A Finite Mixture Model for Clustered Bivariate Binary Data - Application to Ophthalmologic Data Structures

John Kwagyan, PhD

Biostatistics, Epidemiology & Research Design Module

Georgetown-Howard University Center for Clinical & Translational Science

Howard University College of Medicine, Washington, DC

jkwagyan@howard.edu

Abstract. Ophthalmologic data have traditionally posed challenges for statistical modelling and inference. First there is the potential association between pairs of eyes. Then there is the situation where data is available on one eye for some persons and on both eyes for others. Measurement of multivariate outcomes also occur time and again in ophthalmologic studies, usually because the diseases are related or form a constellation as a syndrome. It is often of clinical interest to model the inter-correlation not only between an outcome and risk factors, but also between different outcomes. We develop a computationally tractable likelihood-based approach that would allow for the detection of correlation between bivariate dichotomous outcomes, modeled simultaneously with the between-eye correlation with and without covariate effects.

Keywords: Clustered Bivariate Binary Data, Random Effect Models, Finite Mixture Models, Ophthalmologiuc Data, Zero-inflated models

1. Introduction

The nature of ocular measurements poses a long-standing question to researchers when developing joint inference on the paired nature of data. Newcombe & Duff [(1)], Murdoch et al [(2)] and others have highlight on the methodological challenges. Generally, vision research studies involve the examination of eyes for a variety of clinical signs or measurements and unlike other paired organs such as the kidneys or lungs, both eyes are easily accessible for assessment and are frequently assessed. However, in some situations data from one eye may be unavailable or "ineligible" for several reasons, resulting in datasets with information on only one eye for some persons and on both eyes for others. If information is available on both eyes, the findings in one eye are likely to be similar to those in the fellow eye. However, the degree of similarity between pairs of eyes are likely to vary considerably from person to person for a multitude of reasons, including environmental and genetic factors. For certain extreme situations, the finding in one eye almost perfectly predicts a similar finding in the fellow eye. For example, blepharitis, a condition characterized by chronic inflammation of the eyelid, almost always affect both eyes bilateral 95% of the time [(3)]. At the other extreme, there are conditions that characteristically will occur in only one eye. This tends to be more likely when the disease is rare. An example is choroidal melanoma which occurs in only one eye in 98% of cases [(4,5)]. The majority of ophthalmologic conditions lie between these two extremes. Furthermore, the classification of case definition varies from diseases to disease: whiles some required the presence of a clinical phenotype to be present in one eye for characterization, others may require the presence of a phenotype in both eyes. For example, characterization of Category

4 Age-related Macular Degeration, AMD, require presence of geographic atrophy, GA, and/or neovascular, NV, in at least one eye, whiles classification of severe glaucoma requires the presence of optic nerve damage with loss of vision in both eyes. Besides, measurement of multiple outcomes are frequent in ophthalmologic research, because diseases are related or form a constellation as a syndrome. For example, category 4 AMD, is classified by the presence of GA and/or NV, conditions that are related. In such situations, it is often of scientific interest to model the inter-correlation, not only between an outcome and risk factors, but also between different outcomes.

The use and development of novel statistical methodology is therefore necessary to realize the full potential for analyzing and making appropriate inference for the different data structures that arise from paired ocular measurements, and designs involving observational, clinical trials and longitudinal study to genome wide association studies. Whether data from only one eye or both eyes are used depends on the study hypotheses and clinical relevance. Glynn & Rosner[6] documented across a range of study designs and various scales of measurement of outcome variables, and showed that when the correlation between paired-eyes is ignored, the p-values and width of confidence intervals are under-estimated, with a consequence would be inaccurate inference and improper recommendation. In a perspectives, Fan et al [7], discussed with details various statistical application for ophthalmology research and noted that most studies adopted use of simple analytical methods, many ignoring the correlation structure between pairs of eyes. Karakosta et al[8] noted that, out of 112 published studies, 38% had measurements from both eyes that used inferential techniques. Regrettably, only 7% of such studies used statistical methods appropriate for correlated outcomes and 74%

made no mention of possible correlation structures.

Some of the novel advances for assessing correlation for paired-eye data have been proposed by Rosner [9] and Dallal [10]. Rosner proposed a constant R model that took the intra- correlation between the paired eyes into account. He assumed that the probability of an outcome at the left eye given a response at the right eye was proportional to the disease rate and showed that it was invalid to treat each eye as an independent random variable in the presence of intra-class correlation. Dallal pointed out that, Rosner's model would perform badly if the attribute is almost certain to occur bilaterally. Dallal then proposed that the probability of an outcome at one eye given a response at the other eye to be a constant parameter. However, both Rosner and Dallas methods do not allow for inclusion of eye-specific covariates. Model based approaches for analysis of ophthalmologic data that have been used include the conditional logistic model [11], mixed effects models [12], where the disease outcome is modeled as a function of measurable characteristics of the eye and individual, and the effect of unmeasurable characteristics of the individual which give rise to the between eye correlation is modeled implicitly. Another approach is the use of GEE, where the outcome in each eye is modeled as a function of the risk factors and the correlation with the other eye is modeled separately and explicitly[13,14].

Regression models with the eye as the unit of analysis and consideration of the correlation between paired eyes offers optimal use of available data, enhanced interpretability of covariate-effects, and efficient use of information from persons who contribute only data on one eye to the analyses. However, when data from a combined paired- and single- eyes are present, there are limited likelihood based approaches available, and the existing methodology ignore the dependence struc-

ture between the paired eyes. The significance and novelty of our proposed methodological development is that, it provides a practical joint distribution for the analysis of different data facets encountered in ophthalmologic research with interpretable parameters.

Of the 2415 DNA specimens available from the AREDS trial[15], 940 were from disease-free subjects and 1475 were from subjects with early or intermediate AMD in one or both eyes. Clearly, the data presents a situation in which excess number of zero pairs of responses occur in some persons. The overall distribution of disease occurrence in such data or similar outcomes should appropriately be a mixture.

We proposed a finite mixture latent class model for clustered binary data in which extra probability is given to events in which all binary outcomes in a cluster unit have a zero probability of occurrence[16,17]. We adopt the approach to develop a computationally tractable likelihood-based model that would allow for the detection of correlation between bivariate dichotomous outcomes, modeled simultaneously with the between-eye correlation with and without covariate effects.

2. Modeling Consideration

Suppose data are composed of N clusters, each of size n_i, $i = 1, ..., N$ and a vector of binary responses $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})^T$ measured on it. Let $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N)^T$, 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, suppose 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, $Z_i(Z_i = 0, 1)$, $i = 1, ...N$, such that

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

Suppose further 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},

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

so that

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

With this, we deduce via mixture formulation[16,17] that the joint distribution of the i-th cluster can be based on,

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

We notice that for $\delta_{ij} = P(Y_{ij} = 1 | Y_{ij} = 1), j \neq j'$. And so is simply the conditional probability of the outcome of one member given another member from the cluster has the attribute. We adapt the approach to vision-data and develop a general likelihood framework for bivariate binary outcomes.

3. Models for Paired Data

Consider a collection of N persons each with a pair of "eligible" eyes under study. Here the individual represents a clustering unit and the eyes the unit of analysis. The pair of eyes on a person form a cluster. Let $i = 1, ..., N$ index the i^{th} individual and $j, (j = [L, R])$ be the *Left or Right* eye from the individual. For the ith individual, define Y_{ij} such that

$$Y_{iL}=\begin{cases} 1, & \textit{if the left eye is diseased} \\ 0, & \textit{if the left eye is not diseased} \end{cases}$$

and

$$Y_{iR}=\begin{cases} 1, & \textit{if the right eye is diseased} \\ 0, & \textit{if the right eye is not diseased} \end{cases}$$

Then the disease vector, D_i, for the i^{th} individual is:

$$D_i = \{Y_{iL}, Y_{iR}\} = [(0,0), (0,1), (1,0), (1,1)],\ i = 1, ..., N$$

Let

$$\mu_{ij} = E(Y_{ij}) = \Pr(Y_{ij} = 1),\ i = 1, 2, ..., N;\ \ j = [L, R].$$

Then μ_{ij} is simply a measure of the population disease rate. For the ith person, let

$$\alpha_i = \frac{P(Y_{iL} = 1)P(Y_{iR} = 1)}{P(Y_{iL} = 1, Y_{iR} = 1)},\ i = 1, 2, ..., N$$

If $\alpha_i = 1$, then the outcomes on the left and right eyes are randomly determined. Thus α_i can be viewed as a measure of the between eye correlation.

Let

$$\delta_{ij} = P(Y_{ij} = 1 | Y_{ij'} = 1), j \neq j' = [L, R]$$

then δ_i is simply interpreted as the probability of the outcome on the left (right) eye given the right(left) eye has the attribute. With the above definitions, the joint distribution for the *ith* person can be based on

$$P(Y_{iL}, Y_{iR}) = (1-\alpha_i) \prod_{j=[L,R]} (1-y_{ij}) + \alpha_i \prod_{j=[L,R]} \delta_{ij}^{yij} (1-\delta_{ij})^{1-y_{ij}} \quad (1)$$

Kwagyan[16] showed that the joint distribution is reproducible. That is the marginal distribution of outcomes on an eye has the same form as the joint distribution of the pair of eyes. In other words, the interpretation of the parameters would not be affected regardless of number of eligible eyes under consideration in a person[18,19]. Hence the situation where a person presents data on a single-eye or paired eyes causes no problem should this approach be adopted. For this reason, the methodology can be applied to *data* structures involving single-eyes, paired-eyes or combined unpaired/paired data structures.

We note that, $\mu_{ij} = P(Y_{ij} = 1) = \alpha_i \delta_{ij}$. And so an alternative parameterization of the joint distribution with interpretable parameters in terms of μ_{ij} and $3b1_i$ is given as

$$P(Y_{iL}, Y_{iR}) = (1-\alpha_i) \prod_{j=[R,L]} (1-y_{ij}) + \alpha_i \prod_{j=[L,R]} (\frac{\mu_i}{\alpha_i})^{yij} \left(1 - \frac{\mu_i}{\alpha_i}\right)^{1-y_{ij}} \quad (2)$$

With this parameterization, we can model the mean, μ_{ij}, directly on person-specific covariate, Z, and eye-specific covariates, X, without having to stratify the analysis of response of one eye on the other eye whiles still accounting for between eye correlation. In a similar manner, we can model the correlation parameter, $3b1_i$, directly on person-specific covariates.

4. Distribution of Number of Diseases Eyes

Let $A_i = \sum_{j=1}^{2} y_{ij}$ represent the number of disease eyes in a person. Then the distribution of number of diseased eyes , $a_i = \sum_{j=1}^{2} y_{ij}$, is given as,

$$P(A_i = a_i) = \sum_{A_i = a_i} \left\{ (1-\alpha_i) \prod_{j=[L,R]} (1-y_{ij}) + \alpha \prod_{j=[L,R]} (\frac{\mu_i}{\alpha_i})^{y_{ij}} (1-\frac{\mu_i}{\alpha_i})^{1-y_{ij}} \right\}.$$

In the absence of eye- specific and person -specific covariates, or if they are not of interest, we set $\mu_i = \mu$ and $\alpha_i = \alpha$, as constants. Then

$$P(A_i = a_i) = \begin{cases} 1 - \alpha + \alpha(1-\frac{\mu}{\alpha})^{n_i}, \text{ if } a_i = 0 \\ \binom{n_i}{a_i} \alpha (\frac{\mu}{\alpha})^{a} (1-\frac{\mu}{\alpha})^{n_i - a_i}, \text{ if } a_i > 0 \end{cases} \tag{3}$$

The mean and variance of the number of disease eyes becomes, respectively,

$$\begin{aligned} E(A_i) &= 2\alpha\delta = 2\mu \\ var(A_i) &= 2\mu(1-\mu) + 2\mu^2(\frac{1}{\alpha} - 1) \end{aligned}$$

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 Williams[20] and Moore[21,22]. If $\mu < \alpha < 1$, $var(A)$ exceeds that of the binomial and so we have positive clustering. In the case of independence $\alpha = 1$, and $\delta = \mu$ and so the equations reduces to

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

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.

4.1. Model for Grouped Data

Let a (=0,1,2) be the number of diseased eyes in a person and suppose that data are collected from a random sample of persons and that the distribution of outcomes under study are represented as in **Table 1**.

Table 1: **Distribution of eligible by no. of affected eyes**

	No. Diseased Eyes (a)			
Eligible Eyes (s)	0	1	2	Total
1	n_{10}	n_{11}		n_{1+}
2	n_{20}	n_{21}	n_{22}	n_{2+}
Total	n_{+0}	n_{+1}	n_{+2}	**n**

The quantities n_{sa} denotes the number of persons of s "eligible" eyes with a dieseased: $s = 1, 2$; $a = 0, ..., s$. so that n_{s0} will represent the number of persons with s "qualified" eyes with no affected eye, n_{s1} with 1 affected, and so on. $n = \sum_{s=1}^{2} \sum_{a=0}^{s} n_{sa}$ is the total number of persons under study. $n_{+a} = \sum_{s=1}^{m} n_{sa}$ is the total number of persons with a affected diseases and $n_{s+} = \sum_{a=0}^{s} n_{sa}$ is the total number of persons with s qualified eyes. We wish to find estimators for μ,which is the population disease rate, and α, a measure of between eyes correlation. The number of affected eyes, a in a person are independent and assumed to be identically distributed with the same parameters so that the likelihood of the data can be derived from **Equation 3**. The likelihood function based on the data is given by the product

$$L(.,.|a) = \prod_{s=1}^{2} \left[\{P(a=0)\}^{n_{s0}} \prod_{a=1}^{s} \{P(a>0)\}^{n_{sa}} \right]$$

The likelihood function is

$$L(\mu,\alpha) = \prod_{s=1}^{2} \left\{ 1-\alpha+\alpha(1-\frac{\mu}{\alpha})^{s} \right\}^{n_{s0}} \prod_{a=1}^{s} \binom{s}{a} \left\{ \alpha(\frac{\mu}{\alpha})^{a}(1-\frac{\mu}{\alpha})^{s-a} \right\}^{n_{sa}}$$

The likelihood is analogous to the *constant R model* proposed by Rosner[9] and the *constant parameter model* by Dallal[10]. Inferential procedures of the joint distribution are within the framework of the likelihood theory and can be developed.

4.2. Parametrization For Regression Analysis

Suppose data consists of persons each with possibly a pair of binary responses $Y_j = (Y_L, Y_R)^T$ on the eyes. Suppose the i^{th} person has a vector of person-specific covariates, $Z = (Z_1, Z_2, \ldots Z_p)$ and q eye-specific covariates $X = (X_1, X_2, \ldots, X_q)$ measured on it. The scientific objective is to characterize the dependence of Y_i on Z and X. Because the outcomes are binary it seems natural to consider a logit transform and so will be studied in detail. Using the framework of generalized linear model [23–27], we model the logit of the parameters, as presented in **Equation (1)** as,

$$\log \text{it}[\delta_{ij}(\mathbf{Z})] = \log \text{it}[P(Y_{ij}=1|Y_i=1,\mathbf{Z}_{ij})] = \beta_0 + \boldsymbol{\beta}\mathbf{Z}_{ij} + \gamma_i$$

where $\gamma_i \backsim N(0,\sigma^2)$, 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.

$$\log it[\alpha_i(\mathbf{X})] = \log it[P(Z_i = 1|\mathbf{X}_i)] = \lambda_0 + \boldsymbol{\lambda}\mathbf{X}_i$$

In a similar manner, we can model the logit of the parameters, as presented in **Equation (3)**.

The joint distributions are computationally tractable and can be quickly fitted to data using computer programs for numerical optimization. Inferential procedures are within the framework of the likelihood theory and can easily be developed. Maximum likelihood using Newton-Raphson algorithm can be used to estimate the parameters. One could also use an EM algorithm combined with Gaussian quadrature, but notably, the EM has the disadvantage of not readily providing standard errors of the parameter estimates.

We developed likelihood-based approach that would allow for the detection of correlation between bivariate dichotomous outcomes, modeled simultaneously with the between-eye correlation with and without covariate effects. The model have very attractive properties. The construction of the joint distribution of the model is based on simple analytic formulations. They are likelihood based and as such require the complete specification of the joint distribution. They have broad applications based on fewer restrictive assumptions. In particular they are suitable for analyzing clusters with unequal sizes and for both ordered and unordered data structures. The joint distributions are computationally tractable and can be quickly fitted to data using computer programs for numerical optimization. Inferential procedures are therefore within the framework of the likelihood theory and can be developed.

In this paper, 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 joint distribution (or likelihood function) for the data. Future studies will include simulation studies to evaluate the small sample properties of the likelihood ratio test for evaluating the hypotheses of independence of outcome within a cluster. Others will include assessment of goodness-of-fit of the proposed model. A common feature in the study of clusters is the existence of multiple groups with levels of nesting within it. An example is the study of outcomes of eye patients in hospitals in multicenter clinical trials. Two levels of nesting exist in this data. The eye patients are nested within the hospitals and the hospitals are nested within the centers. A second example is the study of patients with retinitis pigmentosa, an eye disease in families seen in an outpatient clinic (Berson, Rosner and Simonoff, 1980). Two levels of nesting exist in these data. Subjects are nested within families and two eyes are nested within each subject. Thus there are family-specific, patient-specific, and eye-specific covariates. Extensions and applications to two-way hierarchical layout can be considered for future studies. In conclusion, we remark that the proposed finite mixture model for correlated binary data is suitable are computationally tractable for the regression analysis of clusterd binary data with and without covariate effects.

Acknowledgement

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. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Newcombe RG, Duff GR. Eyes or patients? - Traps for the unwary in the statistical analysis of opthalmologic studies. *Br J Ophthalmol. 1987;71:645-646*

2. Murdoch IE, Morris SS, Cousens SN. (1998). People and eyes: statistical approaches in ophthalmology. *Br J Ophthalmol, 82(8), 971-973.*

3. Seal D, Wright P, Ficker L, et al. Placebo controlled trial of fusidic acid gel and oxytetracycline for recurrent blepharitis and rosacea. *Br J Ophthalmol 1995;79:42–5.*

4. Dunn ES, Jaeger EA, Jeffers JB, et al. The epidemiology of ruptured globes. Ann Ophthalmol 1992:24:405–10.

5. Castella AP, Bercher L, Zografos L, et al. Study of the blood-aqueous barrier in choroidal melanoma. *Br J Ophthalmol 1995;79:354–7.*

6. Glynn RJ, Rosner B. Regression methods when the eye is the unit of analysis. *Ophthalmol 1992:24:405-10*

7. Fan Q, Teo Y, Saw W. Application of advanced statistics in ophthalmology. *Investigative Opthalmology & Visual Sc. 2011; 52(9): 6059-6065*

8. Karakosta A, Vassilaki M, Plainis S, Elfadl NH, Tsilimbaris, M, Moschandreas, J. (2012). Choice of analytic approach for eye-specific outcomes: one eye or two?. *Am J ophthalmol, 153(3), 571-579.*

9. Rosner B. Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. *Biometrics 1982; 38:105-114*

10. Dallal GE. Paired Bernoulli trials. *Biometrics 1988; 44:253-257*

11. Connolly, MA and Laing KY (1988). Conditional logistic regression models for correlated binary data. *Biometrika, 75, 501-506*

12. Rosner B. Multivariate methods in ophthalmology with application to paired-data situations. *Biometrics, 1025-1035.*

13. Liang KY, Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73. 13-22*

14. Zeger SL, Liang KY (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics 1986;42:121-30.*

15: Perlee LT, Bansal AT, Gehrs K, et al. (2013). Inclusion of genotype with fundus phenotype improves accuracy of predicting choroidal neovascularization and geographic atrophy. *Ophthalmology, 120(9), 1880-1892.*

16. Kwagyan J. (2001). Further Investigation of the disposition Model for correlated binary outcomes. *Ph.D Thesis, Department of Statistics, Temple University, Philadelphia, PA*

17. Kwagyan, J. , Apprey V. (2015). Finite Mixture Logistic-Gaussian Model for Zero-Inflated Clustered Binary Data. *In JSM Proceedings, Biometrics Section, American Statistical Association. Alexandria, VA. 1788-1802*

18. Rubin DB. (1976). Inference and missing data. *Biometrika, 63(3), 581-592.*

19: Liang, K. L, Zeger S. L, Qaqish (1992). Multivariate regression models for categorical data. *Journal of the Royal Statistical Society. Series B 54: 3-40*

20: Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31, 144-148*

21: Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika, 73(3), 583-588.*

22: Moore, D. F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics, 8-14.*

23: Prentice R. L(1988). Correlated binary regression with covariates specific to

each binary observation. *Biometrics 44, 1033-1048*

24: Gibbons, R. and Hedeker, D. (1997) Random effects probit and logistic regression models for three-level data," *Biometrics 53, 1527–1537, 1997.*

25: Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics, 4(1), 12-35.*

26: Pinheiro, J. C., & Bates, D. M. (2000). Mixed effects models in S and S-PLUS. Springer.

27: Min, Y., Agresti, A., (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling 5, 1–19.*