Published in final edited form as:

Stat Methods Med Res. 2018 September ; 27(9): 2641–2656. doi:10.1177/0962280216683740.

Modeling clustered binary data with excess zero clusters

John Kwagyan, Victor Apprey

Biostatistics, Epidemiology & Research Design Module, Georgetown-Howard University Center for Clinical & Translational Science, Howard University College of Medicine, Washington, DC, USA

Abstract

We establish a zero-inflated (random-effects) logistic-Gaussian model for clustered binary data in which members of clusters in one latent class have a zero response with probability one, and members of clusters in a second latent class yield correlated outcomes. Response probabilities in terms of random-effects models are formulated, and maximum marginal likelihood estimation procedures based on Gaussian quadrature are developed. Application to esophageal cancer data in Chinese families is presented.

Keywords

Clustered binary data; Gaussian quadratures; logistic-Gaussian model; random-effects models; structured zeros; zero-inflated models

I Introduction

Correlated data arise in many application areas including studies of disease occurrence among family members, studies involving repeated measures of outcome on units, and studies involving group randomization. In such contexts, interest is often focused on the characterization of the dependence structure and the response probabilities. Likelihood-based approaches for analyzing correlated binary data are limited and include the works of Bonney,^{1,2} Prentice,^{3,4} Connolly and Liang,⁵ Bowman and George,⁶ George and Bowman,⁷ and Fitzmaurice and Laird.⁸ Developments for the regression analysis of correlated data have been concerned with methods based on specifying desired marginals and correlation structures, and fitting them using estimating equations. These approaches have attracted much attention owing to the works of Liang and Zeger,⁹ Zeger and Liang,¹⁰ Prentice and Zhao,¹¹ and Liang et al.¹² To account for correlation within clusters, random-effects models have also been proposed and have been used in various applications for correlated binary data.^{13–19}

Corresponding author: John Kwagyan, Biostatistics, Epidemiology & Research Design Module, Howard University College of Medicine, Georgetown-Howard University Center for Clinical & Translational Science, 2041 Georgia Ave, NE, Washington, DC, USA. jkwagyan@howard.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

The presence of binary or count data with excess zeros are also a common phenomenon in a wide variety of disciplines. A literature review²⁰ on this cited examples from a variety of research areas including agriculture, econometrics, epidemiology, public health, medicine, and social work. For example, in laboratory litter studies, it may frequently happen that some animals are unaffected by treatment—the so-called nonresponse phenomena. In correlated grouped-time survival data, some groups of individuals may be immune to the event of interest. Moreover, in genetic studies, it is often suspected that only a small subgroup of patients may have a disease gene that would be linked to a disease marker. Consequently, in studying rare, genetic, or familial diseases, data that are randomly sampled will lead to many families that are largely devoid of individuals with the attribute. In practice, the status of the excess zeros is often not observed and presents a latent nature that may complicate the data analysis. Mixture models²¹ provide a natural framework for unobserved heterogeneity in population studies and the overall distribution of disease occurrence in such data or similar outcomes should appropriately be a mixture.

Approaches for dealing with the excess zero phenomena, notably the zero-inflated and hurdle models, which are based on mixture distribution have been developed.^{22–24} A zero-inflated model assumes that the zero outcomes come from two different processes, “structural” and “random.” Structural zeros can be expected in situations where a subgroup of individuals is not at risk for an outcome and thus will always provide a zero outcome. On the other hand, random zeros assume zero outcomes happen by chance due to sampling variability. A hurdle model assumes that all zero outcomes are from one structural source that is having zeros only from an at-risk subgroup of individuals. Zero-inflated models may be conceptualized as allowing zeroes to arise from at-risk and not-at-risk populations. In this article, we focus on zero-inflated models.

Hall²⁴ introduced zero-inflated binomial model for count data and incorporated random effect to accommodate correlation of outcomes in a repeated measures design. To address both excess zeros and overdispersion, Lewsey and Thomson²⁵ used zero-inflated negative binomial (ZINB) regression models in examining the effect of economic status on decay-missing-filled data in both cross-sectional and longitudinal setting and showed that they provide a better fit than ZIP models. Albert et al.²⁶ gave an excellent overview of ZINB and zero-inflated beta-binomial (ZIBB) models with discussion and application to dental caries. Hur and others²⁷ proposed a model for clustered count data with excess zeros for health outcomes research. Lee et al.²⁸ discuss multilevel zero-inflated Poisson regression modeling for correlated count data with an interesting application to bottle feeding in infants. Recently, Loeys et al.²⁹ gave a more general introduction, where Bayesian alternatives have been proposed. However, relevant references about zero-inflated models for (cross-sectional) hierarchical data are limited. Recently, Fulton et al.³⁰ proposed a mixed model and estimating equation approach for zero inflation in correlated binary response with application to dating violent data.

The current paper builds on works of zero-inflated models and introduces a zero-inflated variance component model for clustered binary data with excess zero clusters. We consider an application of disease occurrence among families for motivation and development of our methods. For most applications in family studies, the outcome is a disease status—affected

or not affected, and excess zero cluster of families is a common phenomena. For example, diseases that are rare, genetic, or familial are more susceptible in certain families than others. In our case study of esophageal cancer in 2951 Chinese families,³¹ 1580 (53%) had no affected family members (see Table I). While models for handling excess zeros for count data have been studied, and while the case study presents excess zero clusters, models that accommodate such data structures for clustered binary data with covariate effects have not been well developed. Our approach considers zero inflation on the “cluster level,” meaning that all binary response of individuals in a cluster are structurally or randomly zero.

In this paper, we (i) introduce a finite mixture likelihood model for clustered binary data with excess zero clusters, (ii) model the response probabilities in terms of random effects based on Gaussian distributional assumption to allow for investigation between cluster and heterogeneity, and (iii) develop approximate maximum (marginal) likelihood procedure using Gaussian quadrature for estimation of parameters. We illustrate our approach with application to esophageal cancer data in Chinese families. Finally, we perform simulation studies to evaluate the effectiveness of the proposed model.

2 Model

Suppose data are composed of N clusters, each of size n_i , $i = 1, \dots, N$ and a vector of binary responses $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ measured on it. If $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N)^T$, then the $\mathbf{Y}_i$'s are independently distributed vectors. Let π_0 and π_1 represent two latent classes where π_0 is the class (of families) whose members do not manifest the attribute under study and π_1 , the class (of families) whose members are susceptible to the attribute under study. In other words, we consider a data situation whose distribution is characterized as resulting from mixing groups from two (latent) populations. We shall call π_0 the *zero-vector state*. Furthermore, given a cluster (or family) from the class π_1 , the probability of an outcome on a member follows a Bernoulli distribution with probability of success, δ_{ij} . We shall call π_1 the *Bernoulli state*.

We define the unobserved random variable, $Z_i (Z_i = 0, 1)$, $i = 1, \dots$, such that

$$Z_i = \begin{cases} 0, & \text{with probability } 1 - \alpha_i \\ 1, & \text{with probability } \alpha_i \end{cases}$$

Suppose that $Z_i = 0$ when $\mathbf{Y}_i$ is generated from the *zero-vector state* and $Z_i = 1$ when $\mathbf{Y}_i$ comes from the *Bernoulli state*. Thus, $1 - \alpha_i$ is the probability that a randomly choosing cluster (family) comes from the *zero-vector state*, π_0 , and α_i the probability that it comes from the *Bernoulli state*, π_1 . Then for the j th outcome in the i th cluster,

$$Y_{ij} = \begin{cases} 0, & \text{with probability } 1 - \alpha_i \\ \text{Ber}(\delta_{ij}), & \text{with probability } \alpha_i \end{cases}$$

so that

$$\delta_{ij} = P(Y_{ij} = 1 | Z_i = 1) \quad (1)$$

Hence, δ_{ij} is simply a conditional probability that describes an individual's tendency to manifest the outcome given: the individual is from a susceptible family.

Also,

$$P\left(\sum_{j=1}^{n_i} Y_{ij} = 0\right) = (1 - \alpha_i) + \alpha_i(1 - \delta_{ij})^{n_i} \text{ and } P\left(\sum_{j=1}^{n_i} Y_{ij} > 0\right) = \alpha_i \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}}$$

Therefore, the joint distribution of the i th cluster is derived via a mixture formulation^{24,29} and from partition theorem in conditional expectations as

$$\begin{aligned} P(\mathbf{Y}_i) &= P(Y_{i1} = y_{i1}, \dots, Y_{in_i} = y_{in_i}) = E[P(y_{i1}, y_{i2}, \dots, y_{in_i} | Z_i)] \\ &= P(Z_i = 0)P(\mathbf{Y}_i | Z_i = 0) + P(Z_i = 1)P(\mathbf{Y}_i | Z_i = 1) \\ &= (1 - \alpha_i)1_{[Y_i = 0]} + \alpha_i P_1(\mathbf{Y}_i) \end{aligned} \quad (2)$$

By assumption,

$$1_{[Y_i = 0]} = \begin{cases} 1, & \text{if } \sum_{j=1}^{n_i} y_{ij} = 0 \\ 0, & \text{if } \sum_{j=1}^{n_i} y_{ij} > 0 \end{cases} = \prod_{j=1}^{n_i} (1 - y_{ij}) \quad (3)$$

Thus $1_{[Y_i = 0]}$ can be thought of as a degenerate (one-point) distribution whose values are localized at 0. In addition, we have by assumption

$$P_1(\mathbf{Y}_i) = P_1(y_{i1}, \dots, y_{in_i}) = \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}} \quad (4)$$

Substituting equations (3) and (4) into equation (2), we establish the joint distribution for the i th cluster as

$$P(Y_{i1} = y_{i1}, \dots, Y_{in_i} = y_{in_i}) = (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}}$$

Thus, the model we obtain is a mixture of a form of a degenerate distribution representing the class (of families) that *does* not manifest the attribute and an independent distribution representing the class (of families) whose members yield correlated outcomes. In effect, we have established a finite mixture model for clustered binary data in which members of

clusters in one latent class have a zero-response vector with probability one, and clusters in the other latent class yield correlated outcomes.

2.1 Features of the joint distribution

1. From equation (5), we find the first moment, the mean of Y_{ij} as

$$\mu_{ij} = E(Y_{ij}) = P(Y_{ij} = 1) = \alpha_i \delta_{ij}$$

and the second moment, the variance of Y_{ij} as

$$\sigma_{ij} = \text{Var}(Y_{ij}) = P(Y_{ij} = 1)\{1 - P(Y_{ij} = 1)\} = \alpha_i \delta_{ij}(1 - \alpha_i \delta_{ij}) = \mu_{ij}(1 - \mu_{ij})$$

Therefore, an alternative parametrization of the joint distribution in terms of μ_{ij} and α_i will also provide interpretable parameters.

The covariance of Y_{ij} and $Y_{ij'} (j \neq j')$, is given by

$$\begin{aligned} \sigma_{ij} &= \text{cov}(Y_{ij}, Y_{ij'}) = P(Y_{ij} = 1, Y_{ij'} = 1) - P(Y_{ij} = 1)P(Y_{ij'} = 1) \\ &= \alpha_i(1 - \alpha_i)\delta_{ij}\delta_{ij'} = \frac{1 - \alpha_i}{\alpha_i} \mu_{ij}\mu_{ij'}, \quad j \neq j' \end{aligned}$$

It follows that the correlation between Y_{ij} and $Y_{ij'}$ is

$$\rho_{jj'} = \frac{\sigma_{jj'}}{\sqrt{\sigma_{jj}\sigma_{jj'}}} = \frac{\alpha_i(1 - \alpha_i)\delta_{ij}\delta_{ij'}}{\sqrt{(1 - \alpha_i\delta_{ij})(1 - \alpha_i\delta_{ij'})}} = \frac{1 - \alpha_i}{\alpha_i} \frac{\mu_{ij}\mu_{ij'}}{\sqrt{(1 - \mu_{ij})(1 - \mu_{ij'})}} \quad (6)$$

This would allow us to write the joint distribution in terms of means and correlations, if desired. In the absence of cluster-level and individual-level covariates, we set $\alpha_i = \alpha$ and $\mu_{ij} = \mu_0$. Let ρ_0 denote the correlation in the absence of individual-level covariates. Then substitution in equation (6) gives

$$\alpha = \left\{ 1 + \frac{\rho_0}{\mu_0(1 - \mu_0)} \right\}^{-1}$$

So, α is determined by the ratio of the baseline intraclass correlation to the population odds. We note an inverse relationship between ρ_0 and α . In particular, as $\rho_0 \rightarrow 0$, $\alpha \rightarrow 1$.

1. If $\alpha_i \rightarrow 1$, $\delta_{ij} \rightarrow \mu_{ij}$, the joint distribution, equation (5), of the i th cluster reduces to

$$P(Y_{i1} = y_{i1}, \dots, Y_{in_i} = y_{in_i}) = \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}} \rightarrow \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1 - y_{ij}}$$

This is the joint distribution of uncorrelated Bernoulli random variables, which in most applications is the null hypothesis of independence of outcomes within cluster.

Moreover, it can easily be shown from equations (1) and (5) that for the i th cluster,

$$P(Y_{ij} = 1, Y_{ij'} = 1) = \frac{1}{\alpha_i} P(Y_{ij} = 1 | Y_{ij'} = 1), \quad j \neq j'$$

Clearly, $\alpha_i = 1$, implies independence of outcomes. Thus, α_i may be interpreted as a measure of cluster dependence. We shall term the parameter, α_i , the relative cluster dependence parameter.

1. The joint distribution is interchangeable, that is, the order of the response does not matter. This property makes the models suitable for analyzing both ordered and unordered data structures.
1. The joint distribution is reproducible. That is, the marginal distribution of any subset of $(Y_1, \dots, Y_n)$ has the same form as the joint distribution. In other words, the interpretation of the parameters is the same regardless of cluster size.³² Hence, the situation where a subset of the cluster is missing causes no problem should this approach be adopted, provided the missing process is ignorable.¹²
1. Let $A_i = \sum_{j=1}^{n_i} y_{ij}$ represent the number of successes of the outcomes in a cluster of size n_i . Then the distribution of $a_i = \sum_{j=1}^{n_i} y_{ij}$, is given as

$$P(A_i = a_i) = \sum_{A_i = a_i} \left\{ (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}} \right\}, \quad a_i = \sum_{j=1}^{n_i} y_{ij}$$

The mean and variance are, respectively,

$$E(A_i) = \alpha_i \sum_{j=1}^{n_i} \delta_{ij} \quad \text{and} \quad \text{var}(A_i) = \alpha_i \left\{ \sum_{j=1}^{n_i} \delta_{ij} (1 - \delta_{ij}) + (1 - \alpha_i) \left(\sum_{j=1}^{n_i} \delta_{ij} \right)^2 \right\}$$

In the absence of individual specific covariates, or if they are not of interest, we set

$\delta_{ij} = \delta = \frac{\mu}{\alpha}$, a constant. Then,

$$P(A_i = a_i) = \begin{cases} 1 - \alpha + \alpha(1 - \delta)^{n_i}, & \text{if } a_i = 0 \\ \binom{n_i}{a_i} \alpha \delta^{a_i} (1 - \delta)^{n_i - a_i}, & \text{if } a_i > 0 \end{cases}$$

Because $\mu = \alpha \delta$, we can use an alternative parametrization in terms of μ and α as

$$P(A_i = a_i) = \begin{cases} 1 - \alpha + \alpha \left(1 - \frac{\mu}{\alpha}\right)^{n_i}, & \text{if } a_i = 0 \\ \binom{n_i}{a_i} \alpha \left(\frac{\mu}{\alpha}\right)^{a_i} \left(1 - \frac{\mu}{\alpha}\right)^{n_i - a_i}, & \text{if } a_i > 0 \end{cases}$$

The mean and variance of the number of successes of outcome becomes, respectively,

$$E(A_i) = n_i \alpha \delta = n_i \mu \quad \text{var}(A_i) = n_i \mu (1 - \mu) + n_i (n_i - 1) \mu^2 \left(\frac{1}{\alpha} - 1\right)$$

The first term of the variance may be thought of as the binomial component (the independent component) and the second term the extra-binomial variation due to dependence within the clusters analogous to that described in the studies by Williams³³ and Moore.^{34,35} If $\mu < \alpha < 1$, $\text{var}(A)$ exceeds that of the binomial and so we have positive clustering. In the case of independence $\alpha = 1$, $\delta = \mu$ and so the equations reduce to

$$P(A_i = a_i) = \binom{n_i}{a_i} \mu^{a_i} (1 - \mu)^{n_i - a_i}, \quad \text{if } a_i = 0, 1, 2, \dots, n_i$$

We obtain the classical binomial distribution with probability of success μ . The variance reduces to that of the binomial, so that α may be interpreted as a measure of overdispersion with respect to the corresponding binomial distribution.

1. Let the number of successes, $0 < a_i = \sum_{j=1}^{n_i} y_j < n_i$. Then, the conditional distribution given at least one success is

$$P(Y_1 = y_1, \dots, Y_{n_i} = y_{n_i} | a_i > 0) = \frac{\sum_{A_i = a_i} \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}}}{1 - \prod_{j=1}^{n_i} (1 - \delta_{ij})}$$

We obtain a form of a weighted distribution, where the relative dependence parameter, α , appropriately drops off. The conditional distribution is analogous to the conditional likelihood model first described by Cox^{36,37} for binary outcomes.

2.2 Parametrization for regression analysis

Suppose the j th subject has a vector of p individual-specific covariates, $\mathbf{X}_{ij} = (x_{ij1}, \dots, x_{ijp})$.

Let the i th cluster has q cluster-specific covariates

$\mathbf{W}_i = (w_{i1}, \dots, w_{iq})$ and $Z_i (Z_i = 0, 1)$, $i = 1, \dots$, an unobserved random variable. The scientific objective is to characterize the dependence of $\mathbf{Y}_{ij}$ on $\mathbf{X}_{ij}$ and $\mathbf{W}_i$. The observations may be taken to be independent between clusters, but certain within-cluster and within-individual correlations are anticipated, which can be modeled explicitly through random effects attached to the linear predictors. Therefore, to account for excess unobserved variation, not accounted for by cluster-specific covariates, we introduce a random-effect term. The logit

model with Gaussian random effects has been studied extensively^{18,19,27,38–40} and will be adopted. Thus, we model the logit of the parameter, δ_{ij} as

$$\delta_{ij}(\mathbf{X}) = P(Y_{ij} = 1 | Z_i = 1, \mathbf{X}_{ij}) = \frac{1}{1 + \exp[\beta_0 + \beta X_{ij} + \gamma_i]}$$

where $\gamma_i \sim N(0, \sigma^2)$ is an unobservable random effect assumed to have a Gaussian distribution with mean zero, and variance σ^2 , to account for excess variation across cluster. In a similar manner, we could model the parameter α_i via a logit link as a function of cluster-specific covariates, $\mathbf{W}_i$ but without random effects. That is,

$$\alpha_i(\mathbf{W}) = [P(Z_i = 1 | \mathbf{W}_i)] = \frac{1}{1 + \exp[\lambda_0 + \lambda \mathbf{W}_i]}$$

This is necessary, because α_i has an embedded property to measure within-cluster dependence, and heterogeneity is accounted for in the modeling for δ_{ij} . Hall²⁵ suggested similar parametrization in a random-effects zero-inflated count model for a repeated measures design.

In the absence of cluster-specific covariates, $\mathbf{W}_i = 0$, thus, we have

$$\alpha_i(\mathbf{W}) = [P(Z_i = 1 | \mathbf{W}_i)] = \frac{1}{1 + \exp[\lambda_0]}, \lambda_0 \text{ constant}$$

In this case, we shall model $\alpha_i = \alpha$, a constant, without loss of generality. The constant relative dependence model can be regarded as the analog of the desirable homoscedastic model in general linear models and to the equal correlation assumption made by Prentice⁶ in regard to the derivation of the extended beta-binomial distribution. An alternative approach would be to leave α_i unrestricted (dependent on i) by simply imposing a distribution on it for estimation purposes.

3 Parameter estimation

Different methods for estimating parameters in random-effects models have been proposed by several authors.^{40–43} Pinheiro and Bates¹⁹ gave a detailed description of approximate log-likelihood functions for mixed-effect models, comparing their computational and statistical properties and concluded the Gaussian quadrature with a reasonable number of abscissas are quite accurate and efficient. When the dimension of the random effects is one or two, numerical integration techniques can be implemented reasonably easily and will be used.^{21,44,45} Our specific application does not include cluster-specific covariates, and, therefore, we shall model $\alpha_i = \alpha$, a constant, without loss of generality. As indicated in the previous section, the constant relative dependence model can be regarded as the analog of the desirable homoscedastic model in general linear models.

Let $\theta = \{\alpha, \beta_0, \beta, \sigma^2\}$ be the parameters to be estimated, then the conditional distribution of $\mathbf{Y} = (Y_1, \dots, Y_N)^T$, $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_N)$ is

$$P(\mathbf{Y} \mid \gamma, \theta) = \prod_{i=1}^N \left\{ (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} [1 - \delta_{ij}]^{1 - y_{ij}} \right\}, \gamma_i \sim N(0, \sigma^2) \quad (7)$$

The (marginal) distribution of $\mathbf{Y}$ is found by integrating out the conditional distribution (equation (7)) with respect to the unobserved random variable γ_i and is given by

$$P(\mathbf{Y} \mid \theta) = \prod_{i=1}^N \left\{ (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} [1 - \delta_{ij}]^{1 - y_{ij}} f(\gamma_i; \sigma^2) \right] d\gamma_i \right\}$$

For convenience, we consider $V_i = \gamma_i / \sigma$, then $V_i \sim N(0, 1)$, and so

$$P(\mathbf{Y} \mid \sigma, \theta) = \prod_{i=1}^N \left\{ (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \alpha_i \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}} \phi(V_i) \right] dV_i \right\} \quad (8)$$

where $\phi(\cdot)$ is the standard normal density and

$$\delta_{ij}(\beta_0, \boldsymbol{\beta}, \sigma) = \frac{1}{1 + \exp[-(\beta_0 + \boldsymbol{\beta}X + \sigma V_i)]}$$

Since the integrals, $\int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}} \phi(V_i) \right] dV_i$, are over Gaussian densities,

Gaussian quadrature would be appropriate and is used. Using an M-point Gaussian quadrature, an integral of the form

$\int_{-\infty}^{\infty} f(v) \phi(v) dv$, where $f(v) = \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}} (1 - \delta_{ij})^{1 - y_{ij}} \phi(V_i)$ and $\phi(v)$ is the standard normal density approximated by the weighted sum

$$\int_{-\infty}^{\infty} f(v) \phi(v) dv \approx \frac{1}{\sqrt{\pi}} \sum_{m=1}^M w_m f(\sqrt{2}v_m)$$

where v_m are the Gaussian quadrature points and w_m the associated weights. The terms v_m and w_m are available from Abramowitz and Stegun.⁴⁶ Applying this to the log-likelihood of equation (8), we have

$$\log L(\boldsymbol{\beta}, U \mid \mathbf{Y}) \approx \sum_{i=1}^N \log \left\{ (1 - \alpha_i) \prod_{j=1}^{n_i} (1 - y_{ij}) + \frac{\alpha_i}{\sqrt{\pi}} \sum_{m=1}^M w_m \left[\prod_{j=1}^{n_i} (\delta_{ijm})^{y_{ij}} (1 - \delta_{ijm})^{1 - y_{ij}} \right] \right\}$$

where

$$\delta_{ijm}(\beta_0, \beta, \sigma) = \frac{1}{1 + \exp\{\beta_0 + \beta X_{ij} + \sqrt{2}\sigma v_m\}}$$

Maximum likelihood using Newton–Raphson algorithm can be used to estimate the parameters. Im and Gianola⁴⁴ recommend using a small number of quadrature points as possible. In principle, one could also use an EM algorithm combined with Gaussian quadrature,^{44,47} but notably, the EM has the disadvantage of not readily providing standard errors⁴⁴ of the parameter estimates, and convergence is usually slow in the absence of close form solutions and therefore will not be discussed further in this application.

4 Model selection

Akaike's information criterion⁴⁸ (AIC), $AIC = -2 * \log(\text{Likelihood}) + 2 * (\text{number of parameters})$, values was used to compare model performance. Smaller values of AIC indicate a better fit. For nested models, the likelihood-ratio test was used to compare whether one model was statistically superior to the other at the significant level of 0.05. For non-nested fitted models that had similar AIC, the Vuong⁴⁹ test was employed to further test for significant differences. The Vuong statistics⁴⁹ has been widely used for model performance in many fields^{22,50} and is defined as

$$V = \frac{LR_n(\hat{\theta}_1, \hat{\theta}_2)}{\sqrt{n\varpi}}$$

where n is the number of observations, $\hat{\theta}_1, \hat{\theta}_2$ represent fitted parameter vectors from two comparable models, and ϖ is the standard deviation of the log-likelihood difference between two models, which is given by

$$\varpi^2 = \text{var}[\log L^1(\hat{\theta}_1) - \log L^2(\hat{\theta}_2)]$$

where L^1 and L^2 are the two likelihood functions for the two comparable models and

$$LR_n = \sum_{i=1}^n \left(\log \frac{L_i^1}{L_i^2} \right) = \sum_{i=1}^n [\log L_i^1(\hat{\theta}_1) - \log L_i^2(\hat{\theta}_2)]$$

The Vuong statistics, V , asymptotically follows a standard normal distribution and is implemented in the R program. Therefore, at a 5% significance level, $|V| < 1.96$ will suggest no significant difference between the two models. In this application, the ZIBB model is the baseline for comparison with the proposed models. Therefore, $V < -1.96$ will suggest the model is significantly better than the ZIBB, and $V > 1.96$ will favor ZIBB.

5 Application to esophageal cancer data in Chinese families

This application involves the study of esophageal cancer in 2951 nuclear families collected in Yangcheng County, Shanxi Province, Peoples Republic of China.³¹ Esophageal cancer, a gastrointestinal cancer, is one of the deadliest cancers worldwide and more commonly seen in China than other parts of the world.^{50–52} The main objective of the study was to assess the presence of familial aggregation of esophageal cancer. In this analysis, we consider as a cluster, the family unit, and assess aggregation of the disease adjusting for measured risk factors. Table I summarizes the distribution of the number of affected individuals by family size. Of the 2951 families, 1371(47%) had at least one affected member and 1580(53%) had no affected members—presenting a substantial number of “excess zero” clusters. The respondent within a family has correlated outcomes which are influenced in part or wholly by the cluster as well as the variables on the individual respondent. The following covariates are available: *SEX* is coded as 0 for *female* and 1 for *male*, *Age* is years centered at 50, *Smoke* is coded as 0 for *nonsmoker* and 1 for *smoker*, and *Alcohol* is coded 0 for *nondrinker* and 1 for *drinker*. The outcome variable, *Y*, is coded 1 if an individual is affected with esophageal cancer and 0 otherwise.

Assuming a single random effect, the general (full) model for predicting an individual's response to esophageal cancer, accounting for potential family to family heterogeneity while adjusting for the excess zero clusters is given as:

$$\begin{aligned} \text{logit}[\delta_{ij}|Z_i = 1, \mathbf{X}, \lambda_i] &= \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Age} + \beta_3 \text{Smoke} + \beta_4 \text{Alcohol} + \lambda_i, \\ \lambda_i &\sim N(0, \sigma^2), \text{ and } 0 < \alpha \leq 1. \end{aligned} \quad (9)$$

We shall call equation (9) the zero-inflated logistic-Gaussian (ZILG) model. Table 2 describes specific fitted derived models. In addition, we fit the data using a traditional zero-inflated approach, ZIBB,²⁶ and compare the results and conclusions to that obtained using the proposed models for clustered binary outcomes. The ZIBB model is appropriate because the total number of cancer cases for a family is bounded by the family size.

Computations of the proposed models were performed using computer programs we developed which was linked with likelihood optimization routines using the R-package (Version 3.2.5). Appendix 1 provides the R-code. For our proposed models, the results did not change much for quadrature points, $M > 9$ and so for computations, $M = 9$ was employed to complete the analysis. In all of the models we have considered, the algorithm converged in less than 38 iterations.

Results of the fitted models are presented in Table 3. Within family dependence is described by the magnitude of the relative dependence parameter, α , and excess familial heterogeneity by the magnitude of, σ , the variance component parameter.

Compared with the standard logistic (independence) model, the likelihood-ratio chi-square statistics show significance of the standard logistic-Gaussian (SLG) model ($p < 0.01$), zero-inflated logistic (ZIL) (fixed effect) model ($p < 0.0001$) and the ZILG model ($p < 0.0001$). Both zero-inflated models (models III and IV) fit the data better than the SLG model (model

II) based on the AIC. The likelihood-ratio test indicates that the ZILG (random-effects) model fits the data better than the ZIL (fixed effect) model ($p < 0.001$), indicating significant variation of outcomes across families. It should be noted that the use of the likelihood-ratio test for testing variance components has been called into question, with some advocating halved p values for such testing of variance component parameters.^{53,54} Therefore, we apply the correction of doubling the p values. In this analysis, we also notice that the difference in log-likelihood values between the zero-inflated fixed-effect model and random-effects models is large, relative to the number of degrees of freedom, so that the preference of the random-effects model is unequivocal.

The ZILG model was therefore the best fitted model which was also confirmed by the Vuong statistic (see Table 3) and thus selected. For this model, the maximum likelihood estimate of the relative dependence parameter, α , was 0.619 and a computed confidence interval (CI) of (0.582, 0.849): This suggests that the data were sampled from a population where the aggregation of esophageal cancer is higher than that from the general population. Similarly, the estimated, the variance component parameter, is 0.771 with CI of (0.605, 0.939) which excludes 0. Thus the data further suggests some degree of heterogeneity of outcomes across families.

At the individual covariate level, *Sex* and *Age* have positive significance while *Alcohol* has negative significance. Smoking was not significant at the 5% level. We conclude that, given a family member has esophageal cancer, the males were at a higher risk (more susceptible) of getting esophageal cancer than the females of the same family and that the older the family member, the more susceptible they become. The negative effect of *Alcohol* means that, it has a propensity to lower the risk of esophageal cancer. Although the amount (number of drinks) of alcohol drank was not available for this analysis, we can speculate that the significance could be due to moderation in drinking.

In comparison, the parameter estimate of the mean risk, β_0 , is much larger in the logistic (independence) model and the logistic-Gaussian model than the zero-inflated models. The parameter estimates in both fixed- and random-effects zero-inflated models are relatively close for sex, age, and smoke. However, the parameter estimate for alcohol in the random-effects model is appreciably smaller (in absolute terms) compared to the estimate in the fixed-effects model. The estimates from ZIBB give estimates that are not much different from the standard logistic regression. The approach seems to underestimate the parameters in the model. This may be due to the fact that the data were not randomly sampled from the population but from families where there is a high risk of aggregation of the occurrence of esophageal cancer. In summary, we conclude that esophageal cancer aggregates in the families sampled.

6 Simulation studies

To evaluate the efficiency and robustness of the proposed method, we generate a correlated binary data that emulates the real data situation for simulation studies. We consider data comprise of $N = 400$ clusters each of cluster size, $n_i = 5$, $i = 1, \dots, N$. Similar to the approach by Fulton et al.,³⁰ we generated two continuous individual-level covariates, (X_1, X_2) from a

standard normal distribution and two binary categorical covariates, (X_3, X_4) . We assume absence of cluster-specific covariates, so that the zero-inflation probability, $\alpha_i = \alpha$, a constant. Consequently, the zero-inflation indicator, Z_i , was generated from Bernoulli(α), with $\alpha = 0.6$. To induce a correlated data structure, the binary outcome, Y_{ij} , was generated from a probit random-effects model³⁰:

$$P(Y_{ij} = 1 | Z_i = 1, X_{ij}, \gamma_i) = \Phi \{ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \gamma_i \}, \gamma_i \sim N(0, \sigma^2)$$

We fixed the regression parameters such that

$$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (0.0, 0.5, -1.4, 1.2, 0.2).$$

The variance component σ^2 was taken to be 0.5^2 to reflect a moderate within-cluster correlation. The simulations were repeated 100 times to compare the performance of the ZILG, ZIL, SLG, and ZIBB models. We calculated the average (Mean) of the estimated parameters, and average of the estimated standard errors (SEs). Furthermore, the mean square error, MSE, of the estimated parameters for each model is evaluated. The simulation results are shown in Table 4. Both the ZILG and ZIL models perform reasonably well, in terms of small MSE compared to the ZIBB model. As expected, the ZILG was almost unbiased for estimating s . The SLG model over estimated s more than two-fold. The ZIBB model tend to understate the true parameter values.

In estimating, α , the zero-inflated probability, the SE of ZILG is the smallest. The estimates of the ZILG model were closest to the true value and the most efficient as they make use of the full distributional assumptions of the observed data.

7 Conclusion

This article has been concerned with development of a finite mixture likelihood formulation for clustered binary data with excess zero clusters, a data structure in which all members of clusters in one latent class have a zero response with probability one, and clusters in the other latent class yield correlated outcomes. The development, albeit being straightforward and based on simple analytic formulation, is novel and well suited for areas of application including public health and biomedical research. For example, in correlated grouped-time survival data,⁵⁵ some groups of individuals may be immune to the event of interest. Through the simple simulation studies and real data analyses, we demonstrate that, the proposed ZILG random-effects model provides a useful tool for analyzing clustered binary data with excess zero clusters.

Often, because of the hierarchical study design or the data collection procedure, zero inflation and lack of independence may occur simultaneously, which render the standard methods inadequate. The relative errors incurred by ignoring the adjustment of excess zeros can be problematic even for a small number of groups with zeros, if traditional modeling methods are used.⁵⁶ The advantage of the proposed model is that it accounts for within-

cluster (family) dependence and provides a good portrayal of cluster (family) differences while adjusting for excess zero clusters.

To our knowledge, the work we have established in this article is the first likelihood formulation for modeling clustered binary data with excess zero clusters where the zero-inflation is at the “cluster level” meaning that all binary responses of individuals in a cluster are structurally zero. It is possible, however, that closer scrutiny, practical considerations, and numerical studies for numerical stability would suggest modifications and/or refinements to the methods discussed. In this article, we only intend to accomplish the first stage of Fisher’s paradigm for the development of statistical methods, namely, the specification of the model which is the development of the joint distribution (or likelihood function) for the data. Future studies will include further simulation studies to evaluate the small sample properties of the likelihood-ratio test for evaluating the hypotheses of independence of outcome within a cluster, conduct sensitivity analyses of the models with different zero-inflation probabilities as well as when the zero inflation depends on covariates. Extensions and applications to polytomous outcomes, correlated survival outcomes, and hierarchical outcomes can be considered for future research. In conclusion, we remark that the proposed finite mixture model for correlated binary data is suitable and computationally tractable for the regression analysis of binary data with zero-inflated clusters with covariate effects.

Acknowledgment

The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. We thank anonymous reviewers for their invaluable comments and suggestions.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has been funded in part with Federal funds (UL1TR000101 previously UL1RR031975) from the National Center for Advancing Translational Sciences (NCATS), NIH, through the Clinical and Translational Science Awards Program (CTSA), National Institute on Minority Health and Health Disparities/NIH Award Number G12MD007597, and by NIH/NIDDK award no. 1R21DK100875-01A1.

APPENDIX A: R Code for likelihood evaluation and optimization

```
## PART A: HOUSEA. r Subroutine to read the datafile
Amat1=read.table("c:/DATAESOC/esocg3.txt",header=T,sep=" ")
cmpa5<-function(rwl) {
  uul=1.0 - rwl[,2]
  return(uul)
}
rwt3=adply(Amat1,1,cmpa5)
prody1=ddply(rwt3,.(famid),summarise,pdyl=min(V1))
rm(rwt3)
##Reading Gaussian quadrature nodes and weights; nn = number of quadrature
points
nn=9
```

```

nnp1=nn+1
mat1=matrix(NA,nrow=nrow(Amat1),ncol=nnp1)
mat1=data.frame(mat1)
xxx=gauss.quad(nn,"hermite")
vvm=xxx$nodes
vvw=xxx$weights
mat1[,1]=Amat1[,1]
aal=unique(Amat1[,1])
aa2=length(aal)
mat2=matrix(NA,nrow=aa2,ncol=nnp1)
mat2=data.frame(mat2)
## PARTB: COMPUTEA. r - Subroutine to evaluation the likelihood function
loglike2 <- function(soln){
  cmpwf<-function(rw5){
    sum1=0.0
    for(j in1:nn){
      sum1=sum1+vvw[j]*rw5[,j+1]
    }
    return(sum1)
  }
  cmp1<-function(rw1){
    y1=rw1[,2]
    sum1=soln[1]+
    soln[2]+soln[3]*rw1[,3]+soln[4]*rw1[,4]+soln[5]*rw1[,5]+soln[6]*vvm[j]*sqrt(2
  )
    ## rw1[,kk]reads variables from column kk; 2.To run the zero-inflated
    logistic model set vvm [j] to 0
    if (y1 == 0)
    {
      sum2= -log(1.0 + exp(sum1))
    } else
    {
      sum2= -log(1.0 + exp(sum1))
    }
  }
  for(j in1:nn){
    rwt2=adply(Amat1,1,cmp1)
    mat1[,j+1]=rwt2$V1
  }
  famid=rwt2[,1]
  for(j in1:nn){
    uu20=mat1[,j+1]
    xx99=data.frame(famid,uu20)
    logv1=ddply(xx99,.(famid),summarise,logsum1=sum(uu20))
  }
}

```

```

mat2[,j+1]=logv1$logsum1
}
rm(uu20) famid=unique(Amat1[,1]) mat2[,1]=famid for(j in1:nn){
mat2[,j+1]=exp(mat2[,j+1])
}
rwt9=adply(mat2,1,cmpwf)
famid=rwt9[,1]
nucsum=rwt9$V1
mat3=cbind(famid,nucsum)
# alpha
  alphac1<-function(rw1){
sumbb=1.0 + exp(-(soln[1]+soln[2]))
sumaa=1.0 + exp(-soln[1])
sumcc=sumbb/sumaa
if(sumcc>1)sumcc=1.0
return(sumcc)
}
# SET alpha = sumcc = 1,to run logistic Gaussian model
  alphav1=ddply(Amat1,.(famid),alphac1)
xx1=merge(mat3,alphav1,by="famid")
xx2=merge(xx1,prody1,by="famid")
cmpall<-function(rw1){
K0=1.0 - rw1[,3]
If(rw1[,3] > 1)K0=1
aal= (rw1[,3]*rw1[,2])/sqrt(pi)
aa3=K0*rw1[,4] + aal
if(aa3 <= 0.0){
return(NA)
}
aa2= log(aa3)
return(aa2)
}
logfamid=ddply(xx2,.(famid),cmpall) logvalue=sum(logfamid[,2])
return(logvalue)
}
# PART C: Main program WHOLEA.R. This links relevant subroutines to perform
optimization.
rm(list=ls())
# source("c:/DATAESOCC/wholea.r",echo=T)
# main program wholea.r
# sub programs housea.r and computea.r
# COMPUTEA.R - likelihood function
# HOUSEA.R - read data file and weights of Gaussian Quadrature
# data file - esocg3.txt

```



```

library(maxLik)
library(plyr)
library(gee)
library(lme4)
library(statmod)
source("c:/DATAESOC/HOUSEA.r")
source("c:/DATAESOC/COMPUTE.r")
initiall=c(0.0,0.0,0.0,0.0,0.0)
res2=maxBFGS(loglike2,start=initiall,print.level=1)
aal=solve(-res2$hessian)
aa2= sqrt(diag(aal))

```

References

1. Bonney GE. Regressive logistic models for familiar disease and other binary traits. *Biometrics* 1986; 42: 611–625. [PubMed: 3567294]
2. Bonney GE. Logistic regression for dependent binary observations. *Biometrics* 1987; 43: 951–973. [PubMed: 3427178]
3. Prentice RL. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J Am Stat Assoc* 1986; 81: 321–327.
4. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; 44: 1033–1048. [PubMed: 3233244]
5. Connolly MA and Liang KY. Conditional logistic regression models for correlated binary data. *Biometrika* 1988; 75: 501–506.
6. Bowman D and George EO. A saturated model for analyzing exchangeable binary data: applications to clinical and developmental toxicity studies. *J Am Stat Assoc* 1995; 90: 871–879.
7. George EO and Bowman D. A full likelihood procedure for analysing exchangeable binary data. *Biometrics* 1995; 51: 512–523. [PubMed: 7662840]
8. Fitzmaurice GM and Laird NM. A likelihood based method for analysing longitudinal binary responses. *Biometrika* 1993; 80: 141–151.
9. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13–22.
10. Zeger SL and Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42: 121–130. [PubMed: 3719049]
11. Prentice RL and Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 1991; 47: 825–839. [PubMed: 1742441]
12. Liang KL, Zeger SL and Qaqish BF. Multivariate regression models for categorical data. *J R Stat Soc Ser B* 1992; 54: 3–40.
13. Anderson DA and Aitkin M. Variance component models with binary response: interviewer variability. *J R Stat Soc Ser B* 1985; 47: 203–210.
14. Rosner B. Multivariate methods for clustered binary data with more than one level of nesting. *J Am Stat Assoc* 1989; 84: 373–380.
15. Gibbons R and Hedeker D. Application of random effects probit regression models. *J Consult Clin Psychol* 1994; 62: 285–296. [PubMed: 8201066]
16. Gibbons R and Hedeker D. Random effects probit and logistic regression models for three-level data. *Biometrics* 1997; 53: 1527–1537. [PubMed: 9423267]
17. Cnaan A, Laird NM and Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med* 1997; 16: 2349–2380. [PubMed: 9351170]
18. Pinheiro JC and Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Gr Stat* 1995; 4: 12–35.

19. Pinheiro JC and Bates DM. Mixed-effects models in S and S-PLUS Statistics and Computing Series. New York, NY: Springer, 2000.
20. Ridout ML, Demetrio CGB and Hinde J. Models for count data with many zeros. In: Proceedings of XIXth International Biometric Society Conference, Cape Town, South Africa, pp.179–192. Washington DC, USA: International Biometric Society, 1998.
21. Brillinger DR and Preisler MK. Maximum likelihood estimation in a latent variable problem Studies in econometrics, time series and multivariate statistics. New York: Academic Press, 1983, pp.31–65.
22. Rose CE, Martin SW, Wannemuehler KA, et al. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J Biopharm Stat 2006; 16: 463–481. [PubMed: 16892908]
23. Hu MC, Pavlicova M and Nunes EV. Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am J Drug Alcohol Abuse 2011; 37: 367–375. [PubMed: 21854279]
24. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 2000; 56: 1030–1039. [PubMed: 11129458]
25. Lewsey JD and Thomson WM. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Commun Dentistry Oral Epidemiol 2004; 32: 183–189.
26. Albert JM, Wang W and Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. Stat Methods in Medical Research 2011; 23: 257–278.
27. Hur K, Hedeker D, Henderson W, et al. Modeling clustered count data with excess zeros in health care outcomes research. Health Serv Outcomes Res Methodol 2002; 3: 5–20.
28. Lee AH, Wang K, Scott JA, et al. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. Stat Meth Med Res 2006; 15: 47–61.
29. Loeyts T, Moekerke B, De Smet O, et al. The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. Br J Math Stat Psy 2012; 65: 163–180.
30. Fulton KA, Liu D, Haynie DL, et al. Mixed model and estimating equation approaches for zero inflation in clustered binary response data with application to a dating violence study. Ann Appl Stat 2015; 9: 275–299. [PubMed: 26937263]
31. Kwagyan J Further investigation of the disposition model for correlated binary outcomes. PhD Thesis, Department of Statistics, Temple University, Philadelphia, PA.
32. Rubin DB. Inference and missing data. Biometrika 1976; 63: 581–592.
33. Williams DA. Extra-binomial variation in logistic linear models. Appl Stat 1982; 31: 144–148.
34. Moore DF. Asymptotic properties of moment estimators for overdispersed counts and proportions. Biometrika 1986; 73: 583–588.
35. Moore DF. Modelling the extraneous variance in the presence of extra-binomial variation. Appl Stat 1987; 36: 8–14.
36. Cox DR. The regression of binary sequence J R Stat Soc B 20): 215–242.
37. Cox DR. The analysis of binary data. London: Chapman & Hall, 1970.
38. Anderson DA and Aitkin M. Variance component models with binary response: interviewer variability. J R Stat Soc Ser B (Methodological) 1985; 47: 203–210.
39. Gilmour AR, Anderson RD and Rae AL. The analysis of binomial data by a generalized linear mixed model. Biometrika 1985; 72: 593–599.
40. Zeger SL and Karim R. Generalized linear models with random effects; a Gibbs sampling approach. J Am Stat Assoc 1991; 86: 79–86.
41. Vonesh EF and Carter RL. Mixed-effects nonlinear regression for unbalanced repeated measures. Biometrics 1992; 48: 1–17. [PubMed: 1581479]
42. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 1997; 92: 162–170.
43. Kuhn E and Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. Comput Stat Data Anal 2005; 49: 1020–1038.

44. Im S and Gianola D. Mixed models for binomial data with an application to lamb mortality. *Appl Stat* 1988; 37: 196–204.
45. Crouch AC and Spiegelman E. The evaluation of integrals of the form $\int_0^t e^{-\lambda t} + \exp(-\lambda t^2) dt$: application to logistic-normal models. *J Am Stat Assoc* 1990; 85: 464–469.
46. Abramowitz M and Stegun IA. *Handbook of Mathematical Functions-with Formulas, Graphs, and Mathematical Tables*. New York: Dover publications, 1970.
47. Bock RD and Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 1981; 46: 443–459.
48. Akaike H Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 1973; 60: 255–265.
49. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 1989; 57: 307–333.
50. Li JY. Epidemiology of esophageal cancer in China. *Natl Cancer Inst Monogr* 1982; 62: 113–120. [PubMed: 7167171]
51. Khuroo MS, Zargar SA, Mahajan R, et al. High incidence of oesophageal and gastric cancer in Kashmir in a population with special personal and dietary habits. *Gut* 1992; 33: 11–15. [PubMed: 1740265]
52. Rasool SA, Ganai B, Sameer AS, et al. Esophageal cancer: associated factors with special reference to the Kashmir Valley. *Tumori* 2012; 98: 191–203. [PubMed: 22677984]
53. Self SG and Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* 1987; 82: 605–610.
54. Snijders T and Bosker R. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage, 1999, pp.90–91.
55. Hedeker D, Siddiqui O and HU FB. Random-effects regression analysis of correlated grouped-time survival data. *Stat Meth Med Res* 2000; 9: 161–179.
56. Gupta P, Gupta R and Tripathi R. Analysis of zero-adjusted count data. *Comput Stat Data Anal* 1996; 23: 207–218.

Table I.

Distribution of number of esophageal cancer cases by family size.

Family size	Number with esophageal cancer								Total
	0	1	2	3	4	5	6	7	
3	435	151	34	3					623
4	536	215	52	16	6				819
5	335	203	81	34	8	0			659
6	159	155	59	27	1	4	0		412
7	74	95	46	12	5	3	0	1	232
8	30	54	25	11	2	3	1	0	129
9	6	21	10	3	1	1	0	0	43
10	4	8	6	2	1	1	1	0	23
11	0	1	1	0	1	1	1	0	8
12	0	1	0	1	0	0	0	0	2
13	1	0	0	0	0	0	0	0	1
Total	1580	906	314	110	24	13	3	1	2951

Table 2.

Derived models and description from the proposed ZILG model.

Model	Model expression	Model description
ZILG	$\text{Logit}(\delta_{ij} X, \lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta X_3 + \beta X_4 + \lambda_i$ $\lambda_i \sim N(0, \sigma^2)$ $0 < \alpha \leq 1$	This is a random-effects model that assumes correlation of outcomes within families and further tests for excess random variation across families, adjusting for excess zero clusters
SL	$\text{Logit}(\delta_{ij} X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta X_3 + \beta X_4$ $\alpha = 1$	This is the complete independence model, which is the basic null hypothesis
SLG	$\text{Logit}(\delta_{ij} X, \lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta X_3 + \beta X_4 + \lambda_i$ $\lambda_i \sim N(0, \sigma^2)$ $\alpha = 1$	This is a random-effects model that assumes correlation of outcomes within families arises from unobserved random variation across families
ZIL	$\text{Logit}(\delta_{ij} X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta X_3 + \beta X_4$ $0 < \alpha \leq 1$	This is a fixed-effect model that assumes correlation of outcomes within families, adjusting for excess zero clusters

For the application: X_1 : Sex; X_2 : Age; X_3 : Alcohol; X_4 : Smoke. SL: standard logistic; SLG: standard logistic Gaussian; ZIL: zero-inflated logistic; ZILG: zero-inflated logistic Gaussian.

Table 3.

Estimates and standard errors of the regression analysis of esophageal cancer in Chinese families.

	Model I: SL	Model II: SLG	Model III: ZIL	Model IV: ZILG	Model V: ZIBB
Parameter	Estimate±SE	Estimate±SE	Estimate±SE	Estimate±SE	Estimate±SE
Mean risk: β_0	-5.235±0.116	-4.494±0.109	-4.152±0.109	-4.272±0.141	-4.549±0.792
Individual covariates					
Sex: β_1	0.982±0.052	0.908±0.059	0.834±0.058	0.811±0.059	0.897±0.059
Age: β_2	0.037±0.003	0.038±0.002	0.037±0.002	0.036±0.004	0.0381±0.002
Alcohol: β_3	-1.175±0.171	-1.145±0.173	-1.046±0.169	-0.968±0.146	-1.1098±0.169
Smoke: β_4	0.056±0.066	0.057±0.067	0.058±0.059	0.064±0.053	0.095±0.059
Dependence (mixing) parameter: α	–	–	0.716±0.023	0.619±0.024	0.742±0.161
Dispersion parameter: ϕ					3.129
Variance component: σ		0.732±0.074		0.771±0.085	–
AIC	10,759.08	10,748.06	10,729.32	10,714.16	10,739.95
Vuong (V) ^a	-10.33	1.56	2.56	5.22	
Likelihood ratio $\chi^2(df)$ ^b		13.02(1)	31.76(1)	46.92(2)	
p value		<0.001	<0.0001	<0.0001	

AIC: Akaike's information criterion; SE: standard error; SL: standard logistic; SLG: standard logistic Gaussian; ZIBB: zero-inflated beta-binomial; ZIL: zero-inflated logistic; ZILG: zero-inflated logistic Gaussian.

^aVuong test for comparing indicated model to ZIBB model (model V).

^bLikelihood ratio and p value for comparing the indicated models to the standard logistic regression model (model I), which represent the basic null hypothesis of independence.

Table 4.

Simulation statistics for average (Mean) parameter estimates, SE, and MSE across 100 simulated data sets.

	Parameter	True value (θ)	Mean ($\hat{\theta}$)	SE	$(\theta - \hat{\theta})$	MSE
ZILG	α	0.600	0.648	0.014	-0.048	0.0029
	B0	0.000	0.041	0.001	-0.041	0.0159
	B1	0.500	0.482	0.042	0.018	0.0021
	B2	-1.425	-1.478	0.027	0.053	0.0050
	B3	1.233	1.346	0.046	-0.113	0.0149
	B4	0.200	0.189	0.029	0.011	0.0010
	σ	0.500	0.534	0.062	-0.034	0.0049
ZIL	α	0.600	0.764	0.231	-0.164	0.0275
	B0	0.000	0.004	0.041	-0.004	0.0102
	B1	0.500	0.558	0.063	-0.058	0.0051
	B2	-1.425	-1.493	0.034	0.068	0.0068
	B3	1.233	1.412	0.049	-0.179	0.0142
	B4	0.200	0.177	0.032	0.023	0.0014
SLG	B0	0.000	0.004	0.071	-0.004	0.0105
	B1	0.500	0.688	0.051	-0.188	0.0371
	B2	-1.425	-1.578	0.045	0.153	0.0256
	B3	1.233	1.239	0.061	-0.006	0.0163
	B4	0.200	0.197	0.034	0.003	0.0089
	σ	0.500	1.216	0.278	-0.716	0.5164
ZIBB	α	0.600	0.491	0.052	0.109	0.0125
	B0	0.000	-0.066	0.165	0.066	0.0146
	B1	0.500	0.713	0.049	-0.213	0.0471
	B2	-1.425	-1.278	0.075	-0.147	0.0238
	B3	1.233	1.095	0.146	0.138	0.0212
	B4	0.200	0.317	0.045	-0.117	0.0145

MSE: mean square error; SE: standard error; SLG: standard logistic-Gaussian model; ZIBB: zero-inflated beta-binomial model; ZIL: zero-inflated logistic model; ZILG: zero-inflated logistic-Gaussian model.