SIM 2147
pp: 1--17 (col.fig.: Nil)
PROD. TYPE: COM
ED: Chandra
PAGN: Suresh – SCAN: Raghwo

STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:000–000
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2147

1

Estimating treatment efficacy over time: a logistic regression model for binary longitudinal outcomes

3

Leena Choi[1,*,†], Francesca Dominici[1], Scott L. Zeger[1] and Peter Ouyang[2]

5 [1]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, U.S.A.*
7 [2]*Johnson & Johnson Pharmaceutical Research & Development, L.L.C., Biometrics and Reporting, 920 Route 202, P.O. Box 300, NJ 08869, U.S.A.*

9

SUMMARY

This paper presents a case study in longitudinal data analysis where the goal is to estimate the efficacy 11 of a new drug for treatment of a severe chronic constipation. Data consist of long sequences of binary outcomes (relief/no relief) on each of a large number of patients randomized to treatment (low and 13 high dose) or placebo. Data characteristics indicate: (1) the treatment effects vary non-linearly with time; (2) there is substantial heterogeneity across subjects in their responses to treatment; and (3) there 15 is a high proportion of subjects who never experience any relief (the non-responders).

To overcome these challenges, we develop a hierarchical model for binary longitudinal data with a 17 mixture distribution on the probability of response to account for the high frequency of non-responders. While the model is specified conditionally on subject-specific latent variables, we also draw inferences 19 on key population-average parameters for the assessment of the treatments' efficacy in a population. In addition we employ a model-checking method to compare the goodness-of-fit for our model against 21 simpler modelling approaches for aggregated counts, such as the zero-inflated Poisson and zero-inflated negative binomial models.

23 We estimate subject-specific and population-average rate ratios of relief for the treatment with respect to the placebo as functions of time (RR_t), and compare them with the rate ratios estimated from the 25 models for aggregated counts. We find that: (1) the treatment is effective with respect to the placebo with higher efficacy at the beginning of the study; (2) the estimated rate ratios from the models for 27 aggregated counts appear to be similar to the average across time of the population-average rate ratios estimated under our model; and (3) model-checking suggests that the hierarchical and zero-inflated 29 negative binomial model fit the data best.

If we are mainly interested to establish the overall efficacy (or safety) of a new drug, it is appropriate 31 to aggregate the longitudinal data over time and analyse the count data by use of standard statistical methods. However, the models for aggregated counts cannot capture time trend of treatment such as the 33 initial treatment benefit or the development of tolerance during the early stage of the treatment which

*Correspondence to: Leena Choi, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, U.S.A.
†E-mail: lchoi@jhsph.edu

Contract/grant sponsor: Johnson & Johnson Pharmaceutical Research & Development

Received 2 December 2003
Accepted 1 November 2004

may be important information to physicians to predict the treatment effects for their patients.

KEY WORDS: binary longitudinal; random effect; mixture; zero-inflated Poisson regression; zero-inflated negative binomial; model checking

1. INTRODUCTION

Inclinical trials, repeated measurements are often taken over time to evaluate efficacy and safety of a new drug for treatment of a chronic disease. In some instances, outcomes are binary indicators of relief of a symptom or of occurrence of an adverse effect. This paper is motivated by a case study in longitudinal data analysis where the goal is to estimate the efficacy of a new drug for treatment of a severe chronic constipation from a phase III clinical trial. Data consist of long sequences of repeated binary outcomes (relief/no relief) on patients randomized to treatment (low and high dose) or placebo. The data are characterized by: (1) non-linearity of the treatment effects over time; (2) heterogeneity among subjects in their responses to treatment; and (3) high frequency of subjects who never experience any relief (non-responders). These data characteristics render standard methods for the analysis of longitudinal data less suitable for an appropriate assessment of the treatment effect which needs to take into account all sources of uncertainty.

Statistical methods for analysis of longitudinal binary data have been rapidly developed during recent years. Diggle *et al.* [1] and Cox *et al.* [2] provide the detailed review. One alternative to longitudinal data analysis is to aggregate the repeated binary observations for each subject over time and then analyse the total number of responses using Poisson regression. Since the repeated measurements for each person are likely to be correlated, we expect extra-Poisson and/or extra-binomial variation [3] which would result in a poor fit to the standard generalized linear model [4]. A number of extensions have been proposed for aggregations of correlated binary outcomes; the readers can find an annotated bibliography in Reference [5].

In this paper we develop a hierarchical model for the analysis of binary longitudinal data having a mixture distribution on the probability of response. Our modelling approach allows estimation of the subject-specific and population-averaged rate ratios of relief (response) for the treatment with respect to the placebo as a smooth function of time taking into account the high frequency of subjects who achieve no relief (non-responders) as well as the heterogeneity of subjects who do experience some relief (responders).

To check our model and compare it with alternatives, we simulate data sets under each fitted model and graphically display departures of the simulated data sets from the observed data. We implement our model checking method to compare goodness-of-fit of our hierarchical model for longitudinal data *vs* the following models for aggregated counts data: (1) the negative binomial (NB) [6]; (2) zero-inflated Poisson (ZIP) [7]; and (3) zero-inflated negative binomial (ZINB) [8].

In Section 2, we describe the case study and the data set. In Section 3, we introduce our hierarchical model for longitudinal analysis and illustrate the estimation approaches of the subject-specific and population-average rate ratios of relief by Monte-Carlo Markov chain (MCMC) methods. In Section 4, we briefly review models for aggregated counts and illustrate

our model-checking method. In Section 5, we summarize the results. A discussion of the results and modelling strategies is in Section 6.

2. DATA

The motivating data of this paper are obtained from a phase III double-blind randomized clinical trial evaluating the efficacy of a new drug for treatment of severe chronic constipation. The primary efficacy outcome is a relief of constipation defined as the occurrence of 'spontaneous complete bowel movement'. A bowel movement (defecation) was defined as 'spontaneous' only if the subject recorded 'No' to the diary question 'Did you take laxatives in the 24 hours preceding that bowel movement?'. A bowel movement was considered 'complete' only if the subject recorded 'Yes' to the diary question 'Did the stool make you feel like you completely emptied your bowels?' From these two items in the patients' diary recorded daily, the primary outcome variable is constructed as a binary response 1 if the subject experience 'spontaneous' and 'complete' bowel movement on each day, 0 otherwise. Subjects who have less than 2 days of relief of constipation per week during a 2-week drug-free run-in period were eligible. After the run-in period, 641 subjects were randomized to one of three groups: placebo, and two doses of the drug (low and high dose). Patients were treated once daily with an oral preparation for 12 consecutive weeks (84 days). In summary, the primary outcome is a binary response Y_{ij} indicating a relief of constipation (response) for subject i on day j. Predictors of interest are time (t_{ij}) which ranges from 1 to 84 days, and treatment indicators ($x_{1i}=1$ if low dose and 0 otherwise; $x_{2i}=1$ if high dose and 0 otherwise).

Table I summarizes the total number of subjects (N_k) for each treatment group (k) where $k=P, L, H$ are indices for placebo (P), low dose (L) and high dose (H); the median and the range of follow-up days across subjects (n_i); the range of the total number of responses during follow-up days ($Y_i=\sum_{j=1}^{n_i} Y_{ij}$); and the average fraction of days with response ($\sum_{i=1}^{N_k} P_i/N_k$, where $P_i=Y_i/n_i$). The average fractions of days with response are 0.138, 0.218 and 0.216 for the placebo, low and high dose groups, respectively, suggesting a possible treatment benefit.

Figures 1 and 2 show the daily relief rate (observed per cent of responders, $P_j=\sum_{i=1}^{n_j} Y_{ij}/n_j$, where n_j is the number of subjects at time j) for each treatment group plotted against follow-up time. For both the high and low dose group, the relief rate of treatment is higher than for placebo for most of days. The relief rate on treatment decreases over time more rapidly at first, the relief rate for placebo increases with time.

Table I. Total number of subjects (N_k) where $k=P, L, H$ are indices for placebo (P), low dose (L) and high dose (H), median and range of follow-up days across subjects (n_i), range of the total number of responses during follow-up days ($Y_i=\sum_{j=1}^{n_i} Y_{ij}$), and average fraction of days with response ($P_i=Y_i/n_i$).

	N_k	Median of n_i	Range of n_i	Range of Y_i	Mean of P_i
Placebo	212	83	5–84	0–81	0.138
Low dose	214	84	2–84	0–75	0.218
High dose	215	83	1–84	0–82	0.216

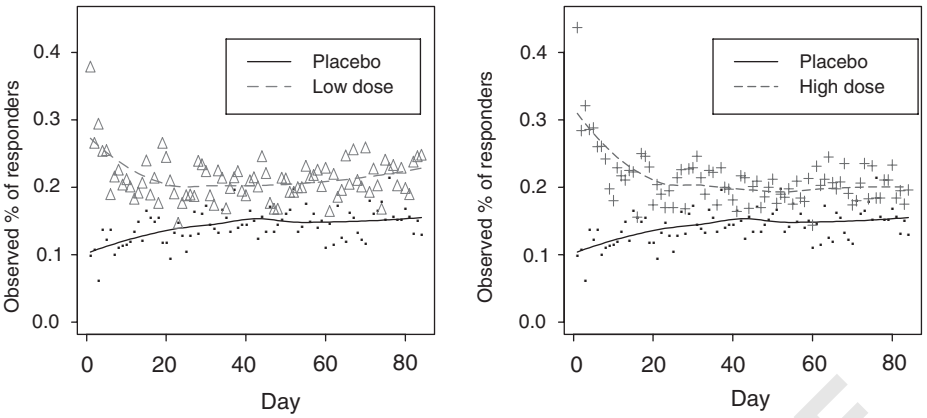

Figure 1. Daily relief rate (observed per cent of responders) over time for treatments (low and high doses of the drug) with respect to the placebo. The points are observed per cent of responders (the dots for placebo, the triangles for low dose and the crosses for high dose) and the lines are smoothing curves over the points of observed per cent of responders for each treatment group.

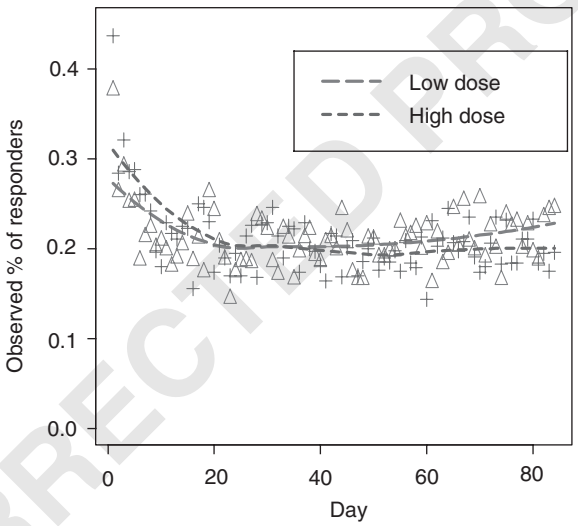

Figure 2. Daily relief rate (observed per cent of responders) over time for low dose with respect to high dose.

We expect that the repeated measurements on the same subject will be correlated. Serial correlation of repeated binary outcomes can be explored by use of the empirical lorelogram [9, 10]. The lorelogram is simply the log odds ratio between observations at each pair of points t_j and t_k defined as

$$\mathrm{LOR}(t_j, t_k) = \log\, \Psi(Y_{ij}, Y_{ik}) \quad \text{where}$$

$$\Psi(Y_{ij}, Y_{ik}) = \frac{P[Y_{ij}=1, Y_{ik}=1]P[Y_{ij}=0, Y_{ik}=0]}{P[Y_{ij}=1, Y_{ik}=0]P[Y_{ij}=0, Y_{ik}=1]}$$

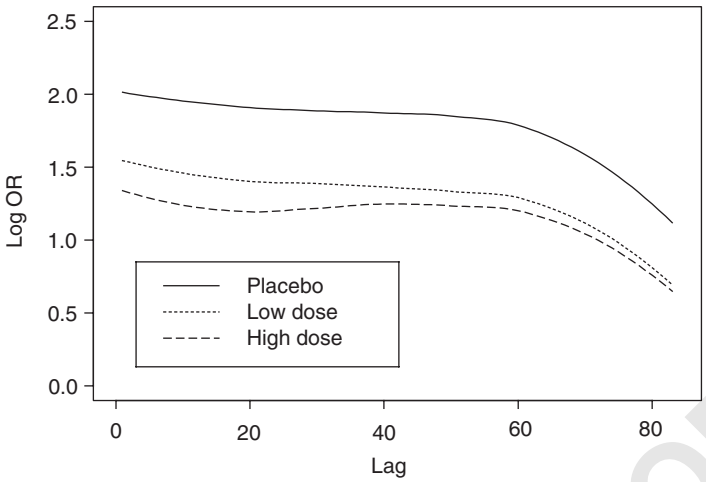

Figure 3. Lorelogram: estimated log odds ratio as function of lag time.

Figure 3 shows the estimated log odds ratio as function of lag time $|t_j - t_k|$ separately for each treatment group. This figure indicates substantial serial dependence (odds ratio >3) for long lags upto 80 days with greater serial dependence for placebo than treatment. This suggests that the treatment diminishes the serial dependence compared to placebo, in other words, the treatment reduces the variability of random effects among subjects in the treatment group compared in placebo. Presence of serial correlation at long lags is consistent with subjects having their own latent propensity of disease that is not fully captured by the covariates. Random intercept model is therefore an appropriate choice to start. Notice that the log odds ratio between observations at very long lag cannot be estimated precisely since there are not much information, for example, there is at most one pair of observations at lag 83 for each subject, which may lead the estimation of lorelogram after 60 days less precise.

Figure 4 shows the histogram of fraction of the total number of days with response ($P_i = Y_i/n_i$) for each treatment group. It shows a high proportion of patients who never experience the relief, for whom the fractions of days with response are zeros. Histograms indicate that there may be two groups of people: the non-responders represented by high proportion of zero fraction of days with response (big piles at zero in the left panel) and the responders who experience at least one response during the follow-up days (normal density shape histogram as log odds scale in the right panel). In addition, the proportions of zero fraction of days with response under the treatment are smaller than under placebo suggesting that the treatment is effective in reducing the proportion of non-responders. Thus, histograms in the log odds scale clearly indicate that the distribution of log odds of a response can usefully be modelled as a mixture distribution with different probability of non-responders for each treatment.

In summary, our exploratory analysis suggests that there are several characteristics of these data that make the evaluation of treatment effect challenging: (1) distinct non-linear time trends for each of the three groups; (2) high serial correlation; (3) evidence of a mixture distribution for the log odds of probability of a response.

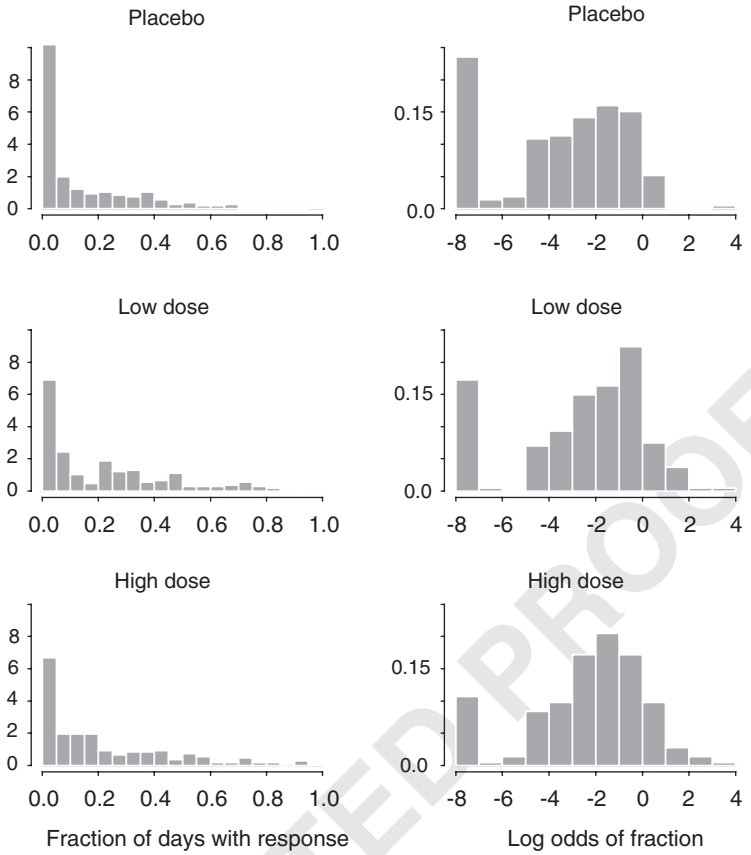

Figure 4. Fraction of days with response, original scale and log odds.

3. LONGITUDINAL DATA ANALYSIS APPROACH

In this section, we describe a Bayesian hierarchical model for longitudinal data analysis of binary outcomes which takes into account the data characteristics illustrated in Section 2. We assume that Y_{ij} (1 for relief for subject i on day j; 0 otherwise) has a Bernoulli distribution with probability p_{ij}. To take into account the high frequency of non-responders, we assume that there are two subpopulations: non-responders who never experience any relief and responders who do experience some relief. We denote by θ_k the probability that a subject taking treatment k belongs to the non-responders group, where $k = P, L, H$ are indices for placebo (P), low dose (L) and high dose (H), respectively. With probability $(1 - \theta_k)$, we specify a logistic regression model where the logit of p_{ij} is modelled as function of covariates including: (1) main treatment effects; (2) a natural cubic spline of time with two knots to allow for non-linear treatment effects; (3) a random intercept (u_i) to approximate the serial dependence.

Table II. Definitions of subject-specific rate ratios of relief ($RR_t(u)$) as function of model parameters under the hierarchical model.

$RR_t(u)_{L:P} = \frac{P_{t(L)}(u_i)}{P_{t(P)}(u_i)}$	$\frac{(1-\theta_L)}{(1-\theta_P)} \frac{e^{\beta_0+u+ns(t;3)+\beta_1+ns(t;3)}}{1+e^{\beta_0+u+ns(t;3)+\beta_1+ns(t;3)}} \Big/ \frac{e^{\beta_0+u+ns(t;3)}}{1+e^{\beta_0+u+ns(t;3)}}$
$RR_t(u)_{H:P} = \frac{P_{t(H)}(u_i)}{P_{t(P)}(u_i)}$	$\frac{(1-\theta_H)}{(1-\theta_P)} \frac{e^{\beta_0+u+ns(t;3)+\beta_2+ns(t;3)}}{1+e^{\beta_0+u+ns(t;3)+\beta_2+ns(t;3)}} \Big/ \frac{e^{\beta_0+u+ns(t;3)}}{1+e^{\beta_0+u+ns(t;3)}}$
$RR_t(u)_{H:L} = \frac{P_{t(H)}(u_i)}{P_{t(L)}(u_i)}$	$\frac{(1-\theta_H)}{(1-\theta_L)} \frac{e^{\beta_0+u+ns(t;3)+\beta_2+ns(t;3)}}{1+e^{\beta_0+u+ns(t;3)+\beta_2+ns(t;3)}} \Big/ \frac{e^{\beta_0+u+ns(t;3)+\beta_1+ns(t;3)}}{1+e^{\beta_0+u+ns(t;3)+\beta_1+ns(t;3)}}$

In summary, we use the following modelling approach:

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit } p_{ij} &= \beta_0 + u_i + \text{ns}(t_{ij}) + \beta_1 x_{1i} + \beta_2 x_{2i} \\
&\quad + x_{1i} \times \text{ns}(t_{ij}) + x_{2i} \times \text{ns}(t_{ij}) \quad \text{w.p. } (1-\theta_k) \\
\text{logit } p_{ij} &= -\infty \quad \text{w.p. } \theta_k \\
u_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

where $\text{ns}(t_{ij})$ is a natural cubic spline of time with two knots at days 20 and 60, x_{1i} and x_{2i} are indicators of the low and high dose treatments, respectively. The two knots at days 20 and 60 were chosen to reflect both our scientific understanding of the drug and empirical consideration as described below. A similar class of drug for treatment of constipation often shows rapid pharmacodynamic action and greater response at the beginning of treatment, but quickly develops tolerance to the treatment during early period of treatment, and finally the treatment effect stabilizes and persists. This time trend could be found in a previous phase III clinical trials for a similar class of drug [11]. As shown in Figures 2 and 3, the exploratory analysis also indicates the change of the treatment response occurred around days 20 and 60. In addition, the two knots are roughly equally dividing the total number of days in the data at days 20 and 60 to ensure flexibility in modelling non-linear time trend.

Parameters of interest are both subject-specific and population-average rate ratio (RR) of relief for treatment (high dose and low dose) with respect to placebo. Table II summarizes the definition of subject-specific rate ratios ($RR_t(u)$) as function of model parameters and random effects. Subject-specific rate ratios of relief are defined as ratios of subject-specific probability of having a relief for treatment (high dose and low dose) with respect to placebo. The subject-specific probability of having a relief for each treatment as a function of time can be calculated by the subject-specific probability of having a relief given being a responder multiplied by the probability of being a responder for each treatment group. For example, the subject-specific probability of having a relief for placebo, $P_{t(P)}(u_i)$, can be calculated by the subject-specific probability of having a relief given being a responder for placebo $e^{\beta_0+u+ns(t;3)}/(1+e^{\beta_0+u+ns(t;3)})$ multiplied by the probability of being a responder for placebo $(1-\theta_P)$ where $\text{ns}(t;3)$ denotes the natural cubic spline of time in the model.

Population-average rate ratios of relief are defined as ratios of population-average probability of having a relief for treatment (high dose and low dose) with respect to placebo. For example, the population-average RR for high dose with respect to placebo is

$$\overline{\mathrm{RR}}_{t(H:P)} = \frac{\int P_{t(H)}(u_i) g(u_i)\,\mathrm{d}u_i}{\int P_{t(P)}(u_i) g(u_i)\,\mathrm{d}u_i} \tag{1}$$

where $\int P_{t(H)}(u_i) g(u_i)\,\mathrm{d}u_i$ and $\int P_{t(P)}(u_i) g(u_i)\,\mathrm{d}u_i$ are the population-average probability of having a relief, and $P_{t(H)}(u_i)$ and $P_{t(P)}(u_i)$ are the subject-specific probability of having a relief for the high dose and placebo, respectively, and $g(u_i)$ is the distribution of random effect u_i, which is normal in this model.

We fit the model under a Bayesian framework using MCMC methods implemented with WinBUGS [12, 13]. We assign non-informative priors on the unknown parameters and we assume that θ_k is U[0, 1], $1/\sigma^2$ is Gamma(0.001, 0.001), and the regression coefficients β are N(0, 10 000). Posterior distributions of the $\mathrm{RR}_t(u)$ can be easily obtained by applying formulas in Table II to the posterior samples of the model parameters [14]. We approximate the posterior distribution of the population-average RR using numerical integration methods. More specifically, let $\boldsymbol{\eta}^{(m)} = (\boldsymbol{\beta}^{(m)}, \boldsymbol{\theta}^{(m)}, \sigma^{2(m)})$ be the mth posterior sample of model parameters. For each m, we:

- simulate the vector of random effects $\mathbf{u}^{l,m}$ from $\mathrm{N}(0, \sigma^{2(m)})$ for $l = 1, \ldots, L = 1000$;
- calculate the subject-specific probability of having a relief $P_{t(k)}(\mathbf{u}^{l,m})$ as a function of $\boldsymbol{\eta}^{(m)}$ and $\mathbf{u}^{l,m}$ where $k = P, L, H$;
- calculate the population-average probability of having a relief by averaging each subject-specific probability with respect to the L random effects;
- calculate the population-average rate ratio $\overline{\mathrm{RR}}_t$ by taking ratio of population-average probability of having a relief for treatment (high dose and low dose) with respect to placebo.

4. STATISTICAL MODELS FOR AGGREGATED DATA AND MODEL COMPARISON

In this section, we briefly describe three modelling approaches for count data which are obtained by aggregating subject-specific binary outcomes over time. The basic structure of each model relies upon a Poisson regression having two main treatment effects (x_{1i} and x_{2i}) defined as in Section 3 and an offset equal to the number of follow-up days (n_i) to take into account the different follow-up time among subjects.

The aggregated data present two major challenges that prevent us from using a standard Poisson regression model. First, they show substantial heterogeneity across subjects which leads to extra-binomial variation [3, 4]. Second, the number of zero counts due to non-responders appears to exceed the predicted number under a Poisson model.

We compare our hierarchical modelling approach for the longitudinal data with the following three modelling approaches for the aggregated counts data: (1) the ZINB model; (2) the NB model; and (3) the ZIP regression model as a special case of ZINB.

Table III. Definitions of rate ratios of relief (RR) as function of model parameters under models for counts data.

	$\mathrm{RR}_{L:P} = \frac{P(Y_i\|L)}{P(Y_i\|P)}$	$\mathrm{RR}_{H:P} = \frac{P(Y_i\|H)}{P(Y_i\|P)}$	$\mathrm{RR}_{H:L} = \frac{P(Y_i\|H)}{P(Y_i\|L)}$
NB	e^{β_1}	e^{β_2}	$e^{(\beta_2-\beta_1)}$
ZIP	$e^{\beta_1}\frac{1+e^{\gamma_0}}{1+e^{\gamma_0+\gamma_1}}$	$e^{\beta_2}\frac{1+e^{\gamma_0}}{1+e^{\gamma_0+\gamma_2}}$	$e^{(\beta_2-\beta_1)}\frac{1+e^{\gamma_0+\gamma_1}}{1+e^{\gamma_0+\gamma_2}}$
ZINB	$e^{\beta_1}\frac{1+e^{\gamma_0}}{1+e^{\gamma_0+\gamma_1}}$	$e^{\beta_2}\frac{1+e^{\gamma_0}}{1+e^{\gamma_0+\gamma_2}}$	$e^{(\beta_2-\beta_1)}\frac{1+e^{\gamma_0+\gamma_1}}{1+e^{\gamma_0+\gamma_2}}$

Let Y_i be the total number of responses for subject i during the study period. Under ZINB we assume:

$$
\begin{aligned}
Y_i \mid u_i &\sim \text{Poisson}(v_i)\\
\log v_i &= \log n_i + \beta_0 + u_i + \beta_1 x_{1i} + \beta_2 x_{2i} \quad \text{w.p. } 1-\theta_i\\
\log v_i &= -\infty \quad \text{w.p. } \theta_i\\
e^{u_i} &\sim \text{Gamma}(1/\alpha,\ 1/\alpha)\\
\text{logit}\,\theta_i &= \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}
\end{aligned}
$$

where $v_i = E(Y_i|u_i)$. Notice that ZINB [8, 15, 16] takes into account over-dispersion due to both extra-zeros and heterogeneity among subjects. More specifically we assume a two-part regression model where with probability $1-\theta_i$ we model Y_i as negative binomial with a random effect u_i to take into account heterogeneity (dispersion parameter, α). In addition we assume that the probability of having zero count (θ_i) varies by treatment (x_{1i} and x_{2i}) with γ_0, γ_1 and γ_2 as regression coefficients.

Notice that under the assumption $\theta_i = 0$, ZINB becomes NB which takes into account over-dispersion by allowing heterogeneity, but assumes everyone is capable of relief. Under the assumption $\text{Var}(u_i) = 0$, then ZINB becomes ZIP which takes into account of over-dispersion by allowing a two-part Poisson regression model to model the extra-zeros. Therefore by assessing the goodness-of-fit of these modelling approaches, we can gain insight into the source of over-dispersion.

Table III summarizes the definitions of rate ratios of relief as function of model parameters under the NB, ZIP, and ZINB models.

4.1. Model checking

In this section, we illustrate an exploratory tool that compares the goodness-of-fit among the three models for aggregated counts data and the hierarchical model for longitudinal binary data. Our method applied half-normal plots [17, 18] to compare simulated data from each

model to the observed data, and it can be described in the following steps:

1. for the hierarchical model, randomly choose 50 samples of parameter estimates from the MCMC, say $\boldsymbol{\eta}^{(m)}$. For each $m = 1, \ldots, 50$, simulate a new data set $Y_{ij}^{(m)}$ from our hierarchical model described in Section 3 with model parameters equal to $\boldsymbol{\eta}^{(m)}$. We calculate the simulated aggregated count $Y_i^{(m)} = \sum_{j=1}^{n_i} Y_{ij}^{(m)}$ and the observed aggregated count $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ to compare the hierarchical model with the models NB, ZIP, and ZINB which are aggregated count models. For the models NB, ZIP, and ZINB, simulate 50 new data sets from the fitted models obtained by setting parameter values equal to their maximum likelihood estimates;
2. obtain $Y_i^{(m,o)}$ by ordering the simulated $Y_i^{(m)}$ and obtain $Y_i^{(o)}$ by ordering the observed data set Y_i;
3. calculate mean and percentiles of $Y_i^{(m,o)}$ across m, and plot these summaries *vs* $Y_i^{(o)}$ for each model and compare these with the line of complete agreement.

5. RESULTS

Figure 5 shows the posterior means and 95 per cent posterior intervals of subject-specific rate ratio (RR) as function of time and for a range of values for random intercept u between -3.5 and 3.5 ($\pm 2\hat{\sigma}$). Also shown is the population-average RR which is obtained by applying equation (1). We found that, on average over the study population, both the high and low doses have a higher rate of relief than placebo (panel at the bottom right). Posterior means of $\overline{\mathrm{RR}}_{t\,(H:P)}$ and $\overline{\mathrm{RR}}_{t\,(L:P)}$ vary over time from 2.7 at $t = 1$ to 1.6 at $t = 84$ and from 2.4 at $t = 1$ to 1.9 at $t = 84$, respectively. That is, at the beginning and after 12 weeks of the treatment, patients who took high dose of the drug are on average 2.7 and 1.6 times more likely to experience relief than patients under placebo. Similarly, at the beginning and after 12 weeks of the treatment, patients who took low dose of the drug are on average 2.4 and 1.9 times more likely to experience relief than patients under placebo. There is little or no evidence of a difference in response between the high and low dose groups. Note that the rate ratio is close to 1.0 over the entire 84 day period.

The subject-specific RR are plotted in correspondence of random effects u equal to $(2\hat{\sigma}, 1\hat{\sigma}, 0, -1\hat{\sigma}, -2\hat{\sigma})$, where $\hat{\sigma}^2$ is the posterior mean of σ^2 under the hierarchical model ($\hat{\sigma} = 1.7$). For larger values of the random effects u, the posterior mean and variance of the subject-specific RR is smaller: if a person has a larger propensity to respond (large random intercept, i.e. $u = 2\hat{\sigma} = 3.5$), then he/she would have higher probability of having a relief whether he/she is treated or not. For those people, the subject-specific RR for the high dose *vs* placebo would be similar to the subject-specific RR for low dose *vs* placebo and the subject-specific RR would not change much over time. On the other hand, when the random effects u is small, we estimate a larger posterior mean and variance of the subject-specific RR: if a person has small propensity to respond (small random intercept, i.e. $u = -2\hat{\sigma} = -3.5$) possibly due to severe disorder, then the treatment effects can be maximized with higher uncertainty depending on the severity of disorder or personal characteristics. An extreme case of small propensity to respond will be a non-responder who might be due to functional disorder in bowel movements or already developed tolerance to similar class of drugs. However, for even

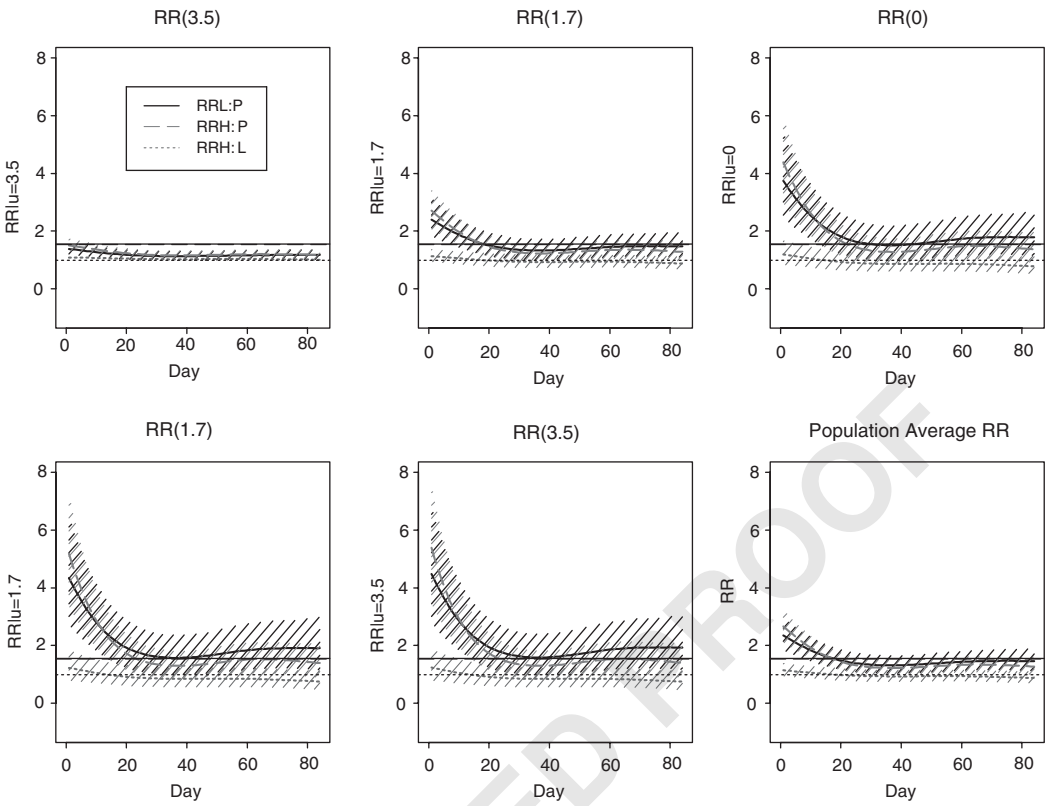

Figure 5. Posterior means and 95 per cent posterior intervals of subject-specific rate ratios (RR(u)) for random intercept $u = (-3.5, -1.7, 0, 1.7, 3.5)$ and population-average rate ratios defined by equation (1) as function of time under the hierarchical model. The estimated rate ratios under ZINB are presented as horizontal lines.

those non-responders, the treatment shows an effect by reducing the probability of being a non-responder compared to placebo.

For the random effect, $u = 0$, the posterior means of $\mathrm{RR}_t(0)_{H:P}$ and $\mathrm{RR}_t(0)_{L:P}$ are estimated to vary over time from 4.4 at $t = 1$ to 1.3 at $t = 84$ and from 3.7 at $t = 1$ to 1.5 at $t = 84$, respectively. Compared to the population-average RR, the subject-specific $\mathrm{RR}_t(0)$ shows larger estimates and much more variability than the population-average RR. In fact, population-average coefficients are in general smaller than subject-specific coefficients, and the degree of the difference depends on the variance of the random effect [19, 20], which is substantial in this case study. The estimated RRs were not sensitive to the exact location of knots at 20 and 60 days.

Figure 6 shows posterior distributions of variance of the random intercept estimated from Bayesian hierarchical model allowing a different random intercept for each treatment group ($\theta_k = 0$ for all k, but use different σ^2 for each treatment in our model) and from Bayesian hierarchical model with a mixture distribution on the probability of response also allowing a

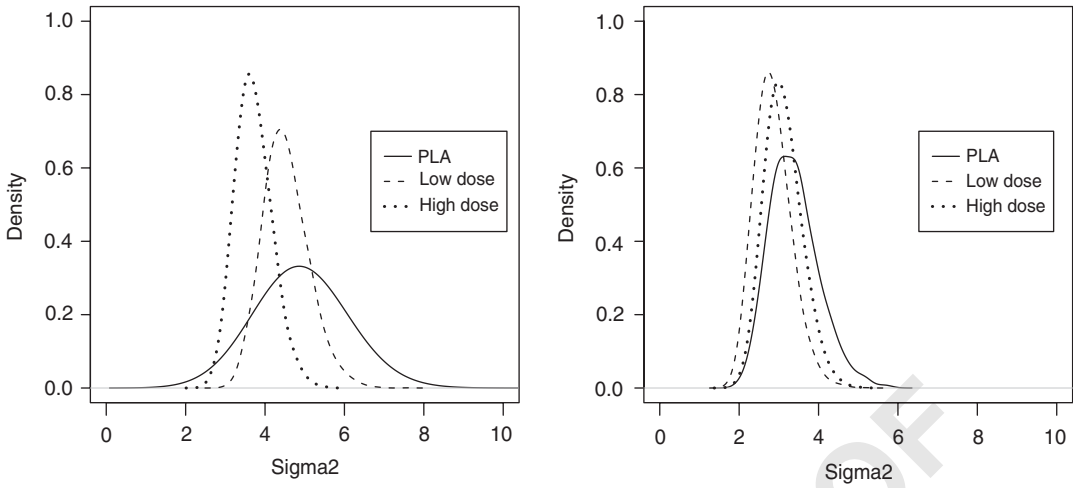

Figure 6. Posterior distributions of variance of the random intercept estimated from Bayesian hierarchical model allowing a different random intercept for each treatment group (left) and from Bayesian hierarchical model with a mixture distribution on the probability of response (right).

Table IV. Maximum likelihood estimates and standard errors* of population-average rate ratios under the models NB, ZIP, and ZINB.

Model	$RR_{L:P}$	$RR_{H:P}$	$RR_{H:L}$
NB	1.56 (0.21)	1.54 (0.21)	0.99 (0.13)
ZIP	1.53 (0.18)	1.50 (0.18)	0.98 (0.10)
ZINB	1.56 (0.20)	1.54 (0.20)	0.99 (0.12)

Between parentheses are denoted the standard errors.
*Obtained with the delta method.

different random intercept for each treatment group (use different σ^2 for each treatment in our model). We can see that the variability of the random intercept from Bayesian hierarchical model with a different random intercept (left) is reduced among subjects in the treatment group compared in placebo as we expected from the lorelogram (Figure 3). However, if we take into account non-responders using a mixture distribution on the probability of response, the variabilities of the random intercept are reduced in all the treatment groups and almost similar across the treatment groups (right), which led to use the same σ^2 for all treatment groups in our final hierarchical model. This indicates that much of the differences in variabilities of random effect among subjects in placebo compared to the treatment was due to non-responders.

Table IV summarizes point estimates and standard errors of the RRs. The standard errors were estimated and compared by using both the delta method and bootstrap. The two methods provided very similar estimates. In Figure 5, the estimated RRs under ZINB are presented as horizontal lines over time for the comparison with the population-average RR and the subject-specific RR. $RR_{L:P}$ and $RR_{H:P}$ overlapped as one solid line due to their similarity. The estimated RRs under ZINB appear to be similar to the average across time of population-

Table V. Maximum likelihood estimates and standard errors of model parameters for NB, ZIP, and ZINB.

	β_0	β_1	β_2	α	γ_0	γ_1	γ_2
NB	−1.981	0.446	0.433	1.837			
	(0.096)	(0.135)	(0.135)	(0.109)			
	(0.089)*	(0.115)*	(0.116)*	(0.108)*			
ZIP	−1.654	0.310	0.233		−1.021	−0.512	−0.890
	(0.021)	(0.027)	(0.027)		(0.157)	(0.238)	(0.261)
	(0.081)*	(0.105)*	(0.108)*		(0.157)*	(0.238)*	(0.261)*
ZINB	−1.806	0.371	0.285	1.367	−1.656	−0.593	−1.895
	(0.098)	(0.131)	(0.129)	(0.139)	(0.296)	(0.432)	(1.143)
	(0.094)*	(0.112)*	(0.115)*	(0.115)*	(0.308)*	(0.430)*	(1.171)*

*Standard errors are estimated by use of a robust variance estimator [21–23].

average RRs obtained under the hierarchical model. Thus, even though the simpler models for the aggregated counts data do not allow to estimate RR as a function of time, the conclusion about the overall treatment effect would remain the same. However, we should notice that ZINB cannot capture the initial treatment benefit or the development of tolerance during the early stage of the treatment.

Table V summarizes the maximum likelihood estimates, standard errors, and robust standard errors [21–23] of the parameters under the models for the aggregated counts data. We notice that under the ZIP model the robust standard errors are about four times greater than the model-based standard errors implying that modelling extra zeros is not enough to take into account over-dispersion. The estimate of the variance of the random effects α under NB and ZINB indicates over-dispersion due to heterogeneity. The estimate of α under ZINB is smaller than under NB, suggesting that part of over-dispersion in ZINB is explained by modelling extra zeros. This finding is consistent with the smaller estimate of the variance of random effect from Bayesian hierarchical model with a mixture distribution as shown in Figure 6. The estimates of γ_0, γ_1 and γ_2 under ZIP are greater than under ZINB, indicating that ZIP captures over-dispersion through modelling extra zeros only, whereas ZINB can take into account over-dispersion also through the variance of random effects.

Figure 7 shows the plots of simulated $(Y_i^{(m,o)})$ which are ordered simulated data $Y_i^{(m)}$ explained in Section 4 *vs* the ordered observed data $(Y_i^{(o)})$ under the hierarchical model and under the models for aggregate counts. Figure 8 shows the same plots under the hierarchical model but separately for the three sub-periods, during 1–4, 5–8 and 8–12 week. In both figures, under the hierarchical model the line of complete agreement lies entirely within the 95 per cent quantile of $Y_i^{(m,o)}$ implying that our hierarchical model is consistent with the marginal distribution of total number of responses. Among the three alternative models, ZINB is the best, although, the plot is skewed in the high value of observed values. Similar phenomenon occurs for NB whereas goodness-of-fit for the ZIP model is very poor. Therefore, in connection with the result in Table V, our method for comparing goodness-of-fit between models suggests that over-dispersion can be mostly explained by the heterogeneity, although modelling the excess in zero counts can improve the fit also. Finally, if the goal is to evaluate overall treatment effect, then ZINB is the best alternative.

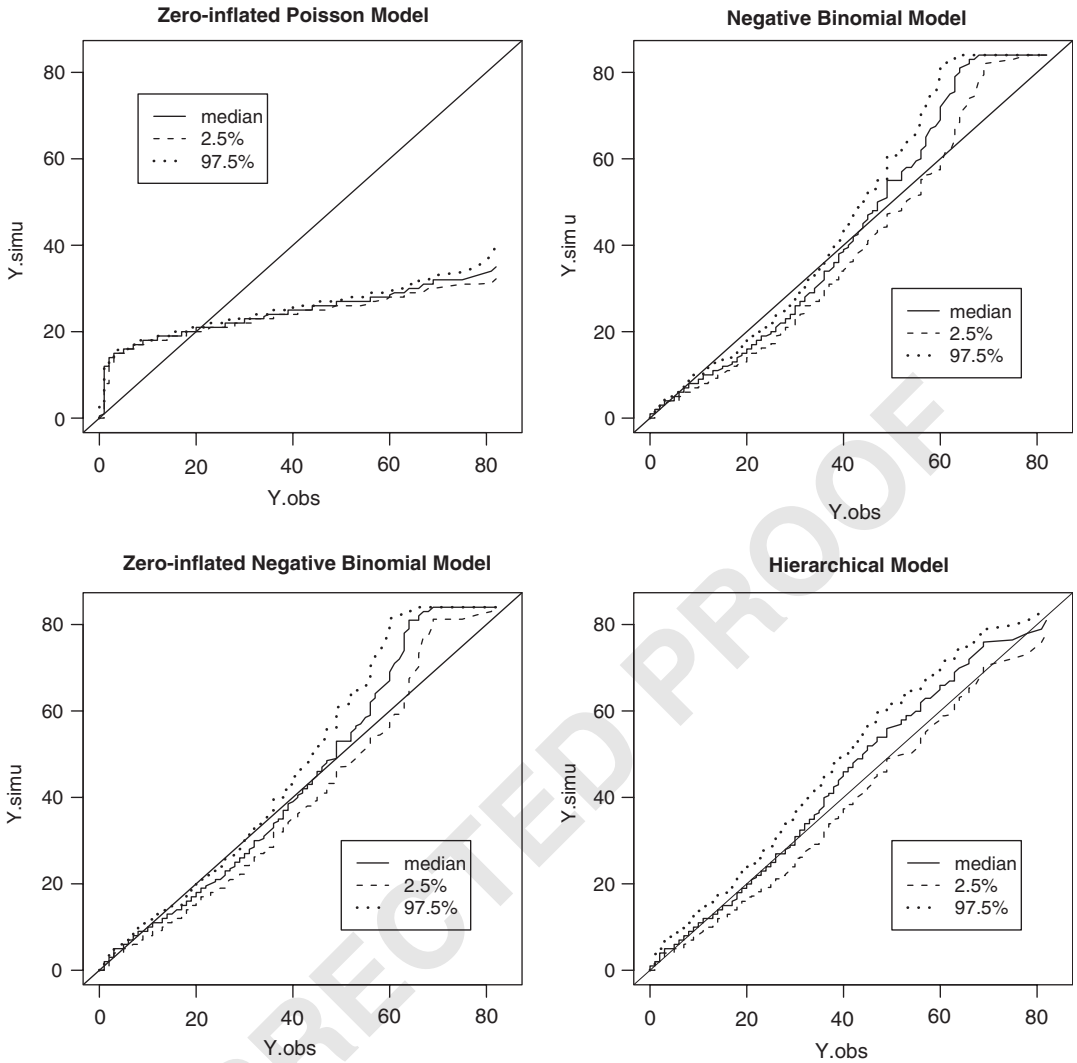

Figure 7. The plots of simulated $Y_i^{(m,o)}$ *vs* observed $Y_i^{(o)}$ for the hierarchical model and models for the aggregated counts data.

6. DISCUSSION

Motivated by a randomized clinical trial of a treatment for a severe chronic constipation, we developed a hierarchical model for longitudinal data analysis of binary outcomes. Our model takes account of: (1) non-linear time trends for the treatment effect in the placebo and treatment groups; (2) a dose-specific parametric contrast between the treatment and placebo groups; (3) serial correlation; and (4) a mixture distribution for the log odds of probability of re-

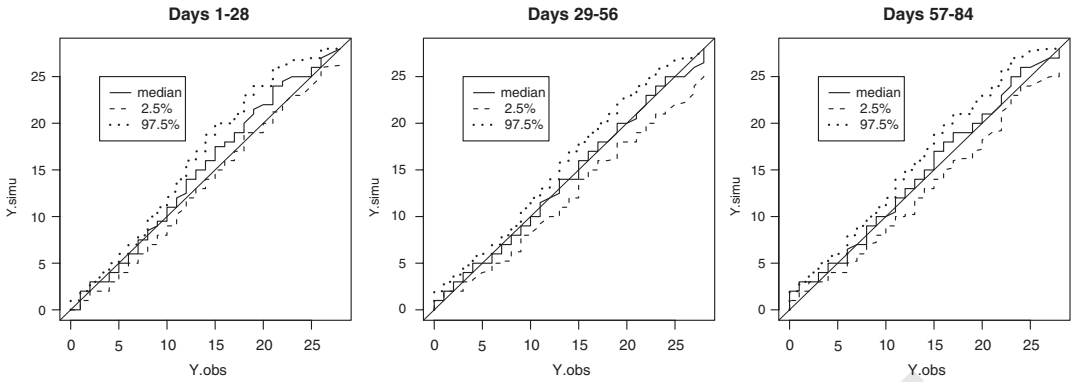

Figure 8. The plots of simulated $Y_i^{(m,o)}$ *vs* observed $Y_i^{(o)}$ for the hierarchical model during three sub-periods of time.

sponse. We estimated posterior distributions of subject-specific and population-average rate ratios of relief for the treatment with respect to the placebo as functions of time (RR_t).

Both subject-specific and population-average rate ratios have their own important interpretations. Drug regulatory authorities would generally be interested in the population-average rate ratios since new drug must show efficacy for the whole population, whereas doctors who should treat individual patient are likely to be more interested in the subject-specific rate ratios. The posterior mean and variance of the subject-specific rate ratios showed that patients with high-risk (low propensity to respond) are more likely to have larger benefit of treatment than patients with low-risk (high propensity to respond) [24, 25]. The estimated rate ratios from simpler modelling approaches for the aggregated counts are similar to the average across time of the population-average rate ratios from our hierarchical model for longitudinal data.

We found that: (1) the treatment is more effective than the placebo throughout the 84 days, having the efficacy decreases at the beginning of the study; (2) there is little difference in the efficacy of the drug between the high and low doses; and (3) even though the simpler models do not produce estimates of the treatment effect as a function of time, the conclusion about the overall treatment effect is similar across modelling approaches.

To compare our hierarchical model for the analysis of binary longitudinal data with simpler modelling approaches, we developed a model-checking method to assess goodness-of-fit. Our approach suggested that the hierarchical model fits the data best. In addition, the comparison of the goodness-of-fit between models of increasing complexity provided a better characterization of the sources of over-dispersion. Our model-checking method suggested that most of over-dispersion in data can be explained by the heterogeneity, but modelling the excess in zero counts improved the fit also.

The statistical analysis of longitudinal binary outcomes adapted in this paper is similar to the one recently proposed by Carlin *et al.* [26]. These authors presented a Bayesian hierarchical model with a mixture distribution to estimate risk of smoking in teenagers as functions of covariates, and they compared their hierarchical formulation with the following alternative modelling approaches: (1) a logistic-regression model with GEE [27]; and (2) a logistic regression model with normal distributions on the random effects. Differently from the

approach adapted by Carlin *et al.* [26], we compared rate ratios from models for aggregated counts *vs* population-average rate ratios obtained by marginalizing over the random effect distribution. In addition, to assess treatment efficacy over time, we estimated rate ratios under the hierarchical model as smooth function of time (RR_t).

Longitudinal data analysis methods and statistical analysis of the aggregated count data have their own pros and cons, and either approach could be legitimate depending on the scientific question. In a longitudinal data analysis, we use the maximum amount of information available in the data. Most importantly, we can estimate the time course of treatment efficacy or adverse effect. In this motivating example, there is clear evidence of changing efficacy with time. Such a result could be valuable to developing guidelines for treatment of individual patients or can be used to plan future clinical trial or drug development for a similar series of new drugs.

On the other hand, if we are mainly interested to establish the overall efficacy (or safety) of a new drug, then it is appropriate to aggregate the longitudinal data over time and analyse the count data using standard statistical methods as done here. Although the models for aggregated data such as ZINB can estimate overall treatment effect and fit the data well, they cannot capture time trend of treatment (e.g. the initial treatment benefit or the development of tolerance during the early stage of the treatment) which may be useful information to physicians to predict the treatment effects for their patients.

ACKNOWLEDGEMENTS

We gratefully acknowledge a grant from Johnson & Johnson Pharmaceutical Research & Development for partial support of this research.

REFERENCES

1. Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data* (2nd edn). Oxford University Press: Oxford, 2002.
2. Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds). *Time Series Models*: *In Econometrics*, *Finance and Other Fields*. Chapman and Hall: London, 1996.
3. Winkelmann R. Duration dependence and dispersion in count-data models. *Journal of Business and Economic Statistics* 1995; **13**:467–474.
4. Haseman JK, Kupper LL. Analysis of dichotomous response data from certain toxicological experiment. *Biometrics* 1979; **35**:281–293.
5. Ashby M, Neuhaus JM, Hauck WW, Bacchetti P, Heilbrow DC, Jewell NP, Segal MR, Fusaro RE. An annotated bibliography of methods for analyzing correlated categorical data. *Statistics in Medicine* 1992; **11**:67–99.
6. Fisher RA. The negative binomial distribution. *Annals of Eugenics* 1941; **11**:182–187.
7. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**:1–14.
8. Greene WH. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *Technical Report*, Department of Economics, Stern School of Business, New York University, 1994.
9. Heagerty PJ. Multivariate multinomial marginal models. *Ph.D. Thesis*, Department of Biostatistics, Johns Hopkins University, 1995.
10. Heagerty PJ, Zeger SL. Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association* 1998; **93**:150–162.
11. Müller-Lissner SA, Fumagalli I, Bardhan KD, Pace F, Pecher E, Nault B, Rüegg P. Tegaserod, a 5-HT_4 receptor partial agonist, relieves symptoms in irritable bowel syndrome patients with abdominal pain, bloating and constipation. *Alimentary Pharmacology and Therapeutics* 2001; **15**(10):1655–1666. DOI:10.1046/j.1365-2036.2001.01094.x
12. Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. *BUGS*: *Bayesian Inference Using Gibbs Sampling*, *Version 0.5*. MRC Biostatistics Unit: Cambridge, U.K., 1996.
13. Albert I, Jais JP. Gibbs sampler for the logistic model in the analysis of longitudinal binary data. *Statistics in Medicine* 1998; **17**:2905–2921. DOI: 10.1002/(SICI)1097-0258(19981230)17:24<2905::AID-SIM911>3.0.CO;2-G

14. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall/CRC: London/Boca Raton, FL, 1995.
15. Cameron AC, Trivedi P. Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1986; **1**:29–55.
16. Welsh AH, Cunningham RB, Chambers R. Methodology for estimating the abundance of rare animals: seabird nesting on north east Herald cay. *Biometrics* 2000; **56**:22–30. DOI:10.1111/j.0006-341X.2000.00022.x
17. Demétrio CGB, Hinde JP. Half-normal plots and overdispersion. *GLIM Newsletter* 1997; **27**:19–26.
18. Vieira AMC, Hinde JP, Demétrio CGB. Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics* 2000; **27**:373–389. DOI: 10.1080/02664760021673
19. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
20. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology* 1998; **147**:694–703.
21. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press: Berkeley, CA, 1967; 221–233.
22. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; **48**:817–830.
23. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**:1–25.
24. Harrell FE. *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer: Berlin, 2002.
25. Knaus WA, Harrell FE, Fisher CJ, Wagner DP, Opal SM, Sadoff JC, Draper EA, Walawander CA, Conboy K, Grasela TH. The clinical evaluation of new drugs for sepsis: a prospective study design based on survival analysis. *Journal of the American Medical Association* 1993; **270**:1233–1241.
26. Carlin JB, Wolfe R, Brown CH, Gelman A. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* 2001; **2**(4):397–416. DOI:10.1093/biostatistics/2.4.397
27. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.

