Statistical Methods in Medical Research 2011; **20**: 175–189

A hierarchical zero-inflated log-normal model for skewed responses

Ning Li Department of Epidemiology and Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL, **David A. Elashoff** Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, CA, USA, **Wendie A. Robbins** and **Lin Xun** School of Nursing, University of California at Los Angeles, Los Angeles, CA, USA

Although considerable attention has been given to zero-inflated count data, research on zero-inflated log-normal data is limited. In this article, we consider a study to examine human sperm cell DNA damage obtained from single-cell electrophoresis (COMET assay) experiment in which the outcome measures present a typical example of log-normal data with excess zeros. The problem is further complicated by the fact that each study subject has multiple outcomes at each of up to three visits separated by six-week intervals. Previous methods for zero-inflated log-normal data are based on either simple experimental designs, where comparison of means of zero-inflated log-normal data across different experiment groups is of primary interest, or longitudinal measurements, where only one observation is available for each subject at each visit. Their methods cannot be applied when multiple observations per visit are possible and both inter- and intra-subject variations are present. Our zero-inflated model extends the previous methods by incorporating a hierarchical structure using latent random variables to take into account both inter- and intra-subject variations in zero-inflated log-normal data. An EM algorithm has been developed to obtain the Maximum likelihood estimates of the parameters and their standard errors can be estimated by parametric bootstrap. The model is illustrated using the COMET assay data.

1 Introduction

Zero-inflated data are commonly encountered in both cross-sectional and longitudinal studies where we observe non-negative outcome measures with excessive zeros accumulated at the origin. The data can be discrete, usually called zero-inflated count data, such as the number of defects in manufacturing and work-related injuries in occupational health, or continuous, such as DNA damage in human sperm cells measured by single-cell electrophoresis (COMET assay), which are typically characterised by highly skewed positive values that form a long right tail.

Excessive zeros in these data lead to lack of fit of commonly utilised parametric distributions, such as Poisson or negative binomial distributions for the zero-inflated count data. Employment of non-parametric approaches based on ranking has been criticised by some authors due to a large number of ties caused by the zero observations

Address for correspondence: Ning Li, Department of Epidemiology and Biostatistics, College of Public Health and Health Professions, University of Florida, P.O. Box 100231, Gainesville, FL 32610-0231. E-mail: nli@phhp.ufl.edu

10.1177/0962280208097372

and the difficulty to predict the response variable.[1] In longitudinal studies with zero-inflated log-normal measures, standard techniques for correlated data, such as linear mixed effects models that assume normality of the outcomes, are obviously inapplicable. Even log-transformed data, after addition of small constants to handle zeros, may not satisfactorily approximate the normal distribution due to the mode generated by the zeros in the original observations. As demonstrated in Section 4.2, analysing the log-transformed data with the linear mixed effects model fails to identify some important covariate effects. One alternative is to categorise the responses into ordinal scales by applying several thresholds and then use the ordinal logit model to analyse the data, but this approach may lose power by discarding considerable outcome information.

The so-called zero-inflated model has been popular to model count data with excess zeros.[2–7] It usually mixes a Poisson, negative binomial, or binomial distribution with a distribution degenerate at zero. For clustered or longitudinal zero-inflated count data, Hall[3] introduced a random intercept in the regression model for the mean of the Poisson (λ) or binomial distribution, and Yau *et al.*[4] extended to auto-regressive random intercepts in the log-linear model for λ in a setting of a time series of counts. Recently Kreuter and Muthén[5] considered both random intercept and slopes for λ in the context of growth and group-based trajectory models. A more complex modelling strategy that includes random effects in both mixture components of the zero-inflated model has also been studied by Min and Agresti,[6] and they proposed to link together the two components by the joint distribution of the random effects. Lee *et al.*[7] also specified cluster- and subject-specific random effects in both components in a three-level hierarchical model, but they assumed zero correlations across the normal–distributed random effects.

In the case of zero-inflated log-normal data, zeros can only occur in the component that degenerates at origin, which is different from the models for zero-inflated count data where zero responses can occur in either mixture components. Therefore, after adding an indicator for component membership for each observed response, the zero-inflated log-normal model becomes a generalised linear mixed model with mixed response types.[8] For uncorrelated observations, Zhou and Tu[9] considered the problem of testing the equality of the means of independent log-normal populations with possibly zero observations. Tian[10] proposed to test and estimate the confidence interval for the mean of zero-inflated log-normal data on the basis of generalised test variables and generalised pivotal quantities. Tooze *et al.*[1] extended these models to repeated measures of zero-inflated log-normal data by introducing random intercepts in both of the logistic regression for the proportion of zero observations and the log-linear model for the non-zero values. The two random intercepts were assumed to follow a bivariate normal distribution. However, it is a two-level model and therefore is not readily applicable for higher level correlated data in which there could be multiple observations for each subject at each visit.

The motivating example of this article is a study conducted at UCLA School of Nursing to assess the effects of cryopreservation on integrity of DNA and the variation of DNA strand breakage in sperm cells within and between men over time. The laboratory measures of sperm cell DNA damage obtained from COMET assay experiment present a typical example of log-normal data with excess zeros (undamaged cells). Each study subject has up to three visits separated by six-week intervals and DNA damage in

50 sperm cells were measured per visit. Therefore, a three-level modelling strategy must be considered to take into account inter- and intra-subject variations. In this article, we propose a hierarchical zero-inflated model for log-normal populations with possible zero observations. We adopt a logistic regression sub-model for the proportion of zeros and a log-linear sub-model to formulate the distribution of the positive observations. Inter- and intra-subject variations are accounted by latent random variables. It is an extension of the model proposed by Tooze *et al.*[1] for two-level data where only one observation is available for each subject at each visit so that intra-subject variation is not considered.

The remainder of this article is organised as follows. The COMET assay and the study design are described in more detail in the next section. The model and the inference procedure are described in Section 3. Section 4 contains simulation studies and application of the model to the COMET data. Some concluding remarks are given in Section 5.

2 Background of COMET assay and the study design

Single-cell gel electrophoresis, also known as the COMET assay, is a technique for detection and quantification of DNA damage in individual cells.[11–19] The damage is represented by migration of DNA fragments out of the cell nucleus towards the anode, and each nucleus and associated tail of damaged strands of DNA resemble a comet. The length and fragment content of the tail are directly proportional to the amount of DNA damage.

In this study, the assay was used to analyse sperm cells in ejaculated semen, one of the most common biological samples used in human reproductive health research, to assess the effects of cryopreservation on integrity of DNA and the variation of DNA strand breakage within and between men over time. A total of 19 men, aged 20–30 years old, were recruited and screened for eligibility. They were asked to provide three specimens: one at baseline, one six weeks later and one twelve weeks later. The specimens were then split into three aliquots: one was used immediately for COMET assay; one was frozen in a -80^{o}C freezer for six weeks before the next assay and the third was frozen for 12 weeks before the next assay. A total of 50 cells were analysed per aliquot of sample. Fresh samples were run alongside frozen (thawed before the analysis) samples according to the time assignments. As summarised in Figure 1, there are three experimental runs: the fresh samples collected at week 0 were analyzed by COMET assay in run 1; the fresh samples at week 6 were analysed alongside with the frozen samples from week 0 in run 2 and the fresh samples at week 12 were analysed alongside with two frozen samples in run 3 – one from visit 1 and the other from visit 2. Sixteen men out of 19 completed all three visits. A detailed description of the experiment can be found in Xun *et al.*[20]

The COMET assay produces three measures of DNA damage: moment, percent DNA in the tail, and moment arm. For all of these measures larger values indicate more DNA damage. The distributions for the measures are heavily right-skewed with a considerable proportion of zeros (13.9% for moment, 10.9% for percent DNA and 12.0% for moment arm), and some extremely large values. In our preliminary analysis, the data were log-transformed to reduce skewness and small constants were added before log-transformation to take care of values at or close to zero. Comparison of the means

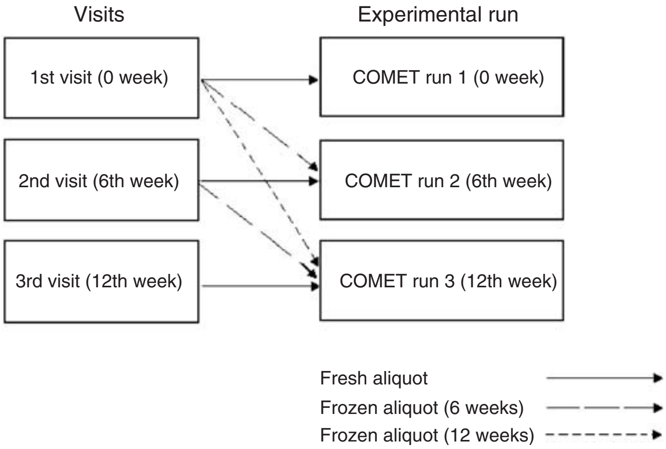

Figure 1 Diagram for the study design.

for fresh samples suggests substantial variation of DNA damage between subjects and between samples within subjects (Figure 2). There were significant correlations between the fresh and the frozen samples (kendall tau correlation = 0.53, $p = 0.0045$ for both moment and percent DNA; kendall tau correlation = 0.47, $p = 0.012$ for moment arm). However, in a few samples extremely high level of DNA damage occurred after freezing (Figure 3, with the 45^o straight line as a reference).

3 The hierarchical zero-inflated model

3.1 The model

The COMET assay was conducted on multiple sperm cells for each sample. Therefore, the sperm cells, visits and subjects are defined, respectively, as level 1, level 2 and level 3 units. Let $Y_{ijk} \geq 0$ be the measurement of the kth sperm cell for the jth visit of the ith subject, where $k = 1, 2, \ldots, K_{ij}$, $j = 1, 2, \ldots, J_i$, and $i = 1, 2, \ldots, I$, with I being the number of subjects, J_i the number of visits for the ith subject, and K_{ij} the number of sperm cells measured for the jth visit of the ith subject. Denote $\boldsymbol{Y}^T = (Y_{111}, \ldots, Y_{1J_1K_{1J_1}}, \ldots, Y_{I11}, \ldots, Y_{IJ_IK_{IJ_I}})$.

Let $\boldsymbol{x}_{ijk}^T = (x_{1ijk}, x_{2ijk}, x_{3ijk})$ be a vector of covariates corresponding to Y_{ijk} where x_{1ijk}, x_{2ijk} and x_{3ijk} are binary indicators for evaluating the design effects of freezing and the effects of experimental run represented in the model as run 2 and run 3, using run 1 as the reference group.

The zero-inflated model has the following two components: (1) the probability of $Y_{ijk} = 0$, denoted as π_{ijk}, is modelled by

$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \beta_0 + \beta_{1i} + \beta_{2ij} + \gamma x_{1ijk} + \phi_{02}x_{2ijk} + \phi_{03}x_{3ijk}, \tag{1}$$

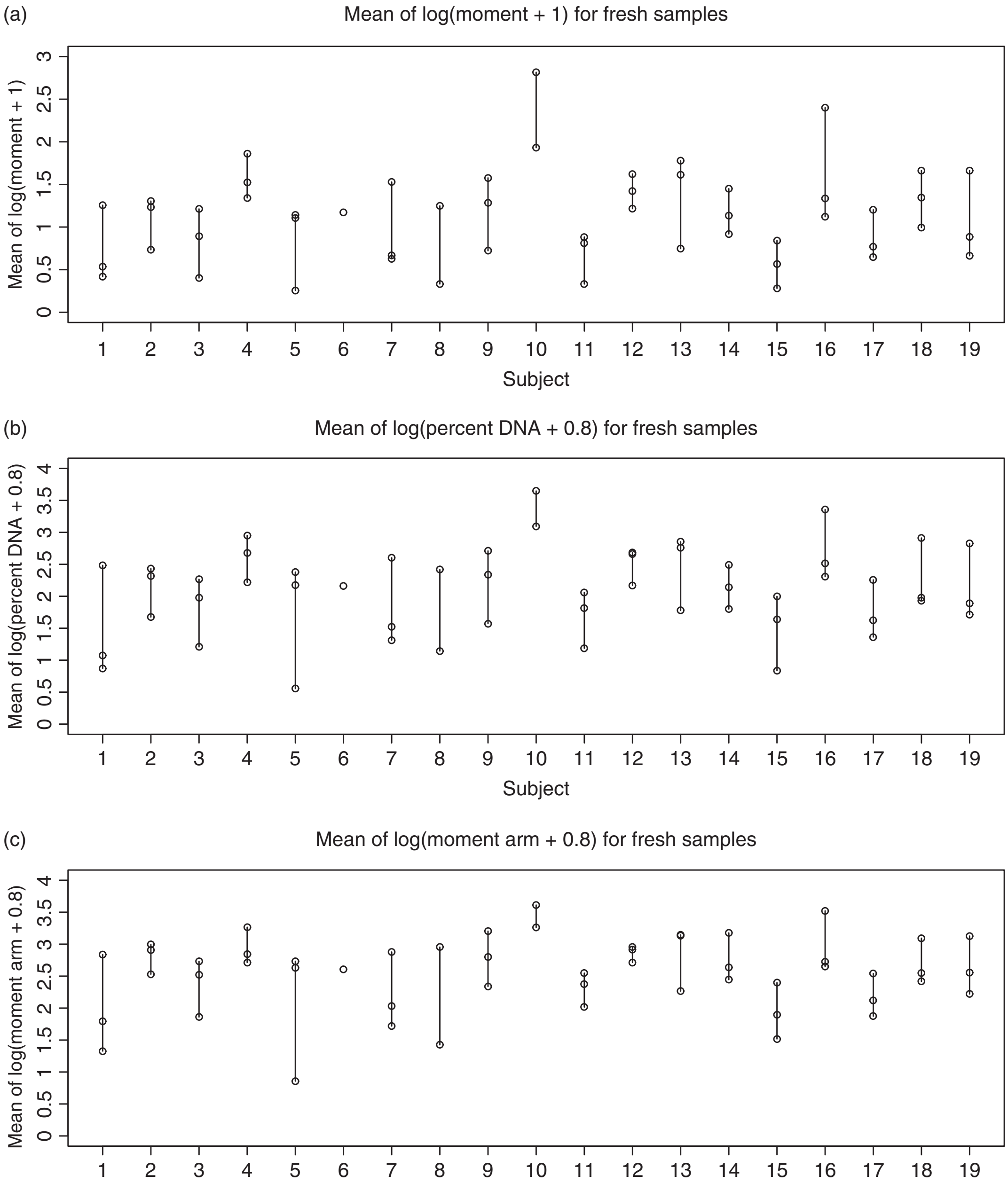

Figure 2 Inter- and intra-subject variations for the fresh samples using log-transformed data.

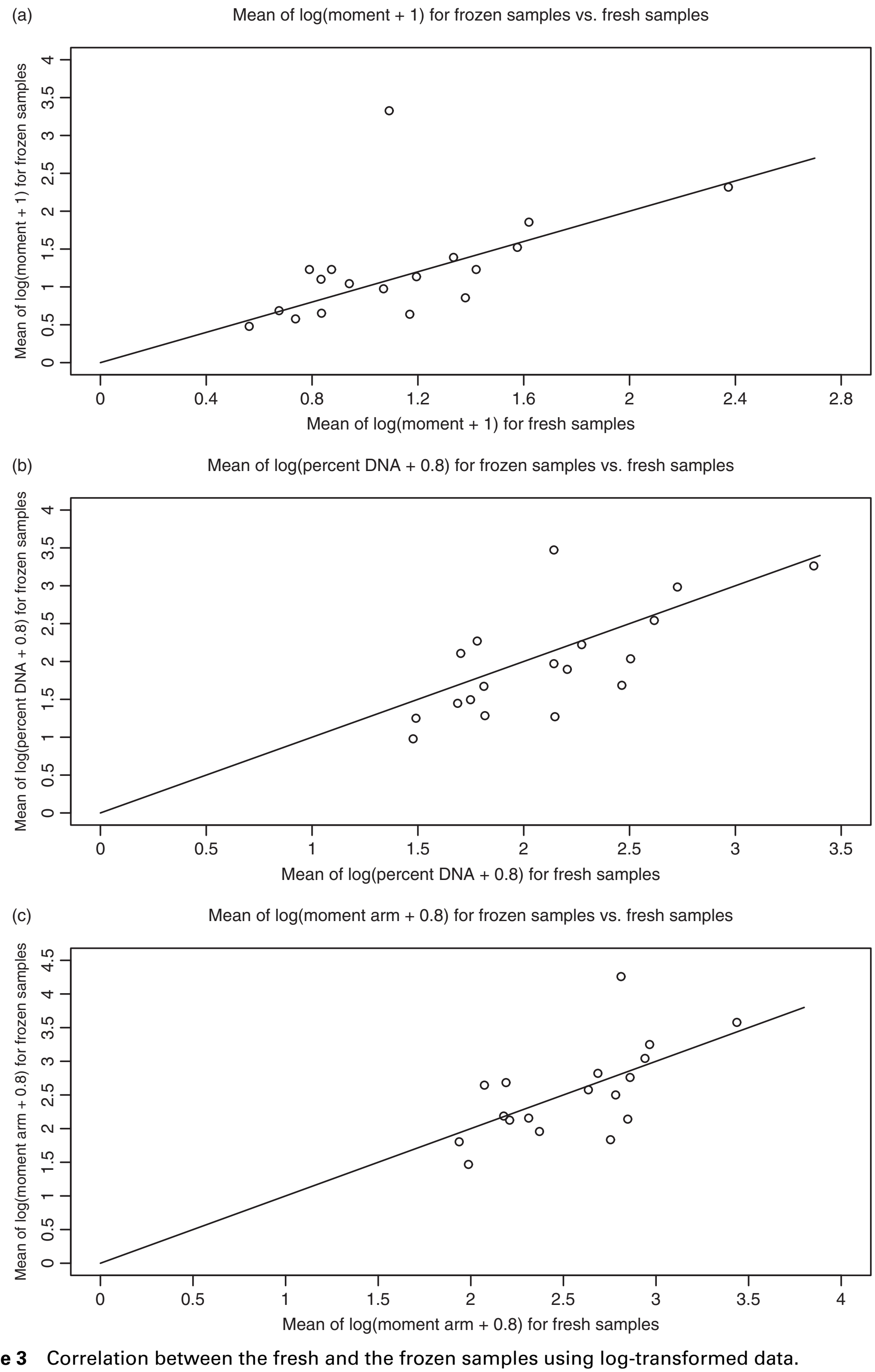

Figure 3 Correlation between the fresh and the frozen samples using log-transformed data.

where β_0, γ, ϕ_{02} and ϕ_{03} are unknown parameters, and $\beta_{1i} \sim N(0, \sigma^2_{\beta_1})$ and $\beta_{2ij} \sim N(0, \sigma^2_{\beta_2})$ represent random subject and visit effects, respectively and (2) for non-zero values of Y_{ijk}, we assume $\log Y_{ijk} | Y_{ijk} > 0 \sim N(\mu_{ijk}, \sigma^2)$, where

$$\mu_{ijk} = \alpha_0 + \alpha_{1i} + \alpha_{2ij} + \eta x_{1ijk} + \phi_{12} x_{2ijk} + \phi_{13} x_{3ijk}. \tag{2}$$

Again, the parameters α_0, η, ϕ_{12} and ϕ_{13} can be interpreted as the corresponding fixed effects, and α_{1i} and α_{2ij} are unobservable random subject effects and visit effects. We assume $\alpha_{1i} \sim N(0, \sigma^2_{\alpha_1})$ and $\alpha_{2ij} \sim N(0, \sigma^2_{\alpha_2})$. We further assume that all the random effects are mutually independent. We denote $\boldsymbol{\alpha}$ to be the collection of α_{1i} and α_{2ij}, and $\boldsymbol{\beta}$ to be the collection of β_{1i} and β_{2ij}, where $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots, J_i$.

In this study, we are not interested in evaluating the effects of length of freezing time on DNA damage, so the coefficients γ and η are estimates of the average effects of freezing for 6–12 weeks. Since x_{1ijk} is a shared covariate in both sub-models, its overall effects on DNA damage could be evaluated by the null hypothesis that the corresponding coefficients are both equal to zero. We note that a mixed model approach can be carried out by utilising generalised linear mixed models (GLMMs) for the zeros and the positive observations separately, but simultaneous evaluation of the freezing effects on both components would become impossible.

3.2 The EM algorithm

Let $\boldsymbol{\theta} = (\beta_0, \gamma, \phi_{02}, \phi_{03}, \alpha_0, \eta, \phi_{12}, \phi_{13}, \sigma^2, \sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \sigma^2_{\beta_1}, \sigma^2_{\beta_2})$ collect all the parameters in (1) and (2). The observed data likelihood of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$ involves integrals with respect to random effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which brings much computation burden when we attempt to obtain maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$. In this article we propose an EM algorithm for parameter estimation that makes use of the following complete-data likelihood:

$$\begin{aligned}
L(\theta | Y, \alpha, \beta) &\propto f(Y|\alpha, \beta, \theta) f(\alpha, \beta|\theta) \\
&= \prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left\{ \pi_{ijk}^{I(Y_{ijk}=0)} (1 - \pi_{ijk})^{I(Y_{ijk}>0)} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right. \right. \\
&\quad \left. \left. \times \exp\left(-\frac{1}{2\sigma^2} (\log(Y_{ijk}) - \mu_{ijk})^2 \right) \right]^{I(Y_{ijk}>0)} \right\} \\
&\quad \times \prod_{i=1}^{I} \prod_{j=1}^{J_i} \left\{ \frac{1}{\sqrt{2\pi\sigma^2_{\alpha_2}}} \exp\left(-\frac{1}{2\sigma^2_{\alpha_2}} \alpha^2_{2ij} \right) \frac{1}{\sqrt{2\pi\sigma^2_{\beta_2}}} \exp\left(-\frac{1}{2\sigma^2_{\beta_2}} \beta^2_{2ij} \right) \right\} \\
&\quad \times \prod_{i=1}^{I} \left\{ \frac{1}{\sqrt{2\pi\sigma^2_{\alpha_1}}} \exp\left(-\frac{1}{2\sigma^2_{\alpha_1}} \alpha^2_{1i} \right) \frac{1}{\sqrt{2\pi\sigma^2_{\beta_1}}} \exp\left(-\frac{1}{2\sigma^2_{\beta_1}} \beta^2_{1i} \right) \right\}.
\end{aligned} \tag{3}$$

The EM algorithm iterates between an E-step in which the expected logarithm of the complete-data likelihood (3) is computed conditional on the observed data and the current estimate of the parameters, and an M-step in which the parameter estimates are updated by maximising this expected log-likelihood. This requires that we solve the score equations in the maximisation step. There are no closed-form MLEs for parameters β_0, γ, ϕ_{02} and ϕ_{03}, and as a result we use a one-step Newton–Raphson algorithm. All the parameter estimates in the M-step depend on the conditional expectations of functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that appear in the complete data log-likelihood, which we denote by $E[h(\boldsymbol{\alpha}, \boldsymbol{\beta})]$. We use Gauss–Hermite quadrature to evaluate the integrals over the sample space of the random effects.[6,21] The algorithm iterates between E-step and M-step until the parameter estimates converge.

The standard errors of the components in $\hat{\boldsymbol{\theta}}$ can be obtained by parametric bootstrap with B repetitions from the estimated parametric model (1)–(2). For the bth bootstrap sample, where $b = 1, \ldots, B$, we apply the EM algorithm to derive the MLE $\hat{\boldsymbol{\theta}}^*_b$. The standard error estimates of the components in $\hat{\boldsymbol{\theta}}$ are the diagonal elements of the matrix $\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^*_b - \overline{\boldsymbol{\theta}}^*)(\hat{\boldsymbol{\theta}}^*_b - \overline{\boldsymbol{\theta}}^*)^T/(B-1)$, where $\overline{\boldsymbol{\theta}}^* = \sum_{b=1}^{B}\hat{\boldsymbol{\theta}}^*_b/B$. We used $B = 100$ in the application of the model to COMET assay (see Section 4.2).

4 Numerical results

4.1 Simulation study

We carried out a simulation study to assess the performance of the proposed estimation procedure for our zero-inflated model. The simulation study mimics the design of the COMET assay as shown in Figure 1, and 100 Monte Carlo samples were generated from model (1)–(2). We set $I = 50$ and $J_i = 3$ for all i. Based on the experimental design we have the following covariate patterns for the three visits: (1) for $j = 1$, we have (a) $x_1 = 0, x_2 = 0$ and $x_3 = 0$, (b) $x_1 = 1, x_2 = 1$ and $x_3 = 0$, (c) $x_1 = 1, x_2 = 0$ and $x_3 = 1$; (2) for $j = 2$, we have (a) $x_1 = 0, x_2 = 1$ and $x_3 = 0$, (b) $x_1 = 1, x_2 = 0$ and $x_3 = 1$; (3) for $j = 3$, we have $x_1 = 0$, $x_2 = 0$ and $x_3 = 1$. There were 50 observations simulated for each of the covariate pattern. The true values of the parameters in model (1)–(2), their estimated bias and standard deviation of the estimates across 100 simulations are presented in Table 1. The proposed EM algorithm appears to work reasonably well since the estimates are close to the true values. These results are based on 10 quadrature points in the E-step that were found to approximate the integrals satisfactorily.

Because in model (1)–(2) we assume that $\alpha_{1i}\perp\beta_{1i}$ and $\alpha_{2ij}\perp\beta_{2ij}$, which is a special case of a more general scenario where both $(\alpha_{1i}, \beta_{1i})$ and $(\alpha_{2ij}, \beta_{2ij})$ have bivariate normal distributions. We conducted a second set of simulations to examine the performance of our model when α and β are correlated. We simulated data with a structure similar to Table 1, but $(\alpha_{1i}, \beta_{1i})$ and $(\alpha_{2ij}, \beta_{2ij})$ were generated from bivariate normal distributions with correlation $\rho = 0.9$. The results based on 100 Monte Carlo samples are summarised in Table 2. All the parameters have small estimated bias and the standard deviations of the estimates are comparable to those in Table 1, although slightly larger variances are

Table 1 Simulation results where the data are sampled from model (1)–(2) with $\alpha_{1i} \perp \beta_{1i}$ and $\alpha_{2ij} \perp \beta_{2ij}$

Parameter	True value	Estimated bias	SD of the estimates.
Proportion of zeros			
Fixed effects			
β_0	−3.0	−0.026	0.279
γ	0.4	−0.006	0.170
ϕ_{02}	0.4	0.023	0.198
ϕ_{03}	0.8	0.008	0.255
Random effects			
$\sigma^2_{\beta_1}$	0.5	0.022	0.378
$\sigma^2_{\beta_1}$	0.7	−0.009	0.221
Positive values			
Fixed effects			
α_0	0.7	−0.006	0.031
η	0.3	−0.013	0.064
ϕ_{12}	−0.5	−0.018	0.091
ϕ_{13}	−0.3	−0.013	0.064
Random effects			
$\sigma^2_{\alpha_1}$	0.2	0.006	0.092
$\sigma^2_{\alpha_2}$	0.2	0.004	0.058
σ^2	2.5	0.010	0.053

observed for β_0 and $\sigma^2_{\beta_1}$. Similar results are obtained when the correlation $\rho = -0.9$ for both $(\alpha_{1i}, \beta_{1i})$ and $(\alpha_{2ij}, \beta_{2ij})$.

4.2 Application to COMET assay

We omitted one suspicious sample in the subsequent analysis because the proportion of zeros is over 80% in this sample, which is very unlikely in the COMET assay. Before applying the hierarchical zero-inflated model to the COMET assay, we need to determine a threshold for each measure, so that the sperm cells below the threshold are treated as undamaged (the zeros) and the cells above the threshold are classified as damaged (the non-zeros). One drawback of using natural zero as the thresholds for all the three variables – moment, percent DNA and moment arm – is that, after taking logarithm of the positive outcomes, the measures that are very close to zero will be exaggerated. Moreover, it is reasonable to classify the cells with negligible amount of DNA breaks as undamaged. Our preliminary analysis suggests the thresholds 0.05, 0.5 and 0.9 for moment, percent DNA and moment arm, respectively. The corresponding proportions of zeros are 0.19, 0.14 and 0.12.

The estimated parameters and their standard errors are shown in Table 3, and the significant findings are labelled by '†'. In calculating these estimates, 10 quadrature points were used in the E-step of the EM algorithm. For all the three COMET measures, freezing has no effect on the proportion of undamaged cells (the parameter γ), but increases DNA strand breaks measured by moment and moment arm in already damaged cells (the parameter η). The overall effects of freezing are significant, by testing the

Table 2 Simulation results where the data are sampled from model (1)–(2) except that both $(\alpha_{1i}, \beta_{1i})$ and $(\alpha_{2ij}, \beta_{2ij})$ have bivariate normal distributions with correlation 0.9

Parameter	True value	Estimated bias	SD of the estimates
Proportion of zeros			
Fixed effects			
β_0	−3.0	−0.050	0.307
γ	0.4	0.008	0.163
ϕ_{02}	0.4	0.002	0.197
ϕ_{03}	0.8	0.016	0.236
Random effects			
$\sigma^2_{\beta_1}$	0.5	0.043	0.448
$\sigma^2_{\beta_1}$	0.7	0.019	0.200
Positive values			
Fixed effects			
α_0	0.7	0.001	0.030
η	0.3	0.001	0.063
ϕ_{12}	−0.5	0.001	0.088
ϕ_{13}	−0.3	0.001	0.063
Random effects			
$\sigma^2_{\alpha_1}$	0.2	−0.010	0.095
$\sigma^2_{\alpha_2}$	0.2	−0.003	0.044
σ^2	2.5	0.003	0.050

hypothesis $H_0 : \gamma = \eta = 0$ for moment and moment arm, with $p - value = 0.0002$ and <0.0001, respectively. In addition, there are significant run effects on all the COMET measures (the parameters $\phi_{02}, \phi_{03}, \phi_{12}$ and ϕ_{13}), indicating that the assay results are also influenced by the experiment conditions. Compared to run 1, run 2 and run 3 tend to produce smaller COMET measures and more zeros. The estimates of σ^2_α and σ^2_β indicate around the same amount of inter- and intra-subject variations. As a comparison, we also analysed the log-transformed data using the linear mixed effects model (Proc Mixed in SAS version 9.1). The same log-transformation was carried out as in Figure 3. The method is not able to find significant freezing effects for moment and moment arm (results not shown) and the run effects are not significant for moment arm, although for run 2 and run 3, it does produce point estimates that are in the same direction as the zero-inflated model. As shown in the simulation study in Section 4.1, our zero-inflated model is able to estimate the parameters reasonably well for similar data structures. Those non-significant findings using the linear mixed model is probably due to the lack of fit of the model.

The goodness of fit of log-normal model (2) is evaluated by plotting the histograms and Q–Q plots of the residuals $e_{ij} = \log(Y_{ijk}) - \hat{\alpha}_0 - \hat{\alpha}_{1i} - \hat{\alpha}_{2ij} - \hat{\eta}x_{1ijk} - \hat{\phi}_{12}x_{2ijk} - \hat{\phi}_{03}x_{3ijk}$, where $\hat{\alpha}_{1i}$ and $\hat{\alpha}_{2ij}$ are empirical Bayes estimates of α_{1i} and α_{2ij}.[22] The plots show good agreement with the normality assumption for the residuals.

Table 3 Application of the hierarchical zero-inflated model to moment, percent DNA and moment arm, using the thresholds 0.05, 0.5 and 0.9, respectively (results shown are estimates (standard errors))

Parameter	Moment	Percent DNA	Moment arm
Proportion of zeros			
Fixed effects			
β_0	−2.72 (0.30)†	−3.38 (0.33)†	−3.49 (0.33)†
γ	0.12 (0.18)	0.24 (0.16)	0.24 (0.15)
ϕ_{02}	0.74 (0.20)†	0.81 (0.20)†	0.73 (0.21)†
ϕ_{03}	1.11 (0.26)†	1.23 (0.23)†	1.13 (0.27)†
Random effects			
$\sigma^2_{\beta_1}$	0.55 (0.31)	0.69 (0.39)	0.73 (0.34)
$\sigma^2_{\beta_2}$	0.58 (0.16)	0.62 (0.20)	0.55 (0.21)
Positive values			
Fixed effects			
α_0	0.86 (0.04)†	2.24 (0.03)†	2.82 (0.01)†
η	0.31 (0.08)†	−0.07 (0.06)	0.16 (0.03)†
ϕ_{12}	−0.43 (0.11)†	−0.36 (0.09)†	−0.09 (0.04)†
ϕ_{13}	−0.31 (0.08)†	−0.30 (0.06)†	−0.13 (0.03)†
Random effects			
$\sigma^2_{\alpha_1}$	0.23 (0.12)	0.16 (0.10)	0.03 (0.02)
$\sigma^2_{\alpha_2}$	0.21 (0.05)	0.14 (0.05)	0.08 (0.02)
σ^2	2.36 (0.05)	0.99 (0.02)	0.79 (0.02)
Freezing effects			
p–value for $H_0 : \gamma = \eta = 0$	0.0002	0.1858	<0.0001

Note: † for $p < 0.05$. We did not carry out Z-test for the random components since they are assumed to be positive.

5 Summary

We developed a hierarchical zero-inflated model for longitudinal data from a COMET assay to investigate the influence of cryopreservation on integrity of DNA in human sperm cells and to evaluate inter- and intra-subject variation of DNA strand breaks. The hierarchical structure is able to capture subject-specific random effects and the nested repeated sampling effects since each study subject has up to three lab visits separated by six-week intervals. An EM algorithm is derived to estimate the parameters of interest and the standard errors are obtained by parametric bootstrap. In the model specified in (1) and (2), we assume that the four random components, β_{1i}, β_{2ij}, α_{1i} and α_{2ij}, are mutually independent random normal variables. The model and the EM algorithm could be naturally extended to the situation where the random effects jointly have a multivariate normal distribution with potential correlations. However, it would require integration of the random effects over a higher dimensional space, which undoubtedly imposes much more computation burden on the EM algorithm. We provide some evidence that our model could be robust against correlated random effects in a simulation where α and β were sampled from bivariate normal distributions with correlation ρ = 0.9 (Table 2).

Acknowledgement

This work was supported in part by the University of California Toxic Substances Teaching and Research Program, and the UCLA Center for Occupational and Environmental Health.

References

1 Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* 2002; **11**: 341–55.

2 Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.

3 Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 2000; **56**: 1030–9.

4 Yau KKW, Lee AH, Carrivick PJW. Modeling zero-inflated count series with application to occupational health. *Computer Methods and Programs in Biomedicine* 2004; **74**: 47–52.

5 Kreuter F, Muthén B. Analyzing criminal trajectory profiles: bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology* 2008; **24**: 1–31.

6 Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* 2005; **5**: 1–19.

7 Lee AH, Wang K, Scott JA, Yau KK, McLachlan GJ. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research* 2006; **15**: 47–61.

8 Rabe-Hesketh S, Skrondal A. Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika* 2007; **72**: 123–40.

9 Zhou X, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* 1999; **55**: 645–51.

10 Tian L. Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Statistics in Medicine* 2005; **24**: 3223–32.

11 Hughes CM, Lewis SEM, Mckelvey-Martin VJ, Thompson W. A comparison of baseline and induced DNA damage in infertile man, using a modified comet assay. *Molecular Human Reproduction* 1996; **2**: 613–19.

12 Aravindan RG, Bjordahi J, Jost LK, Evenson DP. Susceptibility of human sperm to in situ DNA denaturation is strongly correlated with DNA strand breaks identified by single-cell electrophoresis. *Experimental Cell Research* 1997; **236**: 231–7.

13 Young KE, Robbins WA, Xun L, Elashoff D, Rothmann S, Perreault SD. Evaluation of chromosome breakage and DNA integrity in sperm: an investigation of remote semen collection conditions. *Journal of Andrology* 2003; **24**: 853–61.

14 Singh NP, Muller CH, Berger RE. Effects of age on DNA double-strand break and apoptosis in human sperm. *Fertility and Sterility* 2003; **80**: 1420–30.

15 Trisini AT, Singh NP, Duty SM, Hauser R. Relationship between human semen parameters and deoxyribonucleic acid damage assessed by the neutral comet assay. *Fertility and Sterility* 2004; **82**: 1623–32.

16 Evenson DP, Wixon R. Environmental toxicants cause sperm DNA fragmentation as detected by the sperm chromatin structure assay. *Toxicology and Applied Pharmacology* 2005; **207**: s532–s537.

17 Chohan KR, Griffin JT, Lapromboise M, De Jonge CJ. Comparison of chromatin assays for DNA fragmentation evaluation in human sperm. *Journal of Andrology* 2006; **27**: 53–9.

18 Migliore L, Naccarati A, Zanello A, Scarpato R, Bramanti L, Mariani M. Assessment of sperm DNA integrity in workers exposed to styrene. *Human Reproduction* 2002; **17**: 2912–18.

19 Van Kooij RJ, De Boer P, De Vreeden-Elbertse MT, Ganga NA, Singh N, Te Velde ER. The neutral comet assay detects double strand DNA damage in selected and unselected human spermatozoa of

normospermic donors. *International Journal of Andrology* 2004; **27**: 140–6.

20 Xun L, Robbins WA, Lim KL, Li N. Influence of cryopreservation on integrity of DNA in human ejaculated spermatozoa over time: study of variation in DNA strand breaks between subjects and within subjects. *In preparation.*

21 Press WH, Teutolsky SA, Vetterling WT, Flannery BP. *Numerical recipes in FORTRAN. The art of scientific computing* (2nd edn). Cambridge University Press, New York; 1992.

22 Weiss RE. *Modeling longitudinal data.* Springer, New York; 2005.

Appendix

The EM Algorithm

In the E-step of the tth iteration, we need to evaluate

$$\begin{aligned} E[h(\alpha,\beta)] &= \int h(\alpha,\beta) f(\alpha,\beta|Y,\theta^{(t)})\mathrm{d}\alpha\mathrm{d}\beta \\ &= \frac{\int h(\alpha,\beta) f(Y|\alpha,\beta,\theta^{(t)}) f(\alpha,\beta|\theta^{(t)})\mathrm{d}\alpha\mathrm{d}\beta}{f(Y|\theta^{(t)})} \\ &= \frac{\int h(\alpha,\beta) f(Y|\alpha,\beta,\theta^{(t)}) f(\alpha,\beta|\theta^{(t)})\mathrm{d}\alpha\mathrm{d}\beta}{\int f(Y|\alpha,\beta,\theta^{(t)}) f(\alpha,\beta|\theta^{(t)})\mathrm{d}\alpha\mathrm{d}\beta}. \end{aligned} \tag{4}$$

The density functions $f(Y|\alpha,\beta,\theta^{(t)})$ and $f(\alpha,\beta|\theta^{(t)})$ have closed-forms as given in (3).

In the M-step, the estimates are updated as follows:

$$\beta_0^{(t+1)} = \beta_0^{(t)} + \frac{\sum_i \sum_j \sum_k (I(Y_{ijk}=0) - E_i[\pi_{ijk}])}{\sum_i \sum_j \sum_k (E[\pi_{ijk}] - E[\pi_{ijk}^2])} \tag{5}$$

$$\gamma^{(t+1)} = \gamma^{(t)} + \frac{\sum_i \sum_j \sum_k x_{1ijk}(I(Y_{ijk}=0) - E[\pi_{ijk}])}{\sum_i \sum_j \sum_k x_{1ijk}^2(E[\pi_{ijk}] - E[\pi_{ijk}^2])} \tag{6}$$

$$\phi_{02}^{(t+1)} = \phi_{02}^{(t)} + \frac{\sum_i \sum_j \sum_k x_{2ijk}(I(Y_{ijk}=0) - E[\pi_{ijk}])}{\sum_i \sum_j \sum_k x_{2ijk}^2(E[\pi_{ijk}] - E[\pi_{ijk}^2])} \tag{7}$$

$$\phi_{03}^{(t+1)} = \phi_{03}^{(t)} + \frac{\sum_i \sum_j \sum_k x_{3ijk}(I(Y_{ijk}=0) - E[\pi_{ijk}])}{\sum_i \sum_j \sum_k x_{3ijk}^2(E[\pi_{ijk}] - E[\pi_{ijk}^2])} \tag{8}$$

$$\begin{aligned}\alpha_0^{(t+1)} = \Bigg\{ \sum_i \sum_j \sum_k I(Y_{ijk} > 0)\{\log(Y_{ijk}) - E[\alpha_{1i} + \alpha_{2ij}] - \eta^{(t)} x_{1ijk} \\ - \phi_{12}^{(t)} x_{2ijk} - \phi_{13}^{(t)} x_{3ijk}\}\Bigg\} \Bigg/ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) \qquad (9)\end{aligned}$$

$$\begin{aligned}\eta^{(t+1)} = \Bigg\{ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) x_{1ijk}\{\log Y_{ijk}) - \alpha_0^{(t+1)} - E[\alpha_{1i} + \alpha_{2ij}] \\ - \phi_{12}^{(t)} x_{2ijk} - \phi_{13}^{(t)} x_{3ijk}\}\Bigg\} \Bigg/ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) x_{1ijk} \qquad (10)\end{aligned}$$

$$\begin{aligned}\phi_{12}^{(t+1)} = \Bigg\{ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) x_{2ijk}\{\log(Y_{ijk}) - \alpha_0^{(t+1)} - E[\alpha_{1i} + \alpha_{2ij}] \\ - \eta^{(t+1)} x_{1ijk} - \phi_{13}^{(t)} x_{3ijk}\}\Bigg\} \Bigg/ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) x_{2ijk} \qquad (11)\end{aligned}$$

$$\begin{aligned}\phi_{13}^{(t+1)} = \Bigg\{ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) x_{3ijk}\{\log(Y_{ijk}) - \alpha_0^{(t+1)} - E[\alpha_{1i} + \alpha_{2ij}] \\ - \eta^{(t+1)} x_{1ijk} - \phi_{12}^{(t+1)} x_{2ijk}\}\Bigg\} \Bigg/ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) x_{3ijk} \qquad (12)\end{aligned}$$

$$\begin{aligned}\sigma^{2(t+1)} = \Bigg\{ \sum_i \sum_j \sum_k I(Y_{ijk} > 0) E[(\log(Y_{ijk}) - \alpha_0^{(t+1)} - E[\alpha_{1i} + \alpha_{2ij}] \\ - \eta^{(t+1)} x_{1ijk} - \phi_{12}^{(t+1)} x_{2ijk} - \phi_{13}^{(t+1)} x_{3ijk})^2]\Bigg\} \qquad (13) \\ \Bigg/ \sum_i \sum_j \sum_k I(Y_{ijk} > 0)\end{aligned}$$

$$\sigma_{\alpha_1}^{2(t+1)} = \frac{1}{I} \sum_i E[\alpha_{1i}^2] \qquad (14)$$

$$\sigma_{\beta_1}^{2(t+1)} = \frac{1}{I} \sum_i E[\beta_{1i}^2] \tag{15}$$

$$\sigma_{\alpha_2}^{2(t+1)} = \frac{1}{\sum_i J_i} \sum_i \sum_j E[\alpha_{2ij}^2] \tag{16}$$

$$\sigma_{\beta_2}^{2(t+1)} = \frac{1}{\sum_i J_i} \sum_i \sum_j E[\beta_{2ij}^2]. \tag{17}$$