary motivation for the Bayesian approach. We can also easily summarize the mean, median, and credible interval, CI_B, e.g., the 2.5th and 97.5th percentiles of the posterior distribution of the RR, which also has the advantage of being interpreted as a probability interval.

Statistical analysis

We used hypothetical data from two published studies to compare the RR and credible intervals estimated using the BLR model with those obtained using stratified MH, log-binomial (Bayesian and frequentist), and Poisson models. These examples were chosen to represent both a simple situation with a large sample size and one categorical confounder, and a more complex situation with a small sample size, and categorical and continuous confounders that led to non-convergence of the log-binomial model.

We used the following guidelines to fit models when using direct modeling methods and for estimating the OR. The logistic standardization by the marginal method was calculated using 5,000 bootstrap samples in SAS 9.3 PROC GENMOD [24, 26]. The standard Poisson model was fit by maximum likelihood estimation using SAS 9.3 PROC GENMOD [26]. The modified Poisson was fit by the generalized estimating equations approach that uses a robust estimator of the SE, available in GENMOD. Log-binomial models were also fit using GENMOD and, if convergence failed, we used COPY [2], IWLS truncated algorithm [19], and IPTW [20]. To implement the COPY method, we use weights of 0.999 for the original data for Y and 0.001 for the 1-Y copy data. To implement the IWLS truncated algorithm method, we use a threshold of 0.9999 and p was set to the minimum of (p, 0.9999) to compute the working residuals and weights for the next iteration. We implemented this procedure using R software [27]. To implement the IPTW method, we used both strata and PS defined weights. The PS weights were estimated using logistic regression. The Bayesian log-binomial model was estimated using SAS 9.3 PROC MCMC with boundary constraints defined by the outer range of the data and the procedure outlined below for the Bayesian logistic regression model [13]. Standard logistic regression was used to estimate the OR for comparison with the estimated RR for common outcomes.

Bayesian logistic regression was implemented using the Bayesian option in SAS 9.3 PROC GENMOD with the following guidelines. We used non-informative priors for $\boldsymbol{\beta}$, one chain of initial parameter values, and a burn-in of 5,000 samples determined using the Brooks-Gelman-Rubin (BGR) convergence criterion [28]. In addition, the BGR criterion was used along with plots of the successive samples to assess autocorrelation and the chain was thinned at 10 to reduce autocorrelation. Each model, after burn-in, used 1,000,000 samples and after thinning resulted in a sample of 100,000. The RR was estimated for each sample and the posterior distribution of the RR was summarized using the mean, median, and 95% CI_B.

Example I: Our first example illustrates that the Bayesian logistic regression model provides point estimates of prevalence and prevalence ratios on a stratum specific basis as well as for a situation in which there is effect modification on the log odds scale but not on the prevalence or log prevalence scales. We use data from a hypothetical cross-sectional study [5] with a common disease outcome for the exposed and unexposed and one confounder (Table 1). The overall prevalence was 47% with stratum I and II specific prevalence for the exposed (80% and 60%) and unexposed (40% and 30%), respectively. The crude OR and RR are 4.44 and 2.03, respectively. The stratum specific ORs are 6.00 and 3.50, which indicates effect modification on the OR scale. In contrast, there is no evidence of effect modification on the RR scale as the RR is 2.00 for both strata (Table 1). We illustrate using the Bayesian logistic model to determine if there is effect modification on the RR scale given effect modification on the OR scale and compare the estimated RR and 95% CI_B to estimates obtained using the other models.

All models considered, stratified MH, logistic standardization marginal, robust and standard Poisson, log-binomial (frequentist and Bayesian), and Bayesian logistic, included the exposure and the stratum variable. They all estimated the RR to be 2.00 with a 95% CI / CI_B of (1.86, 2.16), with the exceptions of the frequentist log-binomial (1.86, 2.15) and standard Poisson (1.79, 2.23). Using our Bayesian logistic regression model, we obtained stratum specific predicted median probabilities (95% CI_B) of disease by stratum for the exposed of 0.80 (0.76, 0.83) and 0.60 (0.56, 0.64), and for the unexposed of 0.40 (0.37, 0.43) and 0.30 (0.27, 0.33). Our predicted difference between the RRs for strata 1 and 2 is 0.0014 (-0.31, 0.30) for the stratum specific RR (Figure 1). This indicates an absence of effect modification on the RR scale so a common RR can be estimated across strata. Our model was re-run ignoring the effect modification on the logit scale to illustrate the differences in predictions. This model has predicted probabilities (95% CI_B) of 0.76 (0.73, 0.79), 0.64 (0.60, 0.67), 0.42 (0.39, 0.45) and 0.28

Exposed				Unexposed				
	D=Yes	D=No	Prevalence	D=Yes	D=No	Prevalence	Odds Ratio	Risk Ratio
Stratum 1	400	100	0.80	320	480	0.40	6.00	2.00
Stratum 2	300	200	0.60	300	700	0.30	3.50	2.00
All	700	300	0.70	620	1,180	0.34	4.44	2.03
Model							Risk Ratio	95% CI
Stratified Mantel-Haenszel							2.00	1.86, 2.16
Logistic Standardization by Marginal Method							2.00	1.86, 2.16
Poisson							2.00	1.79, 2.23
Robust Poisson							2.00	1.86, 2.16
Log-Binomial							2.00	1.86, 2.15
Bayesian Logistic							2.00	1.86, 2.16*
Bayesian Log-Binomial							2.00	1.86, 2.16*
Logistic							4.44†	3.75, 5.25

Notes: *Credible interval, †Odds ratio.

Table 1: Example I: Hypothetical cross-sectional data [5] and the estimated adjusted risk ratio and 95% confidence interval (CI) for each method.

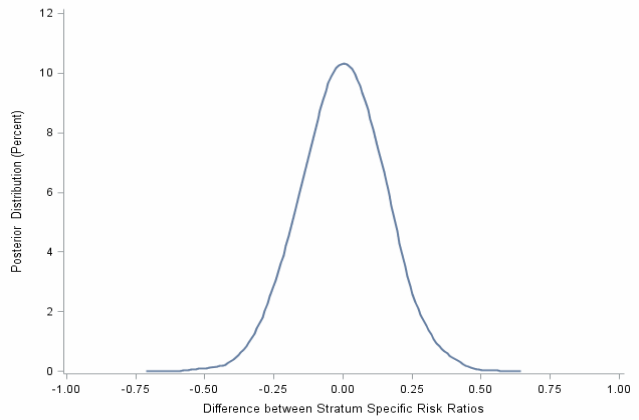

Figure 1: Estimated posterior distribution for the difference between stratum-specific risk ratios estimated using the Bayesian logistic regression model on example I data. The estimated RR (95% credible interval) was 2.0 (1.8, 2.2) for stratum 1 and 2.0 (1.8, 2.3) for stratum 2.

Conventional Therapy				New Therapy			
Subject	**Age (years)**	**EOD***	**Recovery†**	**Subject**	**Age (years)**	**EOD**	**Recovery**
1	20	0	1	21	20	0	1
2	23	0	1	22	24	0	1
3	22	0	0	23	28	0	1
4	26	0	0	24	30	0	1
5	29	0	0	25	32	0	1
6	34	0	0	26	33	0	0
7	32	0	1	27	38	0	1
8	30	0	0	28	36	0	0
9	38	0	0	29	24	1	0
10	37	0	0	30	26	1	1
11	38	0	1	31	29	1	1
12	25	1	1	32	34	1	0
13	24	1	0	33	32	1	0
14	25	1	0	34	34	1	1
15	29	1	0	35	33	1	1
16	32	1	0	36	36	1	0
17	34	1	0	37	38	1	0
18	37	1	0	38	39	1	0
19	40	1	0	39	38	1	1
20	40	1	0	40	40	1	1
Mean	30.75	0.45	0.25	Mean	32.2	0.6	0.6

Notes: EOD* is the Extent of disease (0=moderate, 1=severe) and †Recovery (0=not recovered, 1=recovered).

Table 2: Example II: Raw data for hypothetical clinical trial [3] comparing the risk of recovery in new and conventional therapy groups.

(0.26, 0.31) and an estimated common RR (95% CI_B) = 2.0 (1.85, 2.14), which illustrates that in the absence of effect modification on the RR scale there is no substantial difference in the RR estimated by ignoring the effect modification on the logit scale. All SAS and R programs and data for the presented models are provided in the S1 Appendix.

Example II: Our second example illustrates potential differences among model results. We use data from a hypothetical study from a clinical trial that has categorical and continuous variables [3]. The study purpose was to compare new and conventional therapies on recovery while controlling for age and extent of disease (EOD) among 40 subjects (Table 2). Recovery (0 = not recovered and 1 = recovered) is the outcome, and therapy group (conventional therapy = 0, new therapy = 1) is the exposure of interest. Confounding variables are EOD, which is measured as moderate (0) or severe (1), and age (years).

Subjects that received new therapy were more likely to recover (60.0% vs. 25.0%), slightly older (years) (32.2 vs. 30.8), and have a more severe EOD (60.0% vs. 45.0%) than those who received conventional therapy (Table 2). Subjects having moderate disease or younger age (20–29) were more likely to recover (52.6% vs. 33.3% and 53.3% vs. 36.0%, respectively) (Table 3). The unadjusted OR and RR are 4.50 and 2.40, respectively (Table 3). Model results (Table 4) illustrate several artifacts for models adjusted for the categorical (EOD) and continuous (age) confounders. The Poisson models (standard and robust SE) estimated two probabilities >1.0. The frequentist log-binomial model failed to converge and we implemented the COPY, IWLS truncated algorithm, and IPTW methods to obtain convergence. Using the COPY method, the log-binomial model did converge but estimated two probabilities extremely close to the boundary (> 0.999). In addition, the estimated RR (2.44) using the COPY method was virtually unchanged from the unadjusted RR (2.40, Table 3) and substantially smaller than the RR estimated using most other methods (Table 4). We used a threshold of 0.9999 for the IWLS truncated algorithm method and achieved convergence. However, the estimated

	New Therapy			Conventional Therapy				
	Recovered	Not Recovered	Prevalence	Recovered	Not Recovered	Prevalence	Odds Ratio	Risk Ratio
Extent of Disease								
Moderate	6	2	0.75	4	7	0.36	5.25	2.06
Severe	6	6	0.50	1	8	0.11	8.00	4.50
Age (years)								
20-29	5	1	0.83	3	6	0.33	10.00	2.50
30-40	7	7	0.50	2	9	0.18	4.50	2.75
All	12	8	0.60	5	15	0.25	4.50	2.40

Table 3: Example II descriptive statistics for hypothetical clinical trial [3] comparing the risk of recovery in new and conventional therapy groups.

Model	Risk Ratio	95% CI	CI Width
Stratified Mantel-Haenszel[a]	2.75	1.11, 6.84	5.73
Logistic Standardization by Marginal Method	2.79	1.32, 6.17	4.85
Poisson	2.73	0.95, 7.82	6.87
Robust Poisson	2.73	1.27, 5.87	4.60
Log-Binomial	**Did not converge**		
Log-Binomial (COPY)	2.44	1.09, 5.46	4.37
Log-Binomial[b] **(IWLS truncated algorithm)**	2.67	1.29, 5.49	4.20
	2.54	1.20, 5.41	4.21
	2.50	1.16, 5.36	4.20
Log-Binomial (IPTW strata weights)	2.65	1.03, 6.82	5.79
Log-Binomial (IPTW PS weights)	2.85	1.17, 6.91	5.74
Bayesian Logistic	2.65	1.31, 6.63[c]	5.32
Bayesian Log-Binomial	2.67	1.11, 6.38[c]	5.27
Logistic	7.93[d]	1.55, 40.44	38.89

Notes: IWLS = iterative weighted least squares; IPTW= inverse-probability-of-treatment-weighted, PS = propensity score. [a]The estimated RR and 95% CI were calculated using age in years. Results were 2.91 (1.22, 6.93) using the two age groups in Table 3. [b]Three sets of starting values were chosen when using IWLS: 1) using the fitted MLE values for the model without age and starting the coefficient for age at zero, 2) estimates from the raw proportions and starting the coefficient age at zero, and 3) starting all coefficients at zero. [c]Credible interval and [d]Odds ratio.

Table 4: Example II results for hypothetical clinical trial [3] comparing the risk of recovery in new and conventional therapy groups. The risk ratio and 95% confidence interval (CI) for recovery adjusted for age and extent of disease was estimated using different models.

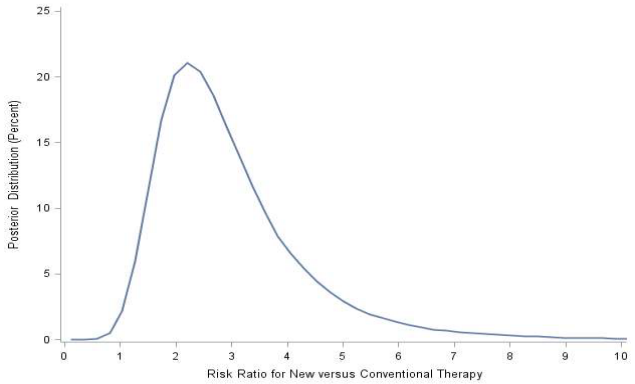

Figure 2: Estimated posterior distribution for the risk ratio for recovery comparing new therapy with conventional therapy estimated using the Bayesian logistic regression model on example II data. The estimated RR (95% credible interval) was 2.65 (1.31, 6.63)

RR using this algorithm was sensitive to starting values (Table 4). Three sets of starting values were chosen by 1) using the fitted values from the log-binomial model without age and starting the coefficient for age at zero, 2) using log-transformed stratum defined proportions for the coefficients for therapy, EOD, and overall (intercept), and starting the coefficient for age at zero, and 3) starting all coefficients at zero. For the IPTW method we used strata weights defined by age and EOD and estimated PS weights by fitting a logistic regression model with therapy as the outcome and EOD (categorical) and age (continuous) as the predictor variables. The IPTW method converged using both the strata and PS weights. The Bayesian log-binomial model converged but the parameter space was constrained by using all the outer ranges of the data as constraints.

The Bayesian logistic model, as well as the logistic standardization using the marginal method, avoided all of these artifacts and model-fitting challenges. We obtained the posterior distribution of the RR and the 95% CI_B, 2.65 (1.31, 6.63) (Figure 2). The stratified MH, logistic standardization marginal, Poisson, and robust Poisson produced results similar to those of the Bayesian logistic model, with estimated RRs (95% CI) of 2.75 (1.11, 6.84), 2.79 (1.32, 6.17), 2.73 (0.95, 7.82), and 2.73 (1.27, 5.87), respectively. The log-binomial model estimated the RR to be 2.44 (1.09, 5.46), 2.65 (1.03, 6.82), and 2.85 (1.17, 6.91) using the COPY and IPTW (strata and PS weights), respectively. The log-binomial IWLS truncated algorithm produced estimated RR and 95% CI for our three sets of starting values of 2.67 (1.29, 5.49), 2.54 (1.20, 5.41), and 2.50 (1.16, 5.36). Lastly, the Bayesian log-binomial RR was 2.67 and 95% CI_B (1.11, 6.38). Although the Poisson, robust Poisson, logistic standardization marginal, and Bayesian logistic models produced similar estimated RR values, their 95% intervals were substantially different: (0.95, 7.82), (1.27, 5.87), (1.32, 6.17), and (1.34, 6.86), respectively. The standard Poisson model was the only model that resulted in a non-significant RR. The log-binomial model (COPY and IWLS truncated algorithm methods) and Poisson model with robust standard error produced the narrowest 95% CIs. The log-binomial (strata, PS weights, and Bayesian) and Bayesian logistic methods produced 95% CIs of similar width.

Results and Discussions

In epidemiology, the RR comparing the exposed to the unexposed is the preferred measure of effect rather than an OR when the study design is appropriate for estimation of an outcome risk. Obtaining a standardized marginal RR is straightforward using logistic regression. However, MLE does not produce a RR CI (11, 12). We have demonstrated the ease of using the RR posterior distribution to calculate a 95% CI_B. We illustrated the feasibility using two examples of Bayesian logistic regression as an alternative to standard methods to estimate the regression standardized marginal RR and 95% CI_B.

The logistic regression estimated OR should not be considered

an approximation to the RR when the prevalence is common [14]. However, the logistic regression model itself may be a viable and better fitting model than the log-binomial or Poisson models. The RR may be obtained using a standardization method coupled with the bootstrap or the delta method to obtain the CI [3, 9, 11, 12]. A recent study [9] illustrated that the marginal method using logistic regression to estimate the RR performed well in simulations that included a single binary exposure variable and a single binary confounder variable. The standardization method using logistic regression is capable of estimating the RR for multiple exposure variables of interest and can be easily implemented in SUDAAN [29] for simple random samples as well as complex survey data [24].

The log-binomial model is often used to directly estimate the RR and generally produces an unbiased estimate of the adjusted RR but the CI may be too narrow [4, 30, 31]. In addition, while the log-binomial model may fail to converge using standard maximum likelihood for several reasons [32], convergence may be achieved by using any one of several methods (COPY, IWLS truncated algorithm, IPTW) or using better coefficient starting values [2, 14]. However, even when reasonable starting values or the COPY method are used non-convergence of the log-binomial model may remain an issue. Moreover, a simulation study comparing several methods for estimating the RR demonstrated that the COPY method did not converge for all simulated datasets [9]. We illustrated that a viable alternative to the frequentist log-binomial is to use a constrained Bayesian log-binomial model [13]. However, if the model has many potential predictor variables then determining all the outer range constraints may be tedious and the constraints will have to be modified when removing variables from the model.

Our example II used the COPY, IWLS, IPTW, and Bayesian methods to fit the log-binomial model and the estimated RR using COPY was 2.44, which is lower than the Poisson (2.73), logistic standardization marginal (2.79), Bayesian log-binomial (2.67), and Bayesian logistic (2.65) models. The IWLS truncated algorithm and IPTW (stratified and PS) methods estimated RR as 2.67, 2.65, and 2.85, respectively. Simulations have shown the IPTW performs adequately but the method is generally limited to studies investigating one exposure variable [9]. Our example II illustrates the IWLS truncated algorithm performed adequately but estimated parameters were sensitive to parameter starting values. Further investigation is warranted into the substantial discrepancy between results from the log-binomial model using the COPY method and RRs produced by other models as well as the sensitivity of the IWLS truncated algorithm to parameter starting values.

Poisson models using a robust SE generally produced a reasonable standard error, but the Poisson model may result in some estimated probabilities >1.0 (example II). Estimated probabilities >1.0 may not be a concern if the focus of the study is on the RR and not on predicted risks, in which case the robust Poisson method is a good alternative. Poisson regression's advantage over log-binomial regression is that it is not prone to non-convergence. Moreover, simulations illustrate the Poisson model using a robust SE adequately estimates the RR and 95% CI with relatively low percent relative bias compared with other methods.

We used published data to illustrate the viability of the Bayesian logistic regression model using the marginal method to estimate the RR and 95% CI_B. Although our study is limited in scope to these hypothetical data sets, the results are consistent with a simulation study that illustrates logistic regression using the marginal method performs well in estimating the RR and 95% CI [9]. In addition, Bayesian logistic regression provides a flexible framework for obtaining and summarizing the RR and other statistics using the posterior distribution. Bayesian modeling is becoming more widespread in epidemiology and the ease of performing Bayesian analysis within standard software packages is increasing. Although beyond the scope of this study future follow-up studies using simulation will provide guidance on the performance of the Bayesian logistic model or other Bayesian models relative to frequentist methods.

Conclusion

In summary, using logistic regression within the Bayesian paradigm allows us to obtain the posterior distribution for the indirectly estimated marginal RR as well as other quantities of interest that are functions of the adjusted risks. The Bayesian logistic regression modeling approach has several practical advantages. A major practical advantage of Bayesian modeling is that credible intervals for adjusted risks, risk differences, and risk ratios can be interpreted as probability intervals. Unlike frequentist confidence intervals, Bayesian credible intervals do not rely on large-sample approximations and are therefore appropriate for small sample sizes [33]. Another advantage is that available prior information about the regression coefficients can be incorporated in the Bayesian model. Last, multilevel data or models are particularly suited to the hierarchical structure of Bayesian modeling.

Disclaimer

The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Acknowledgements

We thank Lillian Lin whose comments on a previous version of this paper significantly improved framing of the Bayesian logistic regression modeling approach and description of the methods for estimation of RR.

References

1. Schmidt CO, Kohlmann T (2008) When to use the odds ratio or the relative risk. Int J Public Health 53: 165-167.
2. Deddens JA, Petersen MR (2008) Approaches for estimating prevalence ratios. Occup Environ Med 65: 481, 501-506.
3. Lee J (1981) Covariance adjustment of rates based on the multiple logistic regression model. J Chronic Dis 34: 415-426.
4. McNutt LA, Wu C, Xue X, Hafner JP (2003) Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol 157: 940-943.
5. Thompson ML, Myers JE, Kriebel D (1998) Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med 55: 272-277.
6. Wilcosky TC, Chambless LE (1985) A comparison of direct adjustment and regression adjustment of epidemiologic measures. J Chronic Dis 38: 849-856.
7. Yelland LN, Salter AB, Ryan P (2011) Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. Int J Biostat 7: 1-31.
8. Cummings P (2009) The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med 163: 438-445.
9. Dwivedi AK, Mallawaarachchi I, Lee S, Tarwater P (2014) Methods for estimating relative risk in studies of common binary outcomes. Journal of Applied Statistics 41: 484-500.
10. Flanders WD, Rhodes PH (1987) Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. J Chronic Dis 40: 697-704.
11. Localio AR, Margolis DJ, Berlin JA (2007) Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. Journal of Clinical Epidemiology 60: 874-882.

12. Santos CA, Fiaccone RL, Oliveira NF, Cunha S, Barreto ML, et al. (2008) Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. BMC Medical Research Methodology 8: 80.

13. Chu H, Cole SR (2010) Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. Epidemiology 21: 855-862.

14. Petersen MR, Deddens JA (2008) A comparison of two methods for estimating prevalence ratios. BMC Med Res Methodol 8: 9.

15. Spiegelman D, Hertzmark E (2005) Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol 162: 199-200.

16. Jewell NP (2004) Control of Extraneous Factors, In: Jewell, NP. Statistics for Epidemiology, Boca Raton, FL: Chapman and Hall/CRC.

17. Greenland S (1998) Introduction to Regression Models. In: Rothman KJ, Greenland S. Modern Epidemiology. (2nd Edition), Lippincott-Raven Publishers, Philadelphia.

18. Zou G (2004) A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 159: 702-706.

19. Dai L, Li Y, Shen Y (2012) Truncated Estimate in Log-Binomial Model: Algorithm and Simulation. American Journal of Biostatistics 2: 20-25.

20. Savu A, Liu Q, Yasui Y (2010) Estimation of relative risk and prevalence ratio. Stat Med 29: 2269-2281.

21. Rosenbaum PR, Rubin DB (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70: 41-55.

22. Barros AJD, Hirakata VN (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.

23. Zou GY, Donner A (2013) Extension of the modified Poisson regression model to prospective studies with correlated binary data. Stat Methods Med Res 22: 661-670.

24. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis. (2ndedn), Chapman and Hall, London.

25. Bieler GS, Brown GG, Williams RL, Brogan DJ (2010) Estimating model-adjusted risks, risk differences, and risk ratios from complex survey data. Am J Epidemiol 171: 618-623.

26. SAS/STAT 9.3, SAS Institute Inc., USA.

27. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0.

28. Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7: 434-455.

29. Research Triangle Institute (2012) SUDAAN Language Manual. Research Triangle Park, NC: Research Triangle Institute.

30. Skov T, Deddens J, Petersen MR, Endahl L (1998) Prevalence proportion ratios: estimation and hypothesis testing. Int J Epidemiol 27: 91-95.

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

32. Williamson T, Eliasziw M, Fick GH (2013) Log-binomial models: exploring failed convergence. Emerg Themes Epidemiol 10: 14.

33. Dunson DB (2001) Commentary: practical advantages of Bayesian analysis of epidemiologic data. Am J Epidemiol 153: 1222-1226.

Supplemental Material

S1 Appendix. The SAS programs and data for all models used in Examples I and II are presented.

Example I

```
data temp1;
input s e response count;
datalines;
1 0 1 320
1 0 0 480
1 1 1 400
1 1 0 100
2 0 1 300
2 0 0 700
2 1 1 300
2 1 0 200
;
run;


/* Data preparation for robust Poisson SE, creating a subject level variable */
data temp2;
set temp1;
sub=_N_;
run;

/* Stratified Mantel-Haenszel*/
procfreqdata=temp2;
table s*e*response / chisqrelriskcmh;
exactpchior;
weight count;
run;


/* Logistic Marginal Method */
data temp3;
set temp2;
do i=1to count;
output;
end;
```

```
drop i count sub ;run;

odstraceon;
sasfile temp3 load;
procsurveyselectdata=temp3 out=outboot
seed=017246
method=urs
samprate=1
outhits
rep=2000;
strata s;
run;
sasfile temp3 close;

procsortdata=outboot;
by replicate;
run;
odstraceoff;

procgenmoddata=outbootdescending;
by replicate;
class e s;
model response = s e / d=b link=logit;
estimate'Odds ratio for Exposed vs Non-Exposed'e -11 / exp;
outputout=betaxxbeta=eta;
odsoutputparameterestimates=parms;
run;
odsoutputclose;



databetax;
setbetax;
_NAME_="Estimate";
run;
data estimates;
setparms;
by replicate;
if level1=""then level1="1";
```

```
keep replicate parameter level1 estimate;
run;

proctransposedata=estimates out=parms2;
by replicate;
id parameter level1;
run;

data temp4;
mergebetax parms2;
by replicate _NAME_;
if s=1thendo;
pe0 = 1/(1+exp(-1*(intercept1+s1+e0)));
pe1 = 1/(1+exp(-1*(intercept1+s1+e1)));
end;
if s=2thendo;
pe0 = 1/(1+exp(-1*(intercept1+s2+e0)));
pe1 = 1/(1+exp(-1*(intercept1+s2+e1)));
end;

drop scale1;
run;

procmeansdata=temp4;
by replicate;
var pe0;
outputout=temp5 mean=pe0;
run;

procmeansdata=temp4;
by replicate;
var pe1;
outputout=temp6 mean=pe1;
run;

data temp7;
merge temp5 temp6;
by replicate;
 RR=pe1/pe0;
```

```
drop _type_ _freq_;
procprint;
run;



procunivariatedata=temp7;
varrr;
outputout=final pctlpts=2.5, 50.0, 97.5pctlpre=rr;
procprintdata=final;
run;


/* Standard Poisson Model*/
procgenmoddata=temp2;
class e s;
model response = s e / d=plink=log;
estimate'Risk Ratio for Exposed vs Non-Exposed'e -11 / exp;
freq count;
outputout=example p=p;
run;

/* Robust Poisson Model*/
procgenmoddata=temp2;
class e s sub;
model response = s e / d=plink=log;
estimate'Risk Ratio for Exposed vs Non-Exposed'e -11 / exp;
freq count;
repeatedsubject=sub / type=un;
outputout=example p=p;
run;

/* Log-Binomial Model*/
procgenmoddata=temp2 descending;
class e s;
model response = s e / d=b link=log;
freq count;
estimate'Risk Ratio for Exposed vs Non-Exposed'e -11 / exp;
outputout=example p=p;
```

```
run;

/* Bayes logistic model*/
procgenmoddata=temp2 descending;
class e s;
model response = s e / d=b link=logit type3;
freq count;
bayesseed=1234outpost=bayes_probinitialmlenbi=5000seed=8847563nmc=100000thinning=10;
run;

procprintdata=bayes_prob(obs=10);run;

data post; setbayes_prob;
p1=exp(Intercept + s2 + e1)/(1+exp(Intercept + s2 + e1));
p2= exp(Intercept + s2 + e0)/(1+exp(Intercept + s2 + e0));
p3=exp(Intercept + s1 + e1)/(1+exp(Intercept + s1 + e1)) ;
p4=exp(Intercept + s1 + e0)/(1+exp(Intercept + s1 + e0));
r1= (1500/2800)*p1 + (1300/2800)*p3;
r2= (1500/2800)*p2 + (1300/2800)*p4;
RR=r1/r2;

r1_s2= (500/1500)*p1;
r2_s2= (1000/1500)*p2;

run;


*mean estimate;
procmeansdata=post meanmedianminmax ;
var RR r1_s2 r2_s2;
run;

*interval estimate;
procunivariatedata=post noprint;
var RR;
outputout=interval1  pctlpts= 2.5, 97.5pctlpre=lower upper;
procprintdata=interval1;run;
```

```
/* Bayes logistic model*/

procgenmoddata=temp2 descending;
class e s;
model response = s|e / d=b link=logit type3;
freq count;
bayesseed=1234outpost=bayes_probinitialmlenbi=5000seed=8847563nmc=10000 ;*thinning=10;
run;

procprintdata=bayes_prob(obs=10);run;

data post; setbayes_prob;
p_s2e1 = exp(Intercept + s2 + e1 + e1s2)/(1+exp(Intercept + s2 + e1 + e1s2));
p_s2e0 = exp(Intercept + s2 + e0 + e0s2)/(1+exp(Intercept + s2 + e0 + e0s2));
p_s1e1 = exp(Intercept + s1 + e1 + e1s1)/(1+exp(Intercept + s1 + e1 + e1s1)) ;
p_s1e0 = exp(Intercept + s1 + e0 + e0s1)/(1+exp(Intercept + s1 + e0 + e0s1));

RR_s2 = p_s2e1/p_s2e0;
RR_s1 = p_s1e1/p_s1e0;

Diff = RR_s1 - RR_s2;

run;

*mean estimate;
procmeansdata=post meanmedianminmax ;
var RR_s1 RR_s2 p_s1e1 p_s1e0 p_s2e1 p_s2e0 diff;
run;

*interval estimate;
procunivariatedata=post noprint;
var RR_s1 RR_s2;
outputout=interval1  pctlpts= 2.5, 97.5pctlpre=lower upper;
run;

odsgraphicson / width=3.25reset=allborder=offoutputfmt=gifimagemap=on;
odshtmlfile="RR_Thompson.html"style=journal
gpath="XXX: \Bayesian_Logistic\Document\program\Final"
;
```

```
proctemplate;
definestylemystyle ;
parent=styles.htmlblue;
stylegraphwalls from graphwalls / frameborder=off;
end;
run;
odshtmlstyle=mystyle;
procsgplotdata=post noautolegend ;
density diff / type=kernelscale=percent;
xaxisvalues=(-1.0to1.0by0.25) label="Difference between Stratum Specific Risk Ratios";
yaxisvalues=(0to12by2) label="Posterior  Distribution (Percent)";
title;
run;

odshtmlstyle=htmlblue;
odsgraphicson / reset=all;
odsgraphicsoff;

/*Bayes, Lob-Binomial*/

procmcmcdata=temp1 nbi=5000nmc=1000000thin=10propcov=quanewdiag=(mcseess) outpost=mcmc_outseed=1234;
parms (alpha0 alpha1 alpha2) -0.5;
prior alpha0 alpha1 alpha2 ~normal (0,var=10000);
     p=exp(alpha0+alpha1*s+alpha2*e);
model y ~binomial(n,p);

run;

data temp2;
setmcmc_out;
RR = exp(alpha2);
run;
procunivariatedata=temp2;
varrr;
outputout=temp3 pctlpts=2.5, 50.0, 97.5pctlpre=rr;
run;
procprintdata=temp3;
run;
```

```
/* Logistic Regression Model*/
procgenmoddata=temp2 descending;
class e s;
model response = s e / d=b link=logit;
freq count;
estimate'Odds ratio for Exposed vs Non-Exposed'e -11 / exp;
outputout=example p=p;
run;
```

Example II

```
data lee;
input therapy eod age y;
datalines;
0 0 20 1
0 0 23 1
0 0 22 0
0 0 26 0
0 0 29 0
0 0 34 0
0 0 32 1
0 0 30 0
0 0 38 0
0 0 37 0
0 0 38 1
0 1 25 1
0 1 24 0
0 1 25 0
0 1 29 0
0 1 32 0
0 1 34 0
0 1 37 0
0 1 40 0
0 1 40 0
1 0 20 1
1 0 24 1
1 0 28 1
1 0 30 1
```

```
1 0 32 1
1 0 33 0
1 0 38 1
1 0 36 0
1 1 24 0
1 1 26 1
1 1 29 1
1 1 34 0
1 1 32 0
1 1 34 1
1 1 33 1
1 1 36 0
1 1 38 0
1 1 39 0
1 1 38 1
1 1 40 1
;
run;


/* Recoding the 0 and 1 for the therapy outcome variable */
data Leeb;
set lee;
therapy2=1-therapy;
eod=1-eod;
run;



/* Standard Poisson Model: Example II Table 4 Results */
Title1 "Standard Poisson Regression using Lee's Data";
procgenmod data=drew descending;
class therapy eod;
model y = therapy eod age/ d=plink=log type3 ;
estimate 'New vs Old Therapy' therapy -11 / exp;
outputout=logit p=p;
run;
```

```
/*Stratified MH approach*/
procfreqdata=Leeb;
table age*eod*therapy2*y / listmissingrelriskcmh;
exactpchior;
run;


/* Standard Logistic Model */
Title1"Standard Logistic Regression using Lee's Data";
procgenmoddata=lee descending;
class therapy eod;
model y = therapy eod age/ d=b link=logit type3 ;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logit p=p;
run;

/* These blocks of code are for estimating the 95% CI using the marginal method
and logistic regression */
/* This is the code for the Logistic RR reaults using bootstrapping */
/* Create the bootstrap datasets for Table 4 Lee's data example*/
odstraceon;
sasfile lee load;
procsurveyselectdata=lee out=outboot
seed=17246
method=urs
samprate=1
reps=5000
outhits;
strata therapy;
run;
sasfile lee close;

procsortdata=outboot;
by replicate;
run;


odstraceoff;

/* Logistic Marginal Regression Model Table 4 Example
```

```
      The marginal estimate for RR comes from the original
data and CI from the bootstrap samples */
procgenmoddata=outbootdescending;
by replicate;
class therapy eod;
model y = therapy eod age / d=b link=logit;
outputout=betaxxbeta=eta;
odsoutputparameterestimates=parms;
run;
odsoutputclose;

/* Beta linear predictor output, creating _name_ for merging */
databetax;
setbetax;
_NAME_="Estimate";
run;

/* Parameter estimate for estimating the marginal probabilities for therapy new and conventional */
data estimates;
setparms;
by replicate;
if level1=""then level1="1";
keep replicate parameter level1 estimate;
run;

proctransposedata=estimates out=parms2;
by replicate;
id parameter level1;
run;

/* Merge data with parameter estimates to estimate p for therapy new and conventional */
data temp4;
mergebetax parms2;
by replicate _NAME_;

pe0 = 1/(1+exp(-1*(intercept1+therapy0+eod0*(1-eod)+age1*age)));
pe1 = 1/(1+exp(-1*(intercept1+therapy1+eod0*(1-eod)+age1*age)));
drop scale1;
run;
```

```
/* creating a dataset to determine if there are any missing y 0 / 1 for the combinations of therapy and eod */
procfreqdata=temp4 noprint;
tables y*therapy*eod / listmissingout=new;
by replicate;
run;

procmeansdata=new noprint;
by replicate;
outputout=new2 n=total;
data new2;
set new2;
keep replicate total;
run;

data new3;
merge new2 new;
by replicate;
run;

data temp4b;
merge temp4 new3;
by replicate;
if total ne 8thendelete;
run;

procmeansdata=temp4b noprint;
by replicate;
var pe0;
outputout=temp5 mean=pe0;
run;

procmeansdata=temp4b noprint;
by replicate;
var pe1;
outputout=temp6 mean=pe1;
run;

data temp7;
```

```
merge temp5 temp6;
by replicate;
 RR=pe1/pe0;

drop _type_ _freq_;
*proc print;
run;

/* The final bootstrap 95% CI, the estimate is 2.79 using the original data */
procunivariatedata=temp7;
varrr;
outputout=final pctlpts=2.5, 50.0, 97.5pctlpre=rr;
procprintdata=final;
run;



/* Standard Poisson Model */
Title1"Standard Poisson Regression using Lee's Data";
procgenmoddata=lee descending;
class therapy eod;
model y = therapy eod age/ d=plink=log type3 ;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logit p=p;
run;


/* Preparing data for Poisson robust SE model */
data lee2;
set lee;
sub=_N_;
run;

/* Robust Poisson Model: Matches Lee's Results */
Title1"Robust Poisson Regression using Lee's Data";
procgenmoddata=lee2;
class therapy eod sub;
model y = therapy eod age / d=plink=log ;
estimate'New vs Old Therapy' therapy -11 / exp;
repeatedsubject=sub / type=un;
```

```
outputout=pois2 p=p;
run;



/* Log-Binomial Model that Doesn't Converge */
Title1"Standard Log-Binomial Regression using Lee's Data";
procgenmoddata=lee descending;
class therapy eod;
model y = therapy eod age  / d=b link=log type3 ;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logit p=p;
run;

/* Prepare data for log-binomial COPY method model */
DATA ONE; SET lee; W=.9999;
DATA TWO; SET lee; Y=1-Y; W=.0001;
DATA THREE; SET ONE TWO;
run;

/*Log-Binomial model using COPY method: Example II Table 4 Results */
Title1"COPY Method Log-Binomial Regression using Lee's Data";
PROCGENMODDATA=THREE descending;
WEIGHT W;
class therapy eod;
MODEL Y=therapy eod age/D=BIN LINK=LOG
LRCI;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logbp=p;
run;

####R code for Log-Binomial (IWLS truncated with three sets of initial values)

#loading the data

data<-as.matrix(read.csv("XXX: /Final Manuscript/Final Analysis Programs/Final/Lee_data.csv", header=F))
```

```
x<-data[,1:4]

y<-data[,5]


#IWLS Method 1, using the fitted coefficient values for the COPY model without age and starting the coefficient of age at zero


iter<- 0

n<-40

bhat0<-matrix(c(-0.7706,0.9095,-0.5139,0),4)


repeat{

iter<-iter+1

eta<- x %*% bhat0

etaexp<-exp(eta)

etaexp[etaexp>= 1] <- 0.9999

etaexp[etaexp<=0.0001] <- 0.0001

z0<-eta+(y-etaexp) / etaexp

w0<-diag(as.vector(t(etaexp / (1-etaexp))))

bhat1 <- solve(t(x) %*% w0 %*% x) %*% t(x) %*% w0 %*% z0

if(all((bhat1-bhat0) ^ 2<=10^-6)){break}

bhat0<-bhat1

}


se<-sqrt(diag(solve(t(x) %*% w0 %*% x)))
```

```
exp.bhat1<-exp(bhat1)

exp.se<-exp(se)

lower<-exp(bhat1-1.96*se)

upper<-exp(bhat1+1.96*se)


exp.bhat1

lower

upper



#IWLS Method 2, using the fitted coefficient values for the COPY model without age and starting the coefficient of age at zero


#raw proportions

#Therapy: model p=1, p(recovery=1|therapy=1)=12/20=0.6, ln(p)=beta0+beta1, beta0=ln(p(recovery=1))=ln(17/40)=-0.85567,

#beta1=ln(p)-beta0=-1.9861+0.85567=-1.13043

#EOD:model P=1, when recovery=1, 7/21=p(recovery=1|EOD=1), ln(p)=beta0+beta2, beta0=ln(p(recovery=1))=ln(17/40)=-0.85567,

#so ln(7/21)=-0.85567+beta2, beta2=-0.51083+0.85567=0.34484

#raw proportions

#bhat0<-matrix(c(-0.85567,0.34484,-1.13043,0),4)


iter<- 0

n<-40

bhat0<-matrix(c(-0.85567,0.53063,-0.213812,0),4)
```

```
repeat{
iter<-iter+1
eta<- x %*% bhat0
etaexp<-exp(eta)
etaexp[etaexp>= 1] <- 0.9999
etaexp[etaexp<=0.0001] <- 0.0001
z0<-eta+(y-etaexp) / etaexp
w0<-diag(as.vector(t(etaexp / (1-etaexp))))
bhat1 <- solve(t(x) %*% w0 %*% x) %*% t(x) %*% w0 %*% z0
if(all((bhat1-bhat0) ^ 2<=10^-6)){break}
bhat0<-bhat1
}

se<-sqrt(diag(solve(t(x) %*% w0 %*% x)))

exp.bhat1<-exp(bhat1)
exp.se<-exp(se)
lower<-exp(bhat1-1.96*se)
upper<-exp(bhat1+1.96*se)

exp.bhat1
lower
upper
```

```
#IWLS Method 3, staring all coefficients at zero

iter<- 0

n<-40

bhat0<-matrix(c(0,0,0,0),4)


repeat{

iter<-iter+1

eta<- x %*% bhat0

etaexp<-exp(eta)

etaexp[etaexp>= 1] <- 0.9999

etaexp[etaexp<=0.0001] <- 0.0001

z0<-eta+(y-etaexp) / etaexp

w0<-diag(as.vector(t(etaexp / (1-etaexp))))

bhat1 <- solve(t(x) %*% w0 %*% x) %*% t(x) %*% w0 %*% z0

if(all((bhat1-bhat0) ^ 2<=10^-6)){break}

bhat0<-bhat1

}


se<-sqrt(diag(solve(t(x) %*% w0 %*% x)))


exp.bhat1<-exp(bhat1)

exp.se<-exp(se)

lower<-exp(bhat1-1.96*se)

upper<-exp(bhat1+1.96*se)

exp.bhat1
```

```
lower

upper


/*IPTW Propensity Score Approach: Estimating weights using logistic regression */
procgenmoddata=lee descending;
class therapy eod;
model therapy = eod age / d=b link=logit type3 ;
outputout=logit p=p;
run;

/* Using the logistic regression results to define the PS weights */
data logit2;
        set logit;
        if therapy=1then weight=0.5/p;
                elseif therapy=0then weight=0.5/(1-p);
run;
procprintdata=logit2 noobs; run;

/* IPTW PS Approach estimated RR for Table 4 */
procgenmoddata=logit2 descending;
class therapy eod;
weightweight;
model y = therapy/ d=b link=log type3 ;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logit3 p=p;
run;


/* The IPTW Strata Weights Table 4 Method: Weights calculated using crude weights strata */
***crude weight, based on P(x1|x2,x3);
procfreqdata=lee; tables therapy; where age=40 and eod=1;run;

data lee3;
        input therapy eod age y     p_denominatorweight_crude;
cards;
0       0       20      1       0.5     1
0       0       23      1       1       0.5
0       0       22      0       1       0.5
```

```
0	0	26	0	1	0.5
0	0	29	0	1	0.5
0	0	34	0	1	0.5
0	0	32	1	0.5	1
0	0	30	0	0.5	1
0	0	38	0	0.667	0.749625187
0	0	37	0	1	0.5
0	0	38	1	0.667	0.749625187
0	1	25	1	1	0.5
0	1	24	0	0.5	1
0	1	25	0	1	0.5
0	1	29	0	0.5	1
0	1	32	0	0.5	1
0	1	34	0	0.333	1.501501502
0	1	37	0	1	0.5
0	1	40	0	0.667	0.749625187
0	1	40	0	0.667	0.749625187
1	0	20	1	0.5	1
1	0	24	1	1	0.5
1	0	28	1	1	0.5
1	0	30	1	0.5	1
1	0	32	1	0.5	1
1	0	33	0	1	0.5
1	0	38	1	0.333	1.501501502
1	0	36	0	1	0.5
1	1	24	0	0.5	1
1	1	26	1	1	0.5
1	1	29	1	0.5	1
1	1	34	0	0.667	0.749625187
1	1	32	0	0.5	1
1	1	34	1	0.667	0.749625187
1	1	33	1	1	0.5
1	1	36	0	1	0.5
1	1	38	0	1	0.5
1	1	39	0	1	0.5
1	1	38	1	1	0.5
1	1	40	1	0.333	1.501501502
;
run;
```

```
/* IPTW PS Strata Weight Table 4 Results */
procgenmoddata=lee3descending;
class therapy eod;
weightweight_crude;
model y = therapy/ d=b link=log type3 ;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logit4 p=p;
run;



/* Code below is for the Bayesian marginal method implemented using GENMOD */
/* Analysis for results presented in Table 4 Example II for Bayesian */
Datalee4(rename=(age=age0));
set lee;
id = _N_;
do replicate=1to100000;
output;
end;
procsort; by replicate id;
run;



/* Initial values, only first row, MLE, initial values are used by SAS for computing the posterior
   The other initial values are for computing the Gelman stat */
data bob;
input intercept therapy0 therapy1 eod0 eod1 age scale;
datalines;
3.2293 -2.0701 0 1.0767 0 -0.0984 1.0
0 0 0 0 0 0 1.0
1.0 -1.0 0 0 0 0 1.0
;
run;

odsgraphicson;
/* Bayes logistic model for Thompson Example I
   This is the full saturated model for looking at stratum specific differences
   We process the stratum specific RR to graph the difference in Figure 1 of the difference */
procgenmoddata=lee descending;
```

```
class therapy eod ;*/ param=ref;
model y = therapy eod age/ d=b link=logit;
bayesseed=7854outpost=bayes_probinitial=bob nbi=5000nmc=1000000coeff=jeffreys(conditional)
        thinning=10samp=gamermanplotsdiagnostics=(autocorressgelman(n=3));
run;
odsgraphicsoff;

data lee5;
set lee4;
sub=replicate;
procsort; by replicate;
run;

databayes_prob;
setbayes_prob;
sub=_N_;
run;

/* Merge the data with the Bayes parameter estimates to estimate the probabilities of new
and conventional therapy by subject */
data post1;
mergebayes_prob lee3;
by sub;
p1=exp(Intercept + therapy0 + eod0*eod + age*age0)/(1+exp(Intercept + therapy0 + eod0*eod + age*age0));
p2=exp(Intercept + therapy1 + eod0*eod + age*age0)/(1+exp(Intercept + therapy1 + eod0*eod + age*age0));
run;

procsortdata=post1;
by replicate;
run;

procmeansdata=post1 noprint;
by replicate;
var p1;
outputout=post1a mean=p1;
run;

procmeansdata=post1 noprint;
by replicate;
```

```
var p2;
outputout=post1b mean=p2;
run;

data post1c;
merge post1a post1b;
by replicate;
 RR=p2/p1;
drop _type_ _freq_;
*proc print;
run;

/* Median RR and 95% credible interval */
procunivariatedata=post1c;
var RR;
outputout=final pctlpts=2.5, 50.0, 97.5pctlpre=rr;
procprintdata=final;
run;


odsgraphicson / width=3.25reset=allborder=off/*outputfmt=gif*/imagemap=on;
odshtmlfile="RR_Lee.html"style=journal gpath="XXX:\Estimation of RR\IJE Manuscript\Final Manuscript\";

proctemplate;
definestylemystyle ;
parent=styles.htmlblue;
stylegraphwalls from graphwalls / frameborder=off;
end;
run;

odshtmlstyle=mystyle ;

procsgplotdata=post1c noautolegend;
densityrr / type=kernelscale=percent;
xaxisvalues=(0.0to10.0by1.0) label="Risk Ratio for New versus Conventional Therapy";
yaxisvalues=(0to25by5) label="Posterior  Distribution (Percent)";
title;
run;
```

```
odshtmlstyle=htmlblueclose;
odsgraphicson / reset=all;
odsgraphicsoff;


/*Bayesian Log-Binomial*/

procmcmcdata=leenbi=5000nmc=2000000thin=10propcov=quanewdiag=(mcseess) outpost=mcmc_outseed=1234;
parms (alpha0 alpha1 alpha2 alpha3) -0.5;
prior alpha0 alpha1 alpha2 alpha3 ~normal (0,var=10000);
    p=exp(alpha0+alpha1*therapy+alpha2*age+alpha3*eod);
model y ~binary(p);
run;


data post; setmcmc_out;

c1=exp(alpha0+alpha1*0+alpha2*20+alpha3*0);
c2=exp(alpha0+alpha1*0+alpha2*38+alpha3*0);
c3=exp(alpha0+alpha1*0+alpha2*25+alpha3*1);
c4=exp(alpha0+alpha1*0+alpha2*40+alpha3*1);
c5=exp(alpha0+alpha1*1+alpha2*20+alpha3*0);
c6=exp(alpha0+alpha1*1+alpha2*38+alpha3*0);
c7=exp(alpha0+alpha1*1+alpha2*24+alpha3*1);
c8=exp(alpha0+alpha1*1+alpha2*40+alpha3*1);
run;


data post2; set post;
if c1>1 or c2>1 or c3>1 or c4>1 or c5>1 or c6>1 or c7>1 or c8>1thendelete;
run;

*randome select 100000 samples to summarize;
procsurveyselectdata=post2 method=srsn=100000reps=1seed=1234out=SampleRep;
run;

data SampleRep2;
setSampleRep;
RR = exp(alpha1);
```

```
run;

procunivariatedata=SampleRep2;
varrr;
outputout=temp3 pctlpts=2.5, 50.0, 97.5pctlpre=rr;
run;

procprintdata=temp3;run;


/* Standard Logistic Model */
Title1"Standard Logistic Regression using Lee's Data";
procgenmoddata=lee descending;
class therapy eod;
model y = therapy eod age/ d=b link=logit type3 ;
estimate'New vs Old Therapy' therapy -11 / exp;
outputout=logit p=p;
run;
```