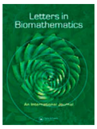
LETTERS IN BIOMATHEMATICS
https://doi.org/10.1080/23737867.2019.1654418



Taylor & Francis
Taylor & Francis Group

Model justification and stratification for confounding of *Chlamydia trachomatis* disease

Sarada Ghosh and G. P. Samanta

Department of Mathematics, Indian Institute of Engineering Science and Technology, Howrah, India

ABSTRACT

This study involves statistical analysis of reported cases of sexually transmitted diseases (STDs) of Chlamydia infection in the United States. The data are collected from 2007 to 2016. The research studies incidence of sexually transmitted diseases and survival among different age groups and gender and race factors which influence the incidence in the target population. In this work, log-binomial, logit model, probit model and complementary log–log model are used to establish a suitable model (using different criteria) that can predict the survival of infected people with STDs based on their age, gender and race. Here we have also focused on stratification: a statistical technique that allows to control for confounding by creating two or more categories. The Mantel–Haenszel formula allows to calculate an overall, unconfounded, that is adjusted, effect estimate for a specific outcome by combining (pooling) stratum-specific relative risks and odds ratios. Simulation is based on *R*-software.

ARTICLE HISTORY
Received 26 September 2018
Accepted 27 July 2019

KEYWORDS
Chlamydia trachomatis; binomial regression; linear models; odds ratio; relative risk

1. Introduction

Nowadays, *Chlamydia trachomatis* (*C. trachomatis*) infections are common among young sexually active people. Sexually transmitted diseases (STDs) are spread out from one people to another through intimate mostly physical contact such as heavy petting and also from sexual activity (i.e. vaginal, oral and anal sex) and it can also be transmitted through other modes such as vertically (from mother to child). Centers for Disease Control (CDC) estimates 20 million new infections occur every year in the United States. There are dozens of STDs. Some STDs: gonorrhea, syphilis and chlamydia are spread through sexual contact. Other diseases, including Zika and Ebola can also be spread sexually but are more often develop through ways other than sex. STDs can mostly be prevented by not having sex without protections. People can lower their risk by using condoms and being in a sexual relationship with a partner who does not have an STD. Since, STDs do not always cause symptoms, therefore, it is possible to have an infection without knowing it. So, it is most important to get tested if anyone having sex. If people are diagnosed with an STD, then all can be treated with medicine and some of them can be cured entirely (Samanta, 2015). But generally it occurs without symptoms and often goes undiagnosed. It is shown that up to

CONTACT Sarada Ghosh ✉ saradaghosha111@gmail.com Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, 711103, India

40% of females with untreated Chlamydia infections develop pelvic inflammatory disease (PID), a condition which can lead to such long-term complications as infertility, ectopic pregnancy and chronic pelvic pain (Scholes et al., 1996). In pregnant women, it may lead to premature delivery and babies born to infected mothers can get infections in their eyes, which is called conjunctivitis or pinkeye, as well as pneumonia. Among men the difficulties from Chlamydia are relatively uncommon but may include epididymitis and urethritis which can cause pain, fever and in rare cases, sterility. Men who are affected in Chlamydia might have a discharge from the penis and a burning sensation when urinating which may range from clear to pussy as symptom. Men might also have burning and itching around the opening of the penis or pain and swelling in the testicles/scrotum or both which can be a sign of epididymitis, an inflammation of a part of the male reproductive system located in the testicles. PID and epididymitis both can result in infertility. In addition, investigations point out that the presence of Chlamydia infection also increases the risk the transmission of HIV. It is also called as a 'silent' disease due to at least 50% of infected men and 75% of infected women have no symptoms. The primary focus of Chlamydia screening efforts among women should be to detect Chlamydia, prevent complications and diagnose (then treat) their partners. Whereas targeted Chlamydia screening in men should only be considered when resources permit, prevalence is high, and such screening does not hinder Chlamydia screening efforts in women. In United States, it is mainly highlighted that trends and distribution of STDs in populations of particular interest to Chlamydia prevention programs in state and local health departments: women and infants, adolescents and young adults, racial and ethnic minority groups and gay and bisexual men and other men who have sex with men (Kimberly and Gail 2015). These populations are most vulnerable to STDs and their consequences and often lack adequate access to health care services. Age was strongly associated with having health insurance in the United States: (i) older adults (65 years and older) and children (19 years and under) were most likely to have health insurance, (ii) working adults (19–64 years) had higher uninsured rates. The Patient Protection and Affordable Care Act (ACA) aims to increase access to sexual and reproductive health services through reforms based on the United States. As per the recommendations of Preventive Services Task Force, Chlamydia screening (for sexually active women under 25 and all higher risk women) should be performed. However, although health insurance coverage has been expanded for most groups, evidence suggests that disparities in health insurance coverage and access to STD services remain.

Many people who have Chlamydia do not develop symptoms, but they can still infect others through sexual contact. Women, especially young and minority women, are hit hardest by *Chlamydia trachomatis* and generally women are most severely impacted by the long-term consequences of untreated Chlamydia. In the United States, the reported Chlamydia case rate for females in 2012–2016 was almost two and half times higher than for males. But before 2012, reported Chlamydia case rate for males in 2007–2011 was almost one and half times higher than for females. From 2012–2016, young females 20–24 years of age had the highest Chlamydia rate followed by females 15–19 years of age and young males from 20 to 24 years had most affected in Chlamydia diseases followed by males 15 to 19 years of age.

In United states, the annual cost for the treatment due to Chlamydia is near about 2.5 billion dollars. Apart from this, another important fact about Chlamydia is that infected individuals can acquire re-infection while recovering from the disease and often proceeds in situations where infected individuals have more than one sex partners. Although Chlamydia-related mortality is negligible in comparison to other STDs such as HIV/AIDS, the aforementioned Chlamydia-associated irreversible complications makes Chlamydia a disease of utmost public health significance (as per link https://www.cdc.gov).

Treating persons infected with *C. trachomatis* prevents adverse reproductive health complications and continued sexual transmission and treating the sex partners can prevent re-infection and infection of other partners. Chlamydia is a common curable STD. But, if left untreated, Chlamydia can make it troublesome for a woman to get pregnant. Treating pregnant women generally prevents transmission of *C. trachomatis* to new born baby during birth. Chlamydia treatment should be provided promptly for all persons testing positive for infection, treatment delays have been associated with complications (e.g. PID) in a limited proportion of women. A meta-analysis of 12 randomized clinical trials of azithromycin versus doxycycline for the treatment of urogenital chlamydial infection demonstrated that the treatments were equally efficacious, with microbial cure rates of 97% and 98%, respectively. These studies were conducted primarily in populations with urethral and cervical infection in which follow-up was encouraged, devotion to a 7-day treatment was effective. It is mentioned that the detection of *C. trachomatis* from an oropharyngeal specimen should be treated with azithromycin or doxycycline. Nowadays retrospective studies have raised concern about the efficacy of azithromycin for rectal *C. trachomatis* infection. But these studies have some limitations, and prospective clinical trials comparing azithromycin versus doxycycline regimens for rectal *C. trachomatis* infection are needed. Statistical model are becoming important tools in analysing the spread and control of infectious diseases. *Chlamydia trachomatis* causes more cases of STDs than any other bacterial pathogen throughout the World, making a major public health problem. Data indicate that performance of NAATs on self-collected rectal swabs is comparable to clinician-collected rectal swabs, and this specimen collection strategy for rectal *C. trachomatis* screening is highly acceptable. Self-collected rectal swabs are a reasonable alternative to clinician-collected rectal swabs for *C. trachomatis* screening by NAAT, especially when clinicians are not available or when self collection is preferred over clinician collection. Although Chlamydia is a disease of significant public health importance, not much has been analysed in terms of using statistical modelling to gain insight into its transmission dynamics at population level (from the link https://www.cdc.gov). In this work, we have fit different types of models of Chlamydia affected people in the United States and choose the appropriate one among those models by distinct criteria. We also have discussed on stratification: a statistical technique that allows to control for confounding by creating two or more categories. The Mantel–Haenszel (MH) formula allows to calculate an overall. The simulation of this work is based on *R*-software. Finally, the last section consists of the general discussion and conclusion of the work.

2. Materials

Nationally notifiable STD surveillance data are collected and compiled from reports sent by the STD control programs and health departments in all 50 states, the District of Columbia, selected cities U.S. dependencies and possessions, and independent nations in free association with the United States to the Division of STD Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and

Prevention (CDC) (Workowski & Bolan, 2015). Included among the dependencies possessions, and independent nations are Guam, Puerto Rico, and the Virgin Islands. These entities are identified as 'outlying areas' of the United States in selected tables as given in the following link: https://www.cdc.gov.

3. Preliminaries

3.1. Basic concepts of regression analysis

The logistic model (or logit model) is a statistical model which is usually taken to apply on a binary dependent variable. The logistic regression is the most important model for ordered categorical response data (Anderson, 1984; McCullagh, 1980). It is used increasingly in a wide variety of applications. It is not only used in biomedical studies but also rapidly used in social science research and marketing in the past 20 years. Apart from this, another area of increasing application is genetics. In statistics, binomial regression is a technique in which the response (often referred to as Y) is the result of a series of Bernoulli trials or a series of one of two possible disjoint outcomes (traditionally denoted as 'success' or 1, and 'failure' or 0). The log-binomial model is simply a binomial generalized linear model (GLM) with a log link function. It is particularly useful (or, popular) in biostatistical and epidemiological applications as an alternative to logistic regression.

For a binary response variable Y and X be explanatory variable, let $\pi(x) = P(Y = 1|X = x)$. The logistic regression model is

$$\pi(x) = \frac{\exp(\theta + \beta x)}{1 + \exp(\theta + \beta x)} \tag{1}$$

Equivalently, the log odds, called the logit, has the linear relationship:

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \theta + \beta x \tag{2}$$

This equates the logit link function to the linear predictor, where θ be the intercept and we want to interpret β in (2). Because the sign of β decides whether $\pi(x)$ is increasing or decreasing when x increases. The rate of ascend or descent increases as $|\beta|$ increases. As $\beta \rightarrow 0$ the curve flattens to a horizontal straight line. When $\beta = 0$, Y is independent of X. $\pi(x)$ approaches 1 at the same rate that it approaches 0 because logistic density is symmetric. If we exponentiate both sides of (2), then the odds become an exponential function of x. It provides a basic interpretation for the magnitude of β. The odds increase multiplicatively by e^{β} for every 1-unit increase in x. There are two alternatives to a logit model: (i) complementary log–log model and (ii) probit model, under the assumption of binary response. Here we fit complementary log–log model to the mentioned data. Both are following the form:

$$\pi(x) = \Phi(\theta + \beta x), \text{ where } \Phi \text{ is continuous cdf.} \tag{3}$$

The complementary log–log model is defined as

$$\log[-\log(1 - \pi(x))] = \theta + \beta x \tag{4}$$

The expression on the left hand side is said to be complementary log–log transformation. It is noted that $\log(1-\pi(x))$ is always a negative number. This is changed to a positive number before taking log at the second time.

Equation (4) can also be written in the following form:

$$\pi(x) = 1 - \exp[-\exp(\theta + \beta x)] \tag{5}$$

To interpret model (4), it is noted that at x_1 and x_2:

$$\log[-\log(1-\pi(x_2))] - \log[-\log(1-\pi(x_1))] = \beta(x_2 - x_1)$$

so that

$$\frac{\log[1-\pi(x_2)]}{\log[1-\pi(x_1)]} = \exp[\beta(x_2 - x_1)] \Rightarrow 1 - \pi(x_2) = [1-\pi(x_1)]^{\exp\beta(x_2-x_1)} \tag{6}$$

For $x_1 - x_2 = 1$, the complement probability at x_2 equals the complement probability at x_1 raised to the power $\exp(\beta)$. From Equation (5), we can also said that $\pi(x)$ approaches 1 at faster rate than approaches 0, i.e. the response of the simple complementary log–log model with one predictor has an S-shaped curve. For the probability of a success, if the complementary log–log model is used, then the log–log model holds for the probability of a failure. Models with log–log links can be fitted using the Fisher scoring algorithm for generalized liner models.

Now we consider the cumulative link model:

$$\Psi^{-1}[P(Y \leq j|x)] = \theta_j + \beta^T x \tag{7}$$

where Ψ^{-1} denotes a link function that is the inverse of continuous cdf Ψ of the latent variable Y^*. The cumulative link model links the cumulative probabilities to the linear predictor. Cumulative link models provide the regression framework familiar from linear models when the response is treating rightfully as categorical (Agresti, 2002, 2007, 2010a, 2010b; Ghosh & Samanta, 2019; Powers & Xie, 2000). While cumulative link models are not the only type of ordinal regression model, they are by far the most popular class of ordinal regression models.

Let the underlying population follows $N(\mu, \sigma^2)$. In GLM form:

$$\Psi^{-1}[\pi(x)] = \theta + \beta x, \text{ where } \theta = -\frac{\mu}{\sigma} \text{ and } \beta = \frac{1}{\sigma} \tag{8}$$

is the probit model. The probit link function is $\Psi^{-1}(\cdot)$.

The logit and probit links are symmetric about 0.5, i.e.

$$link[\pi(x)] = -link[1-\pi(x)]. \tag{9}$$

Now,

$$\begin{aligned} logit[\pi(x)] &= \log\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= -\log\left[\frac{1-\pi(x)}{\pi(x)}\right] \\ &= -logit[1-\pi(x)] \end{aligned}$$

This means that the response curve for $\pi(x)$ has a symmetric appearance about the point where $\pi(x) = 0.5$, so $\pi(x)$ approaches 0 at the same rate it approaches 1. Logit and probit models are inappropriate when this is badly violated (Agresti, 2002, 2010a, 2010b).

In this context, it is mentioned that extreme value distributions are the limiting distributions for the minimum or the maximum of a very large collection of random observations from the same arbitrary distribution. An underlying extreme value distribution for Y^* implies a model in the following form:

$$\log(-\log[1 - P(Y \leq j|x)]) = \theta_j + \beta^T x \tag{10}$$

Sometimes the ordinal model using this link is also called 'proportional hazards model'. The model with complementary log–log link has the following interpretation:

$$P(Y > j|x_1) = P(Y > j|x_2)^{exp[\beta^T(x_1 - x_2)]}. \tag{11}$$

The related log–log link modelled by $\log(-\log[P(Y \leq j)])$ is appropriate when the complementary log–log link holds for the categories listed in reverse order.

A good-fitted model evaluates effects, apart from this it includes relevant interactions and provides smoothed estimates of response probabilities. The residual deviance is a measurement which can decide whether a model is good-fitted or not, further it can also measure 'badness-of-fit' (Blizzard & Hosmer, 2006). In this context, it is mentioned that the log-likelihood is concave for many cumulative link models (including logit, probit, and complementary log–log) (Pratt, 1981). There is an important role of AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) for model selection. Iterative algorithms usually converge rapidly to the maximum likelihood estimates. Estimates of the parameters in the likelihood equation are calculated in an attempt to maximize the likelihood of the observed data. These estimates are then improved once each iteration using a Newton–Raphson method.

3.2. Basic concepts of MH analysis

Now, we have considered a hypothetical case-control study investigating the association between gender and Chlamydia for the mentioned dataset. Stratification is the simplest method for controlling the confounding during data analysis. It also represents the preliminary step for invoking the MH formula and standardization. Here we focus on stratification, a statistical technique that allows to control for confounding by creating two or more categories (strata) in which the confounding variable either does not vary or does not too much (Mantel & Haenszel, 1959). The MH analysis allows to calculate an overall, unconfounded, that is adjusted and also consider the combining (pooling) stratum-specific relative risks (RR) or odds ratios (OR). Stratum-specific RRs or ORs are computed within each stratum of the confounding variable. These are also useful for comparing with the corresponding effect estimates in the whole group (that is, with unstratified RR or OR). 'No confounding' means that the effect estimates are roughly homogeneous across strata and do not differ from that in the whole group. Also, the 'presence of confounding' indicates that the effect estimates are substantially similar across strata but differ from that in the whole group (Maldonado & Greenland, 1993). In this context. it is mentioned that at the time of

Table 1. Data layout for MH analysis (2 × 2 table).

Outcome	Black	White	
Male	a	b	$a+b$
Female	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

comparing between stratified and unstratified effect estimates, epidemiologists consider as relevant a RR or OR difference by more than 10% (Maldonado & Greenland, 1993).

We have considered five steps for assessing confounding through the MH analysis (Mantel & Haenszel, 1959) as follows: (i) compute the crude RR and OR (i.e. without stratifying); (ii) stratify using the confounding variable and compute stratum-specific RR and OR; (iii) determine the homogeneity of the effect estimates across strata and compare stratified and unstratified RRs and ORs; (iv) compute the overall, adjusted RR and OR by the MH formula, if there is homogeneity in effect estimates across strata; and (v) stratum-specific effect estimates should be reported separately, if there is heterogeneity and we have interested on effect modification.

Before computing an MH estimate, it is useful to have a standard layout for the two by two tables in each stratum. We have used the general format as depicted in Table 1.

Using the notations in this table, estimates for a RR and an OR (Mantel & Haenszel, 1959; Miller, 1980) would be computed as follows:

$$\widehat{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

and

$$\widehat{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

To explore and adjust for confounding, a stratified analysis is performed in which we set up a series of two-by-two tables, one for each stratum (category) of the confounding variable. Then the averages for RRs and ORs are computed as follows:

$$\widehat{RR_{MH}} = \frac{\sum_{i=1}^{s} \frac{a_i(c_i+d_i)}{n_i}}{\sum_{i=1}^{s} \frac{c_i(a_i+b_i)}{n_i}}$$

and

$$\widehat{RR_{MH}} = \frac{\sum_{i=1}^{s} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{s} \frac{b_i c_i}{n_i}},$$

where a_i, b_i, c_i and d_i are the numbers of participants in the cells of the two-by-two table in the ith stratum of the confounding variable and n_i represents the number of participants in the ith stratum. For strata i: from 1 to s.

4. The model

In this work, we have considered the age-group distribution for U.S. residents from 2007 to 2016, by race and gender. For gender G ($1=$ female, $0=$ male), race R ($1=$ black, $0=$ white).

It can be easily seen from the dataset that in the United States the proportion of men were affected more than the proportion of women during 2007–2011. After 2011, the proportion of women were affected more than the proportion for men (2012–2016), when race is considered as constant. Given gender, the proportion of blacks were more affected than the proportion for whites throughout 2007–2016. So, based on affected proportion of gender, we have considered year 2007–2011 as Category I and from 2012 to 2016 considered as Category II. Binomial regression is a technique in which the response (often referred to as Y) is the result of a series of Bernoulli trials, or a series of one of two possible disjoint outcomes (traditionally denoted as 'success' or 1, and 'failure' or 0). The log-binomial model is simply a binomial GLM with a log link function. In this analysis the following logit model has been implemented:

$$\text{logit}[P(Y \leq j|G = g, R = r)] = \theta_j + \beta_1 g + \beta_2 r \tag{12}$$

Using Equation (8), the following probit model has been considered:

$$\Psi^{-1}[P(Y \leq j|G = g, R = r] = \theta_j + \beta_1 g + \beta_2 r \tag{13}$$

Lastly, the following cumulative log–log model has been implemented:

$$\log(-\log[1 - P(Y \leq j|G = g, R = r)]) = \theta_j + \beta_1 g + \beta_2 r \tag{14}$$

5. Results

Here we have considered *Sexually Transmitted Disease Surveillance: 2007–16*, which presents statistics and trends for STDs in the United States during this period (Ref. https://www.cdc.gov). The comparison between different regressions are shown in Table 2 (2007–2011) and in Table 3 (2012–2016).

Proceeding next years (from 2012 to 2016):

Table 2. Log-likelihood, residual deviance, AIC, BIC: comparison of log-binomial regression, C log–log regression, Probit regression and Logit regression (for 2007–2011).

	Log-binomial	C log–log	Probit	Logit
Log-Likelihood	−4.11	−291.67	−190.24	−243.13
Residual deviance	5.44	477.2	274.4	380.1
AIC	12.21	595.3	392.5	498.2
BIC	12.37	591.6	388.8	494.6

Table 3. Log-likelihood, residual deviance, AIC, BIC : comparison of log-binomial regression, C log–log regression, Probit regression and Logit regression (for 2012–2016).

	Log-binomial	C log–log	Probit	Logit
Log-Likelihood	−3.38	−192.6	−140.46	−164.63
Residual deviance	3.98	278.9	174.6	222.95
AIC	10.76	397.2	292.9	341.26
BIC	10.91	393.5	289.2	337.57

Table 4. Combinations of RR, OR and MH analysis (stratified by age) used in the simulation study.

	Stratified by age							
	Crude approximation		Age ≤ 29		Age > 29			
	RR	OR	RR	OR	RR	OR	MH adjusted RR	MH adjusted OR
Category I	1.49	1.54	1.30	1.35	1.22	1.26	1.29	1.34
Category II	1.48	1.55	1.31	1.38	1.26	1.32	1.3	1.37

The population under study is stratified into two age categories (below/above 29 years) for confounding effect of age on gender due to the Chlamydia infection. Also, MH analysis has been performed to control for confounding.

The RR of race is homogeneous across strata (RR = 1.30 for infected people having age ≤ 29 years and RR = 1.22 for those aged > 29 years) and unstratified and stratified effect estimates differ by more than 10% (Table 4). To calculate the overall RR adjusting for the confounding effect of age, stratum-specific $\widehat{RR_{MH}}$ are pooled by the MH formula and $\widehat{RR_{MH}}$ =1.29. The RR of race is homogenous across strata (RR = 1.31 for infected people having age ≤ 29 years and RR = 1.26 for those aged > 29 years) and unstratified and stratified effect estimates differ by more than 10% (Table 4). To compute the overall RR adjusting for the confounding effect of age, stratum-specific $\widehat{RR_{MH}}$ are pooled by the MH formula and $\widehat{RR_{MH}}$ =1.30.

The OR for race (=1.54) indicates that the odds of race is significantly higher ($p < .01$) in male than in female (Table 4 for Category I). Similarly, it is also concluded that the odds of race is significantly higher ($p < 0.01$) in female than in male (Table 4 for Category II). The OR of race is homogeneous across strata (OR = 1.35 for infected people having age ≤ 29 years and OR = 1.26 for those aged > 29 years) and unstratified and stratified ORs differ by more than 10% (Table 4). These results also indicate the apparent strong link between gender and race which emerged in the unstratified analysis (OR = 1.54) due to the confounding effect of ageing. To calculate the overall OR of race associated to gender adjusting for the confounding effect of ageing, stratum-specific ORs are pooled by the MH formula and $\widehat{OR_{MH}}$= 1.34. The OR of race is homogeneous across strata (OR = 1.38 for infected people having age ≤ 29 years and OR = 1.32 for those aged > 29 years) and unstratified and stratified ORs differ by more than 10% (Table 4). These results indicate the apparent strong link between gender and race which emerged in the unstratified analysis (OR = 1.55) due to the confounding effect of ageing. To compute the overall OR of race associated with gender adjusting for the confounding effect of ageing, stratum-specific ORs are pooled by the MH formula and $\widehat{OR_{MH}}$= 1.37.

5.1. Stepwise regression

Let us consider stepwise regression, i.e. in each step explanatory variables are incorporated based on some pre-specified criterion. For each underlying model provides better results when the independent variables are considered together. For Category I, when gender or race effect is considered individually, log-binomial model gives large AIC (G: 14.02, R: 15.57), large BIC (G: 14.16, R: 15.59) than the AIC and BIC of the combination effect, i.e. taking both as explanatory variable (given in Table 2). Apart from this, when we take both

Table 5. Simulation for comparison of estimated adjusted RR and standard error using various methods.

Method		Percentage of relative bias	Standard error	95% confidence interval coverage
Mantel–Haenszel	Category I	−0.12	0.43	92.76
	Category II	−0.09	0.33	93.56
Log-binomial	Category I	−0.03	0.46	92.36
	Category II	−0.01	0.34	92.99

as explanatory variables, we get small residual deviance (mentioned in Table 2) than residual deviance of the individual effects (G: 10.001, R: 10.18), and also get high log-likelihood than the individual effect (G: −5.00, R: −7.29). Similarly, for Category II, we get low AIC, BIC and residual deviance and high log-likelihood (when we take both the explanatory variables together) than considering the explanatory variables taking individually. Also for the other models (cloglog, probit, logit), we get better results (low AIC, BIC and residual deviance and high log-likelihood) when taking both independent variables together than taking individuals for both category. Since log-binomial is best among all, so we proceed (in the next section) with suitable simulation of other comparison by taking log-binomial model only.

6. Comparison of simulation study

There are different methods which are used to estimate adjusted RRs in the literature. An MH risk ratio is calculated by taking a weighted average of RRs in strata of covariables, where the weight depends on the size of the strata. Log-binomial regression is a generalized linear model with a log link and a binomial distribution. It is similar to logistic regression, except that the link function is a log link instead of a logit link, hence providing RRs instead of ORs.

Simulation studies are carried out for comparing the potential bias in the estimates of RR. Standard errors (s.e.) of adjusted RR are also compared in this work. From Table 5, relative bias is smaller in MH analysis than log-binomial model and 95% confidence interval coverage portion is larger for MH analysis. Since the relative bias is negative (for both), the estimation is under estimated.

7. Concluding remarks

The statistical performances of the most popular model-based approaches are used in this work. For both Category I and Category II, it can be concluded that log-binomial model performs better than any other models (for the underlying dataset) since it has high log-likelihood and low residual deviance, low AIC and low BIC among any other models (as per Table 2 and Table 3). It can be concluded that between the odds and prevalence ratios (which are the measures of association), prevalence ratios (related to log-binomial model) provide better result since prevalence ratios are more interpretable and easier to communicate. If the log-binomial model is omitted, then probit model is better among three (logit, probit and complementary log–log) since the given dataset is symmetric and also comparing the log-likelihood, residual deviance, AIC and BIC.

The MH analysis is performed to control for confounding. This method combines stratum-specific RRs or ORs. The pooling estimate provides an average of the stratum-specific RRs or ORs with weights proportional to the number of individuals in each stratum. MH method is appropriate in clinical and epidemiological research to remove confounding in studies with relatively large sample size and with a relatively low number of potential confounders.

First, we have got the following conclusions from Category I and Category II: Analysis of data stratified by age shows that there is high significant ($p < 0.01$) excess risk of race associated to gender in both age categories. We can conclude that there exit strong link between gender and race that emerged in the unstratified analysis due to the confounding effect of age. Now, we also get the conclusions that, data analysis stratified by ageing shows that there is high significant ($p< 0.01$) excess probability of race associated to gender both in those aged $\leq$ 29 and in $>$ 29. After data adjustment for the confounding effect of ageing, the OR of race do not differ in male and female. Ageing engenders a positive confounding because it determines an overestimation of the risk associated with gender. After data adjustment for the confounding effect of ageing, the OR of race do not differ in male and female. Ageing engenders a positive confounding because it determines an overestimation of the risk associated with gender.

In this context, it is mentioned that there are two prime limitations to control for confounding by the MH formula: (i) if there is more than a single confounder, the application of this formula is cumbersome due to the higher number of strata and demands a relatively large sample size and (ii) this technique needs continuous confounders to be constrained into a limited number of categories which potentially generating residual confounding (Mantel & Haenszel, 1959).

Our simulation study results (as per Table 5) suggest that the bias is minor in stratified analysis and the coverage of confidence interval is high in stratified MH analysis. In this work, we have reviewed commonly used approaches to estimate RR for observational data. Log-binomial model is used for estimating adjusted RR in the presence of both continuous and categorical confounders but there are some drawbacks in log-binomial model. It may not always converge for any data set (Williamson, Eliasziw, & Fick, 2013).

Our goal is to control this infection and prevent reproductive health problems and also aware the young generation about unprotected sex which can increase the risk of this disease. However, the implicit purpose is to improve young people's sexual health suffered by virtue of being poorly articulated. If there is an association between adverse sexual health and Chlamydia (and also for other STDs like gonorrhea) infection, then these individuals may be benefited from sexual health information, advice contraception, condoms or other STDs testing. It is necessary to promote the step for providing the delivery of treatment to those people testing positive and of interventions to prevent repeat infections. The important components of Chlamydia (and also for other STDs like gonorrhea) control and socio-economic variations in the delivery for these purposes should be formed as a part of future research by taking other risk factors as covariates which effect the STDs mostly or other suitable complex models can be formed which can be decided by suitable criteria. The outcomes of contracting a second STD while any person already actively infected with an STD is called co-infection. The most common use of the term co-infection is that the positive testing of both, i.e. gonorrhea and chlamydia at the same time. The risk of testing positive for both chlamydia and gonorrhea is also increased when the use of drugs or alcohol are reported (Loza et al., 2010). Therefore, the cohort under study is stratified into two categories abusers and non-abusers (for drug or alcohol) for confounding effect of addiction on gender due to the co-infection. Also, MH analysis may be performed to control for confounding.

Nowadays, Chlamydia can lead to serious health sequelae, particularly among women. There are many reasons (e.g. living in disadvantaged areas and a lack of education) which can increase the high risk of Chlamydia. It can also be concluded that there are high prevalence of Chlamydia in young people due to change of partner in higher rates than the general population. Apart from this, the earlier age of first sexual experience or strong desire of sexual relationship are related to a higher risk of STDs in teenagers and young adults since young generation are more active by biologically and hormonally. The results of this work also indicate that there are social variations (most notably by race) in Chlamydia. So, this study suggests that even though early sexual experience makes young generation susceptible to STDs from a younger age, it does not necessarily place them on a trajectory of engaging in sexual behaviour that place them at high risk of STDs when they reach young adulthood. So, in future work, we can consider the region, number of illiterate persons, biologically fitted or sexually desirable persons or persons who have more than one partner or any others covariates (which are closely related to STDs) can be choosen as explanatory variables and construct any properly justified model and their effects in environment may be helpful for decreasing the risk of STDs.

Acknowledgments

The authors are grateful to the anonymous referees and Editor in Chief Prof. Olcay Akman for their careful reading, valuable comments and helpful suggestions, which have helped them to improve the presentation of this work significantly. The first author (Sarada Ghosh) is thankful to the University Grants Commission, India for providing JRF.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York, NY: Wiley.

Agresti, A. (2010a). *The analysis of ordinal categorical data* (2nd ed.). New York, NY: Wiley.

Agresti, A. (2010b). Modeling ordinal categorical data. A link manual at www.stat.ufl.edu/∼aa/ordinal/ord.html.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of Royal Statistical Society, Series B*, *46*, 1–30.

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

Ghosh, S., & Samanta, G. P. (2019). Statistical modelling for cancer mortality. *Letters in Biomathematics*. doi:10.1080/23737867.2019.1581104.

Loza, O., Strathdee, S. A., Martinez, G. A., Lozada, R., Ojeda, V. D., Staines-Orozco, H., & Patterson, T. L. (2010, Jul). Risk factors associated with Chlamydia and gonorrhea infection among female sex workers in two Mexico-U.S. border cities. *International Journal of STD & AIDS*, *21*(7), 460–465. doi:10.1258/ijsa.2010.010018.

Maldonado, G., & Greenland, S. (1993, Dec 1). Simulation study of confounder-selection strategies. *American Journal of Epidemiology*, *138*(11), 923–936.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI: Journal of the National Cancer Institute*, *22*(4), 719–748. doi:10.1093/jnci/22.4.719.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Ser. B*, *42*, 109–142.

Miller, R. G. (1980). *Survival analysis, division of biostatistics*. Stanford, CA: Stanford University.

Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego, CA: Academic Press.

Pratt, J. W. (1981). Concavity of the log likelihood. *Journal of American Statistical Association*, *76*(373), 103–106.

Samanta, G. P. (2015). Mathematical analysis of a chlamydia epidemic model with pulse vaccination strategy. *Acta Biotheoretica*, *63*(1), 1–21.

Scholes, D., Stergachis, A., Heidrich, F. E., Andrilla, H., Holmes, K. K., & W. E. Stamm (1996, May 23). Prevention of pelvic inflammatory disease by screening for cervical chlamydial infection. *The New England Journal of Medicine*, *334*(21), 1362–1366.

Williamson, T., Eliasziw, M., & Fick, G. H. (2013). Log-binomial models: Exploring failed convergence. *Emerging Themes in Epidemiology*. doi:10.1186/1742-7622-10-14.

Workowski, K. A., Bolan, G. A., & Centers for Disease Control and Prevention (2015). Sexually transmitted diseases treatment guidelines. *MMWR Recommendations and Reports*, 64(RR-03):1–137.