The University Of Sheffield.

Department Of Economics.

Sheffield Economic Research Paper Series.

A Semiparametric Bayesian Approach to a New Dynamic Zero-Inflated Model

Kiranmoy Das, Bhuvanesh Pareek, Sarah Brown, and Pulak Ghosh

ISSN 1749-8368

SERPS no. 2017001

January 2017

A Semiparametric Bayesian Approach to a New Dynamic Zero-Inflated Model

Kiranmoy Das, Bhuvanesh Pareek, Sarah Brown*, and Pulak Ghosh

January 14, 2017

Abstract

We develop a dynamic zero-inflated model to analyse the number of hospital admissions within an aging population, which allows for the considerable number of zero hospital admissions at the individual level and occurrence dependence. In addition, certain health conditions may lead to groups of individuals having similar hospital admission rates. We analyse the US Health and Retirement Survey, which includes self-assessed health (SAH), which can be predictive of hospital admissions. Our modelling framework embeds a dynamic hierarchical matrix stick-breaking process to flexibly characterize this dynamic group structure allowing individuals to belong to different SAH groups at different points in time.

Key Words: Bayesian models; Dirichlet process; Dynamic hurdle; Lasso; Matrix stick-breaking process; Zero-inflated data.

JEL Classifications: C11; C14; I12

Kiranmoy Das is Assistant Professor, Indian Statistical Institute, Kolkata, India; Bhuvanesh Pareek is Assistant Professor, Indian Institute of Management, Indore, India; Sarah Brown is Professor of Economics at University of Sheffield, UK; Pulak Ghosh is Professor, Department of Decision Sciences and Information Systems, Indian Institute of Management, Bangalore, India.

*Corresponding Author Details: Sarah Brown; Department of Economics, 9 Mappin Street, Sheffield, South Yorkshire, S1 4DT, UK; Telephone, +44(0)114 2223404; Fax, +44(0)114 2223458.

1 Introduction

The fact that the world population is aging is well-established: as stated by the United Nations (2015), 'virtually every country in the world is experiencing growth in the number and proportion of older persons in their population.' Specifically, their latest estimates predict that between 2015 and 2030, the number of people in the world aged 60 years or over is projected to grow by 56 per cent, from 901 million to 1.4 billion. With respect to increases in life expectancy, the World Health Organisation (2016) reports US life expectancy at 79.3 years in 2015, compared to 66.6 years in 1960 and 46.3 years in 1900. It is apparent that such demographic changes will be associated with an increased demand for health care, as well as, changes in the nature of health care demand, with needs associated with chronic conditions becoming increasingly important. It is crucial, therefore, that we enhance our understanding of the demand for health care.

Hospital admissions are an important aspect of the demand for health care and, furthermore, it is apparent that as chronic conditions become more prevalent in the context of an aging population, understanding the needs associated with inpatient care and hospital visits becomes increasingly important. As aging changes the nature of health care demand with more emphasis on hospitalization and thus more challenges faced by health care systems throughout the world, further analysis into hospital admissions at the individual level seems to be particularly warranted. Thus, in this paper, we contribute to the existing literature which has focused on developing models to analyse the rate of hospital admissions at the individual level. We exploit data drawn from the US Health and Retirement Survey, which provides detailed information following individuals over time for an aging population. A number of analytical challenges associated with modelling hospital admissions at the individual level have been discussed in the existing literature, which serves to highlight the complexities associated with modelling the demand for health care. Such challenges have generally not been addressed within a unified framework. Thus, we aim to fill this gap in the existing literature by conducting a comprehensive study of these issues. We now discuss these analytical challenges in more detail.

Given the considerable amount of zero observations that are observed in measures of hospital admissions at the individual level, existing studies have developed zero inflated approaches for modelling counts of hospital admissions to account for the excess zeros (see, for example, Deb and Trivedi, 1997, Winkelmann, 2004 and Atella and Deb, 2008). We build on this existing literature by proposing a flexible modelling framework in which we

analyse the number of hospital admissions at the individual level in the context of an aging population. Simple distributions such as the Poisson or log normal distributions have been used to model count variables such as the number of hospital admissions. However, such an approach does not allow for the ordering information, which is likely to be inherent in such data: i.e., an individual who was hospitalised 3 times may be regarded as more serious compared to an individual who was hospitalised once in the same year. Hence, in this paper, we propose an ordered logistic regression approach, i.e. a proportional odds model, to analyse the number of hospital visits. Furthermore, as expected, there are a considerable proportion of zero observations in our data, on average, 84%, which is based on the US Health and Retirement Survey (HRS), reflecting the fact that a significant proportion of individuals are not admitted to hospital in a given year. In order to account for such inflation at zero, we develop an approach based on a zero-inflated proportional odds model.

There is an important additional issue associated with such data which is related to the inherent assumption that every year an individual faces the same probability of hospitalization, which may not be the case. Specifically, in the context of an aging population, the probability of hospitalization is likely to be different between an individual who has never been admitted to hospital compared to an individual who has been admitted to hospital. Clearly, the onset of chronic conditions varies across individuals and over time, as does the extent to which individuals are affected by such conditions and ultimately require hospital care. Once an individual has been admitted to hospital, it may well be the case, for example, that further follow-up visits ensue. Indeed, Westbury et al. (2016) analysing Hospital Episode Data for a specific region of the UK, state that 'in the context of hospital admission among older people, it is reasonable to expect that risk of admission will increase with the accumulated number of previous admissions'. Indeed, their argument is supported by their empirical analysis of multiple hospital admissions for a sample of elderly individuals with an average age of 66. In a similar vein, Banerjee et al. (2010) explore persistence in health care utilisation using dynamic panel data models. Their analysis of the 2000-2004 US Medical Expenditure Panel Survey (MEPS) endorses a dynamic modelling approach, with inpatient hospitalization at the initial period found to have a large positive and significant impact on current hospitalization. From a modelling perspective, such observations imply that the distribution of the time to the first hospitalization may have a very different rate to the distribution of times between subsequent hospitalization events, see the recent contribution in this regard by Baetschmann and Winkelman (2016). Thus, we account for 'occurrence dependence' in our modelling approach.

Finally, categories of self-assessed-health (SAH), where respondents rate their own general health on a response scale from say poor to excellent, have been used extensively in the health economics literature. SAH is regarded as a good proxy for health risk since it contains private information on health and health related behaviours that are predictive of future health and are known only to the respondent. Idler and Benyamini (1997), for example, show that SAH is predictive of mortality even after conditioning on objective measures of health. It is often useful to group individuals based on covariate characteristics and in our application it is interesting to group individuals according to their SAH. Furthermore, in the HRS data, these groups can be dynamic with SAH varying over time. Within a time-varying group structure, stronger dependence is expected among the data temporally close to each other. Thus, it is interesting to develop models where one can estimate the group-specific parameters by efficiently borrowing information across the groups dynamically in an automated manner.

Dirichlet Process (DP) priors have been successfully used in a wide range of applications for borrowing information across groups in an automated manner. Some examples include Ferguson (1973), Antoniak (1974), Blei and Jordan (2006), Dunson, Herring, and Engel (2008) and Rodriguez and Dunson (2014). Dunson et al. (2008) developed a matrix stick-breaking process (MSBP) which is an important extension of the usual DP. MSBP is essentially a generalization of the stick-breaking structure of DP (Sethuraman, 1994) where the row and column stick-breaking random variables induce dependent local clustering. In this paper, we first extend the stick-breaking weights into a product of group, time and predictor specific components and then induce extra dependence among the data that are temporally close to each other following Ren et al. (2010). The resultant proposed dynamic hierarchical MSBP (DH-MSBP) is powerful because here we allow: (i) multiple shrinkage of a large set of model parameters; (ii) global and local clustering by borrowing information within and across groups; and (iii) a higher probability of sharing the same set of parameters for temporally close data points.

The rest of the paper is organized as follows. In Section 2, we propose a zero-inflated dynamic hurdle proportional odds model for the count of hospital visits. In Section 3, we propose the DH-MSBP prior for the covariates with time-varying effects on the response. We also discuss the zero-inflated Lasso priors for the covariates with time-invariant effects in this section along with the joint posterior density and computational details. In Section 4, we present the simulation results demonstrating the effectiveness of the proposed approach. The results for the US HRS data analysis are presented in Section 5. Finally Section 6 concludes.

2 A Zero-Inflated Dynamic Hurdle Proportional Odds Model

Here we consider modelling hospital visits as our dependent variable. We distinguish between two types of individual: (i) those who started visiting hospitals from the very first wave (the start of the survey) until the end, and (ii) those who did not visit hospital initially but then started visiting after a few waves.

Define Y_{irt} as the count of hospital visits for the i-th individual in the r-th health status (self-assessed) group at wave t ($t = 1, 2, \ldots, 10$). Health status is based on self-assessment and can be categorized as "poor", "fair", "good", "very good", "excellent". These health states can obviously vary over time. Thus, we have a very general set-up.

Our base model is the following zero-inflated dynamic model for Y_{irt}:

$$Y_{irt} = (1 - \pi_{irt})1_{[Y_{irt}=0]} + \pi_{irt} G(Y_{irt}|Y_{irt} > 0), \tag{1}$$

where, π_{irt} is the probability of hospital admission, i.e., $\pi_{irt} = P(Y_{irt} > 0)$, and $G(Y_{irt}|Y_{irt} > 0)$ is the distribution of the count of hospital visits conditional on hospitalization. In general, when an individual visits hospital in a year, Y_{irt} can take any of the values among $k = 1, 2, \cdots, K$.

2.1 Model for π_{irt}

For modelling the proportion of non-zeros, we consider a Probit model and thus express π_{irt} as the following: $\Phi^{-1}(\pi_{irt}) = \mathbf{x}_i^T \boldsymbol{\delta} + \eta_i$, where Φ denotes the cdf of the standard normal, and $\mathbf{x}$ is the set of covariates (both with time-varying and time-invariant effects on the response) and $\boldsymbol{\delta}$ is the vector of regression coefficients. The individual-specific random effects η_i capture the correlation among the measurements from the same individual over different waves. Thus we allow the probabilities of the non-zero responses (and hence the zero responses) to vary over the waves.

2.2 Modelling $G(Y_{irt}|Y_{irt} > 0)$

The distribution of the count of hospital visits, i.e., $G(Y_{irt}|Y_{irt} > 0)$ can be of two types: (1) individuals who visit the hospital for the first time at the t-th wave and; (2) individuals who visited the hospital before the t-th wave. We argue that the rate of hospitalization should be different for these two types of individual who are exposed to hospital for the first time at t and who were exposed before t, respectively. We use a proportional odds model with a dynamic hurdle component for case 1 following Baetschmann and Winkelmann (2016) and use a simple proportional odds model for case 2.

Case 1

Let t be the wave where the i-th individual visits hospital for the first time. In this case, we follow Baetschmann and Winkelmann (2016) and use a dynamic hurdle Poisson model for modelling G. This is based on the assumption that in this particular wave, for each individual, the time of the first hospital visit and the total number of hospital visits are related to each other. Suppose the first hospital visit for individual i occurs at time, T_i; for $0 < T_i < t$. Therefore, one can write,

$$Pr(Y_{irt} = k, T_i) = Pr(Y_{irt} = k|T_i)f_1(T_i) = Pr(Y_{irt}(T_i, t - T_i) = k - 1)f_1(T_i);$$

where $Y_{irt}(T_i, t - T_i)$ denotes the number of hospital visits for the i-th individual in the t-th wave within the time interval $(T_i, t - T_i)$, and f_1 denotes the pdf of the time to the first hospital visit. The marginal distribution of Y_{irt}, thus, can be obtained as:

$$Pr(Y_{irt} = k) = \int_{t-1}^{t} Pr(Y_{irt}(T_i, t - T_i) = k - 1)f_1(T_i)dT_i. \tag{2}$$

We consider T_i as the time to elapse before the event of interest (hospital admission) and thus model this waiting time by a Weibull distribution. Specifically, we consider f_1 as the density of a Weibull distribution with parameters η_1 and η_2. Thus, $T_i \sim \text{Weibull}(\eta_1, \eta_2)$.

The cdf of $Y_{irt}(T_i, t - T_i)$ is modelled by the following proportional odds model:

$$\text{logit}(Pr(Y_{irt} \leq k)) = \sum_{j=1}^{J} \alpha_{jkr}(t)x_{ij}(t) + \sum_{j'=1}^{J'} \beta_{j'kr}z_{ij't} + b_i, \tag{3}$$

for $k = 1, 2, \ldots, K$. We consider J time-varying covariates; the effect of the j-th covariate on the log odds at time t is $\alpha_{jkr}(t)$, which is also considered to be time-varying. We consider J' covariates with non-time-varying effects. Note that the individual-specific random effects,

b_i, capture the correlation among the measurements from the same individual at different waves.

For modelling the dependence among the zero-response probabilities and non-zero response probabilities, we assume that (η_i, b_i) jointly follow a bivariate normal density with mean vector=$\mathbf{0}$ and $\text{var}(\eta_i) = \sigma_\eta^2$, $\text{var}(b_i) = \sigma_b^2$, $\text{correlation}(\eta_i, b_i) = \rho$.

To model the time-varying coefficient $\alpha_{jkr}(t)$ for the j-th covariate, we use penalized splines (Ruppert, Wand and Carroll, 2003) and model $\alpha_{jkr}(t)$ as follows:

$$\alpha_{jkr}(t) = b_{jkr0t} + b_{jkr1t}t + b_{jkr2t}t^2 + ... + b_{jkrgt}t^{g_j} + \sum_{s=1}^{S_j} c_{jkrst}(t - \mathcal{T}_s)_+^{g_j}, \tag{4}$$

where $(x)_+^g = x^g I(x > 0)$ and $(\mathcal{T}_1 < \mathcal{T}_2 < ... < \mathcal{T}_{S_j})$ is a fixed set of knots.

Case 2

Here the t-th wave indicates the waves after the i-th individual visited hospital for the first time. For such waves, we model G as the proportional odds model. The cdf of $Y_{irt}(T)$ is modelled by the proportional odds model given in equation (3).

3 A Nonparametric Bayesian Model for the time-varying Coefficients

3.1 Dynamic Hierarchical Matrix Stick-Breaking Priors

We have coefficients from multiple groups of SAH and we assume in our setting (based on the data) that the individual may not necessarily be in the same group throughout the study since depending on health and other factors, their SAH will vary over time. Thus, the modelling framework should take account of this dynamic switching of SAH. We aim to model this time-varying structure where stronger dependence may be needed among data that are temporally close to each other. Therefore, it is desirable to develop a prior where one can efficiently estimate group specific parameters by borrowing information across groups in an automated manner, allowing commonality across subsets of the groups over time, and allowing equality of the parameters within a group over time. The matrix stick-breaking prior (MSBP) of Dunson et al. (2008) is somewhat similar but does not allow the dynamicity over

time. The MSBP of Dunson et al. (2008) has practical advantages, in that individuals can be very similar for most of their coefficients, while still allowing distinct local deviations. Here we consider a novel extension of the MSBP where information is shared across the self-reported health status groups (r) over the waves (t) for different numbers of hospital admission (k).

Define the vector of polynomial coefficients from equation (4), $\mathbf{b}_{jkrt} = [b_{jkr0t}, b_{jkr1t}, \ldots, b_{jkrg_jt}]^T$. Here we assume the vector $\mathbf{b}_{jkrt}$ is drawn from an unknown distribution $G^{(b)}_{krt}$, which itself is random, and a Dirichlet process (DP) is placed on the distribution of $G^{(b)}_{krt}$. Thus, we assume,

$$\mathbf{b}_{jkrt} \sim G^{(b)}_{krt} = \sum_{h=1}^{N_b} \pi^{(b)}_{krth} \delta_{\xi^{(b)}_{th}}; \quad j = 1, 2, \ldots, J; \quad k = 1, 2, \ldots, K; \quad r = 1, 2, \ldots, R; \quad t = 1, 2, \ldots, T,$$
$$\xi^{(b)}_{th} \sim G^{(b)}_t, \tag{5}$$

where δ_x denotes a point mass at x. Define $\Xi^b = \left(\xi^{(b)}_{th}\right)$ to be a matrix of order $T \times N_b$, the rows of which correspond to the parameters with the base distribution $G^{(b)}_t$ and the columns correspond to the "clusters". The stick-breaking weights $\pi^{(b)}_{krth}$ are defined as the following:

$$\pi^{(b)}_{krth} = V^{(b)}_{krth} \prod_{s'<h} (1 - V^{(b)}_{krts'}); \quad V^{(b)}_{krth} = U^{(b)}_{kh} Z^{(b)}_{rh} W^{(b)}_{th},$$
$$U^{(b)}_{kh} \overset{\text{iid}}{\sim} Beta(1, \delta^{(b)}_1); \quad Z^{(b)}_{rh} \overset{\text{iid}}{\sim} Beta(1, \delta^{(b)}_2); \quad W^{(b)}_{th} \overset{\text{iid}}{\sim} Beta(1, \delta^{(b)}_3). \tag{6}$$

The stick-breaking weights $\pi^{(b)}_{krth}$ control the dependence among the distributions $G^{(b)}_{krt}$. Note that we partition $\pi^{(b)}_{krth}$ into three components $U^{(b)}_{kh}$, $Z^{(b)}_{rh}$ and $W^{(b)}_{th}$ which allocate the vector of the polynomial coefficients from the k number of hospital visits and the r-th health status group and the t-th wave to the h-th cluster. We need to take $V^{(b)}_{krtN_b} = 1$, for all k, r, t; to make $G^{(b)}_{krt}$ a valid probability measure.

The dynamic nature of our setting requires stronger dependence among the distributions of parameters close to each other temporally. We propose the following hierarchical dynamic DP (Ren et al., 2010) for $G^{(b)}_t$:

$$G^{(b)}_t = (1 - \omega^{(b)}_{t-1}) G^{(b)}_{t-1} + \omega^{(b)}_{t-1} H^{(b)}_{t-1},$$
$$\omega^{(b)}_t | \alpha_\omega, \beta_\omega \sim Beta(\alpha_\omega, \beta_\omega); \quad \alpha_\omega \sim Gamma(\nu_1, \nu_2); \quad \beta_\omega \sim Gamma(\kappa_1, \kappa_2),$$
$$G^{(b)}_1 \sim DP(\alpha^{(b)}_{01}, G^{(b)}_0); \quad H^{(b)}_{t-1} \sim DP(\alpha^{(b)}_{0t}, G^{(b)}_0),$$
$$G^{(b)}_0 \sim DP(\gamma^{(b)}_0, H^{(b)}); \quad H^{(b)} \sim N(\boldsymbol{\mu}, \Sigma_0). \tag{7}$$

In the above prior, we note that $G_t^{(b)}$ is identical to $G_{t-1}^{(b)}$ with probability $1-\omega_{t-1}^{(b)}$ and with probability $\omega_{t-1}^{(b)}$ it is identical to an innovation distribution, $H_{t-1}^{(b)}$. A Dirichlet Process Prior (DPP) is considered for the innovation distribution. For the base distribution $G_0^{(b)}$ we further consider a DPP similar to the hierarchical Dirichlet process (HDP) proposed by Teh et al. (2006). The key point to note is that the above prior not only encourages "information exchange" among the temporally proximate parameters but also for the temporally distant parameters.

Formulations 5, 6 and 7 complete our proposed structure for the dynamic MSBP.

For the set of spline coefficients $\mathbf{c}_{jkrt} = [c_{jkr1t}, c_{jkr2t}, \ldots, c_{jkrS_jt}]^T$ in (4), we similarly specify the following dynamic MSBP prior:

$$
\begin{aligned}
&\mathbf{c}_{jkrt} \sim G_{krt}^{(c)};\ \ j=1,2,\ldots,J;\ \ k=1,2,\ldots,K;\ \ t=1,2,\ldots,T,\\
&G_{krt}^{(c)} = \sum_{h=1}^{N_c} \pi_{krth}^{(c)} \delta_{\xi_{th}^{(c)}};\ \ \pi_{krth}^{(c)} = V_{krth}^{(c)} \prod_{s'<h} (1-V_{krts'}^{(c)});\ \ V_{krth}^{(c)} = U_{kh}^{(c)} Z_{rh}^{(c)} W_{th}^{(c)},\\
&U_{kh}^{(c)} \overset{\text{iid}}{\sim} Beta(1,\delta_1^{(c)});\ \ Z_{rh}^{(c)} \overset{\text{iid}}{\sim} Beta(1,\delta_2^{(c)})\ \ W_{th}^{(c)} \overset{\text{iid}}{\sim} Beta(1,\delta_3^{(c)}),\\
&\xi_{th}^{(c)} \sim G_t^{(c)};\ \ G_t^{(c)} = (1-\omega_{t-1}^{(c)}) G_{t-1}^{(c)} + \omega_{t-1}^{(c)} H_{t-1}^{(c)},\\
&G_1^{(c)} \sim DP(\alpha_{01}^{(c)}, G_0^{(c)});\ \ H_{t-1}^{(c)} \sim DP(\alpha_{0t}^{(c)}, G_0^{(c)}),\\
&G_0^{(c)} \sim DP(\gamma_0^{(c)}, H^{(c)});\ \ H^{(c)} \sim N(\boldsymbol{\mu}_0, \lambda^{-1} I);\ \ \lambda \sim Gamma(\alpha^*, \beta^*).
\end{aligned} \tag{8}
$$

We note that the above prior is very similar to the priors proposed in (5-7) except the prior for $H^{(c)}$. The spline coefficients essentially measure the roughness at the respective knots and one can make the function smoother by shrinking the roughness towards zero. In a Bayesian framework, this is achieved by considering a multivariate normal prior with the covariance matrix $\lambda^{-1}I$ for the base distribution $H^{(c)}$ and then placing a gamma prior on the penalty parameter λ, (Das and Daniels, 2014).

3.2 Properties

Note that the features of our proposed DH-MSBP priors for $\mathbf{b}_{jkrt}$ and $\mathbf{c}_{jkrt}$, for $j=1,\ldots,J$, have the same form. Here we focus on the properties of the prior for $\mathbf{b}_{jkrt}$ only. Details of the other properties of the MSBP and dynamic DP can be found in Dunson et al. (2008) and Ren et al. (2010).

We note that the dynamic structure of $G_t^{(b)}$ can be expressed as: $G_t^{(b)} = (1-\omega_{t-1}^{(b)}) G_{t-1}^{(b)} +$

$\omega_{t-1}^{(b)} H_{t-1}^{(b)} = \tilde{\omega}_{t1}^{(b)} G_1^{(b)} + \tilde{\omega}_{t2}^{(b)} H_1^{(b)} + \ldots + \tilde{\omega}_{tt}^{(b)} H_{t-1}^{(b)}$, where $\tilde{\omega}_{tl}^{(b)} = \omega_{l-1}^{(b)} \prod_{m=1}^{t-1} (1 - \omega_m^{(b)})$, for $l = 1, 2, \ldots, t$.

Proposition 1

Consider sets A in the Borel field of the real line. As $N_b \to \infty$, we have

$E\{G_{krt}^{(b)}(A)\} = \sum_{l=1}^{t} \tilde{\omega}_{tl}^{(b)} H^{(b)}(A)$,

$V\{G_{krt}^{(b)}(A)\} = \sum_{l=1}^{t} \frac{\tilde{\omega}_{tl}^{(b)2}}{1+\alpha_{0l}^{(b)}} \left(\frac{\alpha_{0l}^{(b)} + \gamma_0^{(b)} + 1}{1+\gamma_0^{(b)}} \right) H^{(b)}(A)\{1 - H^{(b)}(A)\} \left[1 - \frac{4}{(\delta_1^{(b)}+2)(\delta_2^{(b)}+2)(\delta_3^{(b)}+2)-4} \right]$

$+ \frac{4}{(\delta_1^{(b)}+2)(\delta_2^{(b)}+2)(\delta_3^{(b)}+2)-4} \sum_{l=1}^{t} \tilde{\omega}_{tl}^{(b)} H^{(b)}(A) \left[1 - \sum_{l=1}^{t} \tilde{\omega}_{tl}^{(b)} H^{(b)}(A) \right]$.

Proof: As $N_b \to \infty$ we have,

$E\{G_{krt}^{(b)}(A)\} = EE\{G_{krt}^{(b)}(A) | G_t^{(b)}\} = E\{\sum_{h=1}^{\infty} \pi_{krth}^{(b)} \delta_{\xi_{th}^{(b)}}(A) | G_t^{(b)}\}$

$= E\{G_t^{(b)}(A)\} = \tilde{\omega}_{t1}^{(b)} H^{(b)}(A) + \tilde{\omega}_{t2}^{(b)} H^{(b)}(A) + \ldots + \tilde{\omega}_{tt}^{(b)} H^{(b)}(A) = \sum_{l=1}^{t} \tilde{\omega}_{tl}^{(b)} H^{(b)}(A)$.

From Dunson et al. (2008), it follows that

$V\{G_{krt}^{(b)}(A)\} = \left[1 - \frac{4}{(\delta_1^{(b)}+2)(\delta_2^{(b)}+2)(\delta_3^{(b)}+2)-4} \right] \left[V\{G_t^{(b)}(A)\} + \left(E\{G_t^{(b)}(A)\} \right)^2 \right]$

$+ \frac{4}{(\delta_1^{(b)}+2)(\delta_2^{(b)}+2)(\delta_3^{(b)}+2)-4} E\{G_t^{(b)}(A)\} - \left[E\{G_t^{(b)}(A)\} \right]^2$.

Since $G_t^{(b)}$ has a dynamic structure, from Ren et al. (2010), we have, $V\{G_t^{(b)}\} = \sum_{l=1}^{t} \frac{\tilde{\omega}_{tl}^{(b)2}}{1+\alpha_{0l}^{(b)}} \left(\frac{\alpha_{0l}^{(b)} + \gamma_0^{(b)} + 1}{1+\gamma_0^{(b)}} \right) H^{(b)}(A)\{1 - H^{(b)}(A)\}$. We also use the fact that $E\{G_t^{(b)}(A)\} = \sum_{l=1}^{t} \tilde{\omega}_{tl}^{(b)} H^{(b)}(A)$. The expression for $V\{G_{krt}^{(b)}(A)\}$ then follows immediately.

Proposition 2

For the proposed DH-MSBP prior:

(a) the probability that for fixed time t, and fixed value k, the health status groups r and r' will have identical coefficients for the polynomial part of (4) is

$Pr(\mathbf{b}_{jkrt} = \mathbf{b}_{jkr't}) = \frac{2}{(\delta_2^{(b)}+1)(\delta_1^{(b)}+2)(\delta_3^{(b)}+2)-2}$;

(b) the probability that the health status group r will have identical coefficients for the polynomial part of (4) at two different times t and t', for a fixed value k is:
$Pr(\mathbf{b}_{jkrt} = \mathbf{b}_{jkrt'}) = \frac{2}{(\delta_3^{(b)}+1)(\delta_1^{(b)}+2)(\delta_2^{(b)}+2)-2}$;

(c) the probability that the health status group r will have identical coefficients at fixed time t for the polynomial part of (4) at two different values k and k' is
$Pr(\mathbf{b}_{jkrt} = \mathbf{b}_{jk'rt}) = \frac{2}{(\delta_1^{(b)}+1)(\delta_2^{(b)}+2)(\delta_3^{(b)}+2)-2}$.

The proofs of the above results are similar to Dunson et al. (2008), and hence are omitted here for brevity.

Note that the prior clustering probabilities in (a), (b), and (c) range between 0 and 1 depending on $\delta_1^{(b)}$, $\delta_2^{(b)}$ and $\delta_3^{(b)}$. Clearly if $\delta_1^{(b)}, \delta_2^{(b)}, \delta_3^{(b)} \to 0$, then the clusters are not different and all the above prior probabilities converge to 1. However, for $\delta_1^{(b)}$ or $\delta_2^{(b)}$ or $\delta_2^{(b)} \to \infty$, none of the sets are clustered together and thus the probabilities in (a), (b) and (c) converge to 0.

3.3 Zero-Inflated LASSO Priors

Since we have a large number of covariates with time-invariant effects on our response variable, we incorporate a shrinkage prior approach. Our approach is similar to the hierarchical Bayes representation of the LASSO proposed by Park and Casella (2008). Recently, Das (2016) proposed a modified Bayesian LASSO with a zero-inflated structure which is relevant to our setting. Define $\boldsymbol{\beta}_{kr} = (\beta_{1kr}, \cdots, \beta_{J'kr})^T$. We add sparsity to the prior and consider the following hierarchical representation:

$$\begin{aligned}
&\beta_{j'kr}|\sigma^2, \tau_{j'}^2, B_{j'} \sim (1 - B_{j'})\delta_0 + B_{j'} N\left(0, \sigma^2\tau_{j'}^2\right), \forall j', k, r, \\
&B_{j'}|\pi_{j'} \sim Bernoulli(\pi_{j'}), \;\; \pi_{j'} \sim Beta(u, v), \\
&\tau_{j'}^2 \overset{\text{iid}}{\sim} \frac{\lambda^2}{2} \exp\left(-\frac{1}{2}\lambda\tau_{j'}^2\right), \;\; \lambda^2 \sim \pi(\lambda^2), \;\; \sigma^2 \sim \pi\left(\sigma^2\right).
\end{aligned}$$

Here δ_0 is a point mass probability measure at 0. Note that the parameters $\pi_{j'}$s are updated from the data and force the coefficients of the insignificant covariates to be exactly equal to 0 and simultaneously shrink the other coefficients towards 0 as LASSO does. The advantage of considering a zero-inflated normal prior for $\beta_{j'kr}$ is discussed in detail in Das (2016). Inverse Gamma priors are taken both for λ^2 and σ^2.

3.4 Joint Posterior and Computational Details

Note that the likelihood function of the longitudinal responses Y_{irt} can be written as follows:

$$L = \left[\prod_{(i,r,t):Y_{irt}=0}(1-\pi_{irt})\right] \times \left[\prod_{(i,r,t):Y_{irt}>0}\pi_{irt}G(Y_{irt}|Y_{irt}>0)\right] \tag{9}$$
$$= \int\int \left[\left(\prod_{(i,r,t):Y_{irt}=0}(1-\Phi(\mathbf{x}_i^T\boldsymbol{\delta}+\eta_i)\right) \times \left(\prod_{(i,r,t):Y_{irt}>0}\Phi(\mathbf{x}_i^T\boldsymbol{\delta}+\eta_i)G(Y_{irt}|b_i)\right)\right] f(\eta_i,b_i)d\eta_i db_i,$$

where $G(Y_{irt}|b_i)$ is the density described in Section 2.2 and $f(\eta_i, b_i)$ is the joint density of the random effects. The log likelihood function can be expressed accordingly. The joint posterior distribution is obtained by multiplying the corresponding prior components to L. Note that for δ (Section 2.1), we consider a multivariate normal prior with mean vector=0 and covariance matrix=$\sigma_\delta^2 I$. Priors for all the other model parameters are discussed in Sections 3.1 and 3.2.

We note that the truncation of the MSBP to finite N_b (and also for N_c) is done using the approximation in Ishwaran and James (2002), such that the truncation value (N_b) makes the expected approximation error smaller than 0.01. For $\mathbf{b}_{jkrt}$, the expected approximation error $= \left[1 - \frac{1}{(1+\delta_1^{(b)})(1+\delta_2^{(b)})(1+\delta_3^{(b)})}\right]^{N_b-1}$, which requires knowing $\delta_1^{(b)}$, $\delta_2^{(b)}$, and $\delta_3^{(b)}$. Following Dunson et al. (2008), we specify independent gamma(1,1) priors for $\delta_1^{(b)}$, $\delta_2^{(b)}$ and $\delta_3^{(b)}$ and we run the MCMC algorithm for about 15% of its total length and use the posterior means of $\delta_1^{(b)}$, $\delta_2^{(b)}$ and $\delta_3^{(b)}$ to determine if the expected approximation error is below 0.01. The same approach is used to choose N_c. Our computations are similar to Das and Daniels (2014), Chatterjee et al. (2016).

For the dynamic hierarchical structure, note that the infinite mixture representations of Dirichlet Process Priors $G_0^{(b)}$, $G_1^{(b)}$, $H_1^{(b)}, \ldots, H_{T-1}^{(b)}$ will be truncated to a level, M say, using the approximation in Ishwaran and James (2002). We consider the same threshold value (0.01) in these cases as well.

4 Simulation Study

We investigate the operating characteristics of the dynamic hurdle model and the dynamic hierarchical MSBP prior through simulation studies. We consider a zero-inflated longitudinal count response, two covariates with time-varying effects and 10 covariates with time-invariant effects on the response.

We simulate data for 100 subjects at 10 evenly spaced time points ($t = 1, 2, \ldots, 10$). Consider 4 groups ($r = 1, 2, 3, 4$) and each subject belongs to one of these 4 groups at each time point. In particular, we consider 30 subjects who are in group 1 until $t = 6$, and then belong to group 2; 10 subjects who change from group 1 to group 4 at $t = 5$. Consider 20 subjects who change from group 2 to group 3 at $t = 7$; 10 subjects who change from group 2 to group 4 at $t = 4$; 10 subjects change from group 3 to group 1 at $t = 5$; 10 subjects change from group 3 to group 4 at $t = 6$, and 10 subjects move from group 4 to group 1 at $t = 8$. We simulate the response Y_{irt} from the model given in equation (1), where the non-zero mixing proportions, π_{irt}s, are simulated from the probit model given in Section 2.2 (the explicit form of which is given in the web-appendix). If for the t-th time point we observe the first non-zero response, then Y_{irt} is generated from the model given in equation (2) with $T_i \sim Weibull(1, 5)$ and the cdf of $Y_{irt}(T_i, t - T_i)$ is obtained from equation (3) with $k = 0, \ldots, 5$, $J = 2$ and $J' = 10$. On the other hand, if the t-th time point is either the time before or after the time of the first occurrence of the event, we then simulate Y_{irt} from the model given in equation (3). We consider the following spline function for $\alpha_{jkr}(t)$:

$$\alpha_{jkr}(t) = b_{jkr0t} + b_{jkr1t}t + b_{jkr2t}t^2 + \sum_{s=1}^{2} c_{jkrst}(t - \mathcal{T}_s)_+^2, \tag{10}$$

with the knots $\mathcal{T}_1 = 3$ and $\mathcal{T}_2 = 7$. We consider both the discrete and continuous covariates with time-invariant effects on the response. These covariates are simulated from Normal and Bernoulli distributions. The details of the parameter values for the simulation are given in the web-appendix. In general, we consider the different parameter values across the groups and times for different k to be somewhat similar but not exactly the same.

We fit three different proportional odds models to the simulated data: (i) a model with a distinct parameter set for each group (the group-specific model); (ii) a model with the same parameter set for all the groups (the common model); and (iii) a model with the DH-MSBP prior in the parameter set. For each of these models, we consider two different specifications for modelling $G(Y_{irt}|Y_{irt} > 0)$: (i) a dynamic hurdle as discussed in Section 2.2; and (ii) a

non-dynamic hurdle, where $G(Y_{irt}|Y_{irt} > 0)$ is modelled for all the subjects at all the time points using the proportional odds model given in equation (3).

We simulate 100 datasets and for each dataset we run the MCMC algorithm for 65,000 iterations, discard the first 5,000 ("burn-in") and thin the remaining 60,000 by keeping every 10-th iteration. The model parameters are estimated using posterior means. We specify $N_b = N_c = 20$, which gives a truncation approximation error less than 0.015.

For the model selection, we use the conditional predictive ordinate (CPO) (Gelfand et al. 1992) defined as $CPO_i = P(Y_i|Y_{-i}) = E_{\boldsymbol{\theta}}\left[P(Y_i|\boldsymbol{\theta}, Y_{-i})\right]$, where Y_{-i} denotes the data excluding Y_i and $\boldsymbol{\theta}$ denotes the set of all model parameters. For the i-th subject, the CPO can be estimated based on the MCMC samples as the following:
$\widehat{CPO}_i = \left[\frac{1}{M}\sum_{m=1}^{M}\frac{1}{P(Y_i|\boldsymbol{\theta}^{(m)})}\right]^{-1}$, where $\boldsymbol{\theta}^{(m)}$ denotes the parameter estimates at the m-th iteration of MCMC. We compute the log pseudo-marginal likelihood (LPML)=$\sum_{i=1}^{n} log\widehat{CPO}_i$, where a greater value of the LPML indicates a better fit.

We compute the LPML across all replications and Table 1 shows the average values for the different models. We note that the DH-MSBP prior with a dynamic hurdle gives the largest LPML value. However, the LPML values for the DH-MSBP prior with the non-dynamic hurdle and the group-specific model with the dynamic hurdle are quite close to this largest value.

Table 2 shows the average estimated bias, average width of the 95% credible intervals (CIs) and the estimated coverage probabilities of a randomly selected subset of the model parameters for the three models with the larger LPML values in Table 2. We note that the proposed DH-MSBP prior with the dynamic hurdle results in the smallest bias and shortest CI with a comparable coverage probability. Thus, our simulation studies demonstrate the effectiveness of the DH-MSBP and the dynamic hurdle model compared to their competitors.

5 Data Analysis

5.1 The HRS Data

We exploit data from the University of Michigan's Health and Retirement Study (HRS), which is a longitudinal survey of Americans over the age of 50, with a follow-up frequency of every two years. The HRS provides multi-disciplinary data to understand the challenges of aging. In this paper, we use data for 10 waves from the 1931-1941 cohort. Baseline observations for this cohort began in 1992 when individuals were aged between 52 and 62, and, hence, were nearing retirement. The HRS is maintained by RAND's Center for the Study of Aging. Our sampling is based on 2630 individuals, who are observed in all 10 waves. For our outcome measure, we use the number of hospital visits made since the previous interview. This variable is derived from the responses to the following question: How many different times were you a patient in a hospital overnight in the last 24 months? As expected, a large proportion of zero observations are observed in the data. Specifically, from waves 1 to 4, 90-94% are zero observations, which falls to 80-86% for waves 5 to 8, and to 75-80% for waves 9-10. In accordance with intuition, the proportion of zero observations declines as the individuals in our sample age. Across all waves, for the non-zero responses, the minimum value of the response variable is 1 and the maximum is 15, with an average of 1.4 visits. With respect to grouping individuals by health status, we have 5 categories of SAH (r); poor (1), fair (2), good (3), very good (4) and excellent (5), which can vary over time.

The HRS provides a rich set of covariates for our analysis. Specifically, we have four covariates with time-varying effects on the response variable, i.e. the number of hospital visits, namely: (i) the body mass index (BMI); (ii) the total value of assets, (iii) the total value of debt and (iv) total household income.

We include BMI in our set of covariates with time varying effects in recognition of the well-documented relationship between obesity and poor health conditions. According to the World Obesity Federation (www.worldobesity.org), the epidemic of obesity is now recognized as one of the most important public health problems facing the world today. Obesity, a condition of excessive body weight in the form of fat, is causally linked to a large number of debilitating and life-threatening conditions. The most commonly used measure to assess whether an individual is obese is BMI: the ratio of the individual's weight to the square of

height. Hence, we include BMI in our modelling framework.

The inclusion of the three financial covariates reflects the large existing literature exploring the relationship between a range of household financial outcomes and health. For example, Adams et al. (2003), Michaud and van Soest (2008) and Hurd and Kapteyn (2003) generally find a positive association between better health and household wealth. Total assets are defined as the summation of the value of: individual retirement accounts; stocks; bonds; checking and saving accounts; certificates of deposit and saving bonds; other saving accounts; the primary residence; transport; net value of any business; and other assets.

With respect to the other side of the household balance sheet, there is a growing literature exploring the relationship between health and debt. For example, Drentea and Lavrakas (2000) find that both credit card debt and stress regarding debt are inversely associated with good health and Brown et al. (2005) find that unsecured debt is inversely related to psychological wellbeing. More recently, Keese and Schmitz (2014) report that a variety of debt measures are strongly correlated with satisfaction with health and mental health. Our measure of total debt includes: all mortgages/land contracts; other home loans; and other debt including credit cards.

With respect to income, a number of studies have explored the relationship between income and health. For example, for the UK, Contoyannis and Rice (2001) report an inverse relationship which poor health and wages for men. For the US, Pelkowski and Berger (2004) use data from the US HRS and find that permanent health problems have a significant effect on labour market participation, wages and hours for both men and women. All monetary variables are entered in natural logarithm form and are expressed in 2010 prices.

For the covariates with time-invariant effects, we have the following 15 covariates. Note that these covariates are all binary variables. We control for being female, education as measured by having General Education Diploma (GED) level education or higher, whether the individual consumes alcohol and whether the individual smokes. In order to capture the effects of long-term health, we include 8 covariates relating to chronic health conditions. Specifically, we have 8 controls for whether the individual has ever had: (i) high blood pressure or hyper-tension; (ii) diabetes or high blood sugar; (iii) cancer or a malignant tumour of any kind; (iv) chronic lung disease except asthma such as chronic bronchitis or emphysema; (v) heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems; (vi) stroke or transient ischemic attack; (vii) emotional, nervous, or psychiatric problems; and (viii) arthritis or rheumatism. Finally, we include 3 controls for

different types of health insurance: (i) having health insurance related to employment; (ii) government insurance; and (iii) other private health insurance.

Some important features of the data are summarized in Figures 1 and 2 (as well as Figures S1 and S2 in the web appendix). In Figure 1, we show the counts of individuals with different numbers of hospital admissions across the 10 waves through a series of bar diagrams. We note that the counts decrease over the waves for zero admission and increase for the number of admissions 1, 2 and 3. For counts 4 and 5, there is no specific pattern evident although the counts are at the maximum for waves 9 and 10. In Figure 2, we show the distribution of individuals across the number of hospital admissions for each SAH category. We note that the distributions are neither the same nor completely different, but rather are similar in pattern for certain waves. This motivates the use of the dynamic MSBP construction as discussed earlier. In Figures S.1 and S.2 in the web appendix, we show the heat map of the average number of hospital admissions across the different SAH categories and waves; and total counts (from all SAH categories) across the number of hospital admissions for different waves.

5.2 Results

In Table 3, we show the distribution of the average number of hospital admissions for the individuals who made their first hospital visit at wave 1, 2, 3, etc. We observe clear differences in the distributions, reflecting the fact that the rates of hospital admission depend on the time of the first observed hospital visit. Under such a scenario, a dynamic hurdle model is more appropriate as demonstrated in Section 4.

In Table 4 we present the effects of all 19 covariates on the probability of non- zero hospital visits. The estimates and the 95% credible intervals are given for the corresponding model parameters. Among the chronic health conditions, cancer and strokes are found to have a significant effect on the probability of zero hospital visits, which accords with intuition since such serious conditions are associated with frequent hospitalization. On the other hand, conditions such as high blood pressure and diabetes, which are often treated by medication rather than hospitalization, are found to exert statistically insignificant effects. This is also the case for emotional, nervous, or psychiatric problems, which also accords with expectations in that only a small number of very severe cases are likely to be associated with hospitalization. It is reassuring, therefore, that the parameter estimates associated

with the health covariates largely accord with our expectations. Health insurance related to employment and other private health insurance are both found to have significant effects, which may reflect better quality health care associated with such insurance relative to government health insurance, leading to more frequent hospital visits for those holding such health insurance. Among the financial covariates, total assets and total household income exert significant effects on the probability of zero hospital visits. This is also not surprising since financially affluent individuals are more likely to be able to access health care thereby potentially leading to better health.

Next, in order to highlight the flexibility of our approach, we summarize the results for the covariates with time-invariant effects from equation (3) for different SAH categories and for different values of k. In Table 5, we present the corresponding coefficient estimates, and 95% credible intervals (based on the MCMC samples) for the people belonging to the "good" SAH group (r=3) with 1 or less hospital admissions. With respect to the chronic health conditions, the findings as above accord with intuition, with the three particularly severe conditions, namely, cancer, heart problems and strokes, found to be statistically significantly related to the response variable, with large parameter estimates. The other controls are characterised by statistically insignificant parameter estimates. The large parameter estimates for education and government insurance are particularly noticeable. Education is clearly related to employment opportunities as well as healthy life styles, (Cutler and Glaeser, 2005), whereas the large parameter estimate for government health insurance may reflect the importance of Medicare and Medicaid in the US. However, these parameter estimates are statistically insignificant in our case, since the 95% Credible intervals contain zero.

We then summarize the results for the two extreme SAH groups, i.e. the "poor" SAH group with the number of hospital admissions 4 or less; and the "excellent" SAH group with 1 or less hospital admissions. Table 6 shows the results for the "poor" group with k equalling 4 or less. Note that here we have some additional important covariates namely, blood pressure, lung problems, arthritis, smoking, alcohol consumption and government insurance. Coefficients for the covariates gender, education and psychological problems are exactly equal to 0, due to the zero-inflated LASSO prior. Table 7 summarizes the results for the "excellent" SAH group with k=1. In addition to the chronic health conditions of cancer, heart problems and stroke, smoking and education are found to be important predictors for this case.

We now turn to the plots (Figures 3-6) for the time varying coefficients for the four

covariates as mentioned earlier, namely: BMI, total assets, total debt and income. We have 5 categories of SAH (r); poor, fair, good, very good and excellent. Across all waves, SAH is distributed as follows: excellent (21.66%); very good (38.78%); good (28.02%); moderate (9.00%); and poor (2.54%). In accordance with expectations, from waves 1 to 10, the proportion in the excellent SAH category falls as individuals age. For example, in wave 1, 32.40% are observed in the excellent SAH category, which falls to 21.53% by wave 5 and 12.85% by wave 10. For each of the four covariates with time varying effects on the number of hospital admissions, we present plots for these five categories of SAH for k=2, 4 and 6 for illustration purposes.

Focusing initially on BMI, which is regarded as one of the leading indicators of health, it is apparent from Figure 3 that, for k=2, the effect of BMI for the fair SAH category increases dramatically over waves 1 to 10. In contrast, the effect of BMI for the poor category is characterised by a less pronounced increase from wave 4 onwards. Interestingly, the effects of BMI fall over the observed waves for the excellent and good SAH categories, yet demonstrate a steady increase over time for the very good health category. Furthermore, the pattern of effects is clearly different for the case when $k = 4$, where less variation in the pattern of the effects is apparent. For all 5 categories, we generally observe an increase in the effects over time, with the increase for the fair health category starting from wave 6 onwards and for very good health, from wave 7 onwards. It is interesting to see that when $k = 6$, i.e. for a very large number of hospital visits, distinct differences in the magnitudes of the effects are apparent across the five categories, with very good health generally characterised by the largest effect and poor health by the smallest. However, over the observed waves, the size of the effects within each SAH category are relatively stable, which clearly contrasts with the case when $k = 2$, where we observe considerable variation over time in the effects of BMI on the outcome variable.

Turning to total financial assets, it is apparent that, when k=2 and k=4, for all five SAH categories, in general, an upwards trend is observed moving from wave 1 to 10, with the size of the effects being most pronounced for the poor and fair SAH categories. The upwards trend is still apparent, yet less pronounced, for the other SAH categories. As in the case of BMI, less variation in the effects of assets over the 10 waves is apparent when k =2 relative to when $k = 4$, i.e. for a larger number of hospital visits. As expected, the poor health category is characterised by the largest size of effect for $k = 6$. Focusing on excellent SAH, it is apparent that across the values of k, a moderate upwards trend can be seen across the waves, with a more pronounced upwards trend discernible from wave 5 onward and the

effect of financial assets on the number of hospital visits within this SAH category being relatively stable over time. This contrasts, in particular, with the lower categories of SAH, specifically, poor and fair health, where the effect of financial assets appears to exhibit much more variation over the waves, especially in the case where $k = 2$.

Interestingly in the case of debt, the observed patterns are generally the opposite to those observed for financial assets. For example, when $k = 2$, a downwards trend is generally apparent for the effect of debt on hospital visits across the five SAH categories, with a particularly pronounced downwards trend apparent for the excellent and good SAH categories. Thus, over waves 1 to 10, the findings suggest that the effect of debt on hospital visits falls for $k = 2$, and $k = 4$. This may reflect the fact that over the life cycle debts generally fall as individuals age: in an early seminal contribution, for example, Ando and Modigliani (1963) hypothesized that individuals may be more comfortable with debt holding when they are young and their income is low, as they expect future income to be much higher, and to be able to pay off the debt at a later stage. Thus, in the context of an aging population, the effect of debt on the number of hospital visits is generally seen to decline over time. Interestingly, in the case of the lowest two SAH categories, poor and fair SAH, a more stable effect is observed across all three cases, suggesting that the effect of debt does not fall for these categories. This may reflect borrowing associated with being in such low SAH states. It is interesting to note that in the case where $k = 6$, all SAH categories, with the exception of very good health, are characterised by an increase in the effect of debt on the number of hospital visits at wave 9.

Finally, with respect to income, the observed patterns tend to follow those associated with financial assets, with an upwards trend generally observed for all five SAH categories across the three values of k. When k =2, the upwards trend is particularly pronounced for the fair SAH category, with the very good health category characterised by the most stable effect over time, which is also the case for the other values of k. It is interesting to note that, although there are some changes in the relative size of the effect across the three values of k, the effect of income on hospital visits seems to be relatively stable over time. This may reflect the possibility that as individuals age, there are less opportunities for increasing income, with many individuals relying on pension income. In contrast, with financial assets, funds can be raised by dis-saving or selling assets if required which may lead to more variation in the effects of financial asset holding over time as compared to the effects of income on the number of hospital visits.

6 Discussion

In clinical trials and biomedical studies, dynamic group structure is not uncommon. For example, with respect to blood pressure an individual might belong to three different groups (high, medium and low) dynamically. Traditional approaches of analyzing clustered data do not work in such situations since the group sizes (and compositions) vary over time. Advanced nonparametric Bayesian methods allow information sharing across groups but the group structure is assumed to be static. We consider a dynamic group structure and develop models for handling the complexities with a zero-inflated longitudinal count measurement.

Our approach, however, suffers from one limitation. We assume no missingness in our data. However, in reality, it is not uncommon to have missing responses and some missing covariates. If the missingness is ignorable (e.g. missing at random), then one can simply add a data-augmentation technique to our method. However, for non-ignorable missingness our method has to be revised substantially following Daniels and Hogan (2008). Also one might have zero-inflated endogenous covariates with time-varying or time-invariant effects on the response. A Bayesian approach of two-stage regression could be developed for handling such cases. We leave these issues for future research.

References

1. Adams, P., Hurd, M. D., McFadden, D., Merrill, A., and T. Ribeiro, 2003, Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status. Journal of Econometrics 112(1), 3–56.

2. Ando, A., and Modigliani, F, 1963, The“ life cycle” hypothesis of saving: Aggregate implications and tests. The American Economic Review 53(1), 55–84.

3. Antoniak, C.E, 1974, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2, 1152–1174.

4. Atella, V., Deb, P, 2008, Are primary care physicians, public and private sector specialists substitutes or complements? Evidence from a simultaneous equations model for count data. Journal of Health Economics 27, 770–785.

5. Baetschmann, G. and Winkelman, R, 2016, A dynamic hurdle model for zero-inflated count data. Communications in Statistics-Theory and Methods (in press).

6. Banerjee, R., Ziegenfuss, J. Y., and Shah, N. D, 2010, Impact of discontinuity in health insurance on resource utilization. BMC Health Services Research 10(1), 1–10.

7. Brown, S., Taylor, K., and Price, S. W, 2005, Debt and distress: Evaluating the psychological cost of credit. Journal of Economic Psychology 26(5), 642–663.

8. Blei, D. and Jordan, M, 2006, Variational inference for Dirichlet process mixtures. Bayesian Analysis 1, 121–144.

9. Chatterjee, A., Venkateswaran, P. and Das, K, 2016, Simultaneous state estimation for clustered based wireless sensor networks. IEEE Transactions on Wireless Communications 15(12), 7985–7995

10. Contoyannis, P. and Rice, N, 2001, The impact of health on wages: Evidence from the British Household Panel Survey. Empirical Economics 26, 599–622.

11. Cutler, D. M. and Glaeser, E, 2005, What explains differences in smoking, drinking, and other health related behaviors? American Economic Review Papers and Proceedings 95(2), 238–242.

12. Das, K, 2016, A semiparametric Bayesian approach for joint modeling of longitudinal trait and event time. Journal of Applied Statistics 43(15), 2850–2865.

13. Das, K. and Daniels, M.J, 2014, A semiparametric approach to simultaneous covariance estimation for bivariate sparse longitudinal data. Biometrics 70, 33–43.

14. Deb, P. and Trivedi, P.K, 1997, Demand for medical care by the elderly: A finite mixture approach. Journal of Applied Econometrics 12, 313–336.

15. Drentea, P. and Lavrakas, P. J, 2000, Over the limit: the association among health, race and debt. Social Science and Medicine 50(4), 517-529.

16. Dunson, D., Herring, A. and Engel, S, 2008, Bayesian selection and clustering of polymorphisms in functionally related genes. Journal of the American Statistical Association 103, 534–546.

17. Dunson, D.B., Xue, Y. and Carin, L, 2008, The matrix stick-breaking process: Flexible

Bayes meta-analysis. Journal of American Statistical Association 103(481), 317–327.

18. Ferguson, T.S, 1973, A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.

19. Gelfand, A.E., Dey, D. and Chang, H, 1992, Model determination using predictive distributions with implementation via sampling based methods (with discussion). Bayesian Statistics 4, Eds: J. Bernardo et al. Oxford University Press, 147–167.

20. Hurd, M. and Kapteyn, A, 2003, Health, wealth, and the role of institutions. Journal of Human Resources 38(2), 386–415.

21. Idler, E. L. and Benyamini, Y, 1997, Self-rated health and mortality: a review of twenty-seven community studies. Journal of Health and Social Behaviour 38(1), 21–37.

22. Ishwaran, H. and James, L.F, 2002, Approximate Dirichlet process computing in finite normal mixtures. Journal of Computational and Graphical Statistics 11, 508-532.

23. Keese, M. and Schmitz, H, 2014, Broke, ill, and obese: Is there an effect of household debt on health? Review of Income and Wealth 60(3), 525–541.

24. Michaud, P. C. and Van Soest, A, 2008, Health and wealth of elderly couples: Causality tests using dynamic panel data models. Journal of Health Economics 27(5), 1312–1325.

25. Pelkowski, J. M. and Berger, M. C, 2004, The impact of health on employment, wages, and hours worked over the life cycle. Quarterly Review of Economics and Finance 44, 102–121.

26. Park, T. and Casella, G, 2008, The Bayesian Lasso. Journal of American Statistical Association 103, 681–686.

27. Ren, L., Dunson, D., Lindroth, S. and Carin, L, 2010, Dynamic nonparametric Bayesian models for analysis of music. Journal of the American Statistical Association 105, 458–472.

28. Rodriguez, A. and Dunson, D, 2014, Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies. The Annals of Applied Statistics 8, 1416–1442.

29. Ruppert, D., Wand, M.P. and Carroll, R.J, 2003, Semiparametric Regression. Cambridge

University Press, New York.

30. Sethuraman, J, 1994, A constructive definition of Dirichlet priors. Statistica Sinica 4, 639-650.

31. Teh, Y.W., Jordan, M., Beal, M. and Blei, D, 2006, Hierarchical Dirichlet processes. Journal of the American Statistical Association 101, 1566–1581.

32. United Nations Department of Economic and Social Affairs: Population Division, 2015, "World Population Aging 2015". United Nations, New York.

33. Westbury, L. D., Syddall, H. E., Simmonds, S. J., Cooper, C. and Aihie Sayer, A, 2016, Identification of risk factors for hospital admission using multiple-failure survival models: A toolkit for researchers. BMC Medical Research Methodology, 1–8.

34. Winkelmann, R, 2004, Health care reform and the number of doctor visits: An econometric analysis. Journal of Applied Econometrics 19, 455-472.

Table 1: LPML values from different model specifications in the simulation study.

Model	LPML value
Dynamic Hurdle model	
Common model	-539.8
Group-Specific model	**-397.2**
DH-MSBP	**-396.6**
Non-dynamic Hurdle model	
Common model	-639.5
Group-Specific model	-463.8
DH-MSBP	**-401.4**

Table 2: Bias, CI width, and CI coverage probability for some of the model parameters in the simulation study.

	Group specific dynamic hurdle		DH-MSBP non-dynamic hurdle		DH-MSBP dynamic hurdle	
Parameter	Bias	width of C.I. (Cov.Prob)	Bias	width of C.I. (Cov.Prob)	Bias	width of C.I. (Cov.Prob)
b_{12101}	0.38	0.87(0.96)	0.35	0.92(0.96)	0.11	0.26(0.94)
b_{23213}	0.33	0.69(0.95)	0.37	0.74(0.96)	0.09	0.24(0.95)
b_{15324}	0.29	0.58(0.95)	0.26	0.55(0.95)	0.10	0.31(0.94)
c_{12121}	0.37	0.73(0.95)	0.34	0.68(0.95)	0.08	0.25(0.95)
c_{23218}	0.35	0.84(0.96)	0.42	0.91(0.96)	0.11	0.28(0.94)
c_{25425}	0.41	0.93(0.96)	0.39	0.86(0.95)	0.13	0.23(0.95)
β_{131}	0.39	0.59(0.96)	0.33	0.64(0.96)	0.06	0.29(0.94)
β_{554}	0.28	0.63(0.96)	0.35	0.58(0.95)	0.08	0.25(0.95)
β_{823}	0.36	0.65(0.96)	0.34	0.67(0.96)	0.10	0.33(0.94)

Table 3: Distribution of the average number of hospital admissions depending on the time of the first hospital visits for the HRS data

	Waves									
Wave for the first visit	1	2	3	4	5	6	7	8	9	10
1	-	0.63	0.90	0.52	0.31	0.18	0.32	0.20	0.28	0.32
2	0	-	0.62	0.40	0.43	0.32	0.28	0.26	0.26	0.38
3	0	0	-	0.57	0.25	0.20	0.18	0.23	0.28	0.27
4	0	0	0	-	0.37	0.19	0.24	0.33	0.21	0.36
5	0	0	0	0	-	0.36	0.22	0.26	0.31	0.27
6	0	0	0	0	0	-	0.38	0.38	0.31	0.41
7	0	0	0	0	0	0	-	0.37	0.34	0.58
8	0	0	0	0	0	0	0	-	0.41	0.42
9	0	0	0	0	0	0	0	0	-	0.49

Table 4: Probit model estimates for all the covariates in the HRS data.

Covariate	Parameter Estimate	95% C.I.
Blood pressure	0.013	(-1.48,0.97)
Diabetes	0.052	(-1.68,2.51))
Cancer*	2.58	(1.43, 4.29)
Lung problem	1.02	(-2.19, 2.47)
Heart problem	0.16	(-1.31,2.51)
Stroke*	5.53	(2.33, 8.16)
Arthritis	0.0027	(-0.91,0.85)
Psychological problem	0.74	(-1.84,2.78)
Employment Insurance*	2.92	(0.49,4.56)
Gov. Insurance	1.62	(-2.20, 3.89)
Other Insurance*	2.56	(1.21,5.73)
Smoking	0.08	(-1.29,1.56)
Alcohol Consumption	1.04	(-2.68,3.77)
Gender	0.94	(-2.68,2.09)
Education level	0.46	(-1.49,2.36)
BMI	0.085	(-0.96,1.88)
Total assets*	2.51	(1.44,5.63)
Total debt	1.16	(-2.06,3.89)
Total household income*	2.75	(-2.41,4.55)

Table 5: LASSO estimates for the covariates with time-invariant effects on the response for the "good" SAH group with k=1, in the HRS data.

Covariate	Parameter Estimate	95% C.I.
Blood pressure	0.064	(-1.33,0.87)
Diabetes	0.041	(-2.02,1.76)
Cancer*	4.58	(2.65, 5.93)
Lung problem	1.16	(-0.29, 2.05)
Heart problem*	3.56	(1.32,5.68)
Stroke*	10.32	(7.96, 13.30)
Arthritis	0.29	(-1.31,2.04)
Psychological problem	0.02	(-2.11,1.19)
Employment Insurance	0.18	(-2.74,1.55)
Gov. Insurance	3.63	(-1.49,6.26)
Other Insurance	0.03	(-1.49,1.14)
Smoking	0.05	(-2.79,0.68)
Alcohol Consumption	0.007	(-0.99,1.06)
Gender	0.58	(-3.61,2.47)
Education level	1.12	(-2.44,2.83)

Table 6: LASSO estimates for the covariates with time-invariant effects on the response for the "poor" SAH group with k=4, in the HRS data.

Covariate	Parameter Estimate	95% C.I.
Blood pressure*	1.89	(0.84,4.02))
Diabetes	0.67	(-1.54,2.51)
Cancer*	3.88	(1.55, 6.01)
Lung problem*	1.53	(1.03, 2.24)
Heart problem*	2.66	(1.13,5.61)
Stroke*	7.19	(3.36, 9.14)
Arthritis*	1.27	(0.59,3.66)
Psychological problem	0	-
Employment Insurance	0.26	(-0.56,0.88)
Gov. Insurance*	2.58	(1.39,4.43)
Other Insurance	0.16	(-1.29,0.94)
Smoking*	1.19	(0.78,2.18)
Alcohol Consumption*	2.34	(1.39,4.11)
Gender	0	-
Education level	0	-

Table 7: LASSO estimates for the covariates with time-invariant effects on the response for the "excellent" SAH group with k=1, in the HRS data.

Covariate	Parameter Estimate	95% C.I.
Blood pressure	0.53	(-2.15,1.73)
Diabetes	0.72	(-1.95,1.23))
Cancer*	5.91	(3.50, 6.99)
Lung problem	1.13	(-2.05, 2.86)
Heart problem*	2.54	(1.12,4.65)
Stroke*	8.11	(5.84, 10.71)
Arthritis	0	-
Psychological problem	0	-
Employment Insurance	0.31	(-0.94,0.75)
Gov. Insurance	0	-
Other Insurance	0.83	(-1.04,1.88)
Smoking*	1.59	(0.96,3.18)
Alcohol Consumption	0.06	(-0.39,0.51)
Gender	0	-
Education level*	1.27	(0.94,2.86)

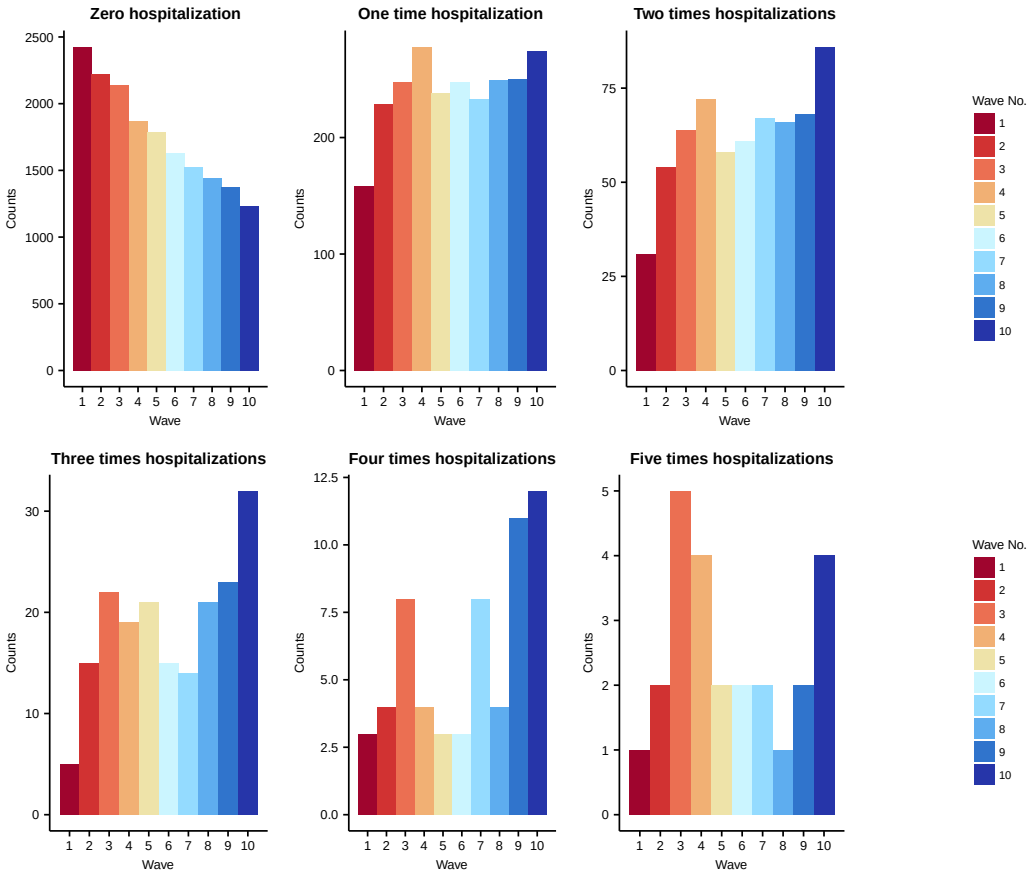

Figure 1: Bar charts comparing the counts of individuals with different numbers of hospitalizations across 10 waves.

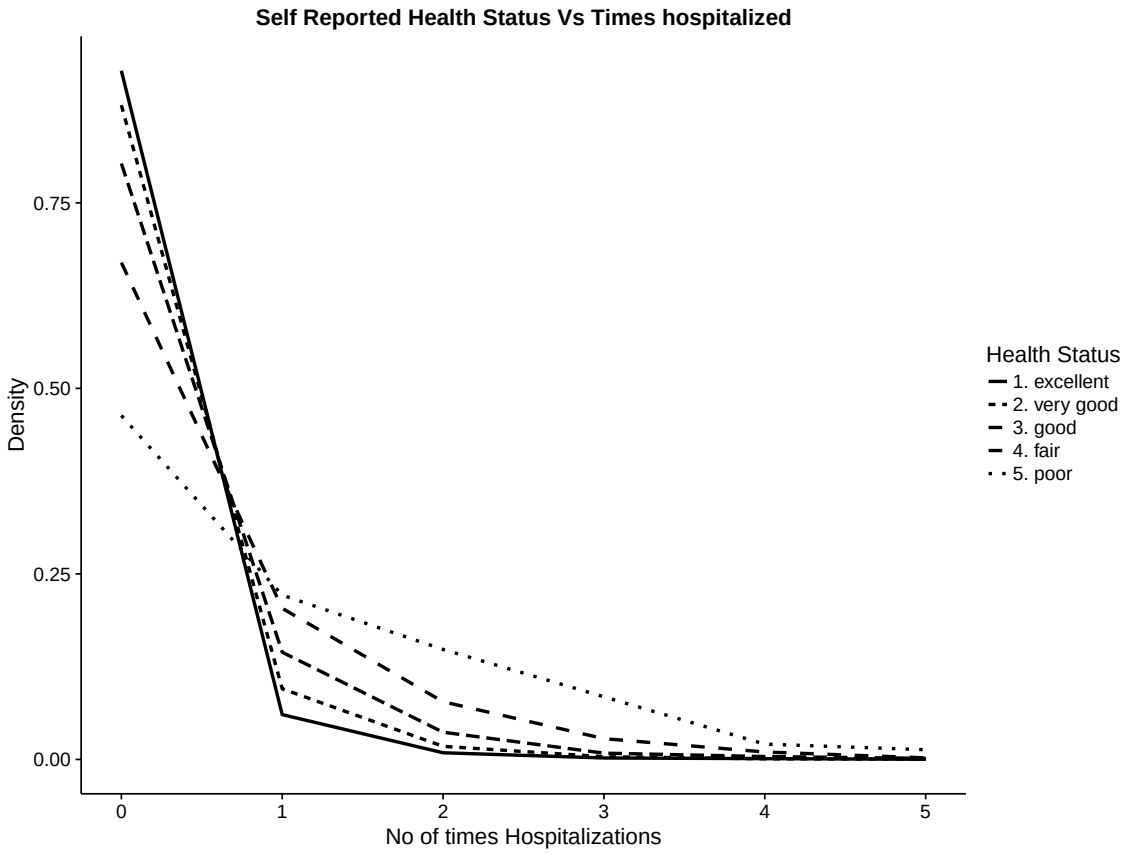

Figure 2: Figure showing the distribution of individuals across the number of hospitalizations, for each self-reported health category.

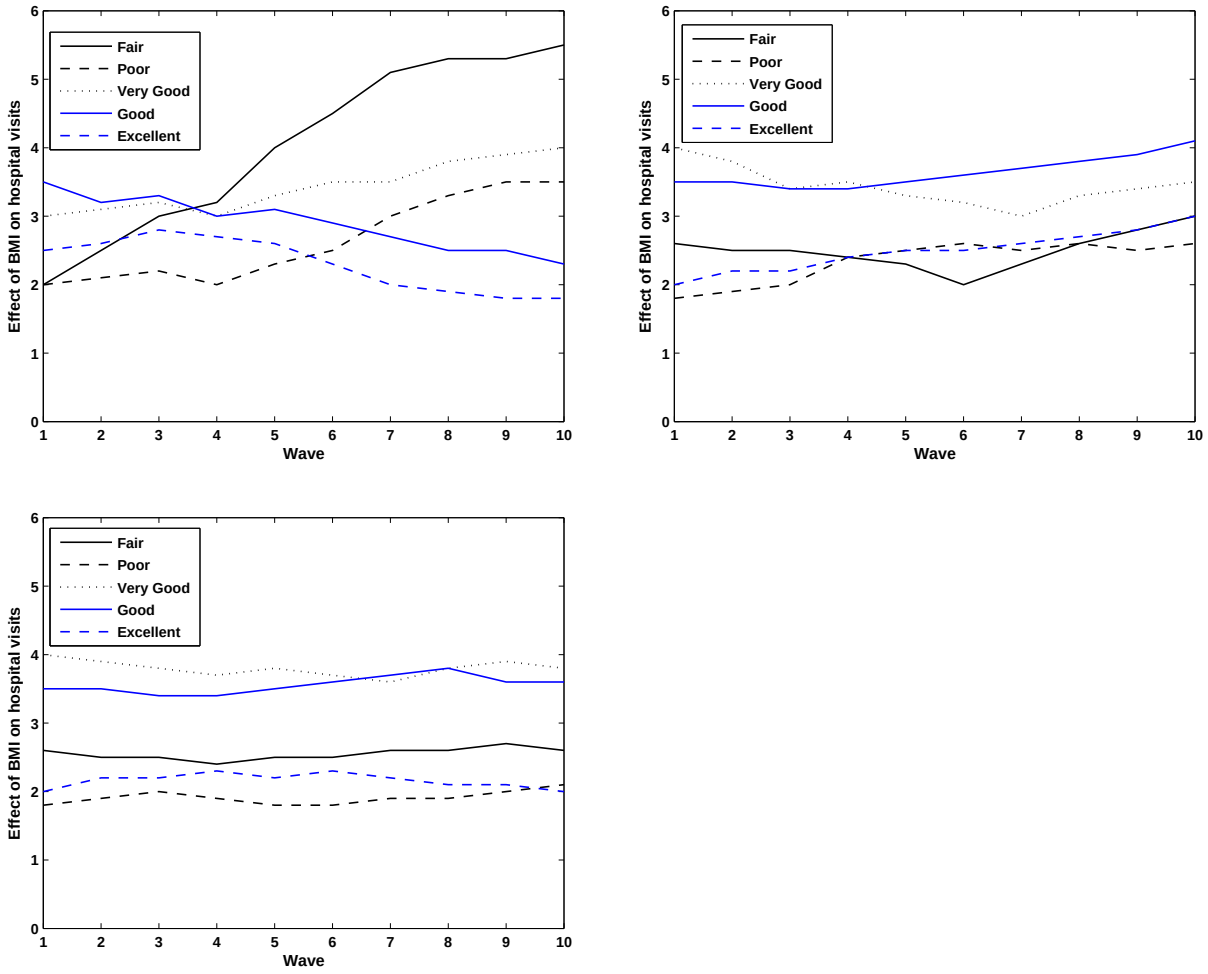

Figure 3: Time varying effect of BMI for 5 different groups for k=2, 4 and 6 respectively.

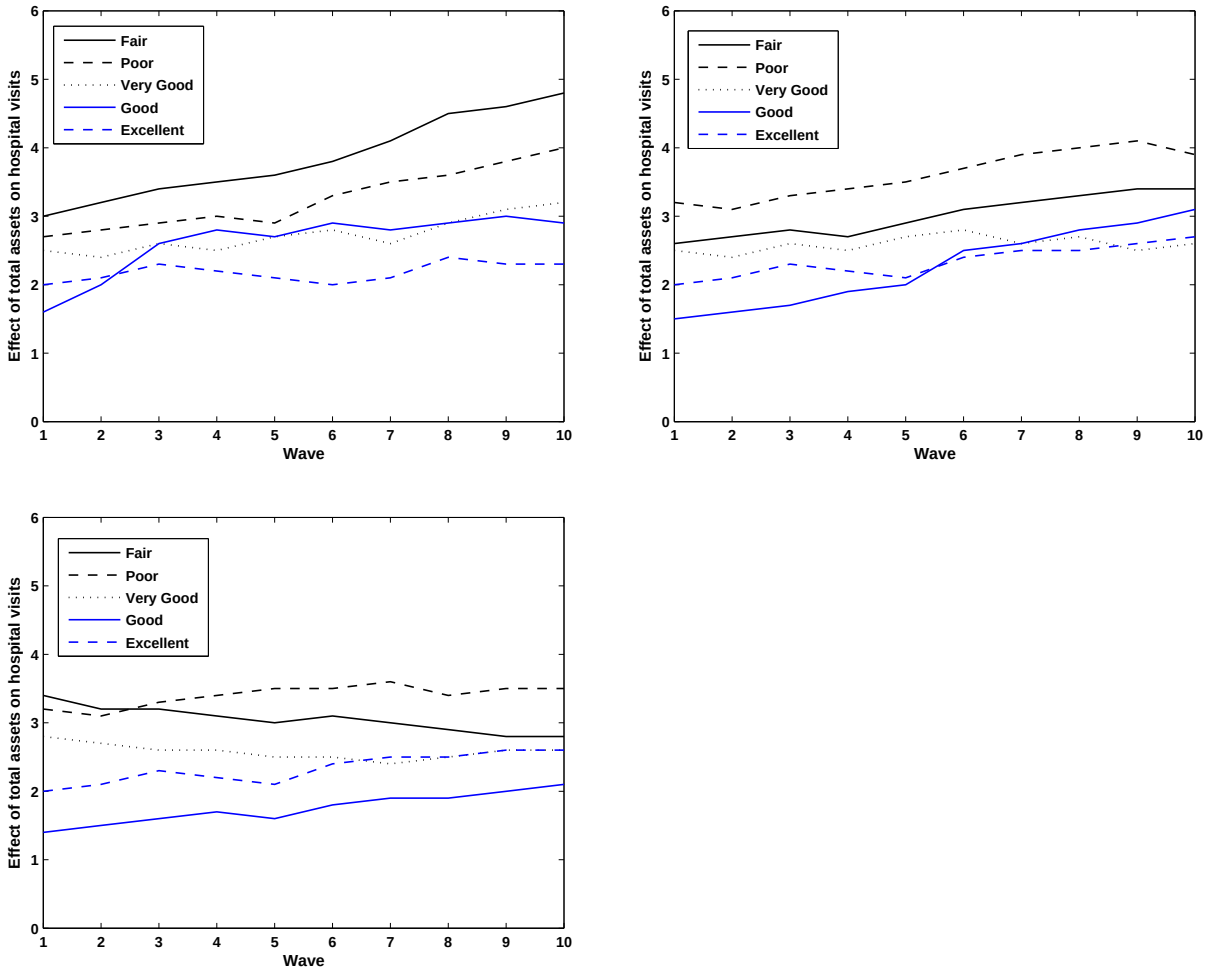

Figure 4: Time varying effect of the total assets for 5 different groups for k=2, 4 and 6 respectively.

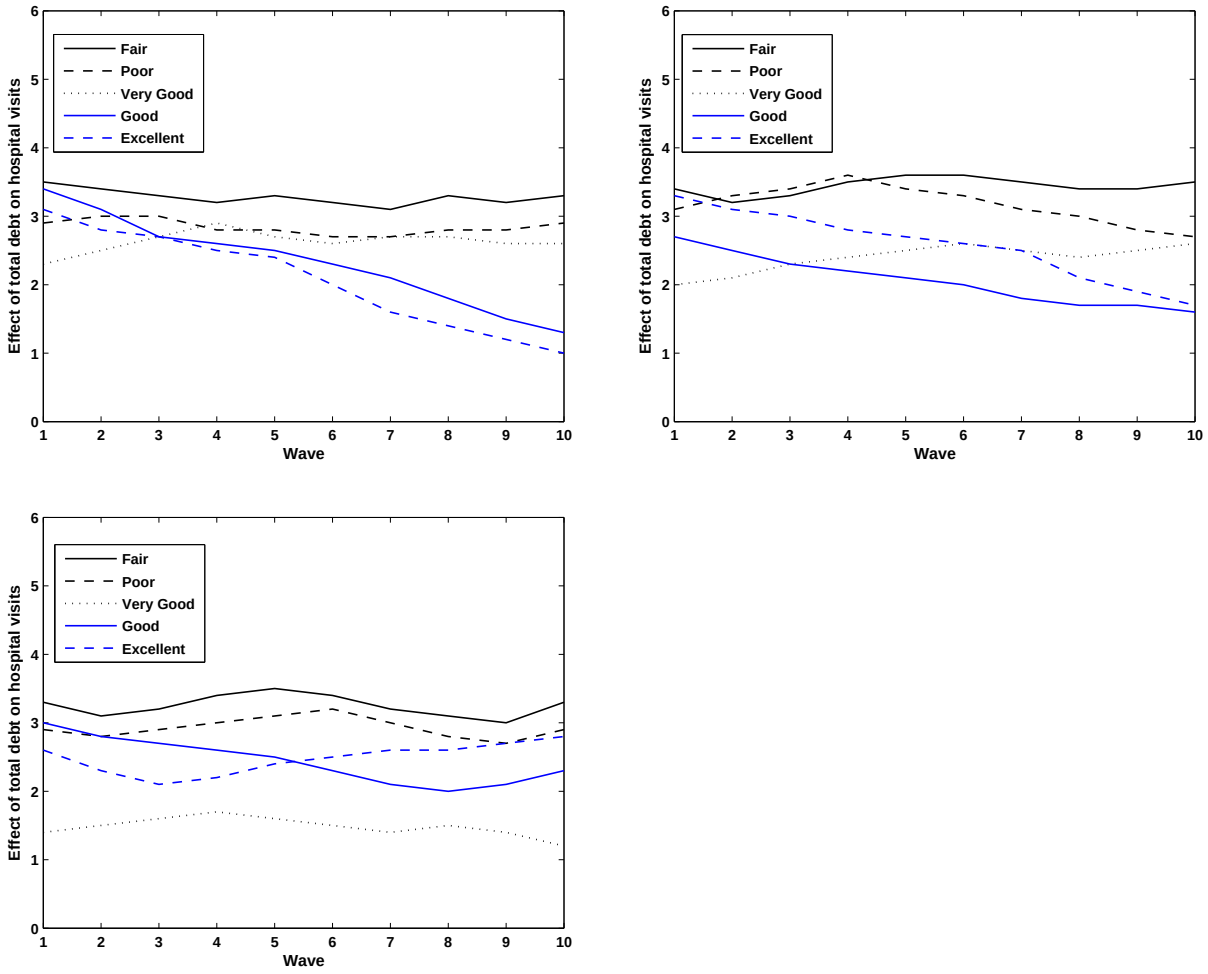

Figure 5: Time varying effect of the total debt for 5 different groups for k=2, 4 and 6 respectively.

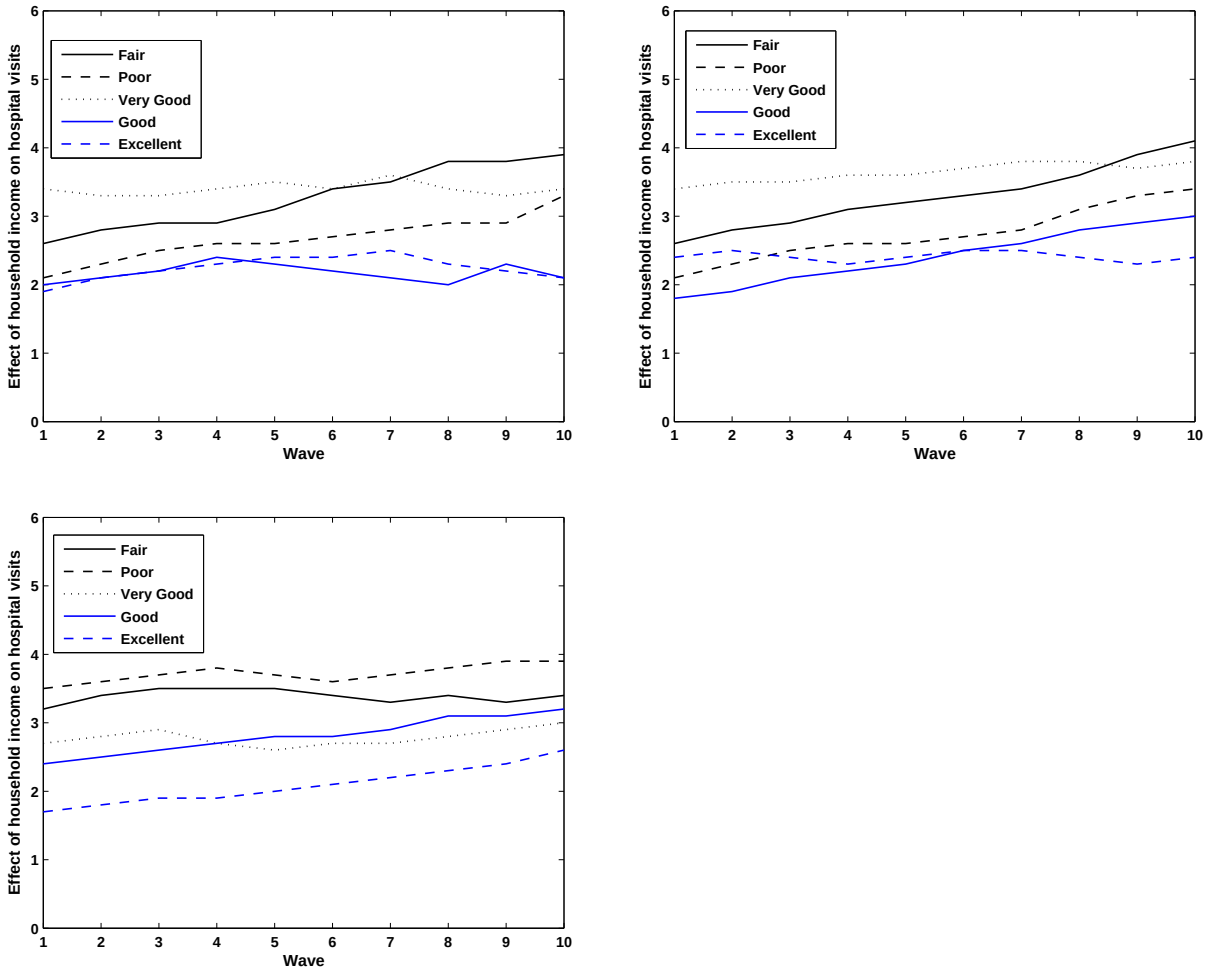

Figure 6: Time varying effect of the total household income for 5 different groups for k=2, 4 and 6 respectively.

Web-Appendix

Simulation Study Details and Additional HRS Data Graphs

We simulate data on 100 individuals belonging to 4 related groups at 10 different time points. Our response is a count variable and we consider 2 continuous predictors with time-varying effects on the response; and 10 predictors with time-invariant effects. Among the predictors with time-invariant effects, there are 8 continuous predictors, and the remaining 2 are categorical in nature.

Let x be the set of all covariates; thus $x = [x_1, x_2, \ldots, x_{12}]^T$. Again, for each predictor x_i, we have measurements for $T = 10$ time points. Hence, $x_i = [x_i(1), x_i(2), \ldots, x_i(10)]^T$. We simulate the predictors x_is, for $i = 1, 2, \ldots, 10$, from a multivariate normal density with mean= $\boldsymbol{\mu}_1 = [1, 3, 5, 4, 5, 6, 3.5, 5.5, 6, 3.8]^T$, and covariance matrix=$\Sigma$, which is the first order auto-regressive structure with $\rho_1 = 0.65$ and $\sigma^2 = 3.6$. The predictors x_{11} and x_{12} are generated at each time point from Bernoulli distributions with p=0.46 and 0.55, respectively.

Next, we generate π_{irt}, the probabilities of non-zero response. We first generate the iid samples of (η_i, b_i) from a bivariate normal density with mean vector=0 and covariance matrix=$\begin{bmatrix} 10 & 5.01 \\ - & 8.6 \end{bmatrix}$. Then the probabilities π_{irt} are generated from the following probit model:

$\pi_{irt} = \Phi(x_i^T\delta + \eta_i)$, with $\delta = [0.04, 1.4, 2.5, 0.005, 3.6, 6.3, -2.56, 0.02, -0.003, 4.36, 1.1, -3.9]^T$. For each individual i, at each time point t, we sample from a uniform (0,1) distribution; and assign a zero value with probability=1-π_{irt}. Note that individuals change their groups as described in Section 4 of the main text.

Next, for each individual i, we find the time corresponding to the first non-zero response. If t is that time, then we generate T_i, the exact time of the first hospital visit for individual i. We generate T_i from a Weibull (1,5) distribution truncated below at $t-1$ and truncated above by t. Next, we sample Y_{irt} for each individual at each time point t as follows.

For the i-th individual, if t is the time for the first non-zero response, then we sample Y_{irt} from the distribution given in equation (2) of the main text, and if t is the time after the first non-zero response, then we sample Y_{irt} using the model given in equation (3) of the main text, with $k = 0, 1, \ldots, 5$. We consider x_1 and x_2 as the predictors with time-varying effects on the response, while the other 10 predictors are treated as the covariates with time-invariant effects on the response. Thus, we have $J = 2$ and $J' = 10$.

Next, we specify the parameter values for the model in equation (3). For our simulation, we take $\beta_{j'kr} = \beta_{j'}$, and consider $\beta = [2.3, 4.5, 3.9, 0.04, 6.1, 3.2, 0.003, -1.65, -0.06, 1.4]$. For the time-varying part, we consider g_j=2, for $j = 1, 2$ and two knots at at time 3 and 7, thus $S_j = 2$. Define $\mathbf{b}_{jkrt} = [b_{jkr0t}, b_{jkr1t}, b_{jkr2t}]^T$, and also $\mathbf{c}_{jkrt} = [c_{jkr1t}, c_{jkr2t}]^T$. We show the parameter values for our simulation study for different choices of k and j. We present selected tables below: the other findings were very similar and, hence, for brevity we do not present them here (they are available on request). Note that the set of parameters for both $\mathbf{b}$ and $\mathbf{c}$ are shared across the groups and times for different values of k and j, thus allowing a shared parameter structure.

Table S.1: Model parameter values for $\mathbf{b}_{jkrt}$ across groups and times for j=1 and $k = 1$.

Time	Group 1	Group 2	Group 3	Group 4
1	(1.74,0.86,1.31)	(1.71,0.82,1.24)	(1.80,0.91,1.31)	(1.75,0.86,1.34)
2	(1.69,0.85,1.42)	(1.80,0.73,1.26)	(1.63,0.81,1.19)	(1.77,0.81,1.34)
3	(1.82,0.93,1.24)	(1.77,0.79,1.14)	(1.69,0.76,1.24)	(1.72,0.88,1.30)
4	(1.77,0.88,1.37)	(1.76,0.81,1.37)	(1.76, 1.02,1.41)	(1.80,0.81,1.38)
5	(1.65,0.78,1.30)	(1.69,0.86,1.42)	(1.79,0.87,1.26)	(1.71,0.87,1.36)
6	(1.73,0.84,1.28)	(1.75,0.84,1.28)	(1.92,0.93,1.33)	(1.80,0.88,1.33)
7	(1.81,0.89,1.32)	(1.68,0.97,1.21)	(2.05,0.77,1.37)	(1.76,0.91,1.35)
8	(1.66,0.73,1.25)	(1.94,0.82,1.04)	(1.78,0.75,1.30)	(1.69,0.74,1.25)
9	(1.76,0.75,1.28)	(1.84,0.82,1.24)	(1.77,0.79,1.30)	(1.70,0.81,1.35)
10	(1.75,0.78,1.33)	(1.88,0.82,1.34)	(1.84,0.85,1.32)	(1.80,0.81,1.37)

Table S.2: Model parameter values for $\mathbf{b}_{jkrt}$ across groups and times for j=1 and $k = 5$.

Time	Group 1	Group 2	Group 3	Group 4
1	(1.84,0.89,1.26)	(1.73,0.87,1.34)	(1.80,0.91,1.31)	(1.85,0.91,1.24)
2	(1.68,0.86,1.44)	(1.85,0.76,1.27)	(1.69,0.89,1.10)	(1.71,0.82,1.33)
3	(1.77,0.91,1.34)	(1.75,0.81,1.24)	(1.66,0.86,1.32)	(1.75,0.83,1.31)
4	(1.71,0.82,1.33)	(1.74,0.95,1.36)	(1.77, 1.08,1.49)	(1.80,0.86,1.34)
5	(1.75,0.78,1.30)	(1.63,0.84,1.44)	(1.69,0.87,1.31)	(1.74,0.88,1.32)
6	(1.73,0.86,1.23)	(1.77,0.88,1.38)	(1.82,0.83,1.23)	(1.80,0.78,1.43)
7	(1.81,0.89,1.32)	(1.78,0.87,1.21)	(2.01,0.77,1.27)	(1.76,0.81,1.25)
8	(1.86,0.73,1.29)	(1.94,0.84,1.44)	(1.76,0.85,1.31)	(1.79,0.84,1.25)
9	(1.71,0.75,1.28)	(1.86,0.82,1.24)	(1.87,0.79,1.30)	(1.70,0.86,1.43)
10	(1.69,0.78,1.43)	(1.98,0.91,1.34)	(1.85,0.88,1.34)	(1.77,0.86,1.39)

Table S.3: Model parameter values for $\mathbf{b}_{jkrt}$ across groups and times for j=2 and $k = 2$.

Time	Group 1	Group 2	Group 3	Group 4
1	(1.54,0.66,1.41)	(1.51,0.62,1.44)	(1.70,0.61,1.36)	(1.65,0.66,1.34)
2	(1.49,0.55,1.42)	(1.60,0.71,1.36)	(1.63,0.51,1.29)	(1.67,0.71,1.31)
3	(1.62,0.63,1.34)	(1.57,0.69,1.44)	(1.69,0.56,1.34)	(1.62,0.68,1.30)
4	(1.77,0.89,1.47)	(1.66,0.61,1.37)	(1.76, 1.02,1.31)	(1.70,0.71,1.28)
5	(1.75,0.88,1.50)	(1.69,0.66,1.32)	(1.79,0.97,1.26)	(1.61,0.77,1.36)
6	(1.73,0.85,1.38)	(1.65,0.74,1.38)	(1.92,0.93,1.33)	(1.80,0.78,1.43)
7	(1.81,0.89,1.32)	(1.69,0.97,1.21)	(1.95,0.87,1.35)	(1.76,0.87,1.35)
8	(1.86,0.93,1.35)	(1.94,0.92,1.24)	(1.78,0.75,1.30)	(1.69,0.74,1.35)
9	(1.86,0.85,1.38)	(1.94,0.91,1.34)	(1.75,0.75,1.27)	(1.80,0.84,1.45)
10	(1.75,0.78,1.33)	(1.78,0.82,1.34)	(1.74,0.85,1.32)	(1.70,0.71,1.30)

Table S.4: Model parameter values for $\mathbf{b}_{jkrt}$ across groups and times for j=2 and $k = 4$.

Time	Group 1	Group 2	Group 3	Group 4
1	(1.64,0.79,1.46)	(1.63,0.81,1.39)	(1.70,0.84,1.36)	(1.75,0.91,1.44)
2	(1.68,0.76,1.44)	(1.65,0.76,1.37)	(1.69,0.89,1.20)	(1.71,0.92,1.43)
3	(1.67,0.81,1.35)	(1.75,0.81,1.34)	(1.76,0.86,1.32)	(1.65,0.81,1.46)
4	(1.74,0.72,1.33)	(1.64,0.75,1.36)	(1.77, 1.01,1.39)	(1.80,0.86,1.34)
5	(1.75,0.78,1.30)	(1.63,0.84,1.34)	(1.69,0.87,1.31)	(1.74,0.84,1.32)
6	(1.63,0.86,1.33)	(1.67,0.88,1.38)	(1.72,0.83,1.33)	(1.70,0.78,1.36)
7	(1.71,0.79,1.32)	(1.78,0.77,1.31)	(2.01,0.77,1.27)	(1.76,0.81,1.35)
8	(1.66,0.73,1.29)	(1.64,0.74,1.34)	(1.77,0.85,1.31)	(1.71,0.74,1.25)
9	(1.71,0.75,1.28)	(1.66,0.82,1.24)	(1.77,0.79,1.30)	(1.70,0.76,1.33)
10	(1.69,0.78,1.33)	(1.68,0.81,1.34)	(1.85,0.88,1.34)	(1.77,0.76,1.31)

Table S.5: Model parameter values for $\mathbf{c}_{jkrt}$ across groups and times for j=1 and k=1.

Time	Group 1	Group 2	Group 3	Group 4
1	(0.74,0.96)	(0.77,0.98)	(0.79,0.95)	(0.81,0.95)
2	(0.79,0.92)	(0.72,0.95)	(0.71,0.91)	(0.76,0.92)
3	(0.81,1.01)	(0.79,0.91)	(0.76,0.94)	(0.77,0.88)
4	(0.77,0.94)	(0.70,0.93)	(0.72,0.97)	(0.79,0.89)
5	(0.74,0.95)	(0.86,0.99)	(0.83,0.90)	(0.74,0.84)
6	(0.75,0.99)	(0.67,0.91)	(0.73,0.99)	(0.76,0.95)
7	(0.68,0.89)	(0.75,0.97)	(0.69,0.92)	(0.69,0.86)
8	(0.64,0.88)	(0.74,0.92)	(0.68,1.12)	(0.69,0.87)
9	(0.73,0.84)	(0.74,0.96)	(0.66,0.86)	(0.79,0.95)
10	(0.74,0.88)	(0.74,0.82)	(0.68,1.02)	(0.69,0.88)

Table S.6: Model parameter values for $\mathbf{c}_{jkrt}$ across groups and times for j=1 and k=5.

Time	Group 1	Group 2	Group 3	Group 4
1	(0.64,0.75)	(0.67,0.73)	(0.59,0.71)	(0.61,0.75)
2	(0.69,0.72)	(0.62,0.75)	(0.61,0.71)	(0.66,0.72)
3	(0.71,0.81)	(0.68,0.71)	(0.56,0.70)	(0.67,0.78)
4	(0.67,0.74)	(0.60,0.73)	(0.62,0.77)	(0.69,0.79)
5	(0.64,0.75)	(0.76,0.75)	(0.63,0.70)	(0.64,0.74)
6	(0.65,0.79)	(0.57,0.71)	(0.63,0.79)	(0.66,0.75)
7	(0.58,0.79)	(0.65,0.77)	(0.59,0.72)	(0.65,0.76)
8	(0.54,0.68)	(0.64,0.70)	(0.58,0.81)	(0.69,0.79)
9	(0.63,0.64)	(0.64,0.76)	(0.66,0.76)	(0.63,0.75)
10	(0.64,0.78)	(0.64,0.72)	(0.68,0.77)	(0.61,0.78)

Table S.7: Model parameter values for $\mathbf{c}_{jkrt}$ across groups and times for j=2 and k=4.

Time	Group 1	Group 2	Group 3	Group 4
1	(0.45,0.55)	(0.47,0.53)	(0.39,0.51)	(0.41,0.55)
2	(0.49,0.52)	(0.42,0.55)	(0.41,0.49)	(0.36,0.52)
3	(0.41,0.51)	(0.38,0.51)	(0.46,0.50)	(0.37,0.48)
4	(0.47,0.54)	(0.40,0.53)	(0.42,0.47)	(0.39,0.49)
5	(0.44,0.55)	(0.46,0.55)	(0.43,0.50)	(0.44,0.54)
6	(0.45,0.49)	(0.47,0.51)	(0.43,0.49)	(0.46,0.55)
7	(0.38,0.59)	(0.45,0.57)	(0.39,0.52)	(0.45,0.56)
8	(0.44,0.48)	(0.44,0.50)	(0.38,0.51)	(0.39,0.49)
9	(0.66,0.84)	(0.64,0.86)	(0.66,0.85)	(0.63,0.80)
10	(0.72,0.68)	(0.74,0.82)	(0.78,0.87)	(0.81,0.79)

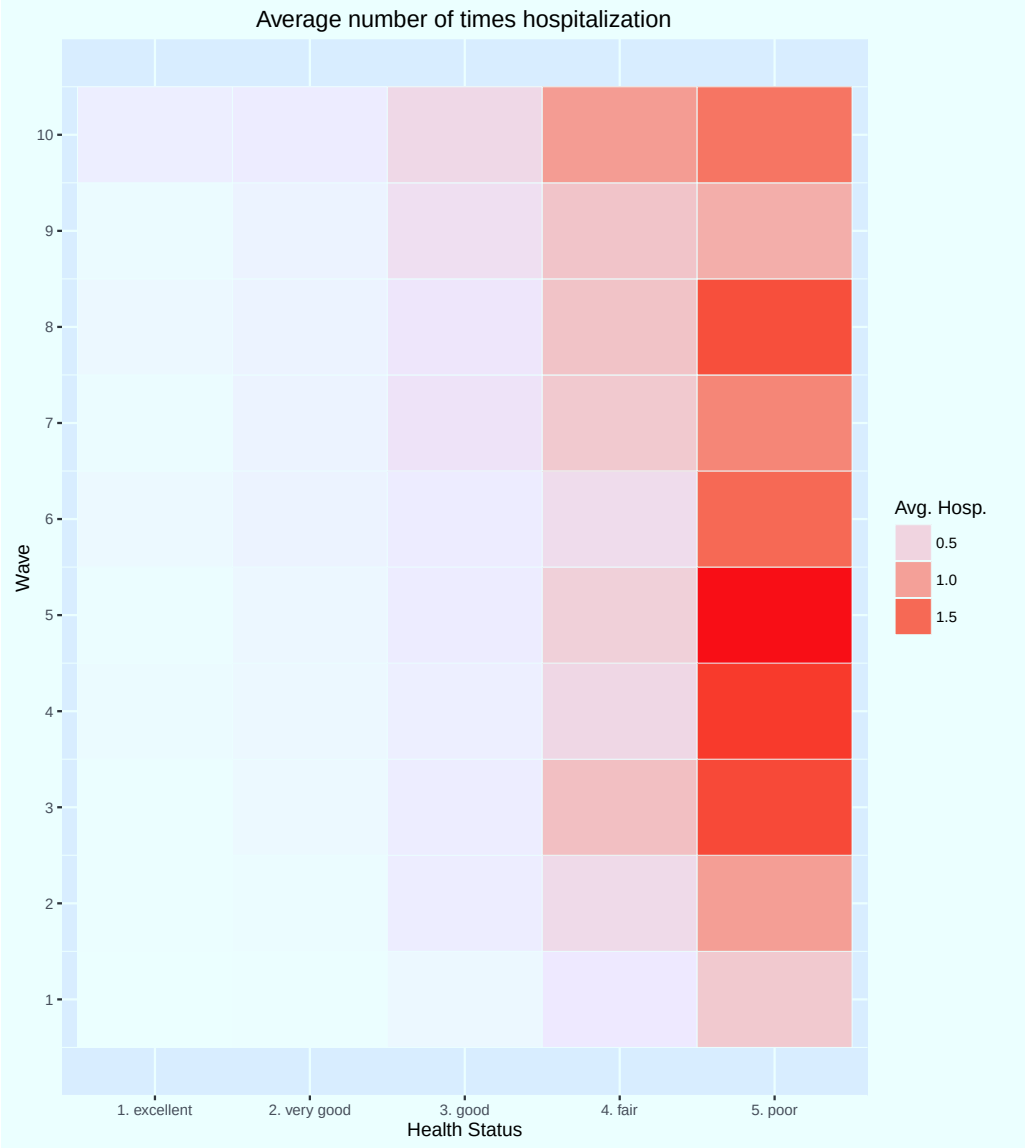

Figure S.1: Heat map displaying the average number of hospitalizations across different waves and self reported health status.

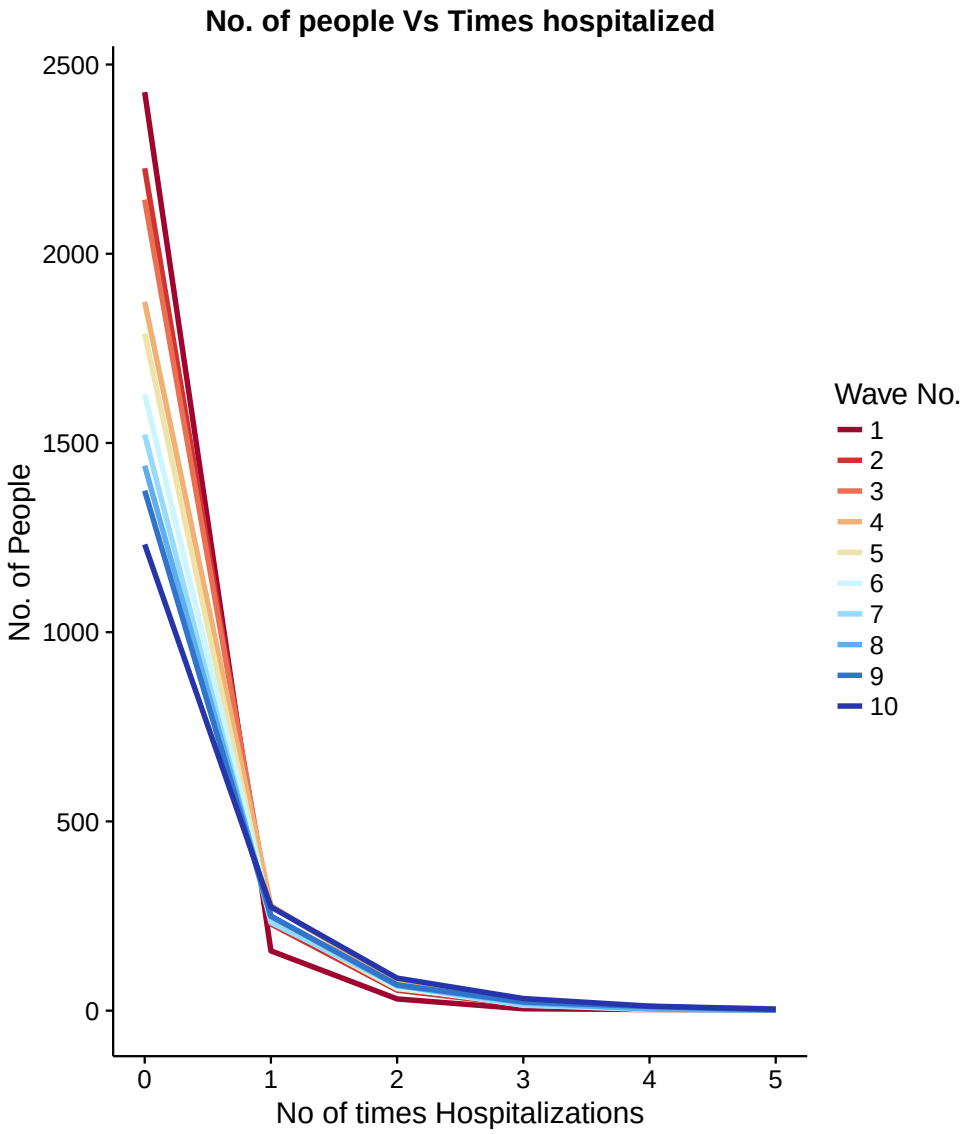

Figure S.2: Plot showing the distribution of individuals across the number of hospitalizations for different waves.