Bayesian multinomial ordered categorical response model for the analysis of length of hospital stay

Lawrence N Kazembe

Length of hospital stay (LOS) is of primary importance in health services research because it is directly related to health care management and cost of health care. In some epidemiological settings the actual length of stay is not directly observed but it is known to have happened in a particular interval or for simple epidemiological interpretation time is categorized into ordered categorical responses. In this paper, we focus our attention on cumulative regression models for ordinal responses to analyze length of hospital stay for children admitted to a paediatric ward for malaria. Such models exploit the ordered scale of the outcomes. We approach our analysis using a Bayesian probit model. Our model incorporated random effects for hospital specific heterogeneity, while simultaneously investigating nonlinear effects in covariates within the general framework of semi-parametric regression models. Findings indicate children who died had relatively shorter LOS, which suggest worse prognosis at admission. Calendar time effects indicated changing seasonal effects with high peaks in wet season and low peak in dry season, largely explained by malaria transmission patterns. Age showed deviation from linearity, and early discharge was associated with much older children than infants.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

Bayesian multinomial ordered categorical response model for the analysis of length of hospital stay

Lawrence N. Kazembe*

Department of Statistics and Population Studies, University of Namibia

P/Bag 13301, 340 Mandume Ndemufayo Avenue, Windhoek, Namibia

January 14, 2016

Abstract

Length of hospital stay (LOS) is of primary importance in health services research because it is directly related to health care management and cost of health care. In some epidemiological settings the actual length of stay is not directly observed but it is known to have happened in a particular interval or for simple epidemiological interpretation time is categorized into ordered categorical responses. In this paper, we focus our attention on cumulative regression models for ordinal responses to analyze length of hospital stay for children admitted to a paediatric ward for malaria. Such models exploit the ordered scale of the outcomes. We approach our analysis using a Bayesian probit model. Our model incorporated random effects for hospital specific heterogeneity, while simultaneously investigating nonlinear effects in covariates within the general framework of semi-parametric regression models. Findings indicate children who died had relatively shorter LOS, which suggest worse prognosis at admission. Calendar

*Corresponding author; E-mail: lkazembe@unam.na

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

time effects indicated changing seasonal effects with high peaks in wet season and low peak in dry season, largely explained by malaria transmission patterns. Age showed deviation from linearity, and early discharge was associated with much older children than infants.

Keywords: Paediatric malaria; Ordinal regression; Bayesian inference; Threshold models

1 Introduction

In heath services research, length of hospital stay (LOS) is of primary importance because it is directly related to health care management [1] and cost of health care [2], and research in this area is ongoing [3–7]. Several methods have been proposed to model LOS and its association with patient characteristics and hospital effects [1–4; 8–11]. These include multiple linear regression where length of stay is assumed to be completely observed for each patient [2]. This model is very basic and ignores skewness and censoring in the data. Models that have been suggested to deals with skewness are gamma, inverse-Gaussian and log-normal models [1; 5; 8]. At times LOS is censored and actual time in hospital is incomplete for some patients. In this case survival models have been fitted [6; 7; 9–11]. Further alternative methods of analysis have used mixture models and multilevel models. For instance, Atienza et al. [4] and Lee et al. [12] used a gamma mixture model to analyze heterogeneity in maternal LOS, while Wang et al. [10] explored use of Poisson mixture regression models to analyze the same data. A study by Carely [13] fitted multilevel models to examine patient and hospital characteristics associated with LOS.

For virulent and infectious diseases, like malaria, the demand and costs for health care are huge and stretched especially during high transmission season (e.g. rainy season). Reducing the amount of time spent in hospital is a priority. Discharge planning emphasize that patients who are not worse-off should be treat and discharge for home management the same day they reported for treatment. Those of worse progno-

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

sis, probably associated with severe stages of the disease, have to be discharged once clinical stability has been reached [14–16]. In this case, the actual duration becomes redundant, but what is recorded is that discharge occurred within some thresholds of interest, often described as short, medium and long term LOS, thus generating some ordered categorical data [17; 18].

Cumulative models, as discussed in [19–21], are widely used regression model for ordinal categorical data. Cumulative models are an alternative approach to modelling waiting time data, in this case length of stay till discharge. The ordinal responses arise by categorizing the continuous outcomes (i.e the interval in days) by adjacent intervals along the continuous scale. The observed response can be regarded as the result of a sequential or cumulative process in which each time point (response category) can be reached successively [19; 20]. Several advantages also justify the choice of cumulative ordinal models to analyse event history data. Firstly, the ordinal categorical model compared to other duration models (e.g. the classical proportional hazards model) avoids the estimation bias introduced by long-term survivors. Secondly, the sequential ordinal model can be used to model non-proportional and non-monotonic hazard functions, and the effect of time-varying covariates can be allowed [20; 21].

Extensions to the cumulative models are many. For example, there are extension to deal with correlated ordinal outcomes, see for extensive discussions [21; 22]. An important correlation in the outcome can be due to spatial dependence or clustering. A popular method for modelling spatial correlation is to introduce random effects. Models with random effects also make it possible to account for unobserved spatial dependence or heterogeneity. Furthermore, continuous variables may exhibit nonlinear effects, and semi-parametric and non-parametric models have been proposed to deal with such [23; 24].

In this paper, we focus our attention on cumulative regression models for ordinal response to analyze length of hospitalization. The use of such models to analyze LOS are few, see for example [17; 18; 20]. In [17; 18] the objective was for effective discharge planning, while in [20] the aim was to exemplify use of ordinal regression model

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

to analyze duration of hospitalization. Inference for these models have proposed use of maximum marginal likelihood or full likelihood, generalized estimating equations, restricted maximum likelihood, penalized quasi-likelihood, empirical Bayesian and fully Bayesian inference [22–24]. We approach our analysis by employing a Bayesian probit model as suggested in literature, see for example [23; 24]. Our model incorporated random effects for area clustering and hospital specific heterogeneity, while simultaneously investigating nonlinear effects in covariates within the general framework of semi-parametric regression models. In the next section 2 we outline the model, while describing the Bayesian framework adopted for the models. Section 4 the model is applied to data on length of hospital stay for children hospitalized for malaria. We conclude with a discussion in Section 6.

2 The Model

2.1 Bayesian multinomial probit model with ordered response

In this section, we consider cumulative models based on threshold approach to model the case of an ordered response Y. Suppose the variable Y has r ordered categories. In additional we have a vector $v = (v_1, \ldots, v_p)'$ of p categorical covariates, the metrical covariate vector $x = (x_1, \ldots, x_p)'$ and further a random effects vector $b_i \in \{1, \ldots, B\}$ for heterogeneity among clusters of observations caused by unobserved or unmeasured covariates for example areal or hospital effects. Now, the observations $(Y_i, x_i, v_i, b_i), i = 1, \ldots, N$ are assumed independent. Then the cumulative regression model relates the cumulative response probabilities $P(Y_i \leq r|x_i, v_i, b_i), r = 1, \ldots, k$ to the covariates by a smooth link function h of the form and the response distribution (y) belongs to the exponential family [23].

A very intuitive approach to this is to postulated that Y is a categorized version of the unobserved latent variable Z [23],

$$Z = \eta + \varepsilon, \tag{1}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

obtained through the threshold mechanism where the density of Z is divided into slices determined by the thresholds

$$Y = r \Leftrightarrow \theta_{r-1} < Z < \theta_r, r = 1, \ldots, k, \tag{2}$$

with thresholds $-\infty = \theta_0 < \theta_1 < \ldots < \theta_k = \infty$. Assuming the error variable ε has the distribution function F, then Y obeys a cumulative model.

$$P(Y \leq r) = F(\theta_r - \eta) \tag{3}$$

where η is the predictor. The common choice for ε is the logistic or standard normal leading to either cumulative logit or probit model respectively. Here we assumed that ε follow a standard normal distribution, hence we have a cumulative probit model or a multinomial probit model with ordered categories,

$$P(Y \leq r) = \Phi(\theta_r - \eta) \tag{4}$$

For identifiability one of the thresholds is set to zero and an intercept is included in the model as a fixed effect. Normally the last category is set to zero, i.e $\theta_k = 0$. The predictor η is specified for each child i as

$$\eta_i = w'\gamma + \sum_{j=1}^{p} f_j(x_i) + h_i + s_i. \tag{5}$$

where w' is a vector of linear fixed covariates, f_j is (probably smooth) function of continuous or metrical variable x_i, and h_i is the hospital specific effects, $h \in (1, \ldots, H), H < n$, where child i was referred from or some unstructured spatial effect of an area. Some of these spatial effects may be structured, thus forcing similarities for neighbouring wards. This similarity in risk is captured by $s_i, s \in 1, \ldots, S$. A rationale is that a spatial effect is usually a surrogate of many unobserved influences, some of them may obey a heterogeneity that be present only locally, such as clustering of cases arising from the same clinic. If clustering is ignored in the model, estimates may be unstable [2]. Such models are common in spatial epidemiology. We propose implementing model (5) using a full Bayesian approach.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

3 Bayesian Inference

3.1 Prior assumptions

For a full Bayesian inference, all the unknown functions and fixed effects are assumed random variables and suitable prior assumptions must be specified, in an additional stage of hierarchy. We follow the prior distributions proposed by Fahrmeir and Lang [23].

The fixed effects are assumed to have a diffuse prior, $p(\gamma) \propto constant$. Similarly the threshold parameters are also assumed to have a diffuse prior distribution. Highly dispersed Gaussian priors are another suitable choice.

For the metrical covariates, priors are based on local smoothness priors. For equally-spaced observations $u = 1, \ldots, m$, say, of a metrical covariate, we assign first and second order random walk models, i.e.,

$$f(u) = f(u-1) + \xi(u) \qquad \text{or} \qquad f(u) = 2f(u-1) - f(u-2) + \xi(u) \tag{6}$$

respectively, with Gaussian errors $\xi(u) \sim N(0, \tau_j^2)$ and diffuse priors $f(1) \propto const$ or $f(1)$ and $f(2) \propto$ constant, for initial values, respectively. Both specifications act as smoothness priors that penalize too rough functions of f.

For the unstructured component the prior assumed is a Gaussian with i.i.d. effects with

$$h \sim N(0, \tau_h^2), h = 1, \ldots, H. \tag{7}$$

Finally, for the spatial components s_i, we assign a Markov random field (MRF) prior [25]. This is analogous to random walk models. The conditional distribution of s, given adjacent areas r, is a univariate normal distribution with mean equal the average r values of s_i's neighbouring areas and variance equal to τ_s^2 divided by the number of

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

adjacent areas. This leads to a joint density of the form

$$p(s|\tau_s^2) \propto \exp\left(-\frac{\tau_s^2}{2}\sum_{i\sim j}(s_i - s_j)^2\right) \tag{8}$$

where $i \sim j$ denotes that area i is adjacent to j, and assumes that parameter values s_i and s_j in adjacent areas are similar. The degree of similarity is determined by the unknown precision parameter τ_v^2.

By writing $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, $\boldsymbol{h} = \boldsymbol{Z}_k\boldsymbol{\beta}_k$ and $\boldsymbol{s} = \boldsymbol{Z}_l\boldsymbol{\beta}_l$, for a well defined design matrix $\boldsymbol{Z}$ and a (possibly high-dimensional) vector of regression parameters $\boldsymbol{\beta}$, all different priors (Equations 6-8) can be expressed in a general Gaussian form

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right) \tag{9}$$

with an appropriate penalty matrix $\boldsymbol{K}_j$. Its structure depends on the covariate and smoothness of the function. In most cases, $\boldsymbol{K}_j$ is rank deficient and hence the prior for $\boldsymbol{\beta}_j$ is improper. For the variances τ_j^2 we assume inverse Gamma priors $IG(a_j, b_j)$, with hyperparameters a_j, b_j chosen such that this prior is weakly informative.

All these priors can be equivalently rewritten in form of the general prior

$$f|\tau^2 \propto \exp(-\frac{1}{2\tau^2}f'Kf) \tag{10}$$

$$\tau^2 \sim IG(a, b) \tag{11}$$

where K^- is an appropriate penalty matrix. Since K is often not of full rank, then $f|\tau^2$ follows a partially improper Gaussian prior

$$f|\tau^2 \sim N(0, \tau^2K^-)$$

where K is a generalized inverse of the penalty matrix K. The parameter τ^2 is the variance parameter that controls the amount of smoothness. For τ^2 we choose a highly dispersed, but proper, inverse Gamma prior as given in equation (11) with $a = 1$ and $b = 0.005$ or $a = b = 0.001$.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

3.2 Posterior inference

Regression tool for full Bayesian inference is based on MCMC techniques. MCMC techniques are used to draw samples from the posterior distribution which are analytically intractable because of highly dimensional nature of the distributions. Assuming Gaussian errors, we obtain multicategorical probit models with latent semiparametric Gaussian models. The sampling scheme has been developed on the basis of the latent variable mechanisms (2). Now for $\alpha = (f, b)$, τ is a vector of all variance components and γ is a vector of all fixed effects, then the posterior is given by

$$P(\alpha, \gamma, \tau, Z|y) \propto p(y|Z)p(Z|\alpha, \gamma)p(\alpha|\tau)p(\tau)p(\gamma). \tag{12}$$

where $p(y|\alpha, \gamma, Z)$ is the likelihood function for the data given parameters and $p(Y|Z) = \prod_i p(Y_i|Z_i)$. A number of sampling schemes are available. For a Gaussian response, one can use Gibbs sampling by drawing samples from the Gaussian full conditionals. Efficiency is guaranteed by Cholesky decomposition for band matrices [26].

The full conditional $\gamma|\cdot$ for fixed effects with diffuse priors is Gaussian with mean

$$E(\gamma|\cdot) = (U'CU)^{-}U'C(y - \tilde{\eta})$$

and covariance matrix

$$Cov(\gamma|\cdot) = \sigma^2(U'CU)^{-}$$

where U is the design matrix of fixed effects and $\tilde{\eta}$ is the part of the additive predictor associated with the other factors in the model such as nonparametric terms. Similarly, the full conditional for the regression functions f_j is Gaussian with mean

$$E(f_j|\cdot) = \left(\frac{X'CX}{\sigma^2} + \frac{K}{\tau^2}\right)^{-} + \frac{1}{\sigma^2}X'C(y - \tilde{\eta})$$

and covariance matrix

$$Cov(f_j|\cdot) = \left(\frac{X'CX}{\sigma^2} + \frac{K}{\tau^2}\right)^{-}.$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

4 Applications to paediatric length of hospital stay

4.1 The Data

The data presented here were from 3937 patients, diagnosed of malaria, in a paediatric ward in Zomba hospital, Malawi. A detailed survey and data description can be found in [27]. The response was time spent in the hospital measured in days, categorized into three: (1) if length of stay was within one day; (2) if length of stay was up to three days and (3) if the length of stay was more than three days. The covariates available for smooth modelling were age at admission, calendar time between between 1/1/2002 and 31/1/2004, measured in months since 1/1/2002 and distance to the hospital. Categorical covariates were: sex (male=1, 0=otherwise), hospital management (1=mission hospital, 0=otherwise), season (1=wet, 0= dry), day admitted (1=weekend, 0=weekday), hospitalization outcome (1=died, 0=discharged), hospital type (1=dispensary, 2=rural hospital, 3=clinic). Summaries of these variables are given in Table 1. The mean and median LOS were 3.04 and 3.00 days respectively. Variability of LOS by season, day, referral and outcome are evident in the Table.

We proposed use of ordinal probit regression models, where our interest was to model the duration of stay being "within 1 day" or "between 2 days up to 3 days" or "more than three days". By segmenting duration of stay we realized a triple response ordered categories variable, y_i, i.e.,

$$y_i = \begin{cases} 1 & \text{if length of hospital stay for child } i \text{ was up to 1 day} \\ 2 & \text{if the length of stay was up to 3 days} \\ 3 & \text{if the length of stay was more that 3 days} \end{cases} \tag{13}$$

The predictor of the model for "probability of staying" has the form

M0: $\eta_i = \theta_j - w'\gamma$.

In this model we estimate fixed effects only, and is therefore considered as the null model against which the performance of all other models is compared. The fixed effects in the Bayesian framework are modelled by assuming diffuse priors.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

The second model, $M1$, is a spatial parametric model which adjusts for both spatially structured and unstructured random effects and covariates,

M1: $\eta_i = \theta_j - w_i'\gamma - s(ward_i) - h(ward_i)$

With this model, we assess how much of the total variability is explained by spatial variation in the response. This is achieved by assuming CAR priors (Eq. 8). Further, the model permits unstructured heterogeneity, modelled through Eq. (7).

In the third model $M2$, we fit a flexible model by allowing calendar effects, in additional to the spatial effects and fixed effects for the other covariates

M2: $\eta_i = \theta_j - w_i'\gamma - f(time) - s(ward_i) - h(ward_i)$.

The time components are assumed linear. The last model $M3$, we fit a spatial semi-parametric model with age of the child, distance to the facility and time assumed nonlinear and the rest of the variables assumed fixed,

M3: $\eta_{ij} = \theta_j - w_i'\alpha - f_1(time) - f_2(age_i) - f_3(distance_i) - s(ward_i) - h(ward_i)$

For the nonlinear effects we use a second-order random walk prior (Eq. 6). Model $M3$ investigates the bias of fitting restrictive linear model, $M2$. Implementation of all models were carried out in `BayesX` version 1.4 [28].

4.2 Model comparison and Sensitivity analysis

The four model were compared using the Deviance Information Criterion (DIC) [29]. The DIC is defined as $DIC = D(\bar{\mu}) + p_D$, where $D(\bar{\mu})$ is the posterior expectation of the deviance, and p_D is the effective number of parameters (which is similar, but not equal, to degrees of freedom). The model with a smaller DIC is better than others.

Bayesian cumulative threshold models have an important problem of mixing and convergence, specifically for the threshold parameters. Large MCMC samples have to be taken to realize stable estimates. Varying hyperparameters a and b is often recommended when modelling nonlinear functions, because in some situations, the estimated nonlinear functions may vary considerably because of the choice of hyper-parameters. We therefore carried out sensitivity analysis by assumed three starting values, $a = 0.001, b = 0.001$ or $a = 1, b = 0.005$ or $a = 0.0001, b = 0.0001$ as suggested

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

in [28].

5 Results

First, we carried out a test of parallel lines to assess the proportional odds assumption. This test assumes, under null hypothesis, that the slopes of the coefficients in the ordinal model are the same across response categories. Test of proportional odds assumption showed that the ordinal model was appropriate at p-value of 0.05. We then proceeded to fitting the Bayesian semiparametric models. Model selection results are shown in Table 2. Sensitivity analysis based on the DIC show that the semi-parametric model ($M3$) had a consistently smaller DIC and hence better compared to the other models.

Results of the fixed estimates, based on model $M3$, are given in Table 3. The model only included variables found significantly associated with LOS based on results in 1. Age, distance and calendar time were fitted as nonlinear effects. Included in the table are estimates of the threshold parameters, θ_1 and θ_2. Parameter θ_1 is a threshold for the LOS of 1 day, while θ_2 is that of LOS between 2 and 3 days. These estimates were significant. For interpretation of the results of threshold parameters, higher (lower) values correspond to early/shorter (delayed/longer) LOS. For instance, a negative sign of θ_1 signifies a shift on the latent scale to the right side, yielding a lower probability for category "up to 1 day". Conversely, a positive sign of θ_2 signifies a shift to the left side, yielding higher probabilities for category "between 2 days and 3 days".

Based on model $M3$, length of hospital stay was associated with season, day and discharge outcome. Discharge through death was negatively associated with LOS (γ_1=-1.18, 95%CI: -1.36,-0.98). This implies children who died tend to have a shorter duration of hospitalization. With regards to season, we observe that LOS was much shorter during the dry season than wet season (γ_2=-0.17, 95% CI:-0.31,-0.03). This may suggest that malaria episodes during the dry season may not be of the severe form, hence improved LOS. Similarly LOS is negatively associated with week days

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

(γ_3=-0.52, 95% CI: -0.68,-0.35). This is expected, as ward rounds are often conducted during weekdays, therefore those admitted during the weekend will stay much longer, till the following Monday or Tuesday to be assessed for discharge. For those referred from another facility, they tend to stay longer than those not referred (γ_4=0.020), however, the result is not significant.

The calendar time effects evidently showed changing seasonal effects in length of stay for period of study (Figure 1-top left plot). High peaks are observed in between 33rd and 37th week, as well as in the 89th to 92nd week, whereas low peaks are in the 13th to 15th week, and in the 50th to 53rd week. The effect of age showed slight deviations from linearity, with the posterior means increasing with increasing age, more pronounced after the age of 60 months (Figure 1-top right plot). In other words the probability of early discharge was higher for much older children than infants. For distance, the estimated effects were almost linear (Figure 1-bottom plot). The small variation at distance 5, 10 and 20 kilometres could be due to data heaping, as the distances to Zomba hospital were approximated from the referring hospital centre.

The residual spatial effects are plotted in Figure 2. The estimated smooth geographical effects, with values ranging from -0.18 to +0.27, are varied. Indeed, some of the effects are significant both at nominal values of 80% and 95% (Figure 3). We observe areas of significantly negative effects, which we interpret as areas where children admitted from those areas did not stay long in the hospital. Those children admitted in black areas stayed relatively longer than others in the district. The uncorrelated spatial heterogeneity at health facility level is given by caterpillar plot in Figure 4. There are no clear differences in catchment area specific effects, and most of them have a near zero effect on the probability of staying in hospital. It is clear that the spatially correlated effects are dominant, based on the ratio of variance components, $\phi = \tau_s^2/(\tau_s^2 + \tau_H^2) = 8.701/(8.701 + 0.005) = 0.99$ (Table 3).

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

6 Discussion and Conclusion

This paper considered a Bayesian cumulative probit model with ordered categories for the analysis of length of hospital stay data. The data concerns time spent in a a paediatric ward among children who were hospitalized for malaria, characterized as short, medium and long term. Modelling of categorized LOS is of major importance in health services research [1–3], as it allows for prediction of the probability of LOS falling within any category. Besides the model allowed us to investigate explanatory variables that influence LOS [4–7]. Our study extends use of ordinal categorical models as an flexible alternative to event history models often considered in analyzing LOS [19; 20; 24].

The cumulative model may also be used for timely assignment of an estimated date of discharge [17; 18]. Using the threshold estimates (θ_j), one is able to estimate the probability of discharge within 1 day, or within three day or more (Table 3). Therefore this model can be considered as a tool to determine factors of delayed discharge. In our analysis, discharge time was associated with health outcome, season, day of the week. The fact that LOS was shorter for those who died, it does indicate that these were worse-off at the time of admission. Many times, in rural Africa, treatment seeking is influenced by the severity of disease [30]. The results also displayed strong seasonality and spatial heterogeneity. This seasonality is evidently governed by malaria transmission patterns [16]. The spatial effects are often a surrogate of underlying unobserved information, and may give leads for further epidemiological research or assist in designing malaria interventions. For example, the increased risk in rural areas may be an influence of different factors, such as unavailability or inaccessibility of health facilities resulting in increased risk for such children. These effects may also reflect health seeking behaviour [27; 30].

Although our results did not show significant results for distance to referral hospital, type of facility and owner of facility these play an important effect as they influence hospitalization trends and define hospital heterogeneity [11]. Studies have shown that geographical accessibility of care is directly related to distance to the health facility

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

[2; 11]. The type of facility variable provided an indicator of quality expected at the hospital. Rural hospitals are expected to give quality care than dispensaries and clinics in terms of diagnosis and appropriate treatment. Lack of quality care make people travel long distances to a facility that would provide adequate care. Management of the hospital is another variable whose effect is not well documented. There is a tendency somewhat that mission hospitals are often preferred to government hospitals especially in urban and peri-urban centers. Understaffing and lack of resources in recent years has contributed to this trend.

In our study we adjusted for hospital level clustering, although it was not significant, it shows considerable variation across health facilities. Moreover, this adjustment was important for two reasons. Firstly, because of similarities in practice styles and organization of the hospital, children originating from the same hospital are likely to receive the same type of care than those coming from another hospital. By adjusting for clustering, thus statistical significance in the results is also adjusted for [2; 13]. Secondly, by specifying random intercepts we directly modelled for any unobserved or unmeasured heterogeneity [23].

In conclusion, the development of flexible models for health services research like LOS is essential to unravel all important determinants of LOS for effective discharge planning. Appropriate models used to describe variability in LOS can save costs and resources, else poorly fit models are detriment for decision making [1–3].

References

[1] Stevens A, Hammer K and Buchkremer G (2001). A statistical model for length of psychiatric in-patient treatment and an analysis of contributing factors. *Acta Psychiatrica Scand.*, **103**:203–211.

[2] Lorch, SA, Zhang X, Rosenbaum PR et al (2004). Nonrural Hospitals Equivalent Lengths of Stay of Pediatric Patients Hospitalized in Rural and Nonrural Hospitals. *Peadiatrics*, **114**:400–408.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

[3] Marazzi A, Paccaud F, Ruffieux C, Beguin C (1998). Fitting the distributions of length of stay by parametric models. *Medical Care*, **36**:951-927.

[4] Atienza N, Garcia-Heras J, Munoz-Pichardo JM, Villa R. (2008). An application of mixture distributions in modelization of length of stay. Statistics in Medicine. 27:1403-1420.

[5] Faddy M, Graves N, Pettitt A. (2009). Modelling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. Value Health. 12:309-314.

[6] Khan SK, Rushton SP, Dosani A, Gray AC, Deehan DJ. (2012). Factors influencing length of stay and mortality after first hip ad second hip fractures: an even modelling analysis. J Orthop Trauma. In Press.

[7] Tang X, Luo Z, Gardiner JC. (2012). Modelling hospital length of stay by Coxian phase-type regression with heterogeneity. Statistics in Medicine. 30:1502-1516.

[8] Polverejan E, Gardiner JC, Bradley CJ, Holmes-Rovner, M and Rovner D (2003). Estimating mean hospital cost as a function of length of stay and patient characteristics. *Health Economics*, **12**: 935–947.

[9] Sorensen, J. (1996). Multi-phased bed modelling. *Health Services Management Research*, **9**: 61–67.

[10] Wang, K, Yau, K.K.W, and Lee, A.H. (2002). A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay. *Statistics in Medicine*, **21**:3639–3654.

[11] Picone G., Wilson, R. Mark.,and Chou, S., (2003). Analysis of hospital length of stay and discharge destination using hazard functions with unmeasured heterogeneity, *Health Economics*, **12**:1021–1034.

[12] Lee AH, Wang K, Yau KK, McLachlan GJ, Ng SK. Maternity length of stay modelling by gamma mixture regression with random effects. Biom J. 2007 Aug;49(5):750-64.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

[13] Carely K (2002). Hospital length of stay and cost: A multilevel modelling analysis. Health Services and Outcomes Research Methodology. 3:41-65.

[14] Government of Malawi (2002). *Malaria Policy*. National Malaria Control programme. Community Health Science Unit. Division of Preventive Health Services. Ministry of Health and Population.

[15] Snow RW, Omumbo JA, Lowe B, et al (1997). Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *Lancet*, **349**: 1650–54.

[16] English M, Esamai F, Wasunna, A et al (2004). Assessment of inpatient paediatric care in first referral level hospitals in 13 districts in Kenya. *Lancet*, **363**: 1948–53.

[17] Parakramaweera Sunil Dharmapala. An Ordinal Logistic Regression Model to Estimate Cumulative Odds of Length-of-Stay for Efficient Hospitals in DEA Setting: The Case of Oman. Proceedings of the 10th Annual Conference of Asia-Pacific Decision Science Institute, Taipei, Taiwan. June 28-July 2, 2005.

[18] Ou L, Chen J, Young L, Santiano N, Baramy LS, Hillman K. (2011). Effective discharge planning- timely assignment of an estimated date of discharge. *Australian Health Review*, **35**:357-363.

[19] Läärä E, Matthews JN. 1985. The equivalence of two models for ordinal data. *Biometrika* **72**:206-207.

[20] Albert J, Chib S. 2001. Sequential ordinal modelling with applications to survival data. *Biometrics* **57**:829–836.

[21] Tutz G. 2003. Generalized semiparametrically structured ordinal models. *Biometrics* **59**:263–273.

[22] Liu I, Agresti A. 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. *Test* **14**:1-73.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

[23] Fahrmeir, L. and Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics*, **53**:10–30.

[24] Gieger, C. (1997). Non- and semiparametric marginal regression models for ordinal response. *Discussion Paper no.* , Statistics department, University of Munich, Germany.

[25] Besag J, York J, Mollie A. 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**: 1-59.

[26] Rue H. (2001). Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society, B 63*: 325-338.

[27] Kazembe LN, Kleinschmidt I, Sharp BL. (2006). Patterns of malaria-related hospital admissions and mortality among Malawian children: an example of spatial modelling of hospital register data. *Malaria Journal* **5**:93.

[28] Brezger A, Kneib T, Lang S (2005). *BayesX*: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software* 14:11.

[29] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B* 2002; **64**:1–34.

[30] de Savigny D, Mayombana C, Mwageni E et al (2004). Care-seeking patterns for fatal malaria in Tanzania. *Malaria Journal*, **3**:27, open access at `http://www.malariajournal.com/content/3/1/27`.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

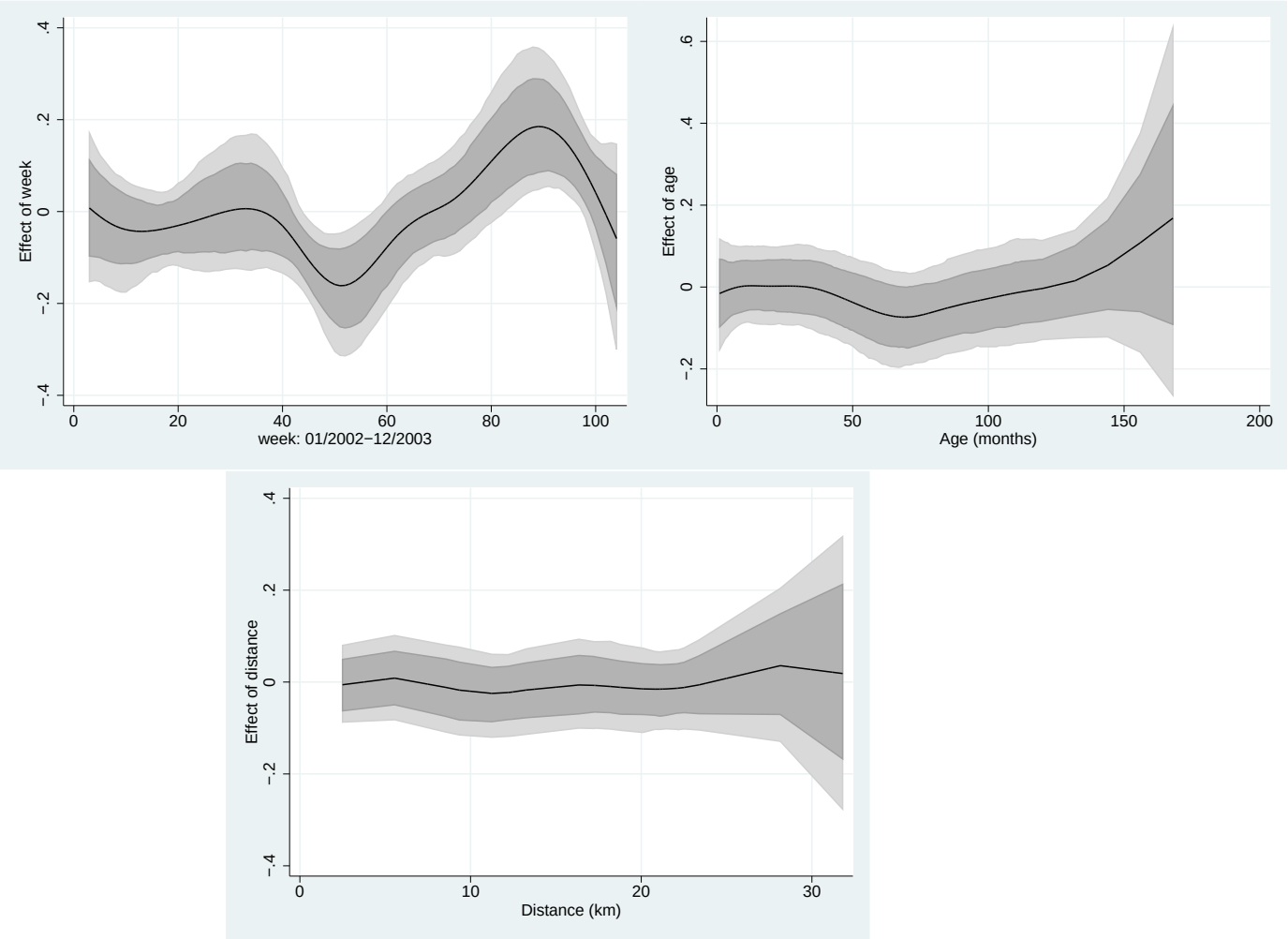

Figure 1: Non–linear effect of calendar time (in weeks), age (in months) and distance (in km), with corresponding 80% and 95% credible bands.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

Figure 2: Structured spatial effects, at ward level, of length of stay (Model M3). Shown are the posterior means.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

Table 1: The proportion of children for the three categories of length of stay. Numbers given are row percentages corresponding to the total given in each category. The p-value is based on χ^2-test.

Variable		Length of hospital stay			Total	p-value
		Same day	1–3days	≥4days		
Age	<1yr	16.1	45.8	38.2	1189	0.32
	1–4yr	17.6	46.3	36.0	2136	
	5+ yr	17.4	42.9	39.8	644	
Sex	Female	17.2	44.6	38.1	1683	0.60
	Male	17.1	46.2	33.6	2286	
Day	Weekday	20.1	39.7	40.5	2418	<0.001
	Weekend	12.4	55.1	32.5	1492	
Season	Wet	22.0	48.5	28.4	2261	0.003
	Dry	22.7	41.7	34.0	1128	
Referred	No	16.5	47.8	35.7	1895	0.032
	Yes	17.9	44.2	37.9	1494	
Distance	≤ 5km	18.0	44.8	37.2	1938	0.49
	>5km	16.4	46.5	37.1	1999	
Hospital type	Dispensary	21.9	47.5	30.6	1869	0.74
	Rural hospital	20.4	51.6	28.0	93	
	Clinic	23.2	46.2	30.5	1975	
Hospital management	Government	22.8	46.6	30.5	3340	0.56
	Mission	20.8	48.9	30.3	597	
Hospitalization outcome	Died	60.9	16.9	13.2	302	<0.001
	Discharged	18.5	49.4	32.1	3667	

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

Table 2: Sensitivity analysis and summary of the DIC of the four models fitted. See text.

Hyperparameter	Diagnostics	Model 0	Model 1	Model 2	Model 3
$a = 0.001$, $b = 0.001$	Deviance ($\bar{D}$)	8049.72	6785.29	6780.49	6778.79
	p_D	23.74	39.99	36.85	40.93
	DIC	8097.20	6865.27	6854.19	6860.66
$a = 1$, $b = 0.005$	Deviance ($\bar{D}$)		6788.33	6782.77	6779.83
	p_D		40.25	36.06	41.78
	DIC		6868.83	6854.90	6863.39
$a = 0.0001$, $b = 0.0001$	Deviance ($\bar{D}$)		6788.38	6788.03	6783.12
	p_D		38.48	33.16	36.54
	DIC		6865.33	6854.35	6860.20

Table 3: Parameter effects in the nonlinear ordinal model (Model 3)

Covariate		Mean	2.5% quant.	97.5% quant.
Threshold parameters				
θ_1		-0.93	-1.14	-0.73
θ_2		0.40	0.21	0.62
Fixed effects				
Outcome	discharged	0		
	died	-1.18	-1.36	-0.98
Day	weekend	0		
	weekday	-0.52	-0.68	-0.35
Season	wet	0		
	dry	-0.17	-0.31	-0.03
Referral	no	0	0	
	yes	0.02	-0.09	0.13
Random effects				
Unstructured effects (τ_H^2)		0.005	0.001	0.019
Spatial effects (τ_s^2)		8.701	2.530	23.896

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

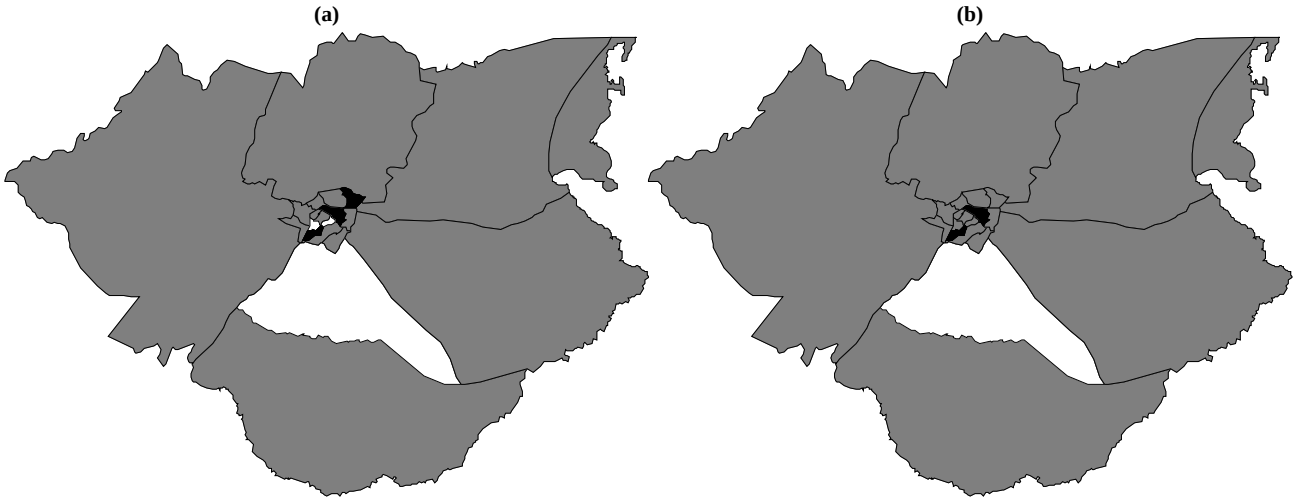

Figure 3: Corresponding posterior probabilities, to structural spatial effects (Figure 2), at (a) 80% and (b) 95% nominal level, white denotes regions with strictly negative credible intervals, black denotes regions with strictly positive credible intervals, and gray depicts regions of nonsignificant effects.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016

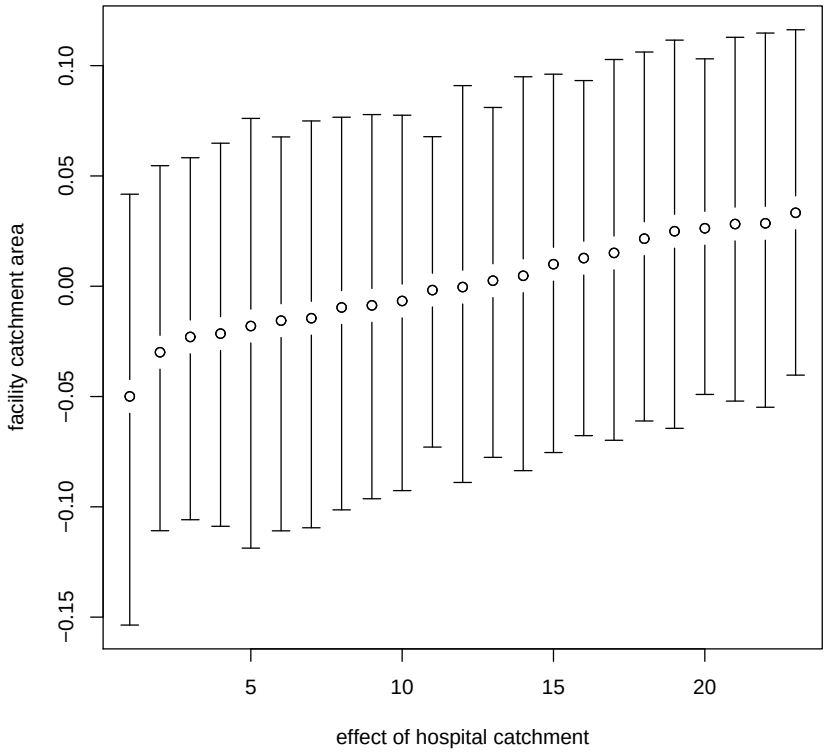

Figure 4: Residual unstructured heterogeneity effects of primary health care facilities. Shown are the caterpillar plots of posterior means (circles), with 80% error bars.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1663v1 | CC-BY 4.0 Open Access | rec: 21 Jan 2016, publ: 21 Jan 2016