Health Serv Outcomes Res Method (2012) 12:302–319
DOI 10.1007/s10742-012-0083-0

# BLUP(REMQL) estimation of a correlated random effects negative binomial hurdle model

**Sung Hee Kim · Chung-Chou H. Chang · Kevin H. Kim · Michael J. Fine · Roslyn A. Stone**

Received: 24 June 2011 / Revised: 22 February 2012 / Accepted: 5 March 2012 /
Published online: 20 March 2012


**Abstract** Bed days is a potentially useful metric of efficiency in clinical studies involving the hospital admission decision. However, this metric involves excess zeros, possible overdispersion, and possible clustering (in multi-site studies). A random effects negative binomial hurdle model can account for each of these issues. We extend this model to include site-level correlation between the two component parts and implement best linear unbiased prediction-type estimation with restricted maximum quasi-likelihood. This approach offers computational advantages over maximum likelihood in a generalized linear mixed model setting. Simulations show that the proposed approach performs well for fixed effects and variance components under a plausible range of bivariate correlation. The Emergency Department Community Acquired Pneumonia study motivates this work and illustrates the methods.

**Keywords** Excess zeros · Overdispersion · Generalized linear mixed model · Maximum likelihood · Best linear unbiased prediction (BLUP)-type estimation · Restricted maximum quasi-likelihood

## 1 Introduction

Community-acquired pneumonia (CAP) is a common, costly, and often fatal illness with more than 4 million episodes in the United States each year (Hsu et al. 2010). Providing

S. H. Kim (✉)
Duke Clinical Research Institute, Duke University Medical Center, Durham, NC 27705, USA
e-mail: sunghee0701@gmail.com

C.-C. H. Chang · M. J. Fine
Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA

K. H. Kim
Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA

R. A. Stone
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

quality and cost-effective care in patients with CAP has important economic and public health implications. The direct medical care costs of treating pneumonia are almost $10 billion per year, with the cost of inpatient treatment being 20 times higher than that of outpatient treatment (Fine et al. 2000; Niederman et al. 1998). Because inpatient cost is comprised mainly of the cost of hospitalizations, reducing the admission rate of low risk patients and reducing the length of stay (LOS) for inpatients with CAP could contribute substantially to medical care cost savings and efficient health care utilization (Fine et al. 2000). Measures of efficiency of care include the probability of outpatient treatment and LOS (Brown et al. 2003). An alternative measure of efficiency that includes both components is "bed days", defined as zero for outpatients and LOS for inpatients, where LOS is the difference between discharge and admission dates (Wang et al. 2002). Bed days has problematic statistical characteristics, including excess zeros and possible overdispersion. In addition, clustering may be present, as in multi-site studies or repeat hospitalizations.

Finite mixture models, including zero-inflated models and hurdle models, commonly are used to allow for excess zeros. In a g-component Poisson mixture model, the number of components must be estimated (Schlattmann et al. 1996; Wang et al. 1996). For bed days, the number of components is known to be two, i.e., inpatients and outpatients. Although the zero-inflated and hurdle models accommodate counts with excess zeros (Cunningham and Lindenmayer 2005; Min and Agresti 2002; Ridout et al. 1998; Welsh et al. 1996), we will not consider zero-inflated models here because the zero-inflated model presumes that zeros can occur in both component distributions. Hurdle models have been used in economic applications and health care services (Arulampalam and Booth 1997; Gurmu 1998; Pohlmeier and Ulrich 1995). A hurdle model is a two-component mixture with a binomial part (probability of passing the "hurdle") and a Poisson or negative binomial part. A hurdle model is more appropriate than a zero-inflated model for the outcome of bed days, because all patients are at risk for hospitalization when they present to a site (hospital), but zero bed days occur only among outpatients.

The hurdle model has been extended to account for clustering (e.g., by site) using maximum likelihood (ML) estimation in a generalized linear mixed model (GLMM) framework. ML estimation integrates out random effects from the joint likelihood using numerical approximations (Min and Agresti 2005). Although efficient, ML involves intensive computing and may not converge. In addition, ML can give biased estimates of variance components for random effects. An alternative method to estimate variance components in the GLMM setting, best linear unbiased prediction (BLUP)-type estimation with restricted maximum quasi-likelihood (REMQL), requires less integration and produces less biased estimates of variance components relative to ML (McGilchrist 1994; McGilchrist and Yau 1995). Although the BLUP(REMQL) approach has been used to estimate random effects in some finite mixture models (including zero-inflated models), it has not been implemented for the random effects hurdle model. A Bayesian approach also has been implemented to reduce small-sample bias and avoid asymptotic approximations or estimation of functions of parameters (Neelon et al. 2010). However, the Bayesian approach still is computationally intensive.

In this paper, we develop BLUP(REMQL) estimation for a correlated random effects hurdle model. We consider Poisson and negative binomial hurdle models and allow the binomial and count components to be correlated at the site level. In Sect. 2, we develop a procedure to estimate the fixed effect parameters. Estimation of the variance components is described in Sect. 3, and the scale parameter estimation is derived in Sect. 4. In Sect. 5, we apply the method to analyze bed days in the multi-site Emergency Department Community Acquired Pneumonia (EDCAP) study (Yealy et al. 2004, 2005). We describe a simulation

study to investigate the validity of the proposed estimation procedure for the negative binomial hurdle model in Sect. 6. Section 7 concludes with a discussion.

## 2 Hurdle model with correlated random effects

Let $Y_j$ ($j = 1, 2, \ldots, n$) be the number of bed days for patient $j$, where the total number of patients is $n$. Because the Poisson hurdle model is a special case of the negative binomial hurdle model, we describe the negative binomial hurdle model here (Pohlmeier and Ulrich 1995).

$$P(Y_j = 0) = p_j, \tag{1}$$

$$P(Y_j = y_j | y_j > 0) = (1 - p_j) \cdot \frac{f(y_j)}{1 - f(0)},$$

where $p_j$ indicates the conditional probability of not passing the hurdle (i.e., not being hospitalized) given patient $j$ is at risk for hospitalization and $f$ is a negative binomial distribution.

This model was extended by Min and Agresti (2005) to include random effects. Let $Y_{ij}$ ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n_i$) be the number of bed days of patient $j$ at site $i$, when $m$ is the number of sites, $n_i$ is the number of patients at site $i$, and the total number ($n$) of patients is $\sum_{i=1}^{m} n_i$. Then, the negative binomial hurdle model with random effects is:

$$P(Y_{ij} = 0) = p_{ij}, \tag{2}$$

$$\begin{aligned} P(Y_{ij} = y_{ij} | y_{ij} > 0) &= (1 - p_{ij}) \cdot \frac{f(y_{ij})}{1 - f(0)} \\ &= (1 - p_{ij}) \binom{y_{ij} + k - 1}{y_{ij}} \frac{t_{ij}^k (1 - t_{ij})^{y_{ij}}}{1 - t_{ij}^k}, \end{aligned}$$

where $p_{ij}$ indicates the conditional probability of not passing the hurdle given patient $j$ at site $i$ is at risk for hospitalization, $\mu_{ij}$ is the mean of the underlying negative binomial distribution, $y_{ij}! = y_{ij} \times (y_{ij} - 1) \times \cdots \times 1$, $t_{ij} = (k/k + \mu_{ij})$, and $k$ is the scale parameter (which is equal to 1/dispersion parameter). Note that the probability ($p_{ij}$) can be modeled by logistic regression and $f(y_{ij})/1 - f(0)$ can be regarded as a truncated negative binomial distribution. When $k$ goes to infinity, the negative binomial hurdle model reduces to the Poisson hurdle model. In the regression setting, both $\text{logit}(p_{ij})$ and $\log(\mu_{ij})$ are assumed to depend on linear functions of covariates. Following notation for the two-component mixture model in Wang et al. (2007), the linear predictors $\xi_{ij}$ and $\eta_{ij}$ are defined by

$$\text{logit}(p_{ij}) = \xi_{ij} = \boldsymbol{w}_{ij}^T \boldsymbol{\alpha} + u_i, \tag{3}$$

$$\log(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i,$$

where $\boldsymbol{w}_{ij}$ and $\boldsymbol{x}_{ij}$, respectively, are vectors of covariates for the logistic and the negative binomial distributions, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding vectors of coefficients. Here, $u_i$ and $v_i$ denote site-level random effects ($i = 1, \ldots, m$), where $\boldsymbol{r}_i^T = (u_i, v_i)^T$ is assumed to be distributed as $N(\boldsymbol{0}, \boldsymbol{D})$ (i.e., a random intercept model). Given the site-level random effects, the two components (binomial part and negative binomial part) are assumed to be independent.

We introduce correlation between the binomial and count components through the covariance matrix $\boldsymbol{D}$ of $\boldsymbol{r}_i^T$ where

$$\boldsymbol{D} = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}, \qquad i = 1, 2, \ldots, m \tag{4}$$

and $\boldsymbol{\rho}$ denotes a bivariate correlation between the random effects. In the case of uncorrelated random effects, $\rho = 0$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are assumed to be independently distributed as N(0, $\sigma_u^2 \mathrm{I}_m$) and N(0, $\sigma_v^2 \mathrm{I}_m$) respectively, where $\mathrm{I}_m$ denotes an $m \times m$ identity matrix. In the uncorrelated case, the logistic regression and negative binomial regression components can be estimated separately. Estimation must be done jointly in the correlated case.

We adapt the framework of McGilchrist (1994) and McGilchrist and Yau (1995) to develop BLUP(REMQL) estimation of the negative binomial hurdle model with random effects and correlated components. The joint BLUP-type loglikelihood of $Y_{ij}$ and $\boldsymbol{r}_i$ can be written as $\ell(\boldsymbol{y}, \boldsymbol{r}) = \ell_1(\boldsymbol{y}|\boldsymbol{r}) + \ell_2(\boldsymbol{r})$, where

$$\begin{aligned} \ell_1(\boldsymbol{y}|\boldsymbol{r}) = & \sum_{i=1}^{m}\sum_{j=1}^{n_i}[\mathrm{I}(y_{ij}=0)\log(p_{ij}) + (1-\mathrm{I}(y_{ij}=0))\log(1-p_{ij}) \\ & + (1-\mathrm{I}(y_{ij}=0))\{\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log(t_{ij}) \\ & + y_{ij}\log(1-t_{ij}) - \log(1-t_{ij}^k)\}], \\ \ell_2(\boldsymbol{r}) = & \text{constant} - \frac{1}{2}\sum_{i=1}^{m}\left[\log(|\boldsymbol{D}(\boldsymbol{\phi})|) + \boldsymbol{r}_i^T\boldsymbol{D}(\boldsymbol{\phi})^{-1}\boldsymbol{r}_i\right], \end{aligned} \tag{5}$$

and $\mathrm{I}(\cdot)$ represents a binary indicator function, $\boldsymbol{y}$ denotes a vector of $y_{ij}$, and $\boldsymbol{r} = (\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, \ldots, \boldsymbol{r}_m^T)$. Here, $\ell_1(\boldsymbol{y}|\boldsymbol{r})$ is the loglikelihood function when the random effects are conditionally fixed and $\ell_2(\boldsymbol{r})$ indicates the penalty function for the conditional loglikelihood. First, coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the linear predictors are estimated for fixed variance components and fixed scale parameter by maximizing (5). Then, the variance component parameters $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$ can be estimated using REMQL estimating equations. The scale parameter $k$, which is assumed to be given in estimation of the regression coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, also is obtained and updated by maximizing a profile loglikelihood with the current estimates. Estimation can be done iteratively via the Newton–Raphson (N–R) algorithm. Suppose $\boldsymbol{\xi} = \boldsymbol{W\alpha} + \boldsymbol{Ru}$ and $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Rv}$ where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{r}^T)^T$ is the vector of unknown parameters of interest, and $\boldsymbol{R}$ is a design matrix for the random components. In the initial step, coefficients in the linear predictor $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{r})$ are estimated given initial values $\boldsymbol{\theta}_0$ by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \boldsymbol{V}^{-1}\frac{\partial\ell}{\partial\boldsymbol{\theta}}, \qquad \boldsymbol{V} = -\frac{\partial^2\ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}, \tag{6}$$

where $\boldsymbol{V}$ denotes the negative second derivatives of the BLUP-type loglikelihood ($\ell$) with respect to $\boldsymbol{\theta}$. Details of these derivations are given in Appendix 1. The inverse of the matrix of negative second derivatives of the BLUP-type loglikelihood $\boldsymbol{V}^{-1}$ can be written as

$$\begin{bmatrix} \boldsymbol{V}_{\boldsymbol{\alpha}}^* & & \\ & \boldsymbol{V}_{\boldsymbol{\beta}}^* & \\ & & \boldsymbol{V}_{\boldsymbol{r}}^* \end{bmatrix}.$$

Asymptotic variances of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are obtained from the corresponding components $\boldsymbol{V}_{\boldsymbol{\alpha}}^*$ and $\boldsymbol{V}_{\boldsymbol{\beta}}^*$ of $\boldsymbol{V}^{-1}$.

## 3 Variance component estimation

When the N–R algorithm was used to estimate linear predictors in Sect. 2, the variance components were assumed to be known. Actually, they need to be estimated and updated in each iteration of the N–R algorithm. The approximate REMQL estimators ($\hat{\boldsymbol{\phi}}_{REMQL}$) of variance components can be obtained by solving the REMQL estimating equation (McGilchrist and Yau 1995) as follows:

$$tr\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\phi}}\right)+tr\left(\boldsymbol{V}_{\boldsymbol{r}}^{*}\frac{\partial \boldsymbol{A}^{-1}}{\partial \boldsymbol{\phi}}\right)+\boldsymbol{r}^{T}\boldsymbol{r}\frac{\partial \boldsymbol{A}^{-1}}{\partial \boldsymbol{\phi}}=0. \tag{7}$$

Note that

$$\frac{\partial \boldsymbol{D}}{\partial \sigma_u}=\begin{pmatrix}2\sigma_u & \rho\sigma_v\\ \rho\sigma_v & 0\end{pmatrix},\qquad \frac{\partial \boldsymbol{D}}{\partial \sigma_v}=\begin{pmatrix}0 & \rho\sigma_u\\ \rho\sigma_u & 2\sigma_v\end{pmatrix},\qquad \frac{\partial \boldsymbol{D}}{\partial \rho}=\begin{pmatrix}0 & \sigma_u\sigma_v\\ \sigma_u\sigma_v & 0,\end{pmatrix} \tag{8}$$

and

$$\begin{aligned}
\frac{\partial \boldsymbol{D}^{-1}}{\partial \sigma_u}&=\frac{1}{\sigma_u^3\sigma_v^2(1-\rho^2)}\begin{pmatrix}-2\sigma_v^2 & \rho\sigma_u\sigma_v\\ \rho\sigma_u\sigma_v & 0\end{pmatrix},\\
\frac{\partial \boldsymbol{D}^{-1}}{\partial \sigma_v}&=\frac{1}{\sigma_u^2\sigma_v^3(1-\rho^2)}\begin{pmatrix}0 & \rho\sigma_u\sigma_v\\ \rho\sigma_u\sigma_v & -2\sigma_u^2\end{pmatrix},\\
\frac{\partial \boldsymbol{D}^{-1}}{\partial \rho}&=\frac{1}{\sigma_u^2\sigma_v^2(1-\rho^2)^2}\begin{pmatrix}2\rho\sigma_v^2 & -(1+\rho^2)\sigma_u\sigma_v\\ -(1+\rho^2)\sigma_u\sigma_v & 2\rho\sigma_u^2\end{pmatrix}.
\end{aligned} \tag{9}$$

After substituting with (8) and (9) in (7), the exact equations for the variance components ($\sigma_u$, $\sigma_v$, $\rho$) are:

$$\begin{aligned}
&\sum_{i=1}^{m}[2\sigma_u^2\sigma_v^2(1-\rho^2)-2\sigma_v^2v_{ii,11}+\sigma_u\sigma_v\rho(v_{ii,12}+v_{ii,21}+2u_iv_i)-2\sigma_u^2u_i^2]=0,\\
&\sum_{i=1}^{m}[-2\sigma_u^2\sigma_v^2(1-\rho^2)-2\sigma_u^2(v_{ii,22}+v_i^2)+\sigma_u\sigma_v\rho(v_{ii,12}+v_{ii,21}+2u_iv_i)]=0,\\
&\sum_{i=1}^{m}[-2\sigma_u^2\sigma_v^2\rho(1-\rho^2)+2\rho\sigma_v^2v_{ii,11}+2\sigma_u^2(v_{ii,22}+u_i^2+v_i^2)-\sigma_u\sigma_v(1+\rho^2)(v_{ii,12}\\
&\quad+v_{ii,21}+2u_iv_i)]=0,
\end{aligned} \tag{10}$$

where $v_{ii}$ denotes the $2\times 2$ block matrix portion of $\boldsymbol{V}_{\boldsymbol{r}}^{*}$ corresponding to $\boldsymbol{r}_i$ and $v_{ii}=\begin{pmatrix}v_{ii,11} & v_{ii,12}\\ v_{ii,21} & v_{ii,22}\end{pmatrix}$. The variance components can be estimated using the N–R algorithm.

## 4 Scale parameter estimation

The estimation via the N–R algorithm in Sect. 2 assumed that the scale parameter $k$ was known. In practice, $k$ is updated and estimated in each iteration in accordance with the updated estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{r}, \sigma_u, \sigma_v$ and $\rho$ by maximizing the profile loglikelihood function:

$$\ell_k = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big[\mathrm{I}(y_{ij}=0)\log(p_{ij}) + (1-\mathrm{I}(y_{ij}=0))\log(1-p_{ij}) + (1-\mathrm{I}(y_{ij}=0))\Big\{\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log(t_{ij}) + y_{ij}\log(1-t_{ij}) - \log(1-t_{ij}^{k})\Big\}\Big]. \quad (11)$$

The asymptotic variance of $\hat{k}$ can be obtained by $Var(\hat{k}) = (-\frac{\partial^2 \ell_k}{\partial k^2})^{-1}$; details are given in Appendix 2.

## 5 Application to the EDCAP study

We illustrate the models by analyzing bed days in the 32-site EDCAP study (Yealy et al. 2004, 2005). EDCAP is a cluster-randomized trial in CT and PA to assess the effectiveness and safety of 3 guideline implementation interventions of low (8 sites), moderate (12 sites), and high (12 sites) intensity to increase the proportion of low risk patients who were treated as outpatients. Risk was ascertained using a validated measure of pneumonia severity, the Pneumonia Severity Index (PSI) (Yealy et al. 2005), with low risk defined as PSI $\leq$ 3 without hypoxemia.

In the EDCAP study, we examine whether the distribution of bed days varies by intervention arm and PSI risk class among 1,877 low risk patients with clinical and radiographic evidence of pneumonia. Among eligible low risk patients, 57 % ($n$ = 1,061) were treated as outpatients and 43 % ($n$ = 816) were treated as inpatients; 37 % of patients had PSI = 1, 37 % of patients had PSI = 2, and 26 % of patients had PSI = 3 (Table 1). Relatively fewer low risk patients at the low intensity intervention sites were treated as outpatients (38 % vs. 62 % at the moderate intensity and 63 % at the high intensity intervention sites).

Figure 1 shows the empirical distribution of bed days for patients in each PSI risk class by intervention arm. The spikes at zero bed days represent outpatients; in each intervention arm, the prevalence of outpatient care decreases with increasing risk class. The distributions

**Table 1** Probability of outpatient, mean and median of inpatients and overall bed days by PSI risk class and by intervention arm for 1,877 eligible low risk patients

| | $n$ | Pr (outpatient) | Bed days | |
|---|---|---|---|---|
| | | | Inpatients mean (median) | Overall mean (median) |
| *PSI risk class* | | | | |
| 1 | 697 | 0.82 | 4.0 (3.0) | 0.7 (0.0) |
| 2 | 691 | 0.51 | 4.6 (4.0) | 2.2 (0.0) |
| 3 | 486 | 0.28 | 5.8 (4.0) | 4.1 (3.0) |
| *Intervention* | | | | |
| Low | 438 | 0.38 | 5.0 (4.0) | 3.1 (2.0) |
| Mod | 748 | 0.62 | 4.9 (4.0) | 1.9 (0.0) |
| High | 691 | 0.63 | 5.1 (4.0) | 1.9 (0.0) |
| Overall | | 0.57 | 5.0 (4.0) | 2.2 (0.0) |

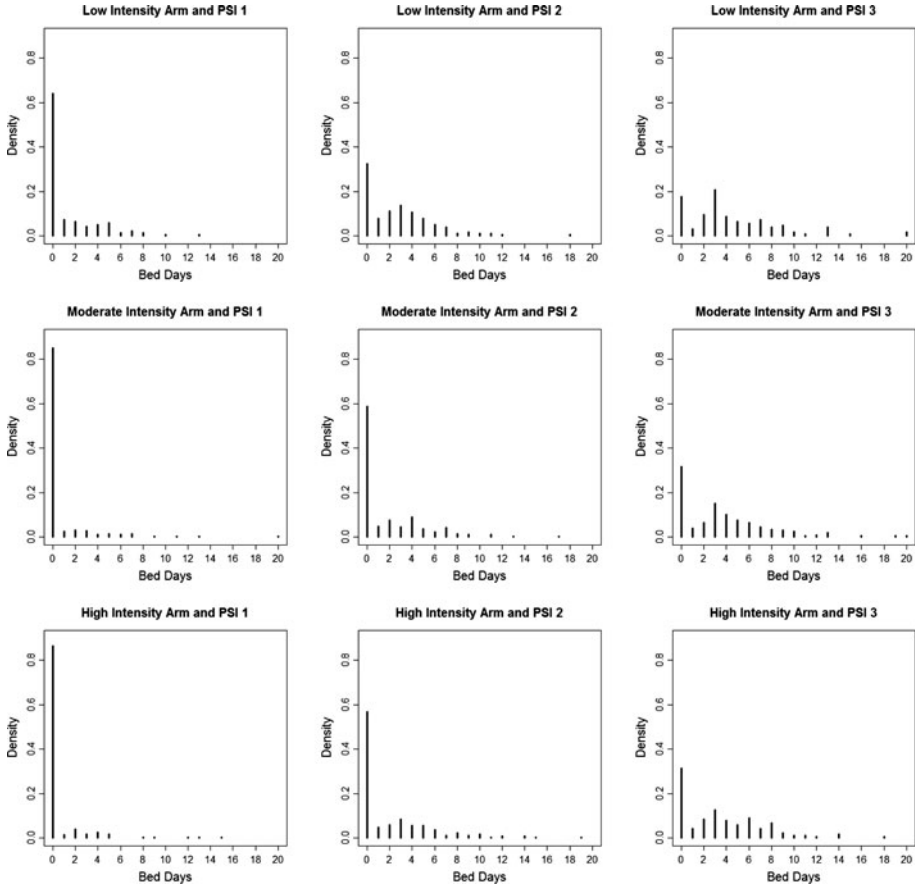


**Fig. 1** Bed days by intervention arm and PSI risk class

at the moderate and high intensity intervention sites are right skewed; both had more outpatients and fewer inpatient bed days than did the low intensity intervention sites.

In the modeling of bed days, the patient-level PSI risk class and the site-level intervention arm were included as dummy variables; with PSI2 = 1 if PSI = 2; 0 else, PSI3 = 1 if PSI = 3; 0 else, Mod = 1 if moderate intensity intervention; 0 else, and High = 1 if high intensity intervention; 0 else. The Poisson/negative binomial hurdle model with random effects is:

$$\begin{aligned} \text{logit}(p_{ij}) &= \alpha_0 + \alpha_1 \cdot \text{PSI2} + \alpha_2 \cdot \text{PSI3} + \alpha_3 \cdot \text{Mod} + \alpha_4 \cdot \text{High} + u_i, \\ \log(\mu_{ij}) &= \beta_0 + \beta_1 \cdot \text{PSI2} + \beta_2 \cdot \text{PSI3} + \beta_3 \cdot \text{Mod} + \beta_4 \cdot \text{High} + v_i, \end{aligned} \tag{12}$$

where $(u_i, v_i)^T$ is assumed to be distributed as $N(\mathbf{0}, \boldsymbol{D})$ when the covariance matrix $\boldsymbol{D}$ is defined in (4) with $i = 1, \ldots, 32$.

Table 2 summarizes the ML and BLUP(REMQL) estimates for the random effects Poisson hurdle model. The fixed effects estimates and standard errors (SE) are almost identical between the two estimation methods for both components of the model. The estimated bivariate correlation between the two components is low (−0.01 for ML; −0.04 for BLUP(REMQL)).

**Table 2** Correlated random effects Poisson hurdle model estimates based on (a) ML and (b) BLUP(REMQL) estimation

| Parameter | (a) ML | | | (b) BLUP(REMQL) | | |
|---|---|---|---|---|---|---|
| | Estimate | SE | *P* value | Estimate | SE | *P* value |
| *Logistic part: Pr (outpatient)* | | | | | | |
| Cons | 0.80 | 0.24 | <.01 | 0.79 | 0.25 | <.01 |
| PSI2 | −1.49 | 0.13 | <.001 | −1.47 | 0.13 | <.001 |
| PSI3 | −2.54 | 0.15 | <.001 | −2.51 | 0.15 | <.001 |
| Mod | 0.98 | 0.29 | <.01 | 0.97 | 0.30 | <.01 |
| High | 0.93 | 0.29 | <.01 | 0.91 | 0.30 | <.01 |
| $\sigma_u$ | 0.54 | | | 0.57 | | |
| *Poisson part: inpatient bed days* | | | | | | |
| Cons | 1.41 | 0.08 | <.001 | 1.41 | 0.09 | <.001 |
| PSI2 | 0.09 | 0.05 | .09 | 0.09 | 0.05 | .09 |
| PSI3 | 0.34 | 0.05 | <.001 | 0.34 | 0.05 | <.001 |
| Mod | −0.06 | 0.09 | .49 | −0.06 | 0.10 | .51 |
| High | 0.02 | 0.09 | .87 | 0.02 | 0.10 | .87 |
| $\sigma_v$ | 0.18 | | | 0.19 | | |
| $\rho$ | −0.01 | | | −0.04 | | |

ML and BLUP(REMQL) estimates for the negative binomial hurdle model are shown in Table 3. Except possibly for $k$, these estimates are quite similar to each other for both components. The log odds ratio (log OR) of outpatient care decreases significantly with increasing risk class, with log ORs of −1.47 and −2.51 for PSI2 and PSI3, respectively, and increases significantly for the moderate and high intensity intervention sites, with log ORs of 0.97 and 0.91, respectively. The scale parameter ($k = 2.71$) indicates significant overdispersion relative to the Poisson distribution. The estimated bivariate correlation is modest (−0.10). The *P* values for the PSI parameters in the count component of the model are less significant in the negative binomial hurdle model than in the Poisson hurdle model, due to the correction for overdispersion.

Figure 2 illustrates the better fit of the negative binomial hurdle model than the Poisson hurdle model to the EDCAP data. To identify unusual sites based on the random effects negative binomial hurdle model, the predicted site-level random effects are plotted for the logistic and the negative binomial parts in Fig. 3. Site 25 appears to be unusual in that it is a moderate intensity intervention site with a low predicted probability of treating low risk patients as outpatients. In addition, Fig. 3 indicates that there is more site-level variation in the logistic part (i.e., hospitalization decision) than in the negative binomial part (i.e., LOS).

## 6 Simulation study

We conducted simulation studies to compare the performance of the proposed BLUP(REMQL) to ML in the correlated random effects negative binomial hurdle model with a plausible range of bivariate correlations. We imitated the unbalanced cluster-randomized structure of the EDCAP data and included patient-level (PSI1, PSI2, PSI3) and

**Table 3** Correlated random effects negative binomial hurdle model estimates based on (a) ML and (b) BLUP(REMQL) estimation

| Parameter | (a) ML | | | (b) BLUP(REMQL) | | |
|---|---|---|---|---|---|---|
| | Estimate | SE | *P* value | Estimate | SE | *P* value |
| *Logistic part: Pr (outpatient)* | | | | | | |
| Cons | 0.80 | 0.24 | <.01 | 0.79 | 0.25 | <.01 |
| PSI2 | −1.49 | 0.13 | <.001 | −1.47 | 0.13 | <.001 |
| PSI3 | −2.54 | 0.15 | <.001 | −2.51 | 0.15 | <.001 |
| Mod | 0.98 | 0.29 | <.01 | 0.97 | 0.30 | <.01 |
| High | 0.93 | 0.29 | <.01 | 0.91 | 0.30 | <.01 |
| $\sigma_u$ | 0.54 | | | 0.58 | | |
| *Negative binomial part: inpatient bed days* | | | | | | |
| Cons | 1.29 | 0.10 | <.001 | 1.30 | 0.10 | <.001 |
| PSI2 | 0.12 | 0.09 | .18 | 0.12 | 0.09 | .18 |
| PSI3 | 0.39 | 0.09 | <.001 | 0.38 | 0.09 | <.001 |
| Mod | −0.04 | 0.10 | .68 | −0.04 | 0.10 | .68 |
| High | 0.03 | 0.10 | .75 | 0.03 | 0.11 | .77 |
| k | 2.56 | 0.26 | <.001 | 2.71 | 0.26 | <.001 |
| $\sigma_v$ | 0.15 | | | 0.17 | | |
| $\rho$ | −0.12 | | | −0.10 | | |

site-level covariates (Low, Mod, High). PSI1 and Low intensity intervention served as the reference levels. For each of the $m = 32$ sites, $n_i$ patients were randomly generated from a Poisson distribution. Based on the estimates in Table 3, $\boldsymbol{\alpha}$ was specified as (0.8, −1.5, −2.5, 1.0, 0.9), $\boldsymbol{\beta}$ was specified as (1.3, 0.1, 0.4, −0.1, 0.1), $k = 2.6$, $\sigma_u = 0.6$, $\sigma_v = 0.2$, and $\rho$ took one of the following values (−0.1, −0.3, −0.5, −0.7). We used 1,000 replications for each of the four simulated settings.

In Table 4, we summarized one randomly chosen simulated dataset to show how well our simulated data replicated the EDCAP data structure summarized in Table 1. Figure 4 confirms that the cumulative distributions of observed and simulated bed days are almost identical.

Results of the simulation studies (Table 5) verify the performance of the proposed BLUP(REMQL) estimation in the negative binomial hurdle model. We report the average bias, the bias relative to the true parameter (Percent), SE, mean square error (MSE), and coverage probability (CP) of the 95 % confidence interval over 1,000 replications for each value of $\rho$ considered. The biases in the estimated fixed effects generally were small ($\leq$3.0 %) for both ML and BLUP(REMQL) for both the logistic and negative binomial components of the model for all values of $\rho$ considered. The exceptions were somewhat larger biases (4.0–8.5 %) in some of the ML and/or BLUP(REMQL) estimated site-level parameters (i.e., Mod or High) in the negative binomial component when $\rho = -0.5$ or −0.7. Biases in the estimated fixed effects generally (but not always) were smaller for ML than for BLUP(REMQL). The SEs and the corresponding MSEs of the estimated fixed effects were similar for ML and BLUP(REMQL), and the CPs were generally at least as good, if not better, for BLUP(REMQL) relative to ML.

The BLUP(REMQL) estimates of the random effects ($\sigma_u$ and $\sigma_v$) and $\rho$ have much smaller biases than the corresponding ML estimates; for example, the percent bias is 1.5 %

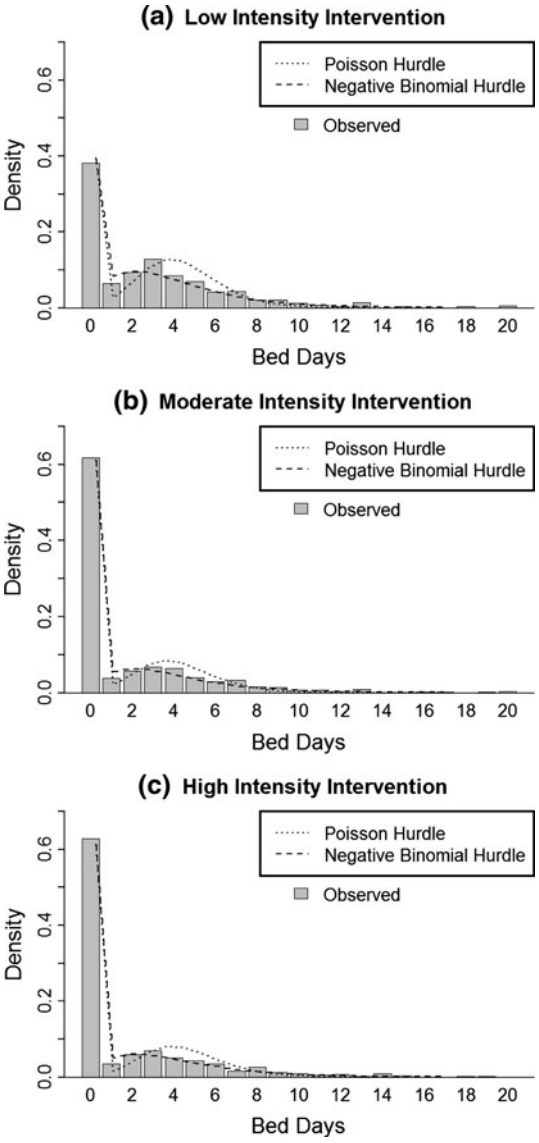


**Fig. 2** Observed vs predicted distribution of bed days by intervention arm. Distributions are predicted based on the Poisson hurdle model (*dots*) and negtive binomial hurdle model (*dashed line*)

vs. 6.5 % for $\sigma_u$, 1.5 % vs. 11.0 % for $\sigma_v$, and 3.0 % vs. 63.0 % for $\rho$, Table 5a. However, the BLUP(REMQL) estimate of the scale parameter ($k$) in the negative binomial component has larger bias than the corresponding ML estimate (e.g., 8.8 % vs. 1.5 % in Table 5a), and poorer CP (e.g., 0.90 vs. 0.96). Similar patterns were observed for the other values of $\rho$.

In summary, these simulation results demonstrate that BLUP(REMQL) estimation in the negative binomial hurdle model with correlated random effects performs well relative to ML for the fixed effects and variance components considered, but not for the scale parameter. All replications converged for BLUP(REMQL), while some did not converge for ML (i.e., 10/1000 replications at $\rho = -0.3$; 26/1000 replications at $\rho = -0.5$; 91/1000

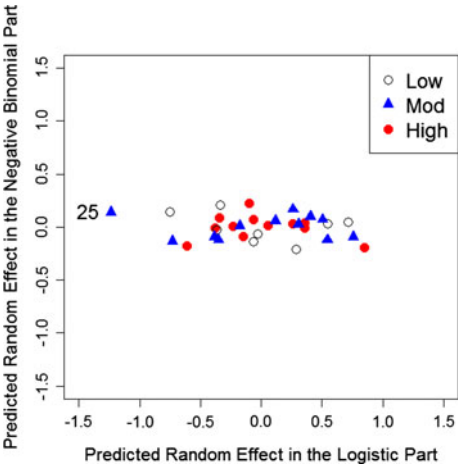


**Fig. 3** Site specific predicted random effects for the logistic and negative binomial parts of the negative binomial hurdle model for the low (*open circle*), moderate (*closed triangle*), and high (*closed circle*) intensity intervention sites

**Table 4** Probability of outpatient, mean and median of inpatients bed days, and mean and median of overall bed days by PSI risk class and by intervention arm for one simulated dataset ($N = 1{,}823$)

| | *n* | Pr (outpatient) | Bed days | |
|---|---|---|---|---|
| | | | Inpatients mean (median) | Overall mean (median) |
| *PSI risk class* | | | | |
| 1 | 654 | 0.79 | 3.9 (4.0) | 0.8 (0.0) |
| 2 | 644 | 0.53 | 4.8 (4.0) | 2.3 (0.0) |
| 3 | 525 | 0.33 | 6.0 (5.0) | 4.0 (3.0) |
| *Intervention* | | | | |
| Low | 532 | 0.42 | 5.0 (4.0) | 2.9 (2.0) |
| Mod | 650 | 0.59 | 4.6 (4.0) | 1.9 (0.0) |
| High | 641 | 0.66 | 6.0 (5.0) | 2.1 (0.0) |
| Overall | | 0.57 | 5.2 (4.0) | 2.2 (0.0) |

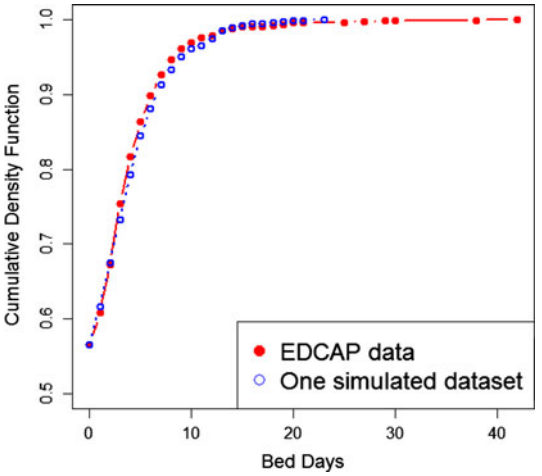


**Fig. 4** Cumulative density function of bed days by EDCAP data (*closed circle*) and one simulated dataset (*open circle*) with $\rho = -0.1$

**Table 5** Simulation results using correlated random effects negative binomial hurdle model based on ML and BLUP(REMQL) estimation with 1,000 replications and a plausible range of bivariate correlation ($\rho = -0.1, -0.3, -0.5, -0.7$)

| Parameter | True | Bias (percent[a]) | | SE | | MSE | | CP | |
|---|---|---|---|---|---|---|---|---|---|
| | | ML | REMQL | ML | REMQL | ML | REMQL | ML | REMQL |
| (a) $\rho = -0.1$ | | | | | | | | | |
| Logistic part: Pr (outpatient) | | | | | | | | | |
| $\alpha_0$ : Cons | 0.8 | −0.007 (0.9) | −0.016 (2.0) | 0.247 | 0.256 | 0.136 | 0.139 | 0.93 | 0.94 |
| $\alpha_1$ : PSI2 | −1.5 | −0.007 (0.5) | 0.010 (0.7) | 0.130 | 0.129 | 0.035 | 0.034 | 0.94 | 0.94 |
| $\alpha_2$ : PSI3 | −2.5 | −0.008 (0.3) | 0.021 (0.8) | 0.150 | 0.148 | 0.046 | 0.045 | 0.96 | 0.95 |
| $\alpha_3$ : Mod | 1.0 | 0.008 (0.8) | −0.004 (0.4) | 0.305 | 0.317 | 0.204 | 0.209 | 0.93 | 0.94 |
| $\alpha_4$ : High | 0.9 | 0.025 (2.8) | 0.014 (1.6) | 0.306 | 0.318 | 0.200 | 0.204 | 0.93 | 0.94 |
| $\sigma_u$ | 0.6 | −0.039 (6.5) | −0.009 (1.5) | | | | | | |
| Negative binomial part: inpatient bed days | | | | | | | | | |
| $\beta_0$ : Cons | 1.3 | −0.002 (0.2) | 0.003 (0.2) | 0.107 | 0.110 | 0.025 | 0.025 | 0.94 | 0.94 |
| $\beta_1$ : PSI2 | 0.1 | 0.002 (2.0) | 0.001 (1.0) | 0.085 | 0.083 | 0.015 | 0.014 | 0.95 | 0.95 |
| $\beta_2$ : PSI3 | 0.4 | 0.004 (1.0) | 0.000 (–) | 0.084 | 0.082 | 0.014 | 0.014 | 0.95 | 0.95 |
| $\beta_3$ : Mod | −0.1 | −0.001 (1.0) | −0.001 (1.0) | 0.114 | 0.119 | 0.029 | 0.030 | 0.92 | 0.93 |
| $\beta_4$ : High | 0.1 | 0.001 (1.0) | 0.000 (–) | 0.113 | 0.118 | 0.028 | 0.029 | 0.93 | 0.94 |
| k | 2.6 | 0.038 (1.5) | 0.229 (8.8) | 0.284 | 0.286 | 0.169 | 0.238 | 0.96 | 0.90 |
| $\sigma_v$ | 0.2 | −0.022 (11.0) | −0.003 (1.5) | | | | | | |
| $\rho$ | −0.1 | 0.063 (63.0) | −0.003 (3.0) | | | | | | |
| (b) $\rho = -0.3$ | | | | | | | | | |
| Logistic part: Pr (outpatient) | | | | | | | | | |
| $\alpha_0$ : Cons | 0.8 | −0.001 (0.1) | −0.010 (1.3) | 0.244 | 0.253 | 0.130 | 0.133 | 0.95 | 0.96 |
| $\alpha_1$ : PSI2 | −1.5 | −0.011 (0.7) | 0.006 (0.4) | 0.130 | 0.129 | 0.034 | 0.033 | 0.95 | 0.95 |
| $\alpha_2$ : PSI3 | −2.5 | −0.002 (0.1) | 0.026 (1.0) | 0.150 | 0.147 | 0.045 | 0.044 | 0.96 | 0.95 |
| $\alpha_3$ : Mod | 1.0 | 0.003 (0.3) | −0.008 (0.8) | 0.302 | 0.314 | 0.197 | 0.202 | 0.93 | 0.94 |

**Table 5** continued

| Parameter | True | Bias (percent[a]) | | SE | | MSE | | CP | |
|---|---|---|---|---|---|---|---|---|---|
| | | ML | REMQL | ML | REMQL | ML | REMQL | ML | REMQL |
| $\alpha_4$ : High | 0.9 | 0.008 (0.9) | −0.002 (0.2) | 0.301 | 0.313 | 0.192 | 0.197 | 0.93 | 0.94 |
| $\sigma_u$ | 0.6 | −0.045 (7.5) | −0.016 (2.7) | | | | | | |
| Negative binomial part: inpatient bed days | | | | | | | | | |
| $\beta_0$ : Cons | 1.3 | −0.002 (0.2) | 0.004 (0.3) | 0.107 | 0.109 | 0.025 | 0.025 | 0.93 | 0.93 |
| $\beta_1$ : PSI2 | 0.1 | 0.002 (2.0) | 0.001 (1.0) | 0.085 | 0.083 | 0.015 | 0.014 | 0.95 | 0.95 |
| $\beta_2$ : PSI3 | 0.4 | −0.002 (0.5) | −0.005 (1.3) | 0.084 | 0.082 | 0.015 | 0.014 | 0.94 | 0.93 |
| $\beta_3$ : Mod | −0.1 | 0.001 (1.0) | 0.001 (1.0) | 0.114 | 0.120 | 0.027 | 0.028 | 0.93 | 0.95 |
| $\beta_4$ : High | 0.1 | −0.001 (1.0) | −0.002 (2.0) | 0.112 | 0.118 | 0.027 | 0.028 | 0.93 | 0.95 |
| k | 2.6 | 0.036 (1.4) | 0.219 (8.4) | 0.282 | 0.284 | 0.159 | 0.222 | 0.96 | 0.91 |
| $\sigma_v$ | 0.2 | −0.023 (11.5) | −0.003 (1.5) | | | | | | |
| $\rho$ | −0.3 | 0.199 (66.3) | 0.016 (5.3) | | | | | | |
| (c) $\rho = -0.5$ | | | | | | | | | |
| Logistic part: Pr (outpatient) | | | | | | | | | |
| $\alpha_0$ : Cons | 0.8 | −0.016 (2.0) | −0.024 (3.0) | 0.242 | 0.253 | 0.127 | 0.132 | 0.94 | 0.95 |
| $\alpha_1$ : PSI2 | −1.5 | 0.002 (0.1) | 0.018 (1.2) | 0.129 | 0.128 | 0.035 | 0.035 | 0.93 | 0.92 |
| $\alpha_2$ : PSI3 | −2.5 | −0.002 (0.1) | 0.024 (1.0) | 0.148 | 0.146 | 0.044 | 0.044 | 0.96 | 0.95 |
| $\alpha_3$ : Mod | 1.0 | 0.003 (0.3) | −0.007 (0.7) | 0.300 | 0.314 | 0.192 | 0.199 | 0.93 | 0.94 |
| $\alpha_4$ : High | 0.9 | 0.010 (1.1) | 0.001 (0.1) | 0.298 | 0.312 | 0.193 | 0.200 | 0.93 | 0.95 |
| $\sigma_u$ | 0.6 | −0.049 (8.2) | −0.018 (3.0) | | | | | | |
| Negative binomial part: inpatient bed days | | | | | | | | | |
| $\beta_0$ : Cons | 1.3 | 0.004 (0.3) | 0.008 (0.6) | 0.107 | 0.11 | 0.024 | 0.025 | 0.94 | 0.94 |
| $\beta_1$ : PSI2 | 0.1 | 0.003 (3.0) | 0.003 (3.0) | 0.083 | 0.082 | 0.014 | 0.014 | 0.95 | 0.95 |
| $\beta_2$ : PSI3 | 0.4 | 0.001 (0.3) | −0.002 (0.5) | 0.083 | 0.081 | 0.014 | 0.013 | 0.96 | 0.95 |
| $\beta_3$ : Mod | −0.1 | −0.006 (6.0) | −0.005 (5.0) | 0.114 | 0.121 | 0.028 | 0.029 | 0.94 | 0.95 |

**Table 5** continued

| Parameter | True | Bias (percent[a]) | | SE | | MSE | | CP | |
|---|---|---|---|---|---|---|---|---|---|
| | | ML | REMQL | ML | REMQL | ML | REMQL | ML | REMQL |
| $\beta_4$ : High | 0.1 | −0.007 (7.0) | −0.008 (8.0) | 0.112 | 0.119 | 0.026 | 0.028 | 0.94 | 0.96 |
| k | 2.6 | 0.051 (2.0) | 0.221 (8.5) | 0.280 | 0.280 | 0.169 | 0.222 | 0.94 | 0.90 |
| $\sigma_v$ | 0.2 | −0.020 (10.0) | 0.000 (–) | | | | | | |
| $\rho$ | −0.5 | 0.326 (65.2) | 0.025 (5.0) | | | | | | |
| (d) $\rho = -0.7$ | | | | | | | | | |
| Logistic part: Pr (outpatient) | | | | | | | | | |
| $\alpha_0$ : Cons | 0.8 | 0.003 (0.4) | −0.004 (0.5) | 0.243 | 0.252 | 0.130 | 0.134 | 0.93 | 0.95 |
| $\alpha_1$ : PSI2 | −1.5 | 0.003 (0.2) | 0.016 (1.1) | 0.130 | 0.129 | 0.034 | 0.034 | 0.95 | 0.94 |
| $\alpha_2$ : PSI3 | −2.5 | 0.003 (0.1) | 0.027 (1.1) | 0.149 | 0.147 | 0.045 | 0.044 | 0.95 | 0.94 |
| $\alpha_3$ : Mod | 1.0 | 0.002 (0.2) | −0.007 (0.7) | 0.302 | 0.314 | 0.200 | 0.206 | 0.93 | 0.94 |
| $\alpha_4$ : High | 0.9 | 0.002 (0.2) | −0.006 (0.7) | 0.300 | 0.312 | 0.195 | 0.201 | 0.94 | 0.95 |
| $\sigma_u$ | 0.6 | −0.047 (7.8) | −0.016 (2.7) | | | | | | |
| Negative binomial part: inpatient bed days | | | | | | | | | |
| $\beta_0$ : Cons | 1.3 | −0.001 (0.1) | 0.002 (0.2) | 0.107 | 0.110 | 0.023 | 0.024 | 0.94 | 0.95 |
| $\beta_1$ : PSI2 | 0.1 | −0.003 (3.0) | −0.003 (3.0) | 0.084 | 0.082 | 0.014 | 0.014 | 0.94 | 0.94 |
| $\beta_2$ : PSI3 | 0.4 | −0.001 (0.3) | −0.004 (1.0) | 0.083 | 0.081 | 0.014 | 0.013 | 0.94 | 0.94 |
| $\beta_3$ : Mod | −0.1 | 0.000 (–) | 0.000 (–) | 0.115 | 0.122 | 0.028 | 0.029 | 0.94 | 0.95 |
| $\beta_4$ : High | 0.1 | 0.004 (4.0) | 0.003 (3.0) | 0.113 | 0.119 | 0.027 | 0.028 | 0.93 | 0.94 |
| k | 2.6 | 0.068 (2.6) | 0.204 (7.8) | 0.284 | 0.280 | 0.179 | 0.212 | 0.95 | 0.91 |
| $\sigma_v$ | 0.2 | −0.018 (9.0) | 0.002 (1.0) | | | | | | |
| $\rho$ | −0.7 | 0.455 (65.0) | 0.035 (5.0) | | | | | | |

For ML, 10 replications did not converge when $\rho = -0.3$, 26 replications did not converge when $\rho = -0.5$, and 91 replications did not converge when $\rho = -0.7$

[a] Relative bias to the true parameter, which is calculated by 100 × abs (bias/true)

replications at $\rho = -0.7$). BLUP(REMQL) ran in about 3/7 the time as ML for these simulated data. We used the SAS procedure NLMIXED to fit the model with ML, and R to obtain the BLUP(REMQL) estimates.

## 7 Discussion

We have proposed a BLUP(REMQL) approach to estimate a negative binomial hurdle model with correlated random effects. We also illustrated the application of this model to a potentially useful efficiency metric in health services studies, bed days. This model appropriately accounts for excess zeros and overdispersion relative to the Poisson distribution, and allows for site-level correlation between the binary and count components of the model. This model gives an overall assessment of the effect on an intervention on two aspects of care, e.g., admission and LOS in the EDCAP study. While the interventions in EDCAP were designed to influence the admission decision and recommended processes of care in the ED, there was no intervention to influence inpatient LOS. Our results confirmed that the intervention was significantly associated with reduced hospitalization but not with LOS. The small negative bivariate correlation indicates some tendency for shorter inpatient LOS at sites with relatively low admission rates for low risk patients. Although not well-illustrated by the EDCAP study, our proposed approach could give more efficient estimates of random effects in a similarly-designed study with interventions that affected both components of the model.

In this paper, we have accounted for correlated random effects at the site level that could be associated with both the hospitalization decision and inpatient LOS. The EDCAP intervention was limited to low-risk patients, so that the predominant factors driving the hospitalization decision in these patients are site-level (e.g., practice patterns, guideline compliance, quality and/or efficiency of care) rather than patient level characteristics. We can extend this model to patient-level correlated random effects by defining a multivariate normal distribution of random effects.

For the scenarios considered, our simulation study indicated that the BLUP(REMQL) approach provides less biased estimates of variance components than ML, and estimates similar to ML for the fixed effects. However, BLUP(REMQL) estimation yields somewhat larger bias in the estimated scale parameter relative to ML. This issue requires further investigation. BLUP(REMQL) estimation remains attractive because it ran faster than ML for these data and had better convergence properties. For either BLUP(REMQL) or ML, the choice of initial values affects convergence rates and computation time. To guarantee the convergence and reduce computation time, we initialized parameter estimates by fitting fixed effect hurdle models. Our simulations mimicked the structure of the EDCAP data and that additional simulations would need to be done to assess the sensitivity of (BLUP)REMQL to initial values.

We can implement ML simply using SAS. In the absence of generally available software, adapted R code is required to obtain the BLUP(REMQL) estimates considered here. A flexible alternative is Bayesian estimation, which can be implemented using WinBUGS (Neelon et al. 2010).

In summary, the proposed BLUP(REMQL) estimation in these hurdle models appears to be promising. The computational advantages may facilitate application of this approach to more complex versions of these models, such as a 3-level model defined by patient, medical provider, and site.

**Acknowledgments** The authors thank K.K.W. Yau and K. Wang for providing the R code for BLUP(REMQL) estimation that was adapted in this work. The R code for the BLUP(REMQL) and the SAS code for the ML estimation are available from the journal website.

## Appendix 1: First and second derivatives of the joint BLUP-type loglikelihood

From the joint BLUP-type loglikelihood, we can obtain:

$$\begin{aligned}
&\frac{\partial \ell}{\partial \boldsymbol{\alpha}} = \boldsymbol{W}^T \frac{\partial \ell_1}{\partial \boldsymbol{\xi}}, \qquad \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \frac{\partial \ell_1}{\partial \boldsymbol{\eta}}, \qquad \frac{\partial \ell_1}{\partial \boldsymbol{u}} = \boldsymbol{R}^T \frac{\partial \ell_1}{\partial \boldsymbol{\xi}}, \qquad \frac{\partial \ell_1}{\partial \boldsymbol{v}} = \boldsymbol{R}^T \frac{\partial \ell_1}{\partial \boldsymbol{\eta}}, \\
&\frac{\partial \ell}{\partial \boldsymbol{r}} = \begin{pmatrix} \frac{\partial \ell_1}{\partial \boldsymbol{u}} \\ \frac{\partial \ell_1}{\partial \boldsymbol{v}} \end{pmatrix} \boldsymbol{G} - \boldsymbol{A}^{-1}\boldsymbol{r},
\end{aligned} \tag{13}$$

where a $2m \times 2m$ matrix $(G)$ satisfies $\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} G = \boldsymbol{r}$ and $\boldsymbol{A} = [\boldsymbol{D}, \boldsymbol{D}, \ldots, \boldsymbol{D}]$ denotes a $2m \times 2m$ block diagonal matrix. Now,

$$\begin{aligned}
&\frac{\partial \ell_1}{\partial \xi_{ij}} = \mathrm{I}(y_{ij} = 0) - \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}}, \\
&\frac{\partial \ell_1}{\partial \eta_{ij}} = (1 - \mathrm{I}(y_{ij} = 0)) \left\{ y_{ij} \frac{k}{k + e^{\eta_{ij}}} - \frac{k(1 - \frac{k}{k+e^{\eta_{ij}}})}{1 - (\frac{k}{k+e^{\eta_{ij}}})^k} \right\}.
\end{aligned} \tag{14}$$

The second derivatives of the joint BLUP-type loglikelihood are obtained as follows:

$$\begin{aligned}
&\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}} = \boldsymbol{W}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} \boldsymbol{W}, \\
&\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \boldsymbol{X}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}} \boldsymbol{X}, \\
&\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}^T} = \boldsymbol{W}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} \boldsymbol{X}, \\
&\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha} \partial \boldsymbol{r}^T} = \left( \boldsymbol{W}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} \boldsymbol{R}, \boldsymbol{W}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} \boldsymbol{R} \right) \boldsymbol{G}, \\
&\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{r}^T} = \left( \boldsymbol{X}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} \boldsymbol{R}, \boldsymbol{X}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \boldsymbol{R} \right) \boldsymbol{G}, \\
&\frac{\partial^2 \ell}{\partial \boldsymbol{r} \partial \boldsymbol{r}^T} = \boldsymbol{G}^T \begin{pmatrix} \boldsymbol{R}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} \boldsymbol{R} & \boldsymbol{R}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} \boldsymbol{R}, \\ \boldsymbol{R}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} \boldsymbol{R} & \boldsymbol{R}^T \frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \boldsymbol{R} \end{pmatrix} \boldsymbol{G} - \boldsymbol{A}^{-1},
\end{aligned} \tag{15}$$

where

$$\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} = \mathrm{Diag}\left[ -\frac{e^{\xi}}{(1 + e^{\xi})^2} \right], \tag{16}$$

$$\frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \mathrm{Diag}\left[-(1-\mathrm{I}(\boldsymbol{y}=0))t(1-t)\left\{\boldsymbol{y} - \frac{k^2(1-t)t^{k-1} - k(1-t^k)}{(1-t^k)^2}\right\}\right],$$
$$\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} = 0.$$

## Appendix 2: First and second derivatives of the profile loglikelihood

Following Lee et al. (2003), suppose $A(k) = \sum_{y_{ij}>0} \log \frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)}$, and $f(\tau) = \#\{y_{ij} \geq \tau, \forall i,j\}$ be the number of patients whose observed count is greater than or equal to $\tau$, then the first and second derivatives of $A(k)$ are derived as:

$$\begin{aligned} A(\dot{k}) &= \sum_{\tau=1}^{max(y_{ij})-1} \frac{f(\tau)}{k+\tau}, \\ A(\ddot{k}) &= -\sum_{\tau=1}^{max(y_{ij})-1} \frac{f(\tau)}{(k+\tau)^2}. \end{aligned} \tag{17}$$

Then, the first and second derivatives of $\ell_k$ can be expressed in terms of $A(\dot{k})$ and $A(\ddot{k})$:

$$\begin{aligned} \frac{\partial \ell_k}{\partial k} &= A(\dot{k}) + \sum_{y_{ij}>0} \frac{B_{ij}}{1-t_{ij}^k} - \frac{y_{ij}t_{ij}}{k}, \\ \frac{\partial^2 \ell_k}{\partial k^2} &= A(\ddot{k}) + \sum_{y_{ij}>0} \frac{\dot{B}_{ij}(1-t_{ij}^k) + B_{ij}^2 t_{ij}^k}{(1-t_{ij}^k)^2} + \frac{y_{ij}t_{ij}^2}{k^2}, \end{aligned} \tag{18}$$

where

$$B_{ij} = \log(t_{ij}) + 1 - t_{ij} \qquad \text{and} \qquad \dot{B}_{ij} = \frac{(1-t_{ij})^2}{k}. \tag{19}$$

## References


Arulampalam, W., Booth, A.L.: Who gets over the training hurdle? A study of the training experiences of young men and women in Britain. J. Popul. Econ. **10**(2), 197–217 (1997)

Brown, K.L., Ridout, D.A., Goldman, A.P., Hoskote, A., Penny, D.J.: Risk factors for long intensive care unit stay after cardiopulmonary bypass in children. Crit. Care Med. **31**(1), 28–33 (2003)

Cunningham, R.B., Lindenmayer, D.B.: Modeling count data of rare species: some statistical issues. Ecology **86**(5), 1135–1142 (2005)

Fine, M.J., Pratt, H.M., Obrosky, D.S., Lave, J.R., McIntosh, L.J., Singer, D.E., Coley, C.M., Kapoor, W.N.: Relation between length of hospital stay and costs of care for patients with community-acquired pneumonia. Am. J. Med. **109**(5), 378–385 (2000)

Gurmu, S.: Generalized hurdle count data regression models. Econ. Lett. **58**(3), 263–268 (1998)

Hsu, D.J., Stone, R.A., Obrosky, D.S., Yealy, D.M., Meehan, T.P., Fine, J.M., Graff, L.G., Fine, M.J.: Predictors of timely antibiotic administration for patients hospitalized with community-acquired pneumonia from the cluster-randomized EDCAP trial. Am. J. Med. Sci. **339**(4), 307–313 (2010)

Lee, A.H., Wang, K., Yau, K.K.W., Somerford, P.J.: Truncated negative binomial mixed regression modelling of ischaemic stroke hospitalizations. Stat. Med. **22**(7), 1129–1139 (2003)

McGilchrist, C.A.: Estimation in generalized mixed models. J. R. Stat. Soc. B **56**(1), 61–69 (1994)

McGilchrist, C.A., Yau, K.K.W.: The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. Commun. Stat. Theory Methods **24**(12), 2963–2980 (1995)

Min, Y., Agresti, A.: Modeling nonnegative data with clumping at zero: a survey. J. Iran. Stat. Soc. **1**(1–2), 7–33 (2002)

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

Neelon, B.H., OMalley, A.J., Normand, S.L.T.: A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Stat. Model. **10**(4), 421–439 (2010)

Niederman, M.S., McCombs, J.S., Unger, A.N., Kumar, A., Popovian, R.: The cost of treating community-acquired pneumonia. Clin. Ther. **20**(4), 820–837 (1998)

Pohlmeier, W., Ulrich, V.: An econometric model of the two-part decision making process in the demand for health care. J. Hum. Resour. **30**(2), 339–361 (1995)

Ridout, M., Demetrio, C.G.B., Hinde, J.: Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, Cape Town, vol. 19, pp. 179–192 (1998)

Schlattmann, P., Dietz, E., Boehning, D.: Covariate adjusted mixture models and disease mapping with the program DismapWin. Stat. Med. **15**(7–9), 919–929 (1996)

Wang, P., Puterman, M.L., Cockburn, I., Le, N.: Mixed Poisson regression models with covariate dependent rates. Biometrics **52**(2), 381–400 (1996)

Wang, K., Yau, K.K.W., Lee, A.H.: A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay. Stat. Med. **21**(23), 3639–3654 (2002)

Wang, K., Yau, K.K.W., Lee, A.H., McLachlan, G.J.: Two-component Poisson mixture regression modelling of count data with bivariate random effects. Math. Comput. Model. **46**(11–12), 1468–1476 (2007)

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., Lindenmayer, D.B.: Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. **88**(1–3), 297–308 (1996)

Yealy, D.M., Auble, T.E., Stone, R.A., Lave, J.R., Meehan, T.P., Graff, L.G., Fine, J.M., Obrosky, D.S., Edick, S.M.: The emergency department community-acquired pneumonia trial: Methodology of a quality improvement intervention. Ann. Emerg. Med. **43**(6), 770–782 (2004)

Yealy, D.M., Auble, T.E., Stone, R.A., Lave, J.R., Meehan, T.P., Graff, L.G., Fine, J.M., Obrosky, D.S., Mor, M.K., Whittle, J. et al.: Effect of increasing the intensity of implementing pneumonia guidelines: a randomized, controlled trial. Ann. Intern. Med. **143**(12), 881–894 (2005)